Stats NZ
Tatauranga Aotearoa

Experimental administrative population census:

Data sources and methods

New Zealand Government

Citation
Stats NZ (2021). *Experimental administrative population census: Data sources and methods*. Retrieved from www.stats.govt.nz.

ISBN 978-1-99-003265-3 (online)

Disclaimer
These results are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI) which is carefully managed by Stats NZ. For more information about the IDI please visit https://www.stats.govt.nz/integrated-data/.

Published in August 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

Experimental administrative population census: Data sources and methods describes a new experimental approach for deriving census information from administrative (admin) data. The paper describes the data and methods used to compile the first version of the experimental administrative population census released in 2021, provides information about quality, and some illustrative results for the variables that have been included.

Summary of key points

The experimental administrative population census (APC) data are **NOT** official statistics. Rather, they are published as an experimental output from research using a different methodology than what is currently used to produce the census or official population statistics.

The APC derives census-type information from linked administrative sources. Stats NZ aims to demonstrate the information available currently and to provide a focus for discussion about the benefits and limitations associated with an admin-based census.

Outputs are released on the Stats NZ website as data tables supported by a graphical and mapping interface, the Experimental administrative population census. Anonymised APC unit-record data is made available in the Integrated Data Infrastructure (IDI) for researchers with microdata access.

Key features of the APC include the merging of census information with official population estimates; so that under this approach, there is only one target resident population measure. Other advantages over a traditional census are that results can be compiled annually, and the underlying unit-record data is inherently longitudinal.

The APC is part of the ongoing census transformation programme looking at the potential for a future census based on administrative data supported by sample surveys. The research and development programme is designed to be iterative. The APC builds on experimental population estimates derived from linked administrative data released in 2016, 2017, and 2018, and extensive and detailed research on the quality of administrative data for census variables, and takes this work one step further. Feedback from customers will help to guide further development.

The APC is based on the admin New Zealand resident population and supplements this with more characteristics of individuals derived from administrative data. The first iteration released in August 2021 is an annual series, from 2006 to 2020, focusing on demographic and identity variables: age, sex, geography, ethnicity, Māori descent, birthplace, and years since arrival in New Zealand. Other topics such as work, income, and qualifications will be added in the 2022 and 2023 releases.

In comparisons for the 2018 year, the APC shows good agreement with the expected population patterns and distributions of census variables, although there is a small net undercount of the total population compared with official population estimates, and missing data for some census variables.

For this first version of the APC, there is no coverage adjustment of the admin population to account for undercoverage and overcoverage, nor is there any imputation for missing values. Further work is planned to address these limitations through reducing the level of missing data, and including statistical estimation models to improve the representativeness of the administrative data.

Administrative data has already been used in the 2018 Census for people who had been missed by the field collection. Administrative data is now purposefully included in the 2023 Census combined census model by design. However, administrative data is still very much a complementary data source, and the success of the 2023 Census depends on achieving high response rates to the field enumeration.

Administrative data of suitable quality is not available for all census variables. Iwi affiliation data is a critical requirement for iwi-Māori and for any future census model. Currently there is a lack of suitable alternative sources for this key population indicator. While not appropriate for iwi affiliation, for some other census topics that have limited information from administrative data such as language spoken and activity limitations, a sample survey will be required as part of any future census based largely on administrative data.

If you would like to provide feedback or want to find out more

Email: censustransformation@stats.govt.nz

We are especially interested in your views on:

- the data sources and methods used to derive the population and variables in the APC, and any suggestions for improvements
- your experiences using the APC data, including benefits for you compared with the traditional census data, and any unexpected patterns
- the structure of the APC unit-record data in the IDI
- what you would like to see in the next iterations of the APC in 2022 and 2023.

Introduction to the experimental administrative population census

The New Zealand 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 most recent censuses were in 2013 and 2018. 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, and is often the only source of detailed information for small communities such as ethnic groups or iwi.

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, it captures a snapshot of who is living in New Zealand at the time.

In 2012, Stats NZ established the census transformation strategy. The strategy has two strands: a short- to medium-term focus on modernising the full field enumeration census, and a long-term focus that is actively working towards a future census based on administrative (admin) data supported by sample surveys. The census transformation programme has built up a considerable body of research on the potential for deriving census-type information from administrative sources. Research has been based on the linked data in the IDI. The IDI now includes a wide range of linked administrative data sources, the 2013 and 2018 Censuses, and Stats NZ household surveys.

The first requirement of a census is to count the population by core demographic variables, nationally and down to small geographic areas. Methods for deriving a New Zealand resident population from linked administrative data have been developed iteratively and released as Experimental population estimates from linked administrative data (Stats NZ 2016a, 2017b, 2018).

Secondly, the census provides information on the social and economic characteristics of people, households and families, and dwellings. We have built up a comprehensive and detailed understanding of the quality of census characteristics that can be derived from administrative data, and identified the types of information collected by census which are unlikely to be obtained from alternative sources (Bycroft et al, 2021). A future census largely based on administrative data will still require a sample survey component. For example, census topics such as language spoken, religious affiliation, and activity limitations will need to be collected through a survey, and other more factual variables such as mortgage payments and number of rooms that are not currently collected by government agencies.

The benefits of administrative data for population statistics are already evident in practice. An admin-derived New Zealand resident population is now included in the IDI regular production process and used by IDI researchers. When the 2018 Census was faced with an unexpected low response rate, methods for using administrative data to count people who had been missed by the field collection were able to be developed rapidly because of the research undertaken for the longer-term admin-based census (Stats NZ, 2019).

Administrative data is now purposefully included in the 2023 Census combined census model by design. However, administrative data is still very much a complementary data source, and the success of the 2023 Census depends on achieving high response rates to the field enumeration. Any shift to a new census model based largely on administrative data and supported by sample surveys could not occur before 2028.

The long timeframes embedded in the five-yearly census cycle make it difficult for Stats NZ and customers alike to gain a sense of reality for changes that might seem a long way off. On the other hand, experience from Nordic countries and more recently Austria and Italy, for example, suggests that transition to an admin-based census takes at least a decade of research and planning, and it is important that customers of census data are kept informed and involved throughout that process. Experimental releases of census-type information based on administrative data are a way to make the idea of an admin-based census more concrete and provide an opportunity for customers to influence further development based on their use of the data.

Aims and scope

The experimental administrative population census derives census-type information from linked administrative sources. The APC is a demonstration of what can be achieved currently with administrative data sources to provide the fine-grained population, social, and economic statistics that are the defining features of a census.

Our aim is to demonstrate the information available and to provide a focus for discussion with customers about the benefits and limitations associated with an admin-based census. Feedback from customers will help guide further development.

Outputs are released on the Stats NZ website as data tables supported by a graphical and mapping interface: Experimental administrative population census. Anonymised APC unit-record data is made available in the IDI for all researchers with microdata access.

The APC is based on the admin New Zealand resident population and expands this with more characteristics of individuals derived from administrative data. The first iteration released in August 2021 is an annual series from 2006 to 2020 and focusses on demographic and identity variables: age, sex, geography, ethnicity, Māori descent, birthplace, and years in New Zealand. We focus first on these variables as they identify key population breakdowns of interest for much of the analysis of other census characteristics. Other admin-derived topics such as work, income, and qualifications will be added in the next iterations in 2022 and 2023.

The APC is derived directly from administrative sources, and some historical data from the 2013 Census. For this first version of the APC, there is no coverage adjustment of the admin population to account for undercoverage or overcoverage, nor is there any imputation for missing values. In future we plan to include statistical estimation models to improve the representativeness of the administrative data.

The APC is limited to information about people as individuals. In future, a similar approach may be used to demonstrate the potential for administrative data to produce census information for dwellings and housing information, and for households and families.

Comparison with other population data

Stats NZ publishes a number of population statistics, including the official estimated resident population (ERP) and the census.

The APC data are **NOT** official statistics. Rather, they are published as an experimental output from research using a different methodology than is currently used to produce the census or official population statistics.

The APC merges aspects of both the ERP and the census.

The APC **target population** is people who usually live in New Zealand at a given date. This is the same target population as that for the ERP. In contrast, the census target population is everyone in New Zealand on census night, which is different from the ERP because it includes visitors from overseas and excludes New Zealand residents who are temporarily overseas on census night. Most census data is disseminated on the basis of the 'census usually resident population count' however, which excludes the overseas visitors.

The APC aims to include a range of **variables** as the census does. The APC includes the same demographic variables as are included in the ERP (age, sex, geographic location, ethnicity, and Māori descent) but also extends to include other census variables where they can be derived from administrative data.

In essence, the APC removes the distinction between census and ERP population counts. In a future admin-based census, customers would have only one estimate of the resident population.

The APC and the 2018 Census

In the 2018 Census dataset, 89 percent of the total number of records for individuals came from census responses, and 11 percent were counted from their administrative records (admin enumerations). Since the APC aims to illustrate the possibilities for providing census information when there is no full-field enumeration census, it does not include 2018 Census responses. Rather, 2018 Census responses are used as an external validation source for the administrative variables.

Information from administrative data and from the previous census in 2013 were used to fill gaps in the 2018 Census when the characteristics of people were missing. The methods used to derive the APC are similar to the methods used for administrative data in the 2018 Census, so there is overlap in data sources for individuals where we do not have responses from the 2018 Census field enumeration.

Data sources

The APC is derived from linked administrative data and the 2013 Census in the Integrated Data Infrastructure.

What is administrative data

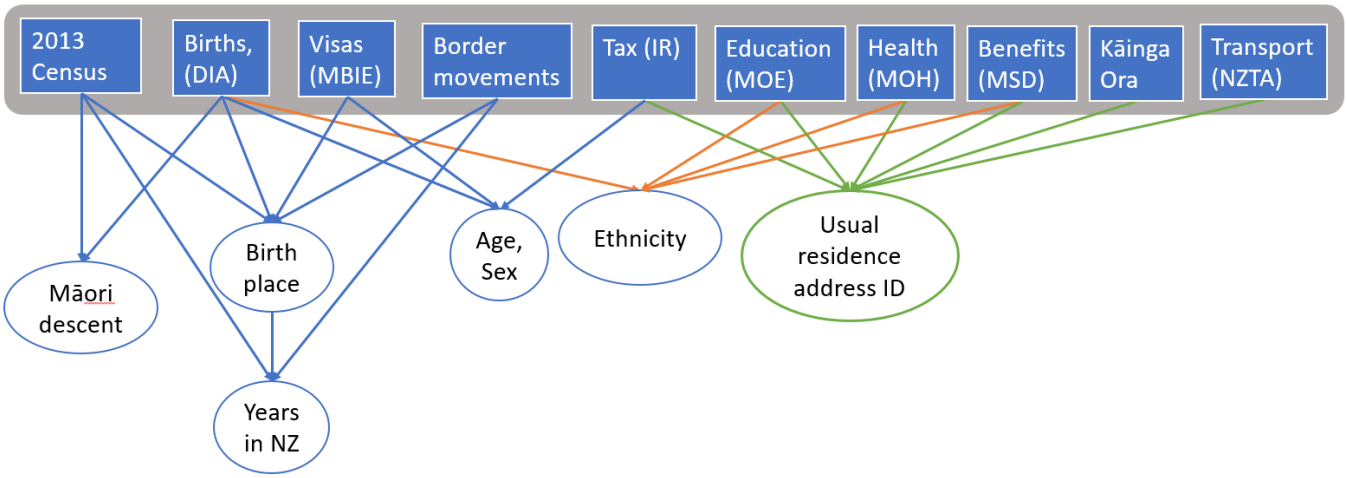
Administrative data is data collected by government agencies or private organisations while conducting their business or services. It is data not collected primarily for statistical purposes. Rather, it is collected for operations such as delivering a 'service' (for example, health or education), or as a legal requirement 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 IDI contains many administrative datasets linked at the individual level. The data includes information on education, work, income, benefit payments, 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 administrative data sources to produce census-type information.

Data sources by variable

The combination of data sources used to derive the census variables in the APC released in 2021 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. Department of Internal Affairs (DIA) birth registrations provide information on Māori descent and birthplace for those born in New Zealand. Ministry of Business, Innovation and Employment (MBIE) visa applications and border movements provide information about migrants' birthplace, and border movements data from Customs is used to derive years since arrival in New Zealand.

Figure 1



Digitisation of government admin systems occurred mainly from the late 1990s. One main reason for lack of coverage in admin sources is the absence of digitised information from earlier periods. For this reason, we also include historical 2013 Census as a data source to provide information for the population missed by the administrative sources. For the APC, 2013 Census is used only where there is no administrative data. As time goes on, the issues with historical data become less relevant. 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.

The APC output tables published in August 2021 are based on the administrative data available in Stats NZ's IDI March 2021 refresh. The APC unit-record data will be available for research within the IDI from the September 2021 refresh, and will be available in subsequent IDI refreshes. The APC can be accessed by all IDI users, and does not require any additional permissions.

Previous releases of experimental population estimates from linked administrative data included the variables age, sex, geography, and ethnicity. Descriptions of the data sources for ethnicity can be found in Reid et al (2016), and for address see Gibb and Das (2015) and Stats NZ (2017b).

We now briefly describe the administrative data sources used for Māori descent, birthplace, and years since arrival. Detailed descriptions of the data sources are available in research reports (appendix 1).

Birth registrations (DIA)

The Department of Internal Affairs manages and maintains birth registration data. New Zealand birth registrations from 1920 onwards are included in the IDI spine. This dataset has a target population of everybody born in New Zealand and children who were born overseas but adopted in New Zealand. Registrations are typically completed within three months of the child's birth. DIA data is provided to the IDI each quarter.

Māori descent has been collected on birth registration forms since 1995.

Border movements (Customs)

The New Zealand Customs Service collects information on international passenger arrivals to, and departures from, New Zealand (border movements). Customs supplies Stats NZ with electronic passport and flight records, and these are combined with information from departure cards (discontinued in October 2018) and arrival cards. The data includes all records of travel journeys across New Zealand's border (migrants, international visitors, and New Zealand-resident travellers).

Stats NZ uses this data to derive international travel and international migration statistics. The new official measure of migration is outcomes-based and uses information from border crossings to measure how long people actually spend in or out of New Zealand. To classify a border crossing as a migrant movement, we potentially need to observe up to 16 months of travel history using the 12/16-month rule. The 12/16-month rule is a way of classifying border crossings as short-term or long-term on the basis of whether travellers spend 12 months (or more) of the following 16 months in New Zealand:

- a migrant arrival is an overseas resident who arrives in New Zealand and cumulatively spends 12 out of the next 16 months in New Zealand
- a migrant departure is a New Zealand resident who departs New Zealand and cumulatively spends 12 out of the next 16 months out of New Zealand.

Border movements data incorporating the indicators of change in resident status based on the 12/16-month rule have been used to improve the previous method for deriving the admin resident population, and to derive years since arrival in New Zealand.

Visa applications and border movements identities (MBIE)

The Ministry of Business, Innovation and Employment provides Immigration NZ data on migrants and international visitors applying for a visa to enter New Zealand to the IDI. The data includes all resident visa applications and those applying for a temporary stay (work, study, and visitor). The visa data is available from July 1997.

Customs also supply border movements data to MBIE. MBIE links visa applications to border movements in the data supplied to the IDI. This information includes a traveller's birthplace. MBIE provides this data to the IDI every six months.

2013 Census

New Zealand's 2013 Census of Population and Dwellings was held on 5 March 2013 after the planned 2011 Census was cancelled due to the Canterbury earthquake. The 2013 Census was a traditional census where paper forms were dropped off and picked up by census enumerators. Online forms were also available and completed by about one-third of respondents. See 2013 Census and Datainfo+ 2013 Census information by variable for more information.

Information from the 2013 Census will not be available for individuals in the APC for a number of reasons:

- those who were not New Zealand residents at the time of the 2013 Census because they lived overseas, were not yet born, or had died before census night
- New Zealand residents who were
    - temporarily overseas on census night
    - did not respond to the census
    - responded to the census but the information for a variable is missing
    - responded to the census, but whose record could not be linked within the IDI.

The 2013 Census dataset included 5 percent substitute records (a unit imputation adjustment for people missed by the census). Item non-response in the census dataset is variable-specific and mainly between 1 and 5 percent. There was no imputation for missing data in the 2013 Census, apart from age, sex, usual residence, and (for non-substitute records) work and labour force status.

Net undercount of the 2013 Census was estimated as 2.4 percent by the 2013 Post-enumeration Survey (PES) (Stats NZ, 2014). Combining all these factors, the overall 2013 Census undercoverage 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.

A further 1.8 percent of New Zealand residents were temporarily overseas on census night and are not included in the 2013 Census by design (Stats NZ, 2013). The linkage rate of the 2013 Census was 94.5 percent in the March 2013 refresh.[1]

[1] IDI linking report, March 2021 refresh.

While response rates were lower for the 2018 Census, administrative enumerations rather than substitute records were used to count people who had been missed. For census characteristics, missing data was mitigated where possible through the use of alternative sources: administrative data, the previous census in 2013, and statistical imputation. However, the low response rates also highlighted the importance of census questionnaire responses for variables where alternative sources were not able to be used. Net undercount in the 2018 Census was measured at 2.6 percent (Stats NZ, 2020b).

The official estimated resident population (ERP)

The official estimated resident population is an estimate of all people who usually live in New Zealand at a given date. National population estimates are produced quarterly, and subnational population estimates are produced annually (reference date at 30 June).

We use the 30 June subnational ERP to make comparisons with our APC resident population at the aggregate level. The subnational ERP is published by a range of geographic boundaries (including regional councils, territorial authority areas, and Statistical Area 2 (suburb) level), five-year age group, and sex. The current methodology for producing the official ERP series relies on a periodic full enumeration census.

The ERP at a given date is derived by:

- updating the census usually resident population count (from the most recent census) to take account of net census undercount (as estimated by the Post-enumeration Survey) and residents temporarily overseas on census night; and
- updating for births, deaths, and net external migration between the census date and the date of the estimate.

In principle, the census-based ERP is most accurate immediately after the census, and accuracy tends to decrease over time the further away from the census. Population estimates remain provisional for up to two years and are revised to incorporate revisions to international (external) migration estimates.

After each census the population estimates for the preceding intercensal period (between censuses) are revised. Population statistics – user guide provides more information about the population statistics Stats NZ produces.

Methods for deriving the APC

We first outline the general features of the APC data that are different from the traditional full-field enumeration census. We then describe the methods used to derive the admin New Zealand resident population and associated core demographic variables, with more detail for each of the additional variables included in the APC published in 2021.

APC data features

The APC is constructed from administrative data which have been collected at different times and then linked in the IDI, while the full-field enumeration census collects data from everyone at the same point in time. An advantage of the APC is that since it is not restricted to a five-yearly collection period, results can be derived every year. The APC is an annual time series beginning in 2006. The APC resident population is selected independently from the IDI spine for each year of the time series.

Use of the same underlying linked data for each year of the time series provides consistency for the values of each individual, and means the data is inherently longitudinal.

Most of the population and identity variables can be thought of as having a single true value that does not change over time: age, sex, country of birth, and Māori descent are determined at birth; and year of arrival in New Zealand is a fixed date from which years since arrival can be calculated. While ethnicity is self-perceived and may change over time, with present methods we are unable to determine inter-ethnic mobility, and so ethnicity is also treated as a fixed concept. For these 'time invariant' variables, a value is derived for each individual in the IDI spine, and that value is used every year that they are a member of the APC resident population.

In practice, time invariance does not always hold. Variables may be reported differently at different times and in different contexts. For example, Māori descent may be reported as 'No' on a child's birth registration, but if new information is discovered later the same person may report 'Yes' as a parent registering the birth of their own child. The methods used to derive each variable include rules for resolving conflicting information.

Even when the reported value is fixed, classifications may change over time. For example, while a person's birthplace does not change, the name or boundaries of a country may change, and this will be reflected in updated classifications.

In contrast, since people move address, the usual residence address is derived independently each year of the APC time series. While the location of an address is fixed, geographic boundary classifications change over time. The same address may be placed in a different meshblock as meshblock boundaries are adjusted. In order to maintain consistency and comparability over the time series, we have applied the most recent classification available in the IDI refresh (March 2021 in this case).

Deriving the administrative resident population

The method for constructing a New Zealand resident population from linked data in the IDI (known as the IDI-ERP) was first published in 2016 (Gibb et al, 2016). The method was refined in the 2016 and 2017 experimental admin population estimates series (Stats NZ, 2016a, 2017b), and when the IDI-ERP was used for the 2018 Census further minor changes were made. While some improvements have been made since the original version in 2016, the basic approach has been consistent. The IDI-

ERP is formed from a list of individuals that make up the IDI spine. The spine is the union of people in three data sources:

- all births registered in New Zealand since 1920
- all visas granted to migrants since 1997 (excluding visitor and transit visas)
- all individuals issued with an Inland Revenue (IR) (tax) number.

The IDI spine in 2021 included around 10 million individuals found in one or more of these sources. Each individual is assigned a unique anonymised identifier (the snz_uid).

The IDI-ERP includes those individuals in the IDI spine who have activity in selected administrative data sources over a two-year period up to the reference date. Those who have died before the reference date are identified by a link to death registrations data and are excluded. International border movements data is used to exclude anyone who was not a New Zealand resident on the reference date, for example a resident who migrates to live overseas, or a short-term visitor to New Zealand.

The 12/16-month rule is used to define migrants only when final migration estimates are available for the entire population, 16 months after the reference date. For this APC released in 2021, the 12/16-month definition is applied for the years 2006 to 2019. As described above, to be considered a long-term migrant, someone must spend 12 of the subsequent 16 months in New Zealand (for migrant arrivals) or outside New Zealand (for migrant departures).

The 12/16-month rule aligns with the official measure of international migration. For 2020, the most recent year of the APC, final migration estimates using the 12/16-month rule are not yet available, and instead the 6/12-month rule is used to exclude non-residents. Using the 6/12-month rule, individuals were classified as non-residents if the total length of time spent overseas is at least six of the 12 months spanning the reference date. This 6/12-month rule approximates the true residence status, and results in some errors due to residents incorrectly removed (a source of undercoverage) and non-residents who were not removed (overcoverage).

For the APC, the IDI-ERP methodology has been refined further with some adjustments to the activity rules, and a step introduced to remove duplicate records. Changes made compared with the 2017 experimental series are:

- Arriving in New Zealand on a residency class or long-term visa has been added as an activity source to cover recent migrants, who might only later be included in the other activity sources.
- Removal of ACC as an activity source. An investigation of the quality of activity source indicators found that ACC was less reliable than other sources. Most ACC activity is associated with activity within the health system. Around 10,000 individuals who were found to be active solely due to ACC claims were removed from the admin resident population.
- A small change in the way IR data is used to better align with IDI-ERP's two-year activity window.
- Addition of Department of Corrections data as an activity source. This adds an activity indicator for individuals residing in prison or are under similar close supervision to the IDI-ERP. It seems reasonable to assume that some of these people would not show any other activity, while at the same time one can be quite certain of them being resident in New Zealand. This results in an increase of around 1,000 for a given reference year.

- Duplicate records are removed. This results in around 20,000 records removed from the admin resident population because there were clear indications of two records for the same person.

These are minor changes at the margins but are expected to reduce the number of individuals wrongly included or wrongly excluded. The net change in total population is less than 1 percent.

Full details of the method for constructing this APC resident population are given in appendix 3.

Relationship with IDI central tables

The IDI has standard processes to derive the variables age, sex, ethnicity, and address. Where the IDI has derived standard variables, we use a similar approach, but not always the same data. We use IDI-derived values for age and sex (from the personal_details table). The methods for deriving ethnicity and usual residence address are similar to the standard processes used in the IDI; however, we do not include any values from the 2018 Census, and 2013 Census is only used when no administrative sources are available.

Methods for deriving APC attributes

In general, we apply the derivation methods developed in earlier research. The main changes are to improve coverage using 2013 Census data when the value is expected to remain constant over time.

Detailed descriptions of the data sources and methods are available in research reports (appendix 1) and are summarised in Bycroft et al (2021).

Multiple sources and conflicting values

There often are multiple sources available for the same variable. A valid admin value is given priority over a 2013 Census response. Conflicting information within the admin sources is resolved in different ways depending on the nature of variable. In general, the approach can be described as follows:

- valid values from administrative sources are used first
    - any conflicting information is resolved using the following approaches as appropriate for each variable:
        - use the highest quality source available
        - select the most recent response
        - prioritise self-response over a proxy response
- if a value is still missing, a valid 2013 Census response is used where available
- otherwise, the variable is missing.

Any 2013 Census residual codes such as response outside scope, or not stated, are set to ‘missing’.

We welcome feedback on how we have chosen to derive variables in this first iteration of the APC. In particular, whether there are any other suitable administrative sources, and the details of how to resolve situations where there are multiple sources with conflicting information.

Address

The most recently updated administrative address prior to the reference date is selected as the usual residence address. However, certain administrative sources were deemed to be of lower

quality, and therefore only used when other address information was not available. See Stats NZ (2017b) for further details. 2013 Census data is used only when no administrative address is available.

Recent improvements (March 2019) in address matching in the IDI have increased the accuracy of address and geography variables over previously reported results.

Ethnicity

Administrative data sources of ethnicity have been ranked by quality. The highest quality available administrative source for each individual is selected to provide ethnicity. See Reid et al (2016) and Stats NZ (2018) for further details. 2013 Census ethnicity data is used for less than 0.5 percent of individuals who have no ethnicity in administrative sources.

Māori descent

The statistical standard for Māori descent defines the concept as: "A person has Māori descent if they are of the Māori race of New Zealand: this includes any descendant of such a person."

The classification has three response options: 'Yes', 'No', and 'Don't know'.

The Māori descent variable for the APC is derived from DIA birth registrations, and from the 2013 Census. Both sources follow the standard and include 'Don't know' as a valid response option.

The birth registration form question on Māori descent is asked of the parents as well as the child. We treat the parent's response as self-reported and the child's as a proxy response provided by the parents. The same person may have reported Māori descent in several birth registration records: they may have been born in New Zealand, later becoming a parent themselves; or have had more than one child in New Zealand.

The statistical standard states that Māori descent is based on a genealogical or biological concept, and therefore we may assume that it does not change over time. In practice, however, the same person can report different descent values. In the logic of the derivation, we prioritise the most recently reported value within birth registrations (including preferring the self-reported value of a parent, over the proxy response of their own birth registration).

APC Māori descent derivation

Derived from DIA birth registrations (for parents and children) and the 2013 Census.

For every unique individual (snz_uid) on the IDI spine:

- valid responses from administrative sources (births) are used first
    - births data can have multiple information for the same snz_uid; in case of conflicting information
        - self-response (that is, parent record) is prioritised over a proxy response (that is, child record)
        - if there are multiple values for the same person, we use the most recent one
- if a value is still missing, a valid 2013 Census response is used where available
- otherwise, the variable is missing.

For Māori descent, 'Yes', 'No' and 'Don't know' are all valid answers.

Birthplace

The statistical standard for birthplace defines the concept as "the country where a person was born".

There are two standard country classifications in use: the four-numeric classification and the two-alpha classification. Stats NZ's standard country classifications are aligned to the countries recognised by the United Nations Statistical Division and the International Organization for Standardization (ISO). The four-numeric classification includes supplementary codes used to process inadequate responses, for example, 'North-East Asia - not further defined (nfd)'.

Birthplace is derived from DIA birth registration records for those born in New Zealand, and birth country from MBIE passport and visa information for those born overseas. 2013 Census is used to fill in gaps where administrative data is missing.

The 2013 Census codes responses to the 4 Numeric classification using the name of the country at the time of the census.

MBIE data is coded according to the 2 Alpha classification. The classification will reflect the country name at the time at which the passport or visa was issued. MBIE codes were converted to the four-numeric country classification at the time of the 2018 Census.

The United Kingdom presents a special case. The 4 Numeric codes used by the census include countries within the United Kingdom as well as the United Kingdom (nfd). While individual countries in the United Kingdom appears in the 2 Alpha coding, only the broader United Kingdom exists as a birthplace in the MBIE data.

DIA births record data is assumed complete from 1920 onwards, therefore birth records are used as the only source for assigning New Zealand-born. A claim of New Zealand as a birthplace in other sources for a person born since 1920 and not found in the birth registrations is assumed to be an error and is not used. For individuals born prior to 1920, claims of New Zealand birth were accepted from MBIE data and 2013 Census responses.

APC birthplace derivation

Derived from DIA birth registrations, MBIE border movements and Visas and the 2013 Census.

For every unique individual (snz_uid) on the IDI spine:

- valid responses from administrative sources are used first
  - all available New Zealand birth registrations are assigned as New Zealand-born
  - the earliest birthplace recorded in the border movements and visas is used to assign overseas country
  - if the individual was born prior to 1920, all border movements and visas claiming New Zealand birth are assigned as New Zealand-born
- If a value is still missing,
  - valid overseas country responses from 2013 Census are used
  - if the individual was born prior to 1920, all 2013 Census responses claiming New Zealand birthplace are assigned as New Zealand-born
- Otherwise, the variable is missing.

Years since arrival in New Zealand

The year of arrival in New Zealand is the year that a person who was born outside New Zealand first arrived in New Zealand as a permanent or long-term resident. The years since arrival is derived by measuring the time elapsed (in completed years) between the first year of arrival in New Zealand and the reference date, irrespective of any intervening absences, whether temporary or long term (Statistical standard for years since arrival in New Zealand).

The administrative data source is the Customs border movements data which starts in September 1997. The date of border crossing for the first year of arrival of a migrant has been determined by Stats NZ using the 12/16-month migration definition. The number of years since arrival in New Zealand is derived from the date of arrival the first time the migrant achieved resident status according to the 12/16-month rule.

2013 Census is used to fill gaps where administrative data is missing. The 2013 Census is the only available source for migrants arriving in New Zealand prior to 1997. Use of 2013 Census data for those who arrived before 1997 also avoids incorrectly using more recent border movements as their date of first arrival. For arrivals between 1997 and 2000 we prioritise 2013 Census data and use an arrival date derived from the 12/16-month rule if no 2013 Census data is available.

Our confidence in the migration estimates derived using the 12/16-month rule is higher after 2000 as it is possible to compare it with the intercensal migration estimate using 2001 and 2006 Census data and good agreement is observed (Stats NZ, 2017a). Therefore, the Customs border movement data is our highest-ranking source for migrant arrivals after 2000.

In cases where the 2013 Census month of arrival is missing, we randomly impute a month, consistent with the derived years since arrival in the 2013 Census.

An edit check is applied to ensure that the date of arrival in New Zealand is later than the date of birth. If not, date of arrival is set to missing.

For those born in New Zealand, years since arrival in New Zealand is not applicable.

APC Years since arrival in New Zealand derivation

Derived from Customs border movements and the 2013 Census.

For every unique individual (snz_uid) on the IDI spine with an overseas-born indicator:

- if date of arrival in 2013 Census is earlier than 2000, the census month and year of arrival is used
- if a 12/16-month migration indicator exists, the date of first arrival is the border crossing date of the earliest migration status change from overseas visitor to resident.
- if a value is still missing, 2013 Census month and year of arrival is used where available
- otherwise, the year of arrival is missing.

The date of arrival must be later than date of birth.

Years since arrival is calculated from the date of arrival, for 30 June of each year of the APC.

Quality measures

Accuracy standards for administrative population estimates

No estimate of the population can be completely accurate, and the question of how accurate do population statistics need to be is a key driver for any approach to census-taking.

Census transformation previously developed a set of quality standards to assess the quality of population estimates produced from administrative data (McNally & Bycroft, 2015). These quality standards were determined through consultation with core customers, and provide a measure of the minimum accuracy acceptable for users. Separate standards were produced for both a survey-based and an administrative-based census model. These standards apply to the final estimates, after implementing statistical models to adjust for any errors in the estimates produced from administrative data alone. Therefore, there is scope for improving estimates described in this paper which are not currently meeting the standards.

Table A2, in appendix 2, summarises these quality standards and shows how far they are achieved for the APC in 2006, 2013, and 2018. The quality standards are specified as the proportion of estimates that should come within a fixed level of error. We use the official ERP as the benchmark. For example, of the national level five-year age group estimates, 90 percent should be within 1.5 percent of the ERP, and all should be within 5 percent.

Quality metrics for census attributes

The APC is trialling a new approach for measuring the quality of statistics derived from multiple sources. The method for deriving the new quality metrics builds on quality measures used in previous census transformation research, the 2018 Census Quality Rating Scale, and techniques used by Statistics Austria for multi-source statistics.

Stats NZ's census transformation programme has evaluated the quality of admin-derived census variables using elements of the Total Error paradigm and described in Stats NZ's quality framework for administrative data (Stats NZ, 2016b). The quality of administrative data for census variables was assessed across the two dimensions of 'representation' and 'measurement'. Representation, which reflects the degree to which administrative data is available for the ideal target population, was mainly quantified through a coverage measure. The 'measurement' dimension reflects whether the intended concept is being measured correctly. Consistency between an admin-derived value and 2013 Census response for the same person was used as a proxy for measurement error. Bycroft et al (2021) provides examples.

The 2018 Census Quality Rating Scale included three metrics capturing different aspects of quality. Metric 1 - 'data sources and coverage' provided a quantitative score indicating the quality of a variable. To calculate a score for a variable, each source that contributes to the output for that variable is rated and multiplied by the proportion it contributes to the total output. The rating for a valid census response is defined as 1.00. This does not account for potential errors in census responses due to, for example, respondent misunderstanding or census processing errors. Ratings for other sources are the best estimates available of their quality relative to a census response. The 2013 Census and administrative source ratings reflect measured consistency with the 2018 Census responses.

The ratings for each source can be thought of as the quality of 'input' data sources, while the final Metric 1 score is a measure of the 'output' variable quality.

Statistics Austria developed a framework for assessing the quality of variables derived using multiple data sources for their first admin-based census in 2011 (Asamer et al, 2016). Statistics Austria produced quality indicators at the input and output stages of the production cycle, which reflect the accuracy of data sources, and of the variables constructed, respectively. Statistics Austria's measure of input quality brings together several factors, including the quality of the data at the source agency and consistency of reported admin values with an external validation source. The measure for output quality used by Statistics Austria extends the approach used for the 2018 Census by considering additional information when several sources are available for an individual.

APC quality measure

For the APC we apply an adaptation of the Statistics Austria approach for the 2018 APC reference period, and for three variables: Māori descent, birthplace, and years since arrival in New Zealand. The sections below describe the method for deriving the quality metrics in more detail.

Each variable in the APC is constructed from one or more administrative data sources and the 2013 Census as described earlier. The quality of the input values for each source variable are evaluated first. The 2018 Census is our external validation source, and as in previous work, the quality rating for each source is the measured consistency with 2018 Census responses for that variable. Some allowance is made for uncertainty in the 2018 Census and the accuracy of the administrative source, depending on the variable, following previous practice (Bycroft et al, 2021; Stats NZ, 2019). Any imputed values are also assigned a quality rating.

However, multiple sources may have information for the same individual, and Statistics Austria used this fact to construct their output quality rating. They applied Dempster Schafer Theory (DST) to assign a level of confidence in the value selected for a given unit. DST is an extension to probability theory to quantify uncertainty; DST is used to derive the belief in a proposition (for example 'the unit of interest was born in New Zealand') based on the evidence (that is data sources) that exist (Beynon et al, 2000). The starting 'belief' is the quality rating derived for the input data source. If two sources have the same value for an individual, the confidence that this is the correct value increases. Conversely, if two sources disagree, confidence in the correctness of the chosen value is weaker. The confidence in a chosen value will change as more data sources become available; allowing for changes in quality ratings as the availability of administrative sources vary over time.

Each individual record is assigned an output quality rating depending on whether the available input data sources agree or disagree. Individuals with missing data for that variable are assigned a quality rating of zero. The overall output data quality metric is derived by taking the mean of belief values over the subject population. Calculating the belief of each value in the unit-record data means that output quality ratings for a variable can be derived for a particular demographic or group of interest.

We now provide two scenarios that illustrate how this approach was implemented for the APC output data quality, using birthplace as an example.

Scenario 1

Table 1 presents a list of four mock records that include information about the units' birthplace from two data sources, the input quality rating for each source, and the final value selected with the updated output quality rating for that unit.

Table 1

Illustration of the calculation output data quality, with two input data sources						
Unit	In Birth Register: (= NZ born)		MBIE		Output data	
	Value	Input quality	Value	Input quality	Value	Output quality
1	NZ	0.8	NZ	0.6	NZ	0.92
2	NZ	0.8	Samoa	0.6	NZ	0.62
3	NZ	0.8	Samoa	0.6	Samoa	0.23
4	NZ	0.8	…	0.6	NZ	0.8
Symbol: … not available						

The belief reflects the amount of evidence that exists to support the final value selected. For example, unit 1 has a high belief because both data sources available are consistent with the final value selected. Unit 4 also has a relatively high belief because the final value selected is consistent with a high-quality data source and does not get penalised for not being recorded in the MBIE data. This may reflect a real-life scenario where an individual was born in New Zealand, but has not travelled outside the country.

Inconsistencies between data sources will reduce the belief in the final value selected. For example, both units 2 and 3 have contradictory evidence regarding their birthplace. Unit 3 is provided as an example to illustrate the case where a lower quality source that contradicts a higher quality source has been selected for the output. The belief for unit 2 is higher than unit 3 because the value selected for unit 2 is consistent with a higher quality data source. That is, there is stronger evidence to support the notion that the value selected for unit 2 is true compared with unit 3.

The output quality rating for birthplace for this group is the average rating across the four individuals:

(0.92 + 0.62 + 0.23 + 0.80) / 4 = 0.64

Scenario 2

Table 2 presents the same list of mock records that now include an additional data source to demonstrate how belief values can change with more evidence.

Table 2

Illustration of the calculation output data quality, with three input data sources								
Unit	In Birth Register: (= NZ born)		MBIE		2013 Census		Output data	
	Value	input quality	Value	input quality	Value	input quality	Value	Output quality
1	NZ	0.8	NZ	0.6	NZ	0.5	NZ	0.99
2	NZ	0.8	Samoa	0.6	NZ	0.5	NZ	0.75
3	NZ	0.8	Samoa	0.6	Samoa	0.5	Samoa	0.45
4	NZ	0.8	…	0.6	…	0.5	NZ	0.8
Symbol: … not available								

The belief in a value will increase if more data sources that are consistent with the final value are introduced. For example, the 2013 Census provides more evidence for the final values selected for Units 1, 2, and 3, and as a result, there is an increase in the belief that the values selected for these records reflects the truth. The belief for unit 4 remains unchanged – although there is no additional evidence to support the record's New Zealand birthplace, there is also no contradictory evidence to reduce the belief in the value selected.

The output quality rating for birthplace, for this group of units, has increased with the introduction of 2013 Census as a data source. The output quality rating for this group of units is:

(0.99 + 0.75 + 0.45 + 0.80) / 4 = 0.75

The quality rating is now higher than in scenario 1 because in this example the third source has provided more evidence for the chosen output values.

Further details of the calculations are provided in appendix 4.

Results

We first provide results for the APC resident population, including time series comparisons against the official estimated resident population (ERP) by age and sex, and for breakdowns by geography and ethnicity. These are compared against quality standards for population statistics based on administrative data.

For the three new census attribute variables (Māori descent, birthplace, and years since arrival in New Zealand) we provide quality measures including coverage of each variable over time, and data quality metrics for the APC in 2018. For each variable we describe the distributions compared with results for the 2018 Census and discuss some data quality issues.

APC resident population

The official ERP is the most natural population to compare with the APC resident population, since both share the same target population of all residents in New Zealand at a given date. We compare the APC for each year with the subnational ERP at 30 June. We note however, that unlike the ERP, at this stage the APC population has not been adjusted to account for any undercoverage or overcoverage. We also compare the APC with census usually resident population counts for 2006, 2013, and 2018. The official ERP is always higher than the census because the census target population excludes residents temporarily overseas on census night, and census counts have not been adjusted for net census undercount.

National population by age and sex

We compare the APC's admin-ERP to both the official ERP and census from 2006 to 2020, for the total population (figure 2a) and by sex (figure 2b).

The APC results suggest a very close alignment between the admin-ERP and the official ERP from 2006–2013 followed by a period of divergence. In 2013, the APC population total is just 0.2 percent lower that the ERP total, whereas in 2018 the gap has increased to 1.9 percent (table 3).

This pattern is due to the interplay between male and female counts which can be seen in figure 2b. We see a relatively consistent lower count for females in the APC compared with the ERP across the entire time series. This is balanced by the higher count of males in the APC between 2006 and 2013. From around 2013 the APC counts for males also gradually become lower than the ERP.

The APC is higher than each census, partly due to the inclusion of residents temporarily overseas in the APC, but not in the census. In 2018, the APC total population is 2.3 percent higher than the 2018 Census.

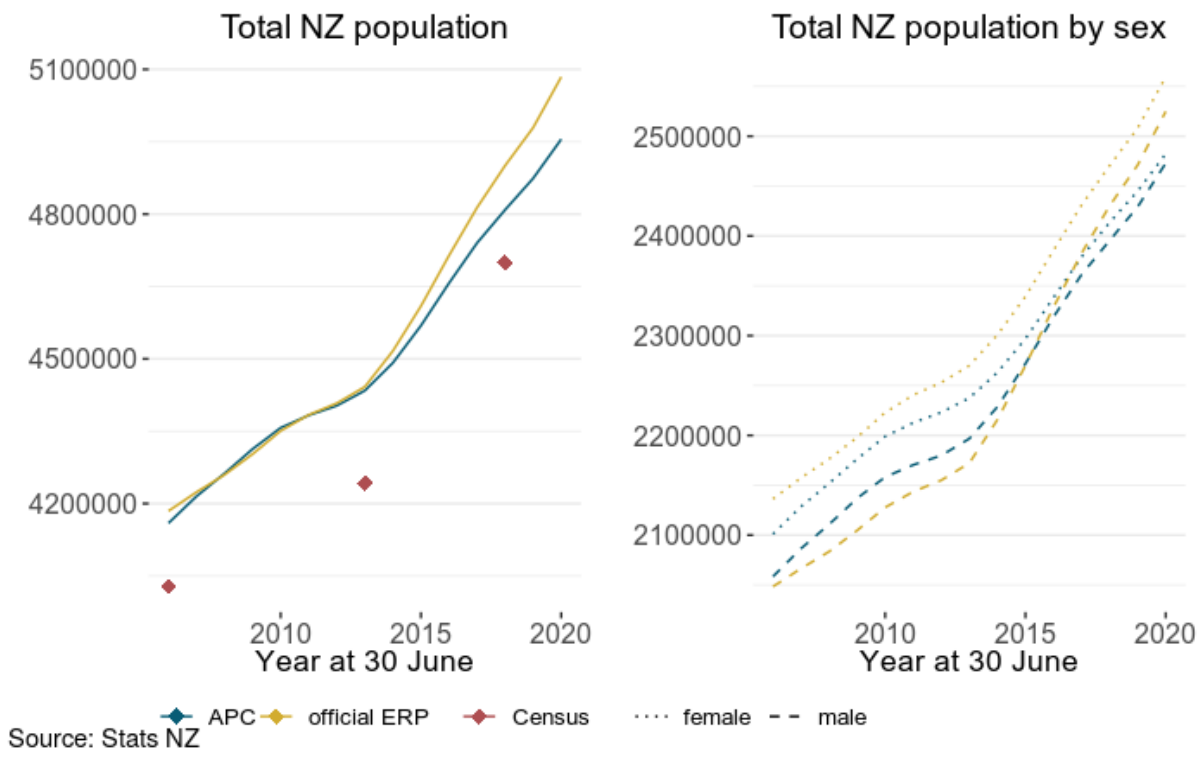
Table 3

Year	Total resident population, comparison between APC, official ERP, and census				
		Count		Percentage difference to APC	
	APC (30 June)	Official ERP (30 June)	Census (7/6/5 March)	ERP	Census
2006	4,159,404	4,184,600	4,027,947	-0.6	3.3
2013	4,434,162	4,442,100	4,242,048	-0.2	4.5
2018	4,808,949	4,900,600	4,699,755	-1.9	2.3

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

Figure 2



The broad pattern of lower counts of the admin resident population compared with the official ERP towards 2020 persists when considering subpopulations. An additional pattern emerges which is visible in most of the rest of this analysis: one of larger differences between groups towards the beginning of the time series and more uniformity towards its end. For example, figure 3 shows percentage differences by 5-year age-groups and sex in 2013 and 2018.

The 2013 pattern of markedly more adult males aged between about 20 to 50 years in the admin resident population compared with the ERP was also seen in previous experimental series (Stats NZ,

2017b). In contrast, for the 2018 comparison, the APC is lower than the ERP across almost all age groups for both males and females, with differences ranging between +0.85 to -3.35 percent.

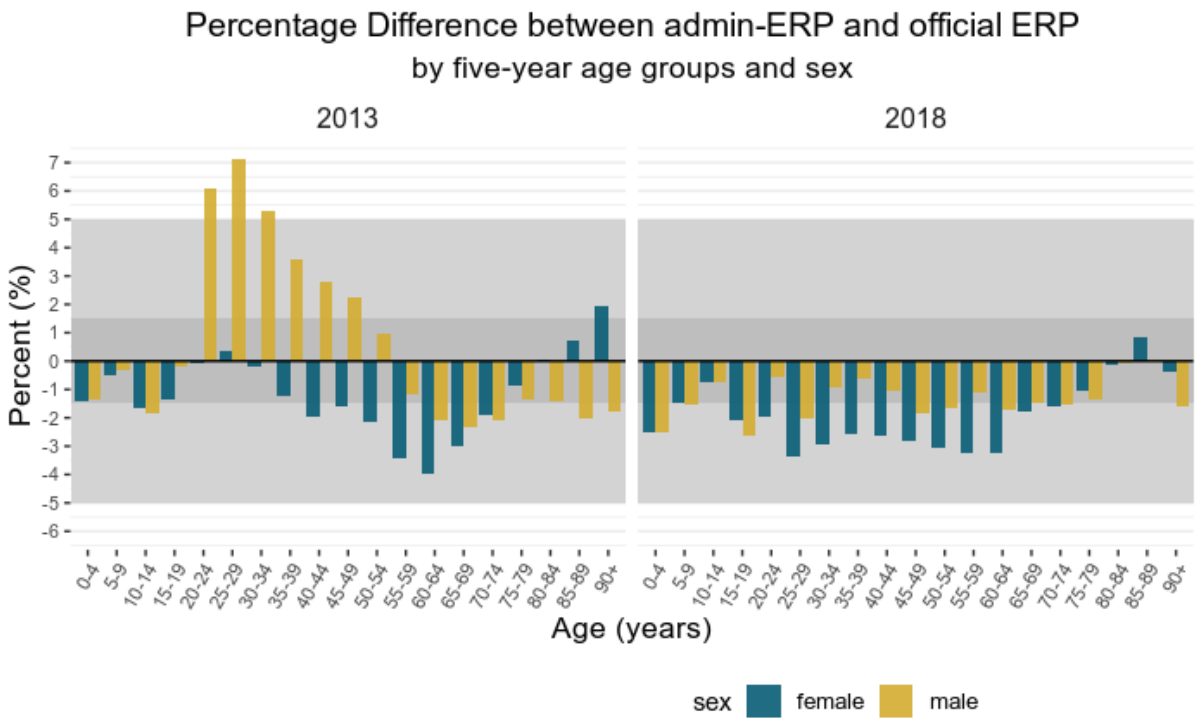
The method for deriving the admin resident population remains constant over the time series (apart from the most recent 2020 year when external migration uses an approximation), and there is considerable redundancy built into the activity sources used to select individuals. Unless there are marked differences in behaviour or significant changes to the administrative sources, we would not expect big differences in the pattern of coverage errors in the APC during the period from 2006 and 2019.

However, investigation into the causes of the recent revision of Māori population estimates suggests that the 2013 ERP age distribution may have too few males aged 15–59 years, partially due to biases in the imputation of age and sex for substitute records in the 2013 Census (Stats NZ, 2021 forthcoming). The effects of any bias in the base 2013 ERP age distributions would partly remain in the ERP time series through to the new 2018 ERP base. The 2018 Census used administrative records to count people who had been missed by the field collection, and used those records for their age and sex characteristics, so no longer relied on imputation for age and sex distributions.

Therefore, comparisons with the 2013 ERP may misrepresent the accuracy of the APC population, and comparisons against the ERP from 2018 may be more realistic.

The grey shading in figure 3 indicates ranges of ±5 percent and ±1.5 percent, respectively. The 2018 APC resident population meets the quality standard requirement of all 5-year age group by sex estimates being within 5 percent of the benchmark value, but does not achieve the requirement of 90 percent within 1.5 percent. This is a good result considering that there has been no coverage adjustment for the APC population.

Figure 3



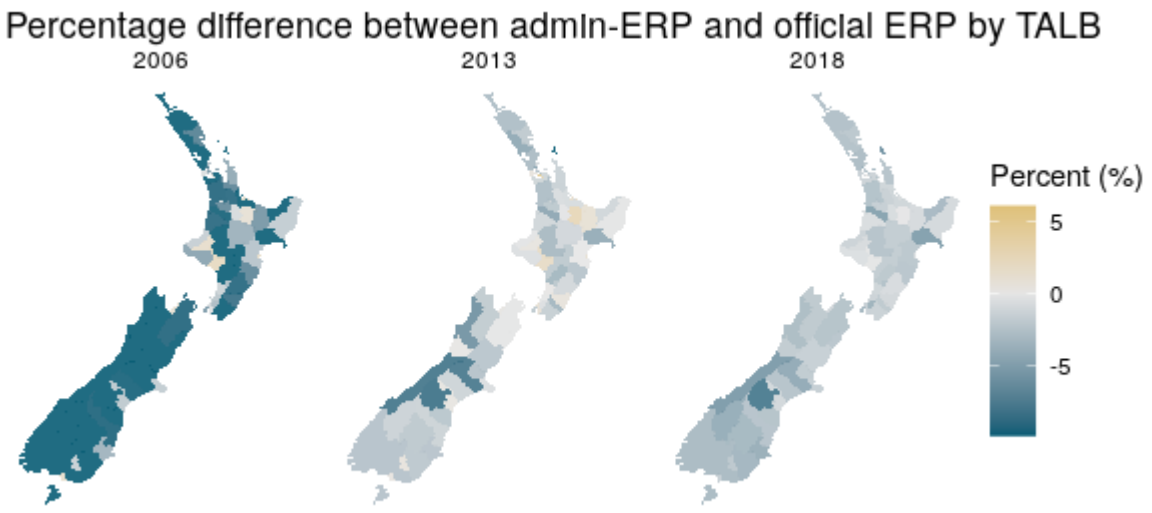
Geography

Subnational estimates depend not only on population counts, but also on the availability and accuracy of address information. In 2006, 5 percent of the APC population is missing an address ID, and so cannot be placed in a subnational geography. From 2013 onwards, an address ID is available for almost everyone.

The same trends as before can be seen in the maps in figure 4 which show the percent difference in counts for territorial authority areas and Auckland local boards (TALB) for each of the census years. For 2006, the many dark blue areas signifying higher undercount in the APC compared with the ERP are partly due to missing addresses. However, there are also some areas with overcount (ochre colour). In 2013 the distribution is more even, though some TALBs still have higher undercount, and some with overcount. By 2018, colouring is more uniform and in lighter shades of blue indicating less extreme differences between the APC and the official ERP, and a consistent undercount. In 2018 there is only a single TALB (Otara-Papatoetoe Local Board Area) with an APC population higher than the ERP (by 0.1 percent).

The 2018 APC TALB total populations do not meet the quality standard accuracy requirement of all being within ±2.5 percent for TALB of 100,000 or more people, and all within ±5 percent for TALB with fewer than 100,000 people (table A2 in appendix 2 provides more detail). In order to meet the accuracy required for official population statistics, coverage adjustment will need to account for variation in coverage across TALB, which will be due to some mis-classification of administrative address information in addition to net undercount of the total population.

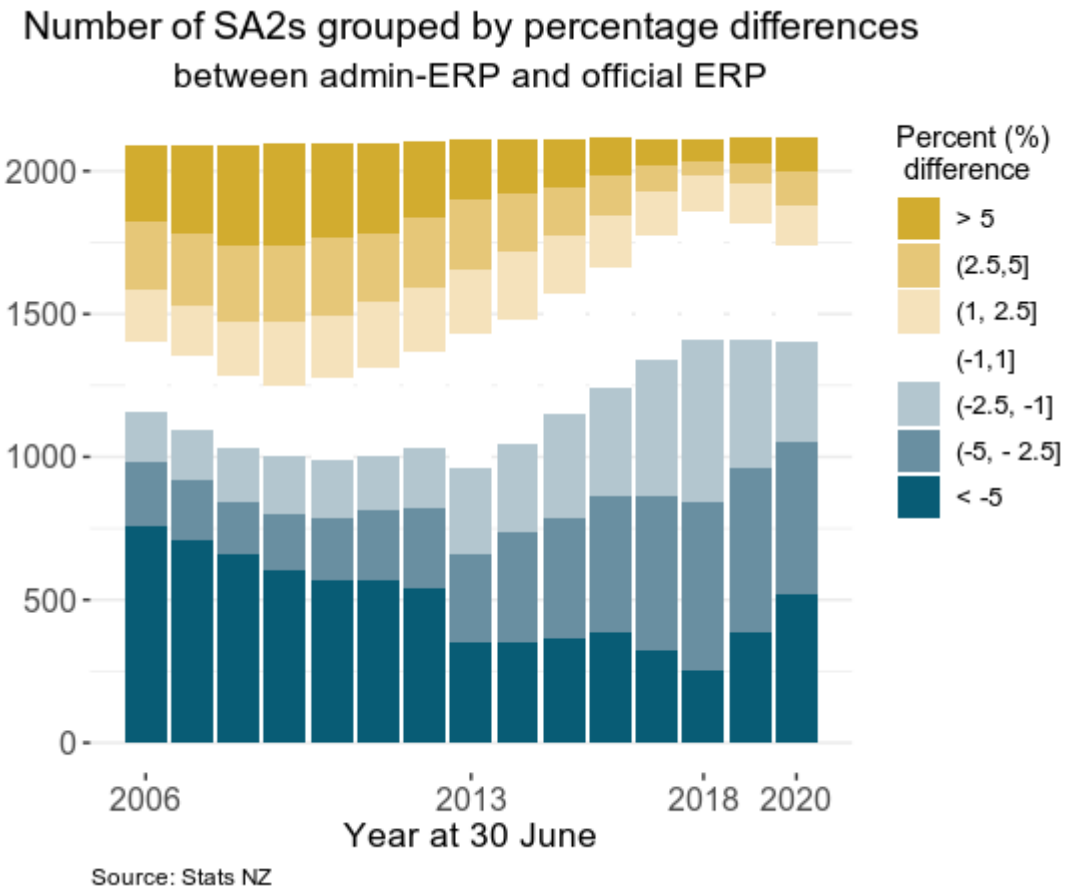
Figure 4



Source: Stats NZ

Percent differences between the admin resident population and the official ERP for the smaller statistical area 2 (SA2) geography are shown in figure 5. Again, we see the pattern of decreasing variability in the differences between the two sources at each successive census year, and a general shift towards APC undercount by 2018. The 2018 results are close to meeting the quality standard for the SA2 geography, with 85 percent of SA2s within 5 percent (the standard is at least 80 percent), and 95 percent within 10 percent (where the standard asks for all SA2s to be within 10 percent). Previous experimental admin population estimates found similar patterns of steadily improving accuracy for subnational geographies from 2007 through to 2013 (Stats NZ, 2017b).

Figure 5



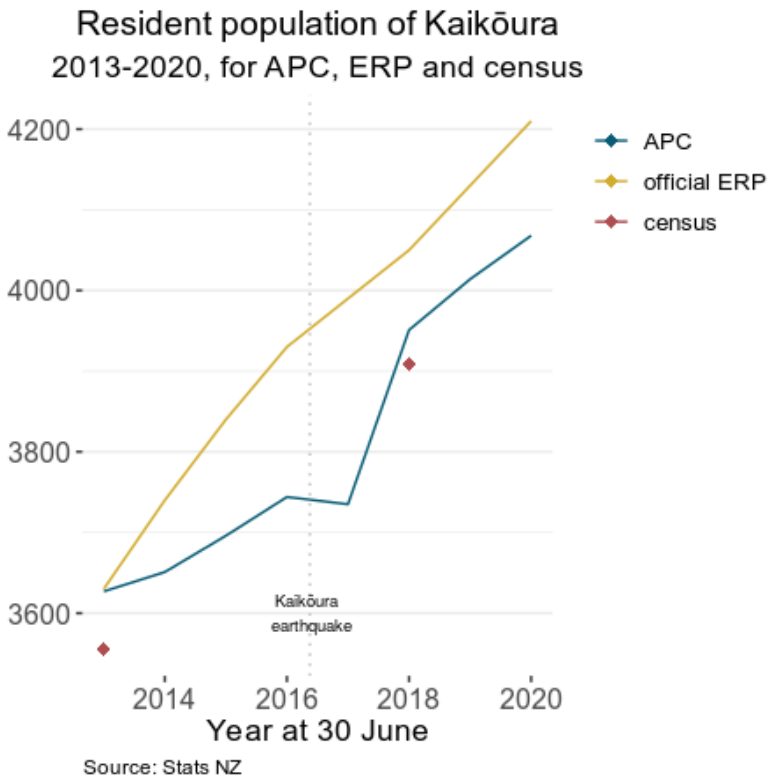
Examining change over time

The method for deriving the admin resident population means that the underlying data is a list of the population each year, and the data is longitudinal. We illustrate the possibilities with an example for Kaikōura. Figure 6 shows the population of the Kaikōura district between 2013 and 2020, comparing APC counts with the official ERP and the unadjusted census counts. This district and time period is of interest because Kaikōura was struck by an earthquake in late 2016, which caused significant disruption over the following year and more.

While the official ERP shows a linear increase in population for that period, the APC based counts show a plateauing population from 30 June 2016 to 30 June 2017, followed by a sharp increase. In the absence of independent counts, it is difficult to decide which measure is more accurate.

It is beyond the scope of this work to further analyse this situation, but we provide the example to highlight that the APC unit-record data available within the IDI enables analysis not just for aggregate counts, but at an individual level. One could, for example, look at the attributes of the cohort of people who have moved away after 2016 and compare them to the cohort who moved to Kaikōura subsequently, or determine the size of the overlap between these two groups, thus answering the question whether people moved away only temporarily or whether an exchange of population occurred.

Figure 6



Ethnicity

Ethnicity is a core demographic variable for describing the New Zealand population. The official ERP published for each census year includes estimates for level 1 ethnic groups, including by subnational area. Official estimates of the national Māori population are updated annually between censuses, while the five-yearly census provides level 4 ethnic information on a unit-record level.

Experimental population estimates for level 1 and level 2 ethnicity were first released in 2018 (Stats NZ, 2018). The APC uses the same method for deriving ethnicity, with the additional use of 2013 Census ethnicity where no administrative data is available. Coverage of ethnicity at levels 1 and 2 of the classification in the APC is close to 100 percent, ranging from 98.8 percent in 2020, to 99.5 percent in 2013.

Level 1 ethnic group estimates from the APC are compared with the official ERP for the census years (table 4). Overall, the APC exhibits very similar patterns to the official ERP. Differences in counts partially reflect differences in the total population, and level 1 ethnic group as a proportion of the total population may be a better reflection of the accuracy of administrative ethnicity derivation compared with the official ERP estimates. The differences in proportions for each level 1 ethnic group are nearly all within 1 percent.

Table 4

Level 1 ethnicities, comparison between APC, and official ERP					
Year at 30 June	Ethnicity	Count		Percent relative to respective base population (APC/official ERP)	
		APC	Official ERP	APC	Official ERP
2006	European or Other	3,106,572	3,213,300	74.7	76.8
	Māori	653,511	624,300	15.7	14.9
	Pacific	307,653	301,600	7.4	7.2
	Asian	354,762	404,400	8.5	9.7
	MELAA	32,706	38,600	0.8	0.9
2013	European or Other	3,217,989	3,312,100	72.6	74.6
	Māori	707,724	692,300	16.0	15.6
	Pacific	366,972	344,400	8.3	7.8
	Asian	503,301	541,300	11.4	12.2
	MELAA	54,552	53,100	1.2	1.2
2018	European or Other	3,337,053	3,441,700	69.4	70.2
	Māori	771,369	816,500	16.0	16.7
	Pacific	404,859	407,700	8.4	8.3
	Asian	706,893	770,600	14.7	15.7
	MELAA	74,853	77,000	1.6	1.6

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

MELAA – Middle Eastern/Latin American/African

Source: Stats NZ

Māori population estimates

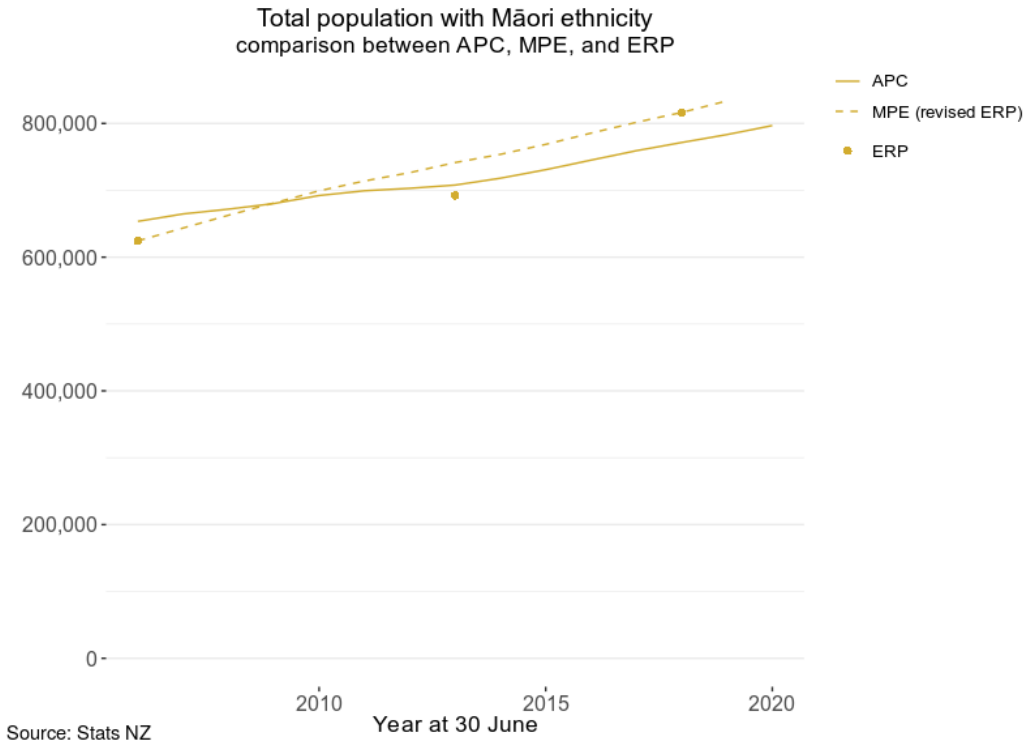
Māori population estimates are national-level annual estimates of the resident population of the Māori ethnic group by age and sex. The estimates are obtained by updating the Māori ethnic group ERP for the base census year for population change during the ensuing period. When Māori population estimates were revised to account for the new ERP at 30 June 2018 it became apparent that there was an under-estimation of the Māori ethnic group ERP at 30 June 2013. The Māori population estimates (MPE) series was then revised back to 2006 (Stats NZ, 2020a). There is now a difference of 49,200 between the revised MPE and the Māori ethnic group ERP for 2013.

The APC provides another time series estimate of the Māori ethnic population. Figure 7 shows the comparison between the APC Māori ethnic group population, the MPE, and the census-year ERP. While the APC counts are below the MPE by about 25,000 for most of the time period, this comparison is between adjusted estimates of the MPE and the official ERP and unadjusted counts for the APC. Nevertheless, the APC counts of 2013 are higher than the official ERP adding further evidence to them being too low.

We do not have an explanation for the pattern between 2006 and 2013 where the APC population is higher than the ERP.

Table A3, in appendix 2, provides the counts underlying the graph.

Figure 7



Attributes: Quality measures and analysis

This section provides information on the quality of the three attribute variables: Māori descent, birthplace, and years since arrival in New Zealand. This includes coverage of the APC population over time, the proportions of data sources used, and formal quality metrics for each variable, for the 2018 year.

Further analysis compares the APC variable distributions with the 2018 Census distributions, and examines situations where quality issues may be affecting results.

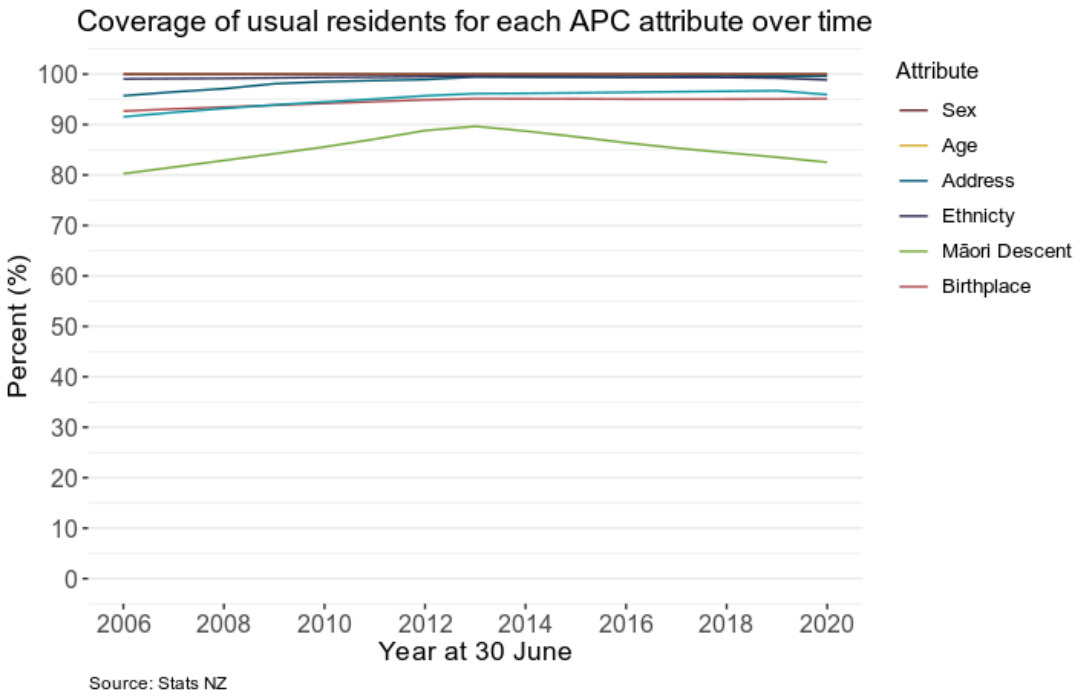
Coverage

Figure 8 shows the level of coverage for each of the APC variables from 2006 through to 2020. Coverage is very high for the main demographic variables. The IDI standard variables age and sex have 100 percent coverage of the APC population throughout the period, ethnicity is close to 100 percent and address reaches close to 100 percent coverage by 2013. Coverage for birthplace and years since arrival in New Zealand both trend slightly upwards over time and plateau at around 95 and 97 percent respectively from 2013. Missing data is largely due to the lack of administrative information for migrants to New Zealand before 1997 when electronic border movements data was first available.

Māori descent has a unique shape which peaks at 90 percent coverage in 2013 and decreases gradually both before and after 2013, down to 83 percent in 2020. This pattern is due to the lack of

administrative sources before 1995 when birth registrations first collected Māori descent. This affects those born in New Zealand before 1995 and those born overseas, unless they have subsequently had children in New Zealand. Consequently, we rely on the 2013 Census for these groups. The use of 2013 Census data significantly improves coverage, but still does not provide information for the whole population, even for the 2013 year. In addition, the 2013 Census will not include information for migrants living in New Zealand before 2013, and who have left the country by the time of the 2013 Census, nor will it include migrants arriving after the 2013 Census. Some migrants will be captured in the DIA births data if they become parents and register their child as born in New Zealand, but the proportion missing Māori descent will continue to increase as we move further away from 2013.

Figure 8



Quality metrics for the APC

Quality metrics for the 2018 year have been derived using the new Statistics Austria method for three variables: Māori descent, birthplace, and years since arrival in New Zealand.

Input data quality

The data sources contributing to each variable were first evaluated separately. Input data quality was calculated using the proportion of consistent values between the admin source and 2018 Census survey responses, using only individuals with valid responses in both sources in the linked data. Scores closer to 1 indicate a higher degree of consistency. Table 5 present the quality of each data source for each variable, along with the percentage contribution of each source to a given variable. This table is the same format as used by the 2018 Census for their output quality Metric 1 score.

Table 5

Quality of each data source by variable, for 2018				
Variable	Data source	Proportion of responses (percent)	Input data quality rating	Derivation of rating score
Māori descent	Birth Register – Child	23	0.91	Exact agreement
	Birth Register – Parent	25	0.94	
	2013 Census	37	0.95	
	Missing	15	0	
	Total	100		
Birthplace	Birth Register	69	1	Exact agreement
	MBIE Border Movements	22	0.92	
	2013 Census	5	0.99	
	Missing	4	0	
	Total	100		
Years since arrival in New Zealand	Customs border movements	66	0.91	Within one year
	2013 Census	31	0.94	
	Missing	3	0	
	Total	100		
Source: Stats NZ				

Māori descent and birthplace are derived using data sources that have a high degree of consistency with 2018 Census survey responses, indicating that the administrative data can generally be relied on as a suitable measure for both variables. While we cannot discount the possibility of some error in the 2018 Census responses, the high degree of consistency also suggests that the census responses are mostly very accurate.

The number of years since arrival in New Zealand is calculated from the month and year of arrival in New Zealand. Year and month of arrival are expected to be of higher quality in administrative records than for census responses. Border crossing data provides exact dates of travel, and the 12/16-month rule applies a formal definition of the first arrival date for a new migrant. Census responses may be less accurate for several reasons: they rely on respondent recall, respondent's interpretation of the time of their first arrival, and the month of arrival can be missing. An incorrectly stated or missing month of arrival could result in the years since arrival being derived as one year too high or one year too low. To account for the limitations of census responses, the input quality rating for years since arrival in New Zealand is calculated as being agreement within one year. With this definition, we also see high consistency for both the border movements and 2013 Census data.

Output data quality

Table 6 presents the output quality ratings for these three variables for the year 2018. Quality ratings for a variable will change based on the:

- number of data sources
- quality of the data sources
- degree of consistency between administrative sources
- amount of missingness in the APC variable.

Table 6

Output quality ratings for Māori descent, birthplace, and years since arrival in New Zealand, for 2018

Variable	Number of data sources	Missingness (percent)	Non-missing quality rating	Output quality rating
Māori descent	3	15	0.95	0.80
Birthplace	3	4	0.98	0.92
Years since arrival in New Zealand	2	3	0.93	0.89

Source: Stats NZ

From these results we see that the average quality ratings for non-missing data is high, and that level of missing data is the main contributor to the overall output quality rating.

To demonstrate the ability to report on the quality of variables at different breakdowns, table 7 presents the quality of each variable for different subsets of the target population. Again, the differences are mainly due to the level of missing data rather than any differences in the quality of the administrative sources. Patterns of missing data are explored further in the following sections.

Table 7

Output quality ratings for Māori descent, birthplace, and years since arrival in New Zealand for different subsets of the population, for 2018

Variable	Subcategory	Missing (percent)	Non-missing quality rating	Output quality rating
Māori descent	Birthplace			
	Born in New Zealand	7	0.96	0.88
	Born overseas	32	0.95	0.66
Birthplace	Age Group			
	0–14 years	2	0.99	0.97
	15–29 years	2	0.98	0.96
	30–64 years	5	0.98	0.92
	65+ years	14	0.98	0.79
Years since arrival in New Zealand	Birthplace			
	United Kingdom	10	0.94	0.85
	Australia	24	0.94	0.83
	India	3	0.93	0.88
	China	4	0.92	0.88
	South Africa	3	0.95	0.88

Source: Stats NZ

Distributions

We now compare 2018 results for the APC against the 2018 Census distributions. Total populations are slightly different since the APC includes more people than the census, and there are missing data in both the census and APC. We present distributions as proportions of each category of those with 'stated' values to remove the effect of population differences.

Consistent distributions provide confidence that both methods are measuring the same concepts, while differences may provide insight into potential quality issues in either census responses or the administrative derivations. We see very good consistency between the APC and the 2018 Census distributions for each of the three variables.

Māori descent

Table 8 shows the counts for Māori descent in the APC and the 2018 Census, and the percent distribution of Māori descent calculated from the stated values. The relatively large number of people with missing data in the APC (nearly 750,000 people) means that the total number of people with Māori descent is clearly too low. The distributions of the total with stated responses are similar to the census, with the APC Māori descent = Yes proportion slightly higher at 19.4 percent compared with the 2018 Census value of 18.5 percent. For comparison, the APC proportion is very close to the official 2018 Māori descent estimated resident population of 19.2 percent (941,200/4,900,600).[2] However, given the large number of missing responses, the APC proportional distribution may simply be a fortuitous result of the particular groups with missing data.

Table 8

Distribution of Māori descent for 2018, APC, and 2018 Census

Māori descent category	APC (30 June 2018)	2018 Census (6 March 2018)	APC (30 June 2018)	2018 Census (6 March 2018)
	Count		Percent of total stated	
Yes	788,829	869,850	19.4	18.5
No	3,176,337	3,715,050	78.2	79.0
Don't know	94,734	114,855	2.3	2.4
Total stated	4,059,900	4,699,755	100	100
Missing	749,049	…	…	…
Total people	4,808,949	4,699,755	…	…

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Symbol: … not applicable

Source: Stats NZ

Missing Māori descent data

We now examine patterns of missing Māori descent over time by broad age group, for New Zealand-born and overseas-born (figure 9). Similar patterns for those who are born in New Zealand were also

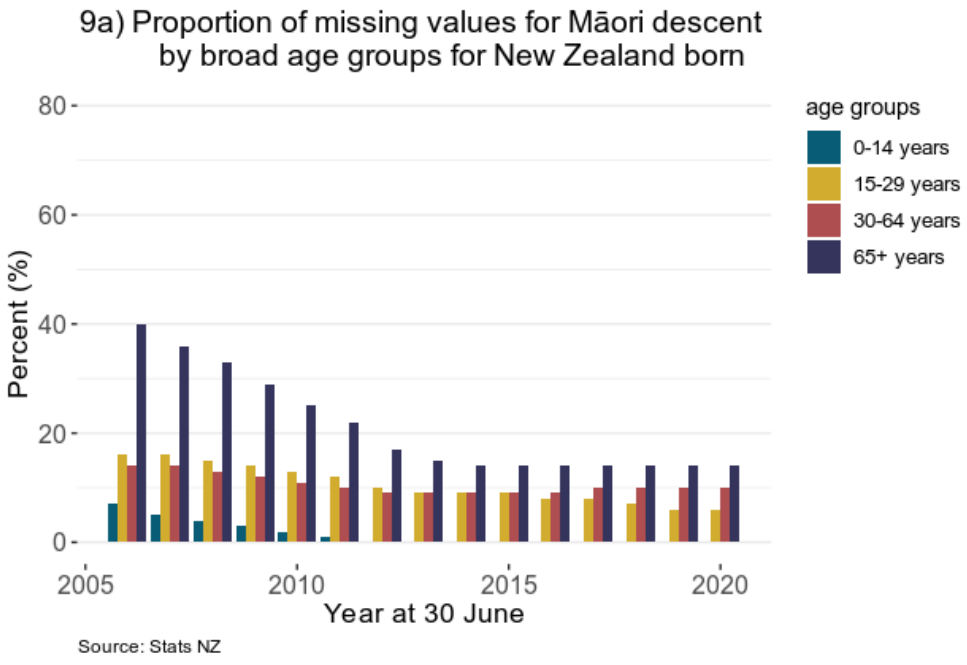
[2] See Estimated resident population (2018-base): At 30 June 2018

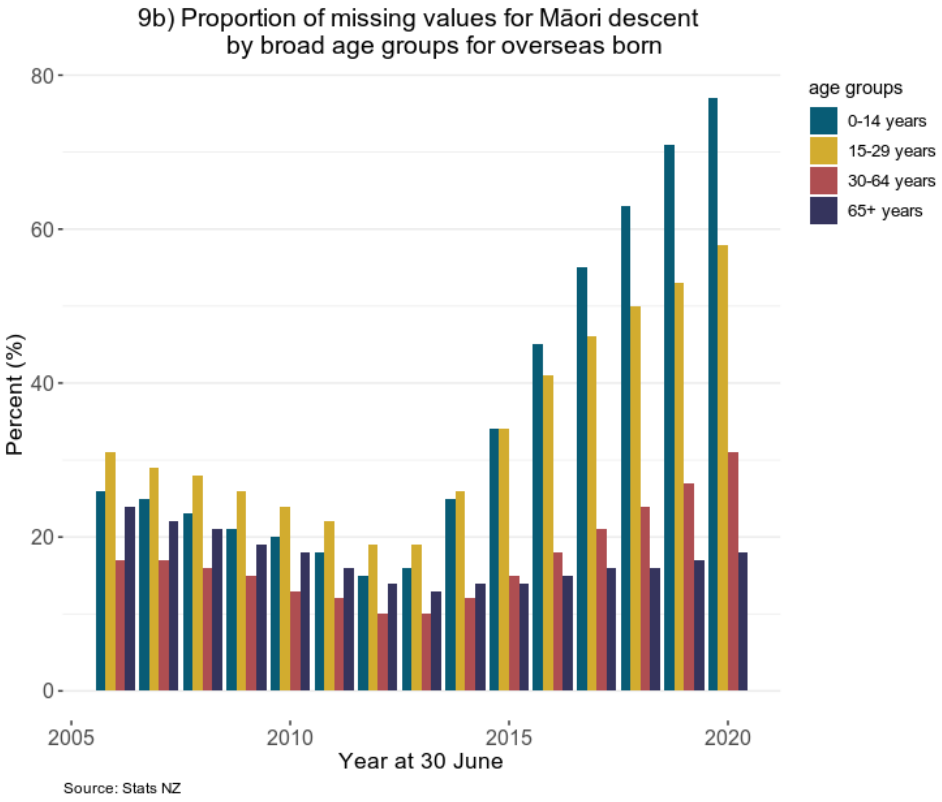
found for those with Māori ethnicity. The earlier period from 2006 shows the highest levels of missing data, decreasing to 2013, when the levels stabilise. Elderly people (65+) have the highest level of missing data decreasing from a maximum of around 40 percent in 2006, and reducing to around 15 percent from 2013 onwards. There is almost no missing data for children born in New Zealand, as expected.

The patterns for those born overseas are quite different. Overall levels of missingness are higher, and while there is also a reduction from 2006 to 2013 when 2013 Census data has the most impact, from 2013 onwards the rate of missing data increases every year. This is especially true for younger groups, reaching 80 percent missing for children aged 0–14 in 2020.

The overseas born population has very low rates of people with Māori descent. Of people stating they were of Māori descent, in the 2013 Census 2.1 percent were born overseas, and in the 2018 Census 2.4 percent were born overseas. Therefore, we can expect that most of those with missing data who were born overseas will report Māori descent = No, while those born in New Zealand will be a mix of responses.

Figure 9





In summary, with the sources currently available for the APC, undercoverage of Māori descent is significant, it varies by age, is higher for those born overseas, and is increasing over time due to the lack of sources for many migrants to New Zealand.

The electoral roll also collects Māori descent and the combination of birth registrations and electoral roll data (those 18 years and older who enrol to vote) would provide high coverage for much of the population. However, the Electoral Act 1993 places restrictions on the use of electoral roll data, meaning it cannot be used as a source for the APC (or the 2023 Census) at present. The Data and Statistics Bill proposes amendments to a number of other Acts to remove legislative barriers that inadvertently restrict or prohibit the provision of data to Stats NZ. For example, proposed amendments to the Electoral Act 1993 would enable Stats NZ to access electoral data.

Death registrations are another source of Māori descent information that could be considered. Even with additional sources, imputation of remaining missing values will be needed – as is the case with the census now. The patterns of missing data are likely to be different from those usually seen in census non-response. Those missing from the administrative data are likely to be migrants under the age of 18 years, and migrant adults and older New Zealand-born people who do not enrol to vote.

Birthplace

Counts for New Zealand-born and overseas-born in the APC and the 2018 Census, and the percent distribution calculated from the stated values are shown in table 9. The distributions are very similar, showing a large proportion of New Zealand residents are born overseas: 27.4 percent in the 2018 Census and 27.6 percent in the APC in 2018. The APC counts are somewhat lower than the 2018 Census counts due to the five percent missing data in the APC.

Table 9

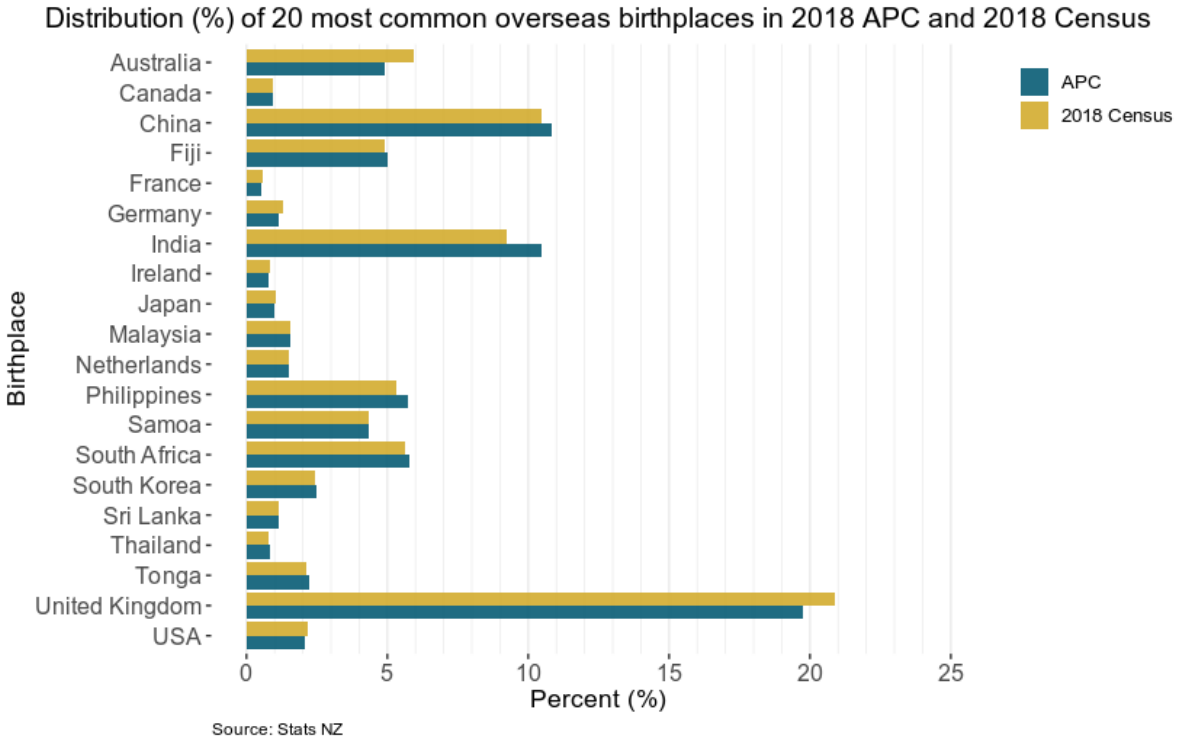
Birthplace results in the APC in 2018 and 2018 Census				
Overseas resident indicator	APC (30 June 2018)	2018 Census (6 March 2018)	APC (30 June 2018)	2018 Census (6 March 2018)
	Count		Percent of total stated	
New Zealand-born	3,307,341	3,370,122	72.4	72.6
Overseas-born	1,263,417	1,271,775	27.6	27.4
Total stated	4,570,758	4,641,897	100	100
Missing	238,191	57,858		
Total people	4,808,949	4,699,755		

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

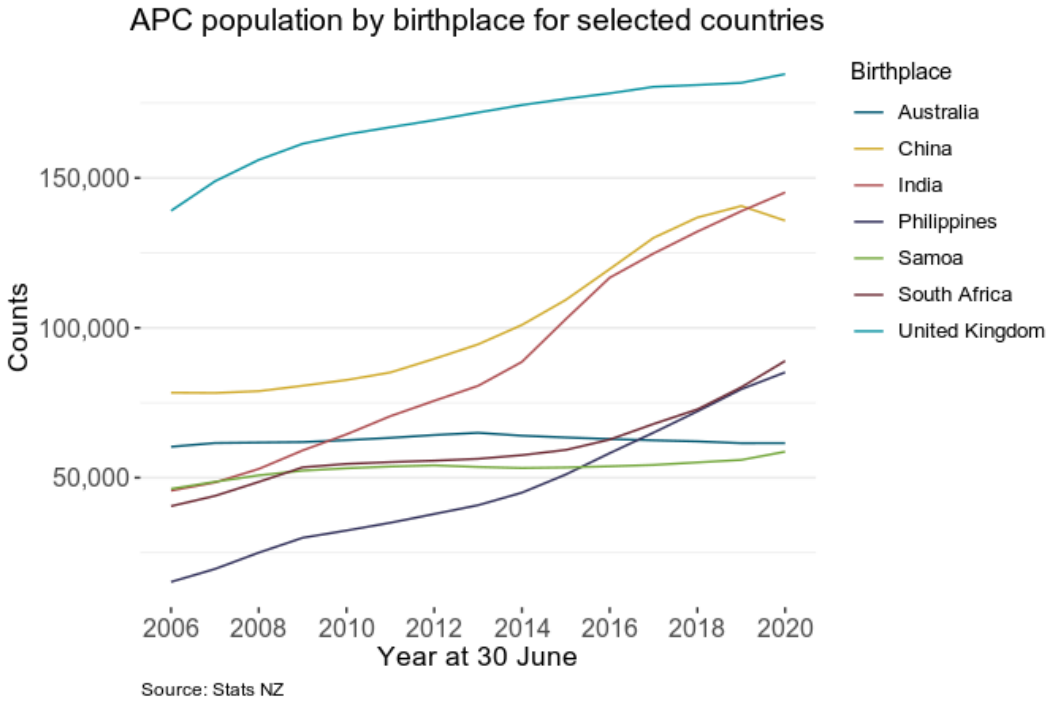
Over 200 different response categories are collected for overseas birthplace. The vast majority of these (>90 percent) each account for less than 1 percent of the total overseas responses. For simplicity, only the distributions of the top 20 countries are shown (figure 10). The distributions are broadly similar, with most having less than half a percentage point difference between APC and the 2018 Census. India and Australia have the largest difference. The APC reports 1.2 percent more individuals from India than the 2018 Census. In contrast for Australia and the United Kingdom, the APC reports 1.0 percent and 1.1 percent fewer individuals than the 2018 Census. Table A4 in appendix 2 provides the counts underlying the graph.

Figure 10



The APC also allows a fine-grained view of annual changes in birthplace compared with the current counts which are only available in census years. As one example of the possibilities, figure 11 shows the number of migrants living in New Zealand from selected countries. There was a rapid increase from countries such as China, India, and the Philippines, where much of the recent growth occurred in the period between the 2013 and 2018 Censuses.

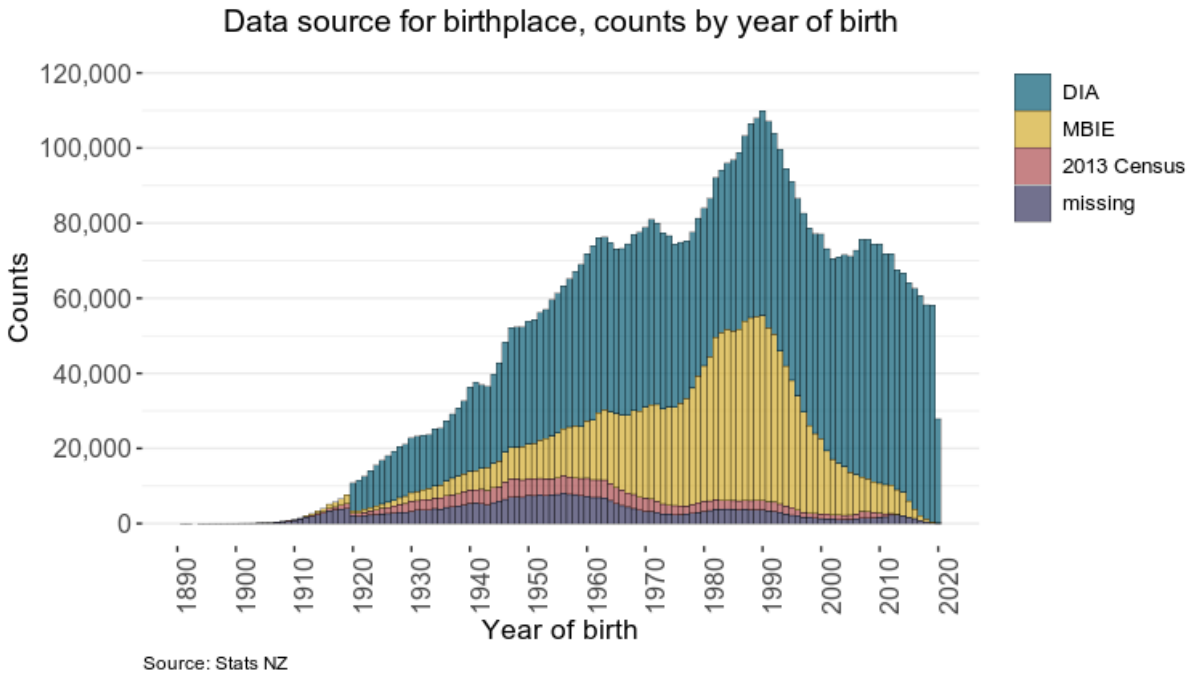
Figure 11



Missing birthplace data

The two administrative sources, DIA birth registrations for New Zealand-born and MBIE border movements data for overseas-born, provide a birthplace for 96 percent of the APC population in 2018. Figure 12 shows the distributions of data source by year of birth. Those born between 1930 and 1970 rely the most on 2013 Census data, or are missing birthplace information, at least partly due to the lack of historical border movements data. For individuals born prior to 1920 there is a dramatic shift in data source since DIA birth registrations are not available. While some data is available from MBIE and the 2013 Census, the rate of missing values remains high for this group.

Figure 12



For those born before 1920, New Zealand birthplace values are accepted as valid from the 2013 Census and MBIE. This resulted in 11,700 individuals born prior to 1920 begin assigned as born in New Zealand. The assumption that DIA birth registrations are complete after 1920 resulted in removing claims of a New Zealand birthplace from other sources. Some of these individuals without a DIA record also have an overseas birthplace in the border movements data, but for around 194,000 people who do not, their birthplace was considered to be missing. It is likely that some of these individuals with missing values in the APC are born in New Zealand and their DIA birth record is either missing or there was an error in IDI linking.

United Kingdom classification

MBIE data does not provide countries within the United Kingdom, whereas census respondents do generally provide a country within the United Kingdom as their birthplace. For example, in the 2018 Census only 14,601 (5 percent) of individuals born in the United Kingdom were coded to the broader United Kingdom (nfd) category. While the APC does at present include United Kingdom country information from the 2013 Census, reliance on MBIE data will increase over time and will result in a loss of detailed information compared with census responses.

Years since arrival in New Zealand

Years since arrival is calculated from the time a new migrant first arrives in New Zealand to live. We compare the total numbers of overseas-born migrants by year and month of arrival in New Zealand for the APC population in 2018 and as reported in the 2018 Census. While exact date of arrival is available from the admin migration data, here we have excluded census responses which are missing month of arrival. Figure 13 shows counts for those who arrived in New Zealand from 2000 to 2018, where the APC source is dominated by Customs border movements data. Both sources reflect the steep rise in migration during this period, and the seasonal patterns are very similar. The APC counts are higher than the 2018 Census, as expected due to the exclusion of census data that is missing month of arrival. Figure 14 shows year and month of arrival from 1960 through to 2018. For the earlier years from 1960 through to 2000, the APC is reliant on 2013 Census data and we see that for this period the APC and 2018 Census pattern of counts are very similar.

Figure 13

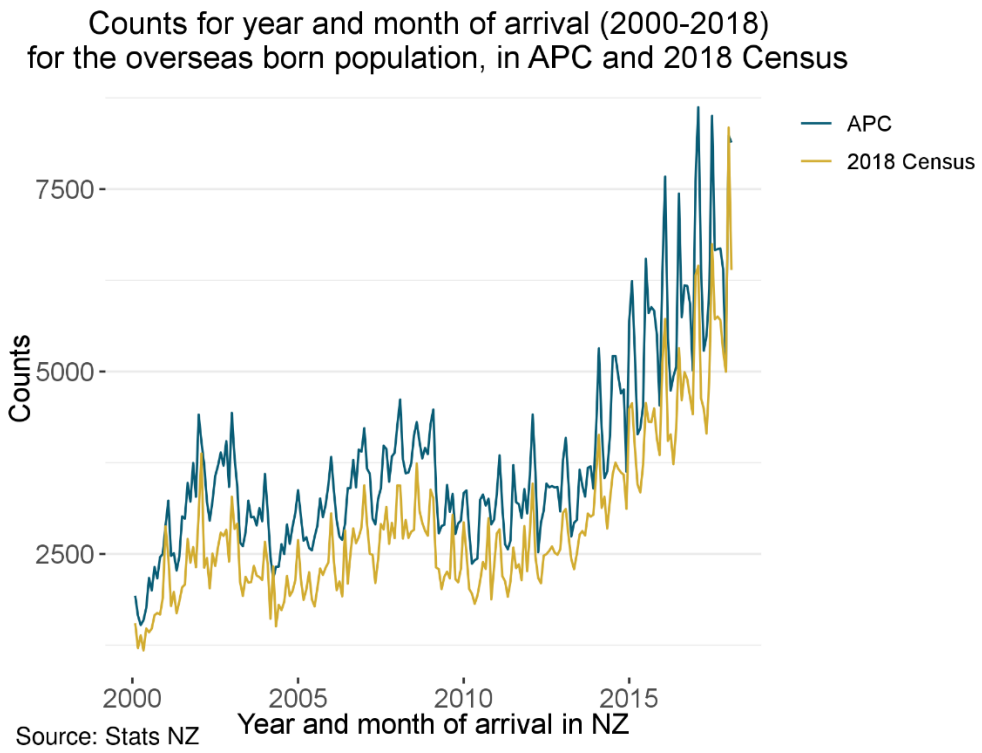
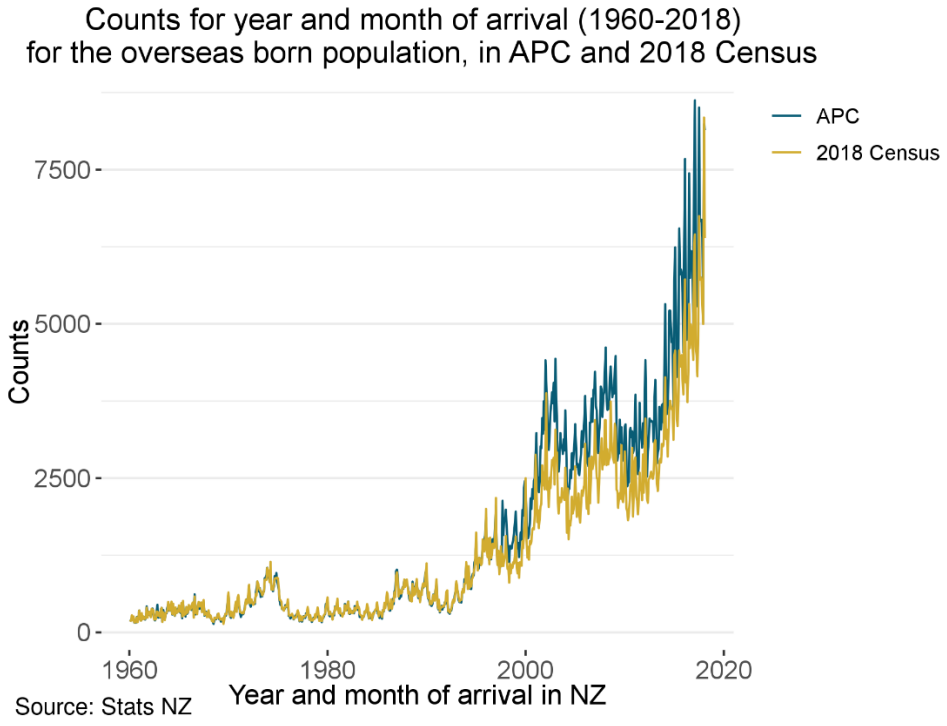
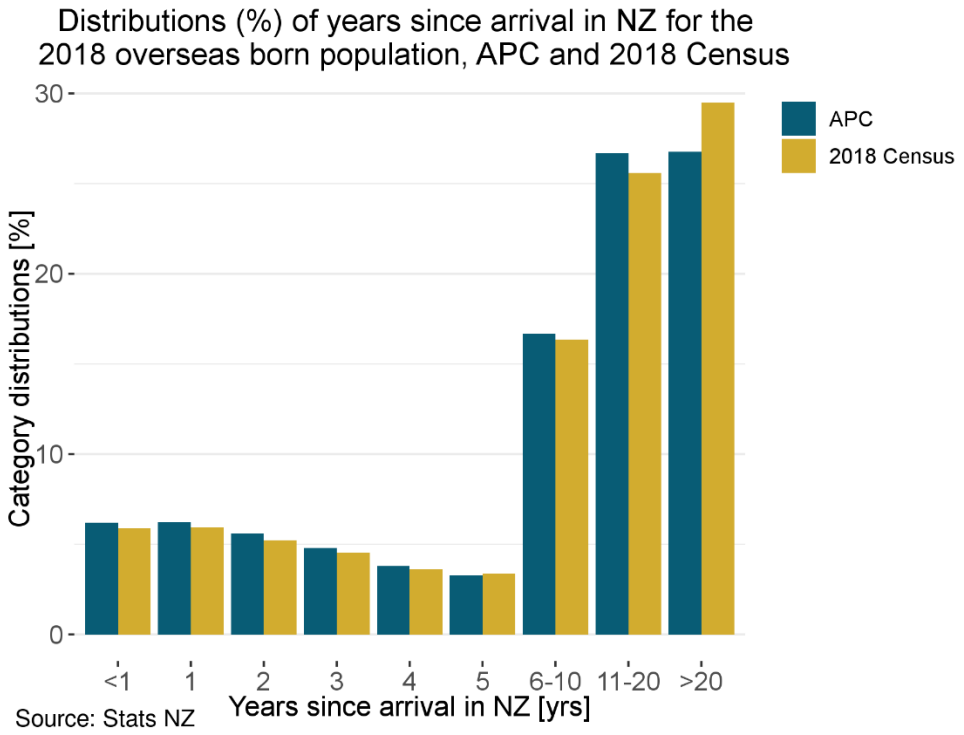


Figure 14



The distribution of years since arrival in New Zealand as a proportion of the 'total stated' is shown in figure 15, and compares APC data for 2018 with the 2018 Census. The 2018 Census derived variable estimates the number of years since arrival for records missing the month of arrival, and so could be up to one year different from the true value for these cases. Distributions are very close, differing by at most 1 percent, except for the 20+ years category where the 2018 Census is 2 percent higher than the APC. This is the only category for which the APC relies solely on 2013 Census data, and the 2013 Census will not include everyone who is in this category in 2018. Counts underlying the figure are provided in table A5 (appendix 2).

Figure 15



Discussion

The experimental administrative population census derives census-type information from linked administrative sources. We aim to demonstrate the information available currently and to provide a focus for discussion about the benefits and limitations associated with an admin-based census.

The APC is part of the ongoing census transformation programme looking at the potential for a future census based on administrative data supported by sample surveys. The long timeframes embedded in the five-yearly census cycle make it difficult to gain a sense of reality for future changes that might seem a long way off. Experimental releases of census-type information based on administrative data provide a stepping-stone to help make the idea of an admin-based census more concrete and to give an opportunity for customers to influence further development based on their use of the data.

The APC is based on the administrative New Zealand resident population and supplements this with more characteristics of individuals derived from administrative data. The first iteration released in August 2021 is an annual series, from 2006 to 2020, focusing on demographic and identity variables: age, sex, geography, ethnicity, Māori descent, birthplace, and years since arrival in New Zealand.

The APC is not a full census, but it is the start of one. More topics such as work, income, and qualifications will be added in the 2022 and 2023 releases. The APC has been compiled with a small fraction of the time and resources needed for a traditional full-field enumeration census. The APC also in some sense offers more than the traditional census. Key features include the merging of census information with official population estimates; so that under this approach, a future admin-based census would have only one estimate of the resident population. Other advantages over a traditional census are that results can be compiled annually, and the underlying unit-record data is inherently longitudinal.

While the APC is clearly not a substitute for recent censuses, it does offer new information. The first release of the APC may be of particular interest to demographers, local councils, and others interested in annual changes in specific populations, for example, Level 1 and Level 2 ethnicities, and birthplaces of migrants. Output tables for these variables are available each year nationally and for TALB and SA2 subnational geographies. The longitudinal nature of the unit-record data allows exploration of gross flows, for example of the number and types of people who move into, and out of different areas.

This paper describes the data sources and methods used for constructing the APC and looks at aspects of data quality. The results for this first release of the APC show good agreement with the expected population patterns and distributions of census variables at high levels of aggregation.

There is a small net undercount of the total population compared with official population estimates, although the APC population is always larger than the corresponding census counts. For the APC released in 2021, there is no coverage adjustment of the administrative population to account for undercoverage and overcoverage. Development of statistical methods for coverage adjustment is in progress and (as with the current census) will be required to provide population estimates of sufficient accuracy for the demands of official population statistics.

A new quality measure has been trialled for three variables: Māori descent, birthplace, and years since arrival in New Zealand. This assessment shows the high quality of the contributing administrative data sources, and that variation in quality for each variable is mainly a factor of missing data. For the variables considered here, missing data is largely driven by migrants who arrived in New Zealand before digitised border movements data was available in 1997, and who do

not appear in the 2013 Census. However, there is no imputation for missing values in the APC yet. Further work is planned to address these limitations through reducing the level of missing data and developing imputation methods. Future access to Electoral Roll data would be essential to fill gaps in Māori descent information.

Administrative data of suitable quality is not available for all census variables. Iwi affiliation data is a critical requirement for iwi-Māori and for any future census model. Currently there is a lack of suitable alternative sources for this key population indicator. While not appropriate for iwi affiliation, for some other census topics that have limited information from administrative data, such as language spoken and activity limitations, a sample survey will be required as part of any future census based largely on administrative data.

Administrative data is now purposefully included in the 2023 Census combined census model by design. However, administrative data is still very much a complementary data source, and the success of the 2023 Census depends on achieving high response rates to the field enumeration. No decisions have been made on the future of census beyond 2023, and considerations will go beyond the availability of suitable data and statistical methods.

Conclusion

The APC demonstrates the substantial promise of the vision of annual population and census data derived from information that has already been provided to government. Some issues are yet to be addressed and can be worked on for future iterations. Improvements can be made at the margins to the methods for deriving the population and variables, and there is a clear need for additional sources to fill a significant information gap for Māori descent. The patterns of missing data are different from those encountered in a survey-based census, but overall, they are no larger than those that occur in a traditional census. Indeed, one of the difficulties in determining the quality of administrative-derived data, is accounting for uncertainty in the census and official population estimates that are used as a comparison. Statistical methods for coverage adjustment and imputation are being developed to adjust for missing data as they have been for the traditional census.

We are excited to provide data users with a tangible example of the possibilities that administrative data might provide. The APC data can be accessed through output data tables and a graphical interface on Stats NZ's website, and anonymised unit-record data is available within the IDI. We hope this data will be a useful resource that will increase understanding of the potential of the administrative data, and we look forward to feedback from customers. The insights gained by those using the data will help to guide further development.

References

Asamer, E-M, Astleithner, F, Cetkovic, P, Humer, S, Lenk, M, Moser, M, & Rechta, H (2016). Quality assessment for register-based statistics - Results for the Austrian census 2011. *Austrian Journal of Statistics*, 45(2), 3–14.

Beynon, M, Curry, B, & Morgan, P (2000). The Dempster-Shafer theory of evidence: An alternative approach to multicriteria decision modelling. *The International Journal of Management Science*, 28, 37-50.

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, 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.

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

Gibb, S, Bycroft, C, & Matheson-Dunning, N (2016). Identifying the New Zealand resident population in the Integrated Data Infrastructure (IDI). Retrieved from www.stats.govt.nz.

McNally, J, & Bycroft, C (2015). Quality standards for population statistics: Accuracy requirements for future census models. 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.

Stats NZ (2013). Estimated resident population 2013: Data sources and methods. Retrieved from www.stats.govt.nz.

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

Stats NZ (2015a). IDI Data Dictionary: 2013 Census data (November 2015 edition). Available on request from info@stats.govt.nz.

Stats NZ (2015b). IDI Data Dictionary: Immigration data (July 2015 edition). Available on request from info@stats.govt.nz.

Stats NZ (2015c). IDI Data Dictionary: Life event data (July 2015 edition). Available on request from info@stats.govt.nz.

Stats NZ (2016a). Experimental population estimates from linked administrative data: methods and results. Retrieved from www.stats.govt.nz.

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

Stats NZ (2016c). IDI Data Dictionary: Immigration data dictionary (July 2016 edition). Available on request from info@stats.govt.nz.

Stats NZ (2017a). Defining migrants using travel histories and the '12/16-month rule'. Retrieved from www.stats.govt.nz.

Stats NZ (2017b). 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 (2019). Overview of statistical methods for adding admin records to the 2018 Census dataset. Retrieved from www.stats.govt.nz.

Stats NZ (2020a). Māori ethnic group population estimates 2006–18: Methods and results. Retrieved from www.stats.govt.nz.

Stats NZ (2020b). Post-enumeration Survey 2018: Methods and results. Retrieved from www.stats.govt.nz.

Stats NZ (2021). Māori population under-estimation in 2013 – analysis and findings. Forthcoming.

Appendix 1: Detail to support data sources

Table A1

Detail of data sources for Māori descent, birthplace, and years since arrival New Zealand variables					
Data sources	Record type	Availability and frequency	Metadata information available	APC variables using the data source	References: Census transformation research
DIA births	Registering a life event	Since 1920. Updated in IDI quarterly	DIA's Notification of birth for registration IDI metadata (Stats NZ, 2015c)	Māori descent, Birthplace, Years since arrival in NZ	Bycroft et al (2021) Bycroft et al (2016) Gath and Das (2019)
MBIE border movement identities	Visas applications NZ border crossings	Since 1997. Updated in IDI every 6 months	IDI metadata on immigration tables (Stats NZ, 2015b, 2016c)	Birthplace	Bycroft et al (2021) Gath and Das (2019)
Customs border movement records	NZ border crossings. 12/16-month resident status indicator	Since 1997, ongoing collection	Stats NZ 12/16-month rule	Years since arrival in NZ	Gath and Das (2019)
2013 Census	Survey	Snapshot in 2013	Census outputs described in DataInfo+ IDI metadata (Stats NZ, 2015a)	Māori descent, Birthplace, Years since arrival in NZ	Bycroft et al (2021) Bycroft et al (2016) Gath and Das (2019)

Appendix 2: Data tables

Quality standards

Table A2

Comparison with the quality standards from McNally & Bycroft (2015)

Geographic area	Population unit	Level of error (within ± percent)	Percent with level of error		APC percentage of counts compared with official ERP[2]		
					2006	2013	2018
National	Total population	0.5	100		0	**100**	0
	5-year age groups (0–4 years, 5–9 years, … 85+ years) by sex	1.5 5	90 100		42 **100**	45 92	39 **100**
Territorial authority areas & Auckland local boards *population 100,000 or more*	Total population	2.5	100		50	67	67
	By broad age group: 0–14 years 15–64 years[1] 65 years and over	 5 5 5	 100 100 100		 69 72 43	 92 89 94	 **100** 93 **100**
	All broad age groups	5	100		66	94	**100**
Territorial authority areas & Auckland local boards *population fewer than 100,000*	Total population	2.5 5	85 100		20 28	74 90	55 89
	By broad age group: 0–14 years 15–64 years[1] 65 years and over	 5 5 5	 85 70/95 80		 26 31 20	 78 90 **86**	 **98** 89 **95**
	All broad age groups	12.5	100		70	98	**100**
SA2s population of 500 or more	Total population	5 10	80 100		52 70	75 91	**85** 96

1 Ranges merged to enable comparisons.

2 Numbers in bold indicate when the quality requirements have been met.

Source: Stats NZ

Ethnicity

Table A3

Total Māori population, comparison between APC, MPE (revised ERP), and ERP						
	Counts			Percent relative to respective base population (APC/ERP)		
Year at 30 June	APC	MPE	ERP	APC	MPE	ERP
2006	653,511	624,330	624,330	15.7	14.9	14.9
2007	664,605	643,590	…	15.8	15.2	…
2008	671,745	662,690	…	15.8	15.6	…
2009	680,115	680,960	…	15.8	15.8	…
2010	692,115	699,220	…	15.9	16.1	…
2011	699,219	713,920	…	16.0	16.3	…
2012	702,903	726,440	…	16.0	16.5	…
2013	707,724	741,470	692,260	16.0	16.7	15.6
2014	718,014	753,780	…	16.0	16.7	…
2015	730,713	768,690	…	16.0	16.7	…
2016	744,987	785,500	…	16.0	16.7	…
2017	759,114	801,680	…	16.0	16.7	…
2018	771,369	816,480	816,470	16.0	16.7	16.7
2019	783,177	833,700	…	16.1	16.7	…
2020	796,692	854,790	…	16.1	16.8	…

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Symbol: … not applicable

Source: Stats NZ

Birthplace

Table A4

20 most common overseas birthplaces in the APC in 2018 and 2018 Census				
Birthplace	Counts		Percent of total born overseas	
	APC (30 June 2018)	2018 Census (6 March 2018)	APC (30 June 2018)	2018 Census (6 March 2018)
Australia	62,115	75,696	4.9	6.0
Canada	11,559	11,928	0.9	0.9
China	136,809	132,906	10.8	10.5
Fiji	63,393	62,310	5.0	4.9
France	6,831	7,593	0.5	0.6
Germany	14,253	16,605	1.1	1.3
India	132,114	117,348	10.5	9.2
Ireland	10,128	10,494	0.8	0.8
Japan	12,261	13,107	1.0	1.0
Malaysia	19,884	19,860	1.6	1.6
Netherlands	18,696	19,329	1.5	1.5
Philippines	72,138	67,632	5.7	5.3
Samoa	55,104	55,512	4.4	4.4
South Africa	72,816	71,382	5.8	5.6
Republic of Korea	31,470	30,975	2.5	2.4
Sri Lanka	14,697	14,349	1.2	1.1
Thailand	10,791	10,251	0.9	0.8
Tonga	27,879	26,856	2.2	2.1
United Kingdom	249,810	265,548	19.8	20.9
USA	26,184	27,678	2.1	2.2
Total born in New Zealand	3,307,341	3,370,122	…	…
Total born overseas	1,263,416	1,271,775	100.0	100.0
Total stated	4,570,758	4,641,897	…	…
Total	4,808,949	4,699,755	…	…

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Symbol: … not applicable

Source: Stats NZ

Years since arrival in New Zealand

Table A5

Years in New Zealand results in the APC in 2018 and 2018 Census				
YSANZ categories	Counts		Percent of total stated	
	APC (30 June 2018)	2018 Census (6 March 2018)	APC (30 June 2018)	2018 Census (6 March 2018)
<1 year	75,648	73,617	6	6
1 year	76,044	74,385	6	6
2 years	68,340	65,526	6	5
3 years	58,551	57,045	5	5
4 years	46,209	45,561	4	4
5 years	40,212	42,252	3	3
6–10 years	203,424	205,155	17	16
11–20 years	325,614	321,042	27	26
>20 years	326,844	369,516	27	29
Total stated	1,220,886	1,254,108	100	100
Missing	42,531	17,667		
Total born overseas	1,263,417	1,271,775		

Note: This data has been randomly rounded to protect confidentiality. Individual figures may not add up to totals, and values for the same data may vary in different tables.

Source: Stats NZ

Appendix 3: How the admin resident population is derived

As mentioned in the main text, the admin resident population is based on the IDI-ERP methodology from Stats NZ (2016a) and refined in Stats NZ (2017b), sometimes referred to as 'IDI-ERP v2' and 'IDI-ERP v3' respectively. Some improvements were implemented when the IDI-ERP was utilised in the 2018 Census, which are also included in the below methodology. While it would make sense to label this version as IDI-ERP v4, we prefer to refer to it as the APC resident population, so as to not introduce an artificial separation between the resident population and its attributes.

To derive the APC resident population we start with the IDI spine as described in the main text.

Inclusions: retain individuals whose presence is indicated by activity

- For ages five years and over, the spine population is restricted to those individuals who had activity in at least one of the following IDI datasets in the two years before the reference date.
    - Inland Revenue tax (employer monthly summary of tax paid at source, or annual tax return data; receipt of taxable benefit payments is included).
    - Ministry of Health (pharmaceutical prescriptions, GP enrolment and attendance, hospital admissions, non-admission hospital visits).
    - Ministry of Education (school enrolment, tertiary enrolment, or attainment).
    - Department of Corrections (custodial sentences).
    - New Zealand Customs Service (recent arrival on a long-term visa).

- For ages under five, a New Zealand birth registration or visa approval (excluding visitor or transit visas) before the reference date is sufficient for inclusion in the population. For this age group there is no additional requirement of activity in the previous two years.

Exclusions: remove non-residents at the reference date and duplicates

- For each reference date (30 June 2006–2020) linked death records are used to identify individuals with a date of death before the reference date.
- Linked migration data is used to identify individuals who were not New Zealand residents on the reference date.
    - For the years 2006–2019 people are excluded who are not resident according to the 12/16-month rule, as described in the main text.
    - For 2020 individuals are excluded who have not spend at least six months in New Zealand in the period 31 December 2019 to 31 December 2020.
- We remove records identified as duplicates caused by missed links when constructing the IDI spine. Duplicates are identified conservatively as those with identical sex, year and month of birth, and address ID. Both records must have only a single spine source, with one coming from birth registrations and one coming from Inland Revenue tax registrations. This combination provides strong evidence that both records refer to the same person and that a link has been missed. Only the record with the most additional links is kept.

Appendix 4: Applying Dempster Shafer Theory

This section outlines the process for using Dempster Shafer Theory to derive the output quality values for individual records. For a more theoretical outline we refer the reader to (Beynon et al, 2000). We continue the birthplace example presented in table 1, which is repeated as table A6 below.

Table A6

Example data used to demonstrate how Dempster Shafer Theory was applied						
Unit	In Birth Register: (= NZ born)		MBIE		Output data	
	Value	Input quality	Value	Input quality	Value	Output quality
1	NZ	0.8	NZ	0.6	NZ	0.92
2	NZ	0.8	Samoa	0.6	NZ	0.62
3	NZ	0.8	Samoa	0.6	Samoa	0.23
4	NZ	0.8	…	0.6	NZ	0.8
Symbol: … not available						

The first step is to create a frame of discernment, which are a finite set of hypotheses. The hypotheses are constructed using the information from the data sources. In the context of the assigning birthplace for our units, these include:

- Hypothesis 1: unit was born in NZ
- Hypothesis 2: unit was born in Samoa.

Each source of evidence is quantified using a basic probability assignment, which reflects the confidence of the information from the data source. This is the input quality metric for a given data source with respect to the attribute of interest (the third and fifth column in the above table).

The evidence in the hypotheses can then be combined. DST provides the terminology and framework for combining evidence. The basic probability assignment assigned to each evidence is referred to as a 'mass'; m_1 and m_2 refer to the mass associated with data source 1 (births register) and 2 (MBIE) respectively. A description of each piece of evidence/mass is provided below for unit 1:

- m_1(born in NZ) = 0.80
- m_1(not born in NZ) = 1-0.80 = 0.20
- m_2(born in NZ) = 0.60
- m_2(not born in NZ) = 1-0.60=0.40

To derive the final values different formulas are used, depending on whether the data sources agree with the final value or not.

Consistency across data sources and final value

The evidence from each data source is used to construct different scenarios that could arise in the observed data. The scenarios, for unit 1, include:

- m_3(is NZ born in both sources) $= 0.80 \times 0.60$
- m_3(is NZ born in births register and NOT NZ born in MBIE data) $= 0.80 \times 0.40$
- m_3(is NOT NZ born in births register and NZ born in MBIE data) $= 0.20 \times 0.60$
- m_3(is not NZ born in any source) $= 0.20 \times 0.40$

All scenarios where a unit could be NZ born are added together to calculate a belief (*Bel*) in the final value that the unit is NZ born.

$$
\begin{aligned}
&Bel(\text{Unit 1 is born in NZ}) \\
&\quad = m_3(\text{is born in NZ in both sources}) \\
&\quad + m_3(\text{is born in NZ births register and NOT born in NZ in MBIE data}) \\
&\quad + m_3(\text{is NOT born in NZ in births register and born in NZ in MBIE data}) \\
&\quad = (0.80 \times 0.60) + (0.80 \times 0.40) + (0.20 \times 0.60) = 0.92
\end{aligned}
$$

Discrepancy between data sources and final value

Where there is a discrepancy between one data source and the final value selected, Dempster's rule of combination is used, which is the case for unit 2 and unit 3.

$$
Bel(\text{Unit 2 is born in NZ, when there are inconsistent values in data sources}) = \frac{Bel(\text{unit is born in NZ})}{1 - (\text{sum of scenarios where data sources disagree})} = \frac{0.80 \times 0.40}{1 - (0.80 \times 0.60)} = 0.62
$$

Notice that for unit 3 the output value is now Samoa, which means we are considering Hypothesis 2.

$$
Bel(\text{Unit 3 is born in Samoa, when there are inconsistent values in data sources}) = \frac{Bel(\text{unit is born in Samoa})}{1 - (\text{sum of scenarios where data sources disagree})} = \frac{0.20 \times 0.60}{1 - (0.80 \times 0.60)} = 0.23
$$

The numerator includes scenarios where the data sources are consistent with the final value; reflecting the belief that the data sources are consistent. Whilst the denominator includes scenarios where data sources disagree with one another. Taken together, the belief in the final value would vary as follows:

- Inconsistency between data sources of high quality would contribute to a lower belief in the final value.
- Inconsistency between a data source of high quality, and other data sources of lower quality, where the value of the former was adopted, would still derive a relatively strong belief in the final value.