



# Linking censuses

New Zealand longitudinal census 1981–2006

Robert Didham, Kirsten Nissen, and Wendy Dobson



**Crown copyright ©**

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

### **Disclaimer**

This report represents the views of the authors. It does not necessarily represent the views of Statistics NZ and does not imply commitment by Statistics NZ to adopt any findings, methodologies, or recommendations. Any data analysis was carried out under the security and confidentiality provisions of the Statistics Act 1975. Unless otherwise stated, results presented are the result of data analysis undertaken by the authors.

### **Liability statement**

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

### **Citation**

Didham, R, Nissen, K and Dobson, W (2014). *Linking censuses: New Zealand longitudinal census 1981–2006*. Available from [www.stats.govt.nz](http://www.stats.govt.nz).

ISBN 978-0-478-42907-7 (online)

### **Published in July 2014 by**

Statistics New Zealand  
Tatauranga Aotearoa  
Wellington, New Zealand

### **Contact**

Statistics New Zealand Information Centre:

[info@stats.govt.nz](mailto:info@stats.govt.nz)

Phone toll-free 0508 525 525

Phone international +64 4 931 4610

[kirsten.nissen@stats.govt.nz](mailto:kirsten.nissen@stats.govt.nz)

[www.stats.govt.nz](http://www.stats.govt.nz)



# Contents

- List of tables and figures ..... 4**
- 1 Purpose and summary ..... 6**
  - Purpose..... 6
  - Summary..... 6
- 2 Introduction ..... 8**
  - Census history ..... 9
  - Metadata ..... 11
- 3 The linking process ..... 12**
  - Census data and key variables ..... 12
  - The deterministic linking process ..... 16
  - Probabilistic linking of residuals..... 18
  - Clerical review ..... 22
- 4 Summary of linking results ..... 24**
  - Evaluation of link rates..... 25
- 5 Variability in link rate for core population characteristics..... 32**
- 6 Preparing for future developments ..... 42**
  - Adding 2013 Census data to the NZLC..... 42
  - Investigating potential for including the 1976 Census..... 42
  - Microdata access development..... 55
- 7 Summary of recommended future developments ..... 57**
- References ..... 58**



# List of tables and figures

## Tables by chapter

<b>2</b>	<b>Introduction .....</b>	<b>8</b>
1.	Census dates and population counts, 1981–2013 .....	9
<b>3</b>	<b>The linking process .....</b>	<b>12</b>
2.	Populations at census (t) available for linking to previous census (t-1), 2006–1986 Censuses .....	13
3.	Linked records from deterministic and probabilistic linking stage, 1986–2006 Censuses .....	18
4.	Combinations of blocking and linking variables, for linking process of 1981–2006 Censuses .....	20
<b>4</b>	<b>Summary of linking results .....</b>	<b>24</b>
5.	Number of records uniquely linked between two censuses, by combinations of last and earliest census observed, 1981–2006 Censuses .....	24
6.	Number of linked records for a given census period, 1981–2006 Censuses .....	25
7.	Link rate of theoretical census population by linking stage, 1981–2006 census pairs .....	27
8.	Census population counts, linked records to previous census, and link rate by ethnic group, 1986–2006 Censuses .....	29
9.	Percentage of links linked probabilistically, by ethnic grouping and ethnic mobility, 2001–2006 Censuses .....	30
10.	Link counts for theoretical populations, by level 1 ethnicity grouping, 1981–2006 Censuses .....	31
<b>6</b>	<b>Preparing for future developments .....</b>	<b>42</b>
10.	Number of migration records in test sample, by passenger type, June 2001 and January 2006 .....	47
11.	Number of births and deaths records linked to the 2006 Census, and link rates for deterministic linking, for births occurrences 2001–2006 and deaths occurrences 2006–2011 .....	50

## Figures by chapter

<b>3</b>	<b>The linking process .....</b>	<b>12</b>
1.	Frequency count of probabilistic linking weight, output by Quality Stage, 2001–2006 census pair .....	21
<b>4</b>	<b>Summary of linking results .....</b>	<b>24</b>
2.	Combined linked, not-linked, and theoretical population, by year of birth, and sex, 1986–2006 Censuses .....	28
<b>5</b>	<b>Variability in link rate for core population characteristics.....</b>	<b>32</b>

3. Age distribution by sex of linked and theoretical populations, Māori and total populations, 2006 Census .....	33
4. Percentage of theoretical population at census t linked to census t-1, by age and sex, 1986–2006 Censuses .....	34
5. Percentage of theoretical populations linked to census t-1, by age group, sex, and ethnic group, 2001–2006 Censuses.....	35
6. Link rate by age group and sex for selected characteristics of the 2006 theoretical population, 2001–2006 Censuses .....	37
7. Link rate by age group, sex, and occupation code stated at the 2006 Census, 2001–2006 census pair .....	38
8. Link rate by age group, sex, and region of residence in 2006, 2001–2006 Censuses .....	39
9a. Link rate by age group, sex, and NZDep2006, NZDep2006 = decile 1–10, 2001–2006 census pair .....	40
9b. Link rate by age group, sex, and NZDep2006, NZDep2006 = decile 1, 5, and 10, 2001–2006 census pair .....	41
<b>6 Preparing for future developments .....</b>	<b>42</b>
10. Linked and not-linked birth occurrences in the five years before a census, and not-linked census records, 1996–2006 Censuses, and birth occurrences 1991–2006 .....	51
11. Number of linked and not-linked death occurrences in the five years following a census, 1991–2006 Censuses, and 1991–2011 death occurrences.....	52
12. Percentage of births and deaths linked to eligible 2006 Census population, by number of months between the event and the 2006 Census, 2001–06 births occurrences, and 2006–11 deaths occurrences .....	53
13. Percentage of births linked to eligible 2006 Census population, by age group of mother at birth of child, 2001–06 birth occurrences .....	54
14. Percentage of deaths linked to eligible 2006 Census, by age group, and ethnicity at death, 2006–11 death occurrences .....	55



# 1 Purpose and summary

## Purpose

This report presents the results of linking data from six recent censuses (1981–2006), and makes recommendations for further work. The linking of census datasets to form a longitudinal data source will enhance our understanding of aspects of population change across time.

## Summary

The report covers the linking of the 1981, 1986, 1991, 1996, 2001, and 2006 Censuses into five linked pairs. In addition to census-to-census linking, we investigated a number of other developments. These included investigations into linking birth and death registration data to their nearest census, assessing the possibility of building a 1976–1981 census pair, and linking international migration data to the New Zealand longitudinal census (NZLC).

This report covers:

- how we refined linking methodologies for the three census pairs covering the period 1991–2006, with improvements to overall link rates, while ensuring a high linkage quality
- how we included the 1986–1991 and 1981–1986 census pairs in the longitudinal census database using the linking methodology consistent with the more recent census pairs
- why we will not include the 1976–1981 census pair in the NZLC at this time
- whether linking records of birth and death registrations to the nearest census theoretical population is feasible
- how international travel and migration data will significantly improve the usefulness of the NZLC dataset, and will provide an opportunity to identify additional links between censuses. However, linking this data is likely only possible from the 2006 Census onwards.

When developing the NZLC in the future, we recommend:

- creating the 2006–2013 census pair. This pair will then be included in the NZLC microdata access product.
- completing the confidentiality rules. Confidentiality rules should be finalised to enable dissemination of longitudinal tabular, analytical, and research outputs in a form that will prevent disclosure of confidential individual information.
- finalising linking of birth and death registrations. This will involve linking birth registrations to the nearest later census and death registrations to the nearest previous census.
- further investigating linking international migration data
- investigating linking to other administrative data sources. Adding ‘cause of death’ for linked death registrations would greatly enhance research into mortality and morbidity topics.
- refining family- and household-level analytical capability. Currently, we have included indicators that identify completely linked and partially linked families. We will review and refine these indicators to enable sophisticated household- and family-level analysis.

- statistically analysing bias and development of weighting methodology. Weighting of census longitudinal data requires in-depth statistical analysis of link bias and coverage. We propose investigating the possibility of creating sets of weights that may be applied in various research scenarios.



## 2 Introduction

This report details the linking of the 1981–2006 Censuses as part of the New Zealand longitudinal census (NZLC). It presents the work we carried out after the initial feasibility investigation for the longitudinal census project was completed. It also provides some background information that will put the census data into context, and assist longitudinal data users in interpreting results. This report sits alongside other documentation such as a data dictionary, concordance information, and database design, which are available to NZLC users in Statistics NZ's Data Lab environment.

Before we completed the census linking process, we completed a feasibility study, [Developing a historical longitudinal census dataset in New Zealand: A feasibility study](#), (Statistics NZ, 2013). This study demonstrated a need for a longitudinal census dataset, and investigated whether this was achievable with the data sources and electronic resources available. This feasibility study was phase 1 of a phased approach to developing a longitudinal census dataset.

Phase 1 established that the most viable way to create a longitudinal census dataset was to link adjacent pairs of censuses (called census pairs), and then to link these census pairs together. Phase 1 resulted in three linked census pairs<sup>1</sup> for the 1991–1996, 1996–2001, and 2001–2006 Censuses.

Phase 2, which is the subject of this report, involved creating five linked pairs covering the six censuses from 1981 to 2006. We refined the deterministic links in the three pairs created under phase 1 of the project, created the two additional pairs to extend coverage back to the 1981 Census, and then added probabilistic links, where possible, for the remaining records in each of the theoretically linkable populations. We have included detailed technical information on the linking methodology in chapter 3 of this report.

The resulting datasets for each linked census pair include the linkable population for the source census (referred to as the theoretical population), the non-linkable population for the source census (referred to as the non-theoretical, or residual population), and the linked records with information from the source census and the target previous census (referred to as the record pairs).

Phase 2 also investigated some developments planned for phase 3, which will involve completing the historical resource to cover the period 1981 to 2006, and identifying what we could develop in the future. Further work also to come in phase 3 includes creating the 2006–2013 census pair, including births and deaths information; investigating further the feasibility of including at least some international travel and migration data; and evaluating options for refining linking methodology, with subsequent marginal improvements to link rates for the census pairs.

Electronic census data is available from 1976 to 2006. Because of technical difficulties with linking 1976 to 1981, we decided to defer final assessment of the feasibility of creating the 1976–1981 linked pair until a later phase of development.

Phase 2 work established that:

- linking methodologies for the three census pairs covering the period 1991–2006, which had been created in phase 1 of this project, were refined with improvements to overall link rates, while ensuring a high linkage quality.

---

<sup>1</sup> In this report the convention we use to identify linked pairs is X-Y where X is the target/earlier census and Y is the source/later census.



- the 1981–1986 and 1986–1991 census pairs are included in the longitudinal census database using the linking methodology consistent with the more recent census pairs.
- the 1976–1981 census pair will not be included in the NZLC at this time.
- including records of birth and death registrations linked to the nearest census is feasible. This draws on existing experience, and tests have shown that we can achieve satisfactory link rates. In assessing the feasibility of linking birth and death registration data to the nearest census, we completed the main steps of the linking process.
- international travel and migration data would significantly improve the usefulness of the longitudinal dataset, and presents an opportunity to identify additional links between censuses. However, data limitations mean that this task will require significant additional information and ongoing work. This suggests that it is likely to be possible only from the 2006 Census onwards. A further limitation is that only international travel events involving permanent and long-term migration with at least one associated census record would be relevant. While this limits potential inclusion of historical data, adding even this limited set of movements in the future does ensure that the future users of the longitudinal information would have an increasingly valuable resource over time.

## Census history

As part of the 2013 Census output, Statistics NZ published [History of the census in New Zealand](#). This short historical summary illustrates how the New Zealand census developed from its early beginnings before the 1851 Census, through to the present day.

See table 1 for the dates of the censuses included in the NZLC and the size of the usually resident population at each census point.

**Table 1**

### Census dates and population counts

1981–2013

Census	Census date	Usually resident population
1981	24 March 1981	3,143,307
1986	4 March 1986	3,263,283
1991	5 March 1991	3,373,926
1996	5 March 1996	3,618,303
2001	6 March 2001	3,737,277
2006	7 March 2006	4,027,947
2013	5 March 2013	4,242,048

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

## Why census questions can change over time

Each census has been developed to meet the needs of its time; therefore, each census has a different set of questions, though with a core set of questions common to all censuses. One of the most important pieces of metadata in any resource of this nature is a set of the questionnaires used to collect the information. Statistics NZ has made electronic copies of [census forms](#) available, including the forms for all of the censuses included in the NZLC. This enables users to see what questions were asked, how they were asked, and where each question sat within the questionnaire.

These details frequently become important in the research process. Not only does the context of a question within the questionnaire matter for that question, it may also have implications for variables derived using that question. In some cases, information is derived from different types of questions in different censuses.

Census questionnaires change to ensure the survey maximises the value at the time. Hence topical questions (such as beekeeping and colour-television ownership in earlier censuses, and Internet access in recent censuses) are introduced or replaced over time.

### **Census questions about educational achievement**

For example, participation in education may be a key analytical item for a particular research theme. In 1981, there were four educational qualification questions (Q26–29 – the final questions on the form), two focusing on secondary education and two on tertiary education. The secondary education questions asked about highest school qualification gained and the highest levels of attendance. The tertiary education questions focused on present and past attendance, and institution type; and tertiary qualification names, subjects, and year gained.

In 1986, these four questions were collapsed into two simple tick-box questions (Q17–18) – one on highest school qualification and the other on qualifications obtained since leaving school. The only indication of current study in the 1986 census form was in the preceding question on ‘main work or activity’ (Q16) where it was possible to specify ‘full-time study’ as a main activity, or ‘still at school’ in the post-school qualification question (Q18).

The omission of part-time study was partially rectified in 1991 with the addition of both full-time and part-time tick-boxes in the activities question (Q20), but this was now placed after the questions on voluntary work and away from the qualification questions (Q16–17).

A separate question on studying (Q30) was introduced in 1996, followed by redesigned qualifications questions (Q31–34). This made it possible to indicate both full-time and part-time study, but the post-school qualification question (Q34) allowed for only two qualifications, although with the additional information on field of study, institution, and year of completion. The latter items were poorly coded and unreliable.

The consequence was that the tertiary qualification question in 2001 (Q24) asked for only one qualification, and the respondent was asked to enter the name of their highest qualification obtained and the main subject. The question on study participation was overlooked and added as a late addition to the unpaid activities question (Q41, a question with a high non-response rate).

Any current research on educational pathways needs to consider the consequences of these changes over time.

### **Census questions about Māori ethnicity, ancestry, and iwi affiliation**

Linked census pairs can add value to neighbouring censuses. In cases where current topics are of high interest, but relevant questions were not asked in earlier censuses, the NZLC can assist in adding new time depth. For example, research involving Māori development may require information on ethnicity, ancestry, and iwi affiliation. While ethnicity has been asked in each census since 1916, the questions used have varied.

However, it is not always immediately clear how, or indeed if, question changes have affected the data. In the case of ethnicity, an analysis of the data (for example, Didham, 2005) has shown that the 1981 ethnic data is approximately comparable with the 1996 to 2006 period. The 1986 data is anomalous. Moreover, ancestry was only collected from 1991 when an iwi question was re-introduced.

Because the correlation between ethnicity and ancestry for Māori is high, the NZLC data will potentially enable us to analyse dynamics of Māori social change and well-being, at an iwi level, a decade further into the past. However, this will only work if the links for members of the relevant iwi or grouping of iwi were sufficiently robust to allow sound historical analysis. To assist with valid analysis on topics requiring ethnicity, ethnic indicators based on the [2005 standard of ethnicity](#) (used for the 2006 Census coding of ethnicity) are included in the linked pairs to provide some measure of consistency.

## Metadata

Beyond question changes are changes to classifications and coding practices. These affect even core questions that appear in every census. Caution should always be exercised when dealing with classification labels that are prone to changes in both definition and content over time.

One of the challenges that the NZLC presents is ensuring that, wherever possible, code files and concordances are made available to enable the longitudinal use of the data. The key concern is to ensure that not only are all code files, descriptors, and concordances available, but users also have information on what the data means and does not mean. The latter becomes important when a descriptor is used for different category content over time.

Metadata also alerts users to inconsistencies in data. One kind of inconsistency relates to accuracy of linking. Deterministically linked records merely indicate that those records had consistent information over time for the blocking variables. There is no absolute certainty that the links refer to the same person since we have neither names nor exact addresses available to use for linking of historic census records. While some clerical checks are possible, these cannot identify all of the erroneous links. Probabilistic links have lower levels of certainty of accuracy. This, in part, implicitly contributes to protection of confidentiality, but also underscores the importance of not over-analysing the data.

Similarly, key variables may not be entirely consistent over time. Changes in definitions and variable quality of data contribute to further inconsistencies. Data users should also be mindful that category labels are sometimes used with quite different names over time. This may be quite clear, as is apparent in occupation coding where there are profound differences between NZSCO68 and NZSCO90.

But it may be more subtle, as we find with the number of children born live question. In 1976, the question was asked only of women who were currently or previously legally married, and the question referred only to children born while the mother was married. The 1981 Census extended the question's scope to cover all children born to all women aged 15 years and over.

Question changes of this nature become important when a topic requires the derivation of apparent rates of change. In addition to this type of information, a good understanding of variable non-response, non-enumeration, accuracy levels of links, coding inconsistencies, misrecognition of responses, and biases inherent in differential missingness is required to fully understand the results.



## 3 The linking process

Linking methodology and procedures have a direct impact on the resulting data. Which records are in and out of scope, which records are linkable and not linkable, and the quality of the links created contributes to the types and levels of bias in the final data. This section outlines the linking of data to create the linked pairs. A description of the data sources is followed by details of the linking process through each of the stages.

### Census data and key variables

#### The census populations used

The theoretical census population (at time  $t$ ) available for linking to the previous census (at time  $t-1$ ) is restricted to those people who have a chance of being linked because they:

- are old enough to have been alive at the last census
- were in New Zealand at the previous census
- had completed a census form at the previous census.

The theoretical population is defined as the usually resident population excluding those people who:

- did not return a census form and, therefore, had a substitute record created, or
- were not born five years ago, or
- were usually resident overseas five years ago.

The portion of the population not available for linking is referred to as the residual or non-theoretical population. This sub-population has almost doubled in size over the census points spanning 1986 through to 2006 (from 379,000 in 1986 to 742,000 in 2006). This is partly because of the increased number of substitute forms created from 1996 onwards, but primarily because of increases in intercensal immigration and return migration.

The proportion of the usually resident population available for linking has consequently declined over the same period (81.6 percent of the usually resident population in 2006, compared with 88.4 percent in 1986). Those not born at the time of the previous census have had no effect on this drop, because the size of the population aged four years and younger has remained fairly stable across the 1981 to 2006 period of interest (table 2).

The theoretical population was split for different stages of the linking process. Linking took place in two broad stages: deterministic linking, followed by probabilistic linking of the residual records. The population eligible as input to the deterministic linking process excluded those census records (at time  $t$ ) where at least one of the following conditions was met:

- age was stochastically imputed
- parts of date of birth information (day, month, year) were not stated
- sex was neither stated nor imputed deterministically (from name, relationship, or, in the case of residents in non-private dwellings, the sex of other occupants)
- address five years ago was not stated or classifiable to an area unit
- or usual residence at census ( $t$ ) was imputed.

Any records eligible for deterministic linking that failed to be uniquely linked were returned to the population eligible for probabilistic linking.

The remaining unlinked records form the residual population available for the probabilistic linking process. These contain the whole theoretically linkable census population except for those uniquely matched by deterministic linking.

The following table summarises the derivation of the 1981–2006 census populations that were available for the deterministic and probabilistic linking processes, respectively, and the number of people who were not available for these linking processes.

**Table 2**

**Populations at census (t) available for linking to previous census (t-1)**  
2006–1986 Censuses

	Number of records at census (t)				
	2006	2001	1996	1991	1986
Usually resident population	4,027,947	3,737,277	3,618,303	3,373,926	3,263,283
Less number of usual residents:					
- recorded by a substitute form	123,780	100,203	101,361	14,157	2,076
- overseas five years ago	343,110	244,098	218,490	156,774	127,914
- not born five years ago	275,076	270,801	279,534	277,146	249,072
Theoretical population available for linking to census t-1	3,285,978	3,122,175	3,018,918	2,925,849	2,884,221
Less number of usual residents:					
- with imputed age, sex, or usual residence, or usual residence five years ago not stated	148,836	45,114	120,288	56,646	54,294
Eligible population available for deterministic linking	3,137,139	3,077,064	2,898,561	2,869,203	2,829,927
Percentage of usually resident population available for linking <sup>1</sup>	81.6	83.5	83.4	86.7	88.4

1. The theoretical population as a percentage of the usually resident population.

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

## Choosing key variables

When choosing the key variables (defined as linking or blocking variables) to be used in the linking process, there were two major considerations:

- How reliable the variable is; ie does it have a high response rate? Is the quality of responses high?
- How consistent and comparable the variable is across censuses. Responses must be consistent over time. For example, while characteristics of individuals such as occupation, qualifications, religion, and ethnicities of respondents change over time, we expect date of birth to remain constant. Secondly, the census question and classification of responses must be comparable over time, as is the case with sex and date of birth.

Sex and date of birth make ideal linking variables. The questions used to capture this information have remained highly comparable over time. They have very high response rates and are unlikely to change in any intercensal period except in very rare cases (or where there has been respondent or processing error).

A geographic variable is also necessary to distinguish between individuals with the same sex and date of birth. Full address would be the natural choice here, with meshblock as second choice. Unfortunately, although the census questionnaire asks for the full address of usual residence and the full address of usual residence five years ago, Statistics NZ only stores the area unit of the usual address five years ago. As a result, area unit of usual residence was the smallest geographic unit available for linking across time. Changes in geographic boundaries in each intercensal period are taken care of by using geographic boundaries rebased to the same classification for each census pair.

Other information, such as names, were not available for the censuses used to build the linked pairs in this stage of the project. All other variables were subject to substantial change over time with the exception of country of birth. This variable was used in the second step of deterministic linking.

### **Comparability of variables over time**

Auckland University's Family and Whānau Wellbeing Project produced [A guide to using data from the New Zealand census: 1981–2006](#) (Errington, Cotterell, von Randow, & Milligan, 2008)<sup>2</sup>, which compared the definitions of selected census variables for censuses between 1981 and 2006. Each of the key variables considered in the linking process for the NZLC had been assessed by the authors of the report as either 'totally comparable' or 'highly comparable' for the 1981–2006 Censuses, with the exception of Māori descent.

### **Area unit of usual residence**

Each census uses different geographic boundaries. For linking purposes, it is essential to compare 'like with like' as much as possible, and this is done by using rebased geographic boundary information, supplemented where necessary by concordance information.

Each census dataset contains area unit information using the geographic boundaries established at the time of the census. In general, datasets for the 1986 to 2006 Censuses have geographic boundaries for location at census and at the time of the previous census rebased to 2006 patterns, and 1986 and 1981 have information on mixed 2001 and 2006 patterns.

While the 2001 and 2006 boundary sets are highly comparable, some precision is lost when boundaries have been nudged (rather than split/merged) between censuses. Nudging refers to moving a boundary between two neighbouring areas without changing the naming of either entity. Nudging is used to align minor changes with higher geographic boundaries. Fortunately, the population affected by nudging is not large. However, rebasing methodology has limitations and for the purpose of linking it was more effective to use the 2006 boundary pattern for recent pairs and 2001 for earlier pairs. The linking process used area units rebased to 2006 geographic boundaries for the 2001–2006 and 1996–2001 pairs, and 2001 rebasing for the 1991–1996, 1986–1991 and 1981–1986 pairs.

### **Area unit of usual residence five years ago**

As for 'area unit of usual residence', we used rebased 'area unit of usual residence five years ago' so that the same geographic boundaries were compared in each census pair. The 'area unit of usual residence five years ago' variable has two main limitations: respondents' recall, and coding issues. The consequence of this is that a number of records have address five years ago coded only to higher geographies, ranging from territorial authority to New Zealand not further defined.

---

<sup>2</sup> Note the database created by the Family and Whānau Wellbeing Project was a series of census snapshots and not a linked dataset.

### **Year of birth, month of birth, and day of birth**

The definition for each of these variables is totally comparable between censuses from 1981 onwards and we would generally expect respondents to consistently report them throughout their lifetime. As with all other variables, proxy responses may reduce the reliability of this variable.

### **Sex**

The definition for this variable is totally comparable between censuses. We would expect the sex of most respondents to remain consistent (and to be consistently reported) throughout a respondent's lifetime.

### **Country of birth**

The definition for the birthplace variable is highly comparable over time. Countries do, however, change names, split, merge, or move over time, and may be reported differently by respondents. Moreover, different classification systems were used in each census so it was necessary to create country concordances, especially for earlier censuses. This was mostly a straightforward exercise but there were a few cases where we lost some potential links because of the difficulty in creating one-to-one concordances.

A common instance of this problem is where a respondent may cite their country of birth in one census as 'United Kingdom' or 'Great Britain', but as one of the component countries of these entities in another census. In this case, grouping the codes is straightforward.

However, this becomes problematic if the response is 'Ireland', which may refer to more than one country, each with a separate code. Similarly, some countries have split (for example, Yugoslavia) or merged (for example, Germany) over the period between censuses, or the country previously reported no longer exists. People may report, at different times, the current name of a country, or the name that was current at the time of their birth. We would expect respondents to consistently report 'country of birth' throughout their lifetime. However, there are a few cases where this does not happen.

### **Māori descent**

Māori ancestry/descent is broadly comparable for the census between 1991 and 2006, albeit with different question wording and different coding practices for each census. Māori descent was not collected in 1986.

The 1981 Census collected ethnic origin but did not specifically ask about Māori ancestry. Responses in 1981 were coded to a pre-set list of combinations of ethnicities approximating to degree of blood, as had been historically collected. The 1986 Census also asked about ethnic origin but the data was collected and coded differently, with up to three responses retained and stored separately.

While the 1981 dataset has a derived variable referred to as Māori descent, this was derived from the ethnicity question and returns identical counts to the total response for Māori ethnicity. In theory, unlike ethnicity, a respondent's Māori descent should not alter throughout their lifetime, but in practice there is known to be considerable variation in reporting.

With this in mind, and because it is not available in all censuses, Māori descent was not recommended as a key linking variable. However, it is used in a minor capacity to identify additional links among the small number of remaining records, or for separating out the 'exact duplicate' links. An exception relates to duplicate records, which represent same-sex twins who we might assume to have the same descent status.

### **Data quality**

Error sources, apart from questionnaire and classification aspects, affect all of these variables. Errors arising during processing may result in linking failures at the

deterministic stages. These include misrecognition of responses, changes to responses during data editing, and miscoding of responses. Some of these types of errors may not prevent a probabilistic link. No re-editing of the data has been done, and inconsistencies may appear between some links as a result.

## Software used

For logistical reasons, we used SAS<sup>®</sup> for the deterministic linking, and QualityStage<sup>®</sup> (QS) software for the probabilistic linking. QS is a standard data integration tool that is widely used in other Statistics NZ data integration projects. It allows the user to specify the amount of uncertainty acceptable for a pair of records to be treated as a link.

We were constrained in our choice of software by what system resources were available in the IT environment. In theory, we could have used QS for each stage of linking but in practice SAS had much quicker processing times. Furthermore, using a combination of SAS and QS eased the pressure on the QS server.

## Terminology

In conversation, the words 'link' and 'match' tend to be used interchangeably. In data integration terminology, they have distinct meanings relative to the nature of the connection made:

- A 'link' refers to a record pair where the connection has been assessed as probable.
- A 'match' refers to a record pair where the connection is true.

In the context of the longitudinal census project, the process of combining records from adjacent censuses is linking and the output datasets are a collection of links. Even in cases where the details appear an exact match we cannot be certain this is so. Nevertheless, the majority of the linked record pairs are expected to be true matches, but it is also expected that a small portion of these will not be true matches. Clerical checking would be needed to attempt to quantify this uncertainty, but this task has been deferred at this stage.

A census pair 't,t-1' refers to a pair of censuses where individual records in census (t) are linked to those of the previous census (t-1). For example, if we are looking at linking records from the 1986 Census to those from the 1981 Census, we will refer to this as the 1981–1986 census pair.

## The deterministic linking process

The phase 2 study (which this paper is about) built on the initial deterministic linking methodology used in the [phase 1 feasibility study](#) of a New Zealand longitudinal census database. The phase 1 report summarised the resulting linked datasets and the match rates, and discussed some characteristics of the non-linked datasets covering the census pairs: 1991–1996, 1996–2001, and 2001–2006.

Following the evaluation of the phase 1 feasibility study, we recommended extending the linked census pairs to 1986–1991 and 1981–1986. This phase 2 study also includes other key variables as linking variables and evaluates the potential improvements to numbers of linked record pairs. The outcome of this evaluation was a two-step deterministic linking process described as follows.



## Deterministic linking: Step 1

In the first step, we attempted to link eligible records in the theoretical population using all of the following key variables:

- sex
- day, month, and year of birth
- area unit of usual residence (as defined in the paragraph below)

Census records are linked from one census to the previous census using the address five years ago in the source census ( $t$ ). This address is then linked with the address of usual residence in the target census ( $t-1$ ). Throughout this report, the term 'area unit of usual residence' refers to area unit of 'usual residence five years ago' collected at census  $t$ , and of usual residence at census  $t-1$ .

We treated two records as a linked pair if all the following conditions were met:

- they matched exactly on all of the linking variables
- the link was unique (that is, no other record was an equally exact link)
- the response for each of the linking variables was complete and valid.

We linked approximately 68 percent of the population eligible for deterministic linking in this first step. This equates to roughly two-thirds of the theoretical population.

The remaining records were sent to the second linking step of the deterministic process.

## Deterministic linking: Step 2

The first deterministic step matched exactly and uniquely approximately 68 percent of the eligible population to the previous census, using the chosen blocking variables. In addition to this, a further 6 percent of the exact matches were non-unique. We subject these non-unique matches to step 2 of the deterministic linking process.

QualityStage (QS) was considered as the tool to distinguish these non-unique exact links. However, early investigations revealed that if two or more records in census  $t$  have an equal chance of linking to the same record in census  $t-1$  (or vice versa), then QS would randomly choose which record pair to assign as the 'true' link. To minimise this problem, we explored different intermediate options for additional linking variables, and included them as subsequent steps in the deterministic linking process.

Step 2 involved a two-step process. The first step (step 2a) linked on sex, date of birth variables, area unit of usual residence, and country of birth. This was effective for the links where one or more of the potential matches was overseas born. The second step (step 2b) involved linking the residuals using sex, date of birth variables, area unit of usual residence, and Māori descent.

The chosen additional linking variables combined the properties of being able to discriminate between records, and being more consistently reported in each census than some other census variables. Some examples of linking variables which we considered but discarded are: year of arrival in New Zealand, years at address code (adjusted for the additional five years between each census), ethnicity, and languages spoken. In each case, the quality of the data and the propensity for the information to change between censuses were limitations.

There are certainly limitations to this approach but using *some* data to determine the correct link is preferable to leaving the choice entirely to chance. It is also important to note that we applied step 2 to a relatively small proportion of records. This step added approximately 2 percentage points to the link rate. We sent the remaining records to the probabilistic linking process.

## Probabilistic linking of residuals

The records we subjected to probabilistic linking comprise those residual census records from the theoretical population which could not be uniquely linked in the deterministic stage of the process. These records comprise:

- non-linked records from deterministic linking
- duplicate or non-unique links
- non-eligible records for the deterministic linking process (see table 3).

**Table 3**

### Linked records from deterministic and probabilistic linking stage 1986–2006 Censuses

Populations	Census t				
	2006	2001	1996	1991	1986
Theoretical population at t for linking to t-1	3,285,978	3,122,175	3,018,849	2,925,849	2,884,221
Eligible population at t for deterministic linking to t-1	3,137,139	3,077,064	2,898,561	2,869,203	2,829,927
Deterministic linking stage ( t, t-1)					
Unique links from Step 1 (date of birth, sex, and area unit)	2,141,913	2,016,147	2,020,476	2,064,087	1,923,846
Unique links from Step 2a (incl. country of birth)	46,632	38,118	33,930	31,218	27,969
Unique links from Step 2b (incl. Māori descent)	30,090	28,245	27,912	22,656	22,002
Unique links from deterministic stage	2,218,635	2,082,507	2,082,318	2,117,961	1,973,814
Residual theoretical population t,t-1	1,067,343	1,039,668	936,531	807,888	910,407
Probabilistic linking stage ( t,t-1)					
Unique links (area unit, sex, date of birth)	92,457	88,179	92,004	102,393	104,613
Total number of unique links ( t,t-1)	2,311,092	2,170,689	2,174,322	2,220,354	2,078,427
Theoretical population at t not linked to t-1	974,886	951,489	844,527	705,495	805,794

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

In the probabilistic linking stage, we compared a record from the set of unlinked records in the source data with the records in the target dataset. We compared selected variables common to both sets of data and assigned a weight that represents the likelihood of the two records belonging to the same person. This stage added approximately 3 percentage points to the link rate.

## Linking and blocking variables

Probabilistic linking using the QS software allows the user to choose both blocking and linking variables, and to set the cut-off weight, whereby links with a weight above the cut-off weight are deemed a true link.

A blocking variable is one whereby every record must have the same values before they will be compared. This is similar to deterministic linking. The deterministic linking stages in SAS effectively blocked on sex, date of birth variables, and area unit.

A linking variable allows comparison of records within blocks. We then evaluate records on how close their linking variables are.

The probabilistic linking stage uses area unit of usual residence and year of birth as blocking variables. Early tests using just one blocking variable (area unit of usual residence) resulted in processing time of over 32 hours for each census pair. Using area unit and year of birth reduced this to a matter of an hour or two. The linking variables are day of birth, month of birth, and sex.

This approach closely mirrored the method used by the Inter-ethnic Mobility Study (IEMS) (Brown and Gray, 2009) except the IEMS used year of birth as a linking (not blocking) variable. We explored using other linking and blocking variables, but manual inspection of the linked records showed that this introduced too many suspected false positive links.

Area unit of usual residence at census t-1 is highly reliable. Year of birth is also highly reliable, consistent, and well reported. Area unit of usual residence five years ago at census t has a greater risk of being misreported (due to respondents' recall) and miscoded (especially when rebased to a different geographic boundary).

We did not take full advantage of the QS software. QS does have the capacity for introducing more 'fuzziness' by either setting a lower cut-off weight or by using match-comparison functions which allow +/- a specified amount (for example, allowing day of birth to be +/- one day either way). This was explored further using some trial datasets but found that any benefit of a higher link rate was easily offset by the drop in link quality.

Passes 1 to 3 (see table 4) refer to the three deterministic stages (step 1, step 2a, and step 2b). The linking process separates the theoretical population into two streams at each pass. Table 4 summarises which variables are used as blocking variables and linking variables at each pass. The linked stream is filtered at each step and only the records failing a pass are progressed to the next pass. Pass 4 is the probabilistic linking process. Records that fail pass 4 remain unlinked in the theoretical datasets and are marked as such by an indicator.

**Table 4****Combinations of blocking and linking variables**

For linking process of 1981–2006 Censuses

Variable	Pass			
	1	2	3	4
Year of birth	B	B	B	B
Month of birth	B	B	B	L
Day of birth	B	B	B	L
Sex	B	B	B	L
Area unit of usual residence (five years ago)	B	B	B	B
Country of birth	..	B	..	..
Māori descent	..	..	B	..

**Symbols:**

B blocking variable

L linking variable

.. not applicable

**Source:** Statistics New Zealand**Input probabilities and cut-off weight**

The m-probability is the probability that a variable agrees with its pair value given that the record pair being examined is a matched pair:

$$m = P(\text{records agree} \mid \text{records are a match}).$$

The m-probabilities were set to match those chosen by the Inter-ethnic Mobility Study:

- sex: m probability of 0.99
- month of birth: 0.97
- day of birth: 0.95

The u-probability is the probability that a variable agrees with its pair value given that the record pair being examined is a non-matched pair:

$$u = P(\text{records agree} \mid \text{records are not a match}).$$

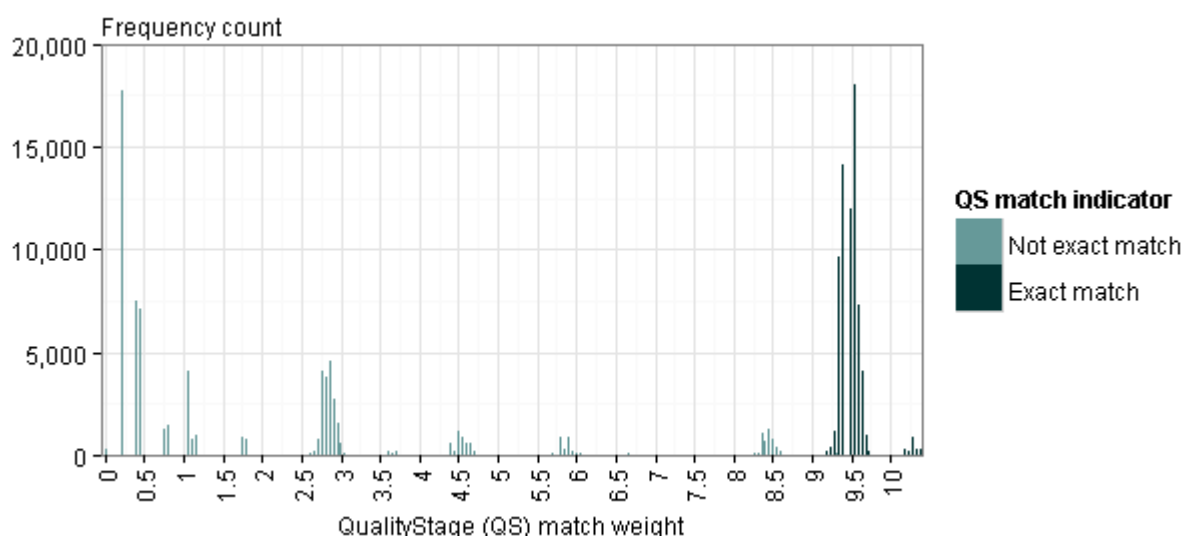
Typically, the u-probability is estimated by the inverse of the number of different values a variable can take. Month of birth, for example, might have an approximate u-probability of  $1/12 = 0.083$ . This was the initial approach for inputting u-probabilities into QS but, ultimately, QS calculates its own u-probabilities based on values of the input records.

We chose the cut-off weight by observing the frequency count of weights of the links created by QS. See figure 1 for output from the 2001–2006 census pair. The weight distributions for the other census pairs were highly comparable.

**Figure 1****Frequency count of probabilistic linking weight**

Output by QualityStage

2001–2006 census pair

**Source:** Statistics New Zealand

Links with a low weight have a small chance of being true matches. Links with a high weight are more likely to be true matches. The chosen cut-off weight was set at 9 (which corresponded roughly with QS's 'exact flag' boundaries). The same cut-off weight was appropriate for each census pair.

Probabilistic linking added approximately 3 percentage points to the link rate. This is discussed further in chapter 4.

**Limitations of the probabilistic linking process**

The methods we used in phase 2 of the NZLC are by no means definitive. They simply offer one way to link censuses and they should not preclude further investigation into linking methods, either for further linking or for later improvements to existing links. International practice for linking censuses includes ongoing additional links and corrections as these are identified. Limitations of the approach used here include:

- address of usual residence five years ago is partially limited by respondents' ability to recall where they lived five years ago. Furthermore, if this variable was missing, or the current usual residence code had been imputed, we were unable to treat it as a valid response. This is particularly important because the variable was used as a blocking variable – that is, we *only* considered links where the area unit of usual residence matched.
- the use of area unit rather than actual address (or meshblock) limits the number of unique links. If possible, linking of future census pairs should be done at a finer level of geography. Currently, only area unit is stored for the usual residence five years ago.
- birthplace classifications change from one census to the next; one-to-one concordances do not exist for all census pairs. Preliminary concordances were used for the census pairs between 1981 and 2001. Subsequent work has substantially refined these concordances.
- country of birth and Māori descent responses may be less reliable than, for example, the sex or year of birth variables. As noted above, using *some*

information to determine the correct link is preferable to leaving the choice entirely to chance.

- the use of the family ID as a blocking variable may potentially identify some residual family members being uniquely linked to the previous census. We have not investigated this additional linking pass in the phase 2 developments.
- the probabilistic linking step did not improve the link rate considerably. Most record pairs linked at this step resulted from QualityStage picking a non-unique exact link at random.

## Clerical review

One outstanding task is a clerical review of the quality of the links achieved.

Two key impediments have forced a postponement of this review. First, we lack essential resources, especially sufficient retained forms and images from previous censuses, to provide a robust investigation. This was exacerbated by inadequate access to those that are available. At least part of the current difficulty lies in locating the resources after the severe disruption of the Canterbury earthquakes. The second is a more proximate constraint, namely the lack of available resources to perform a fairly labour-intensive task.

The two most common quality measures for linked data are the link rate and the false positive rate.

The link rate relates to the proportion of the theoretical population that is linked, and can easily be calculated by taking the number of individual links created and dividing by the theoretical population. For the NZLC, this measure is typically around 70 percent for the overall population, with higher and lower rates for population subgroups.

The false positive rate relates to the accuracy of the links and is more difficult to determine. A false positive is when a link is made between two records, which do not belong to the same person, but have identical linking data. For example, we might link to records in the deterministic matching stage which match exactly on sex, date of birth, and area unit of usual residence but, by chance or as a result of erroneous data, have linked two completely different people. These may be apparent when other variables are compared, but this would require close scrutiny across many variables.

Manually reviewing each and every linked record pair would be impractical so we typically manually review a sample of records. Even this approach isn't entirely robust, as the two records may still not belong to the same person. This becomes important when the data of interest are the trajectories of change across time. A high or not known false positive rate may compromise the quality of the analysis.

Clerical reviews therefore form an important component of quality assurance. Although we have not conducted a formal review, we are confident that, at least for key variables, the quality of the NZLC is adequate. For example, all dates of birth match exactly. However, one indication of the sort of data inconsistency that may exist is that there are 71 records that we linked probabilistically, but which have inconsistent sex for some points across the links. Inconsistencies of this type are an inevitable consequence of probabilistic linking. These links may indeed be correct and the data incorrect for specific points, but this is not proven. Without clerical checking we cannot say with certainty that this assumption is valid; it is clear that in some cases the link is possibly wrong.

In the case of inconsistent sex, the frequency is much less than 1 in 100,000 (there are 7.399 million people appearing at least once in the theoretical populations) and at least some may in fact be a valid change of sex. The impact on the data is very minor. On the other hand, we do not know how many, and what type, of other erroneous links may be in the data.

While we are sure we have a sufficiently robust outcome, we do not have an objective measure of quality of the links. A clerical review remains an outstanding task of high importance. The added advantage of this would be the experience gained which would then contribute considerably to the development of the 1976–1981 pair. For 1976, we do not have date of birth and we are aware of lower quality in the geographic variables – this may require some use of field books and forms from 1981 and 1976 to improve the quality of the result.

## 4 Summary of linking results

Table 3 above provided an overview of the size of the relevant populations and the results of the linking process. In this section we look more closely at the results from two different perspectives.

We can present the number of links in two different ways. Each way answers a different question.

### How many records are only linked between two specified census points?

The answer is displayed in table 5. This shows, for example, that 646,935 records have information linked for the whole period from 1981 to 2006. Similarly, 122,058 records span the 1986–1996 period only. Data users are able to easily extract subsets of records on the basis of time depth, by using two derived variables identifying the earliest and the latest census data available for any record.

**Table 5**

### Number of records uniquely linked between two censuses

By combinations of last and earliest census observed

1981–2006 Censuses

Last census observed	Earliest census observed				
	1981	1986	1991	1996	2001
2006	646,935	234,915	291,201	419,427	718,614
2001	202,779	92,097	103,278	180,051	
1996	303,957	122,058	177,099		
1991	427,491	190,122			
1986	497,265				
1981					

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

### How many linked records are available for a specified study period?

This question is subtly different from the previous question because it also includes records linked beyond either end-point of the period. See the answer to this question in table 6 below. For example, if the period of interest is 1986 to 1996, covering the three census points 1986, 1991, and 1996, we find that there are 2,174,322 records available. This includes the 122,058 records linked for only that period (1986–1996), as above, as well as those that are linked beyond the end-points specified. So, in this case, the period is also covered by records linked from 1981 to 2006, 1986 to 2006, 1986 to 2001, 1981 to 2001, 1986 to 1996, and 1981 to 1996. We can extract these populations using the link indicator variables associated with each pair.

One feature that is immediately apparent from table 6, is that the size of linked populations is relatively stable across similar time spans. For example, very broadly, just over 2 million records are available for a study requiring a span of a decade, regardless which decade in the 1981–2006 period is chosen. Around 1.1 million records cover any given span of two decades. While this also translates into approximately similar link rates, further research is required to identify if this also involves similar inherent bias.



**Table 6**

**Number of linked records for a given census period**  
1981–2006 Censuses

Census						Link rate <sup>1</sup> (percent)
1981	1986	1991	1996	2001	2006	
				2,311,092		70.3
			2,170,689			69.5
		2,174,322				72.0
	2,220,354					75.9
2,078,429						72.1
			1,592,479			54.5
		1,571,208				56.2
	1,602,744					59.4
1,581,162						59.4
			1,173,051			45.4
	1,176,726					47.5
1,153,671						47.5
		881,847				38.6
	849,714					38.3
		646,935				31.5

1. Percentage of the theoretical population at the latest census that is linkable across the censuses observed.

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

## Evaluation of link rates

The proportion of the theoretically linkable population that was successfully linked is just as important as the size of the linked populations, discussed in the previous section. The census pairs linked in phase 2 of the project contain more than 70 percent of the theoretically linkable populations. In this section, we examine some of the high-level features of the link rates.

We need to undertake further statistical research of link bias, because although the populations are of similar size, their age, sex, and ethnic compositions may differ across time in a way that has variable effects on research questions. We strongly advise data users to make a preliminary assessment of this aspect of the data when analysing trends. In this section we consider some of the characteristics of the linked and non-linked records, along with some of the causes of variability in link rates.

### Linkage and non-linkage

There are many complex reasons why some records cannot be linked. The differences between the census usually resident population count and the theoretical population available for linking point indirectly to some of the reasons. People in the usually resident population but not in the theoretical population include people not born five years ago, or who stated that they were usually resident overseas five years ago. The level of non-response or imprecise response to the 'address five years ago' question also contributes

to the residual populations. These characteristics vary markedly between different groupings of ethnicities.

Approximately 82 percent of the total usually resident population and 86 percent of the Māori populations in 2006 were in their respective theoretical populations. The younger and recent migrant populations reduced this percentage for the Pacific and Asian populations to 80 percent and 58 percent, respectively.

For people of Māori and Pacific ethnicities, a larger proportion of the populations was theoretically linkable, but the rate of linking was lower than the total population. At least part of the explanation for this relatively lower success rate is that these two groups are generally younger, more mobile, and more likely not to have been counted at one or other of the censuses being linked. Also, when counted, they are more likely to have data missing from key linking variables.

In contrast, the Asian population had a higher proportion of people living in New Zealand for less than five years (and, therefore, a smaller theoretically linkable population), but the theoretical population was easier to link than the Māori and Pacific populations.

The key reasons why a person will not be eligible for linking from census  $t$  to census  $t-1$  include:

- the person was not counted and was recorded only by a substitute form
- the person was living overseas or not born at census  $t-1$
- the person did not state a usual residence five years ago address.

Similarly, a record may meet the criteria for inclusion in the theoretical population, but fail to be linked because:

- the person was in the theoretical population, but either did not return a census form in the previous census (thus contributing to the census undercount at census  $t-1$ ) or was recorded at census  $t-1$  by a substitute form
- the person provided a usual residence five years ago address that was inconsistent with the usual residence address recorded at census  $t-1$ , and could not be confidently identified as the same person
- there was insufficient information for the key linking variables at either census  $t$  or census  $t-1$ , or both, to be linked uniquely.

## Overall link rates

We linked the majority of the linked records in step 1 of the deterministic stage. This step contributed links for two-thirds of the theoretical populations. The subsequent stages (steps 2a and 2b of the deterministic stage, and the probabilistic stage) added approximately 5 percentage points to the link rates. Table 7 shows the achieved link rates for each census pair at each stage of the linking process.

**Table 7**

**Link rate of theoretical census population by linking stage**  
1981–2006 Censuses

Linking stage	Census t to census t-1				
	1986–1981	1991–1986	1996–1991	2001–1996	2006–2001
	Percent				
Step 1 deterministic	66.7	70.5	66.9	64.6	65.2
Step 2a deterministic	1.0	1.1	1.1	1.2	1.4
Step 2b deterministic	0.8	0.8	0.9	0.9	0.9
Probabilistic	3.6	3.5	3.0	2.8	2.8
Total	72.1	75.9	72.0	69.5	70.3

**Note:** The link rate is the percentage of theoretical population at latest census linked to previous census. Because of rounding the percentages may not add exactly to the stated totals.

**Source:** Statistics New Zealand

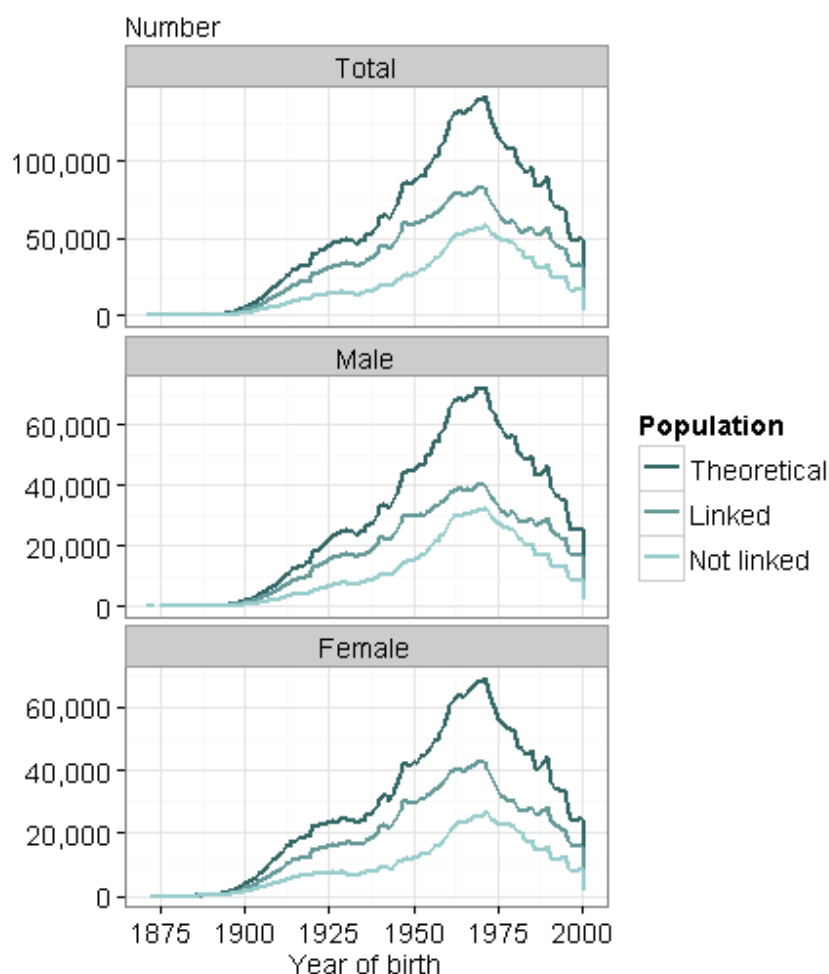
The number and proportion of records linked varied by age of respondent. If we plot the records of the union of all the theoretical populations covering the 1986–2006 Censuses (totalling 7,399,134 individuals who appear in at least one theoretical population) by their year of birth, we see that the general shape of the populations linked and not linked are similar to the total theoretical populations (figure 2).

The proportion of records linked is smallest for those born after around 1970. People born after 1970 were around 16 years of age in 1986 and 36 years of age in 2006. This group covers the people who were young adults for a significant part of the period covered by the NZLC. The divergence of the lines in figure 2 for people born before around 1970 illustrates the greater success in linking. The sharp drop-off in the younger ages is due to declining linking eligibility, because people under five years of age at the time of each census are excluded.

**Figure 2****Combined linked, not-linked, and theoretical population**

By year of birth, and sex

1986–2006 Censuses

**Source:** Statistics New Zealand

**Note:** In this figure, the theoretical population is the union of all individuals appearing in at least one theoretical population for the 1986–2006 Censuses. Individuals may have been linked or not linked to a previous census, and may have been linked across more than two census points.

Linking for males follows a different pattern from females (figure 2). Younger adult (aged 15 and 29 years) males are harder to link than younger adult females, as seen by the relationship between linked and not-linked records. For males, the two populations converge markedly for those born during the 1970s. These are the cohorts most affected by transitions into adulthood, movement from education to work, international travel, and frequent changes of address.

**Link rates by ethnicity**

Overall link rates by ethnicity showed variability consistent with differences in age and sex between the groups, as well as consistent with migration histories and geographic mobility. For the total population, the overall link rate was in excess of 70 percent. Māori and Pacific populations achieved a lower overall rate, generally dropping below 60 percent. Link rates for people of Asian ethnicities were higher, at around 65 percent, although still lower than the total population (table 8). People of European ethnicities are not included in the table below, but for this group the rate was overall a little higher than the total population.

**Table 8**

**Census population counts, linked records to previous census, and link rate by ethnic group**  
1986–2006 Censuses

	Census t				
	1986	1991	1996	2001	2006
Total population					
Census usually resident count at t	3,263,284	3,373,927	3,618,302	3,737,278	4,027,947
Theoretical population at t for linking to t-1	2,884,221	2,925,849	3,018,919	3,122,176	3,285,978
Number of unique links at t, t-1	2,078,428	2,220,355	2,174,322	2,170,688	2,311,093
Link rate (percent)	72.1	75.9	72.0	68.5	70.3
Māori					
Census usually resident count at t	404,776	434,847	523,372	526,281	565,328
Theoretical population at t for linking to t-1	345,190	366,714	442,290	450,813	487,841
Number of unique links at t, t-1	197,530	222,772	265,752	252,877	280,743
Link rate (percent)	57.2	60.7	60.1	56.1	57.5
Pacific					
Census usually resident count at t	130,295	167,071	202,234	231,799	265,974
Theoretical population at t for linking to t-1	97,918	121,159	157,460	178,404	211,853
Number of unique links at t, t-1	54,827	67,721	92,328	99,649	121,533
Link rate (percent)	56.0	55.9	58.6	55.9	57.4
Asian					
Census usually resident count at t	54,020	99,757	173,503	238,177	354,551
Theoretical population at t for linking to t-1	35,709	46,658	88,855	131,957	204,608
Number of unique links at t, t-1	23,382	31,113	57,992	84,396	128,389
Link rate (percent)	65.5	66.7	65.3	64.0	62.7

**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census. Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

While there are biases inherent in probabilistic linking that may result in lower precision of the links for some population subgroups, the proportion of the links affected is small. For example, table 7 shows that we linked only 2.8 percent of the theoretical population in 2006 probabilistically. This is equivalent to saying that probabilistic linking accounted for 4 percent of those records that were linked. This means that 96 percent of the links were achieved at step 1 of the deterministic stage.

Māori and Pacific populations had similar proportions of records linked probabilistically. For example, in the links from 2006 to 2001, we linked 4.0 percent of those with Māori ethnicity in the 2001–2006 census pair probabilistically. Similarly, we linked nearly 1 in 20 (4.8 percent) of successfully linked people that reported ethnicities in the Pacific grouping in 2006 probabilistically. This contrasts with almost 1 in 40 (2.6 percent) for people of Asian ethnicities in 2006.

We must reiterate that we only linked a small proportion of records probabilistically. At 2.6 percent, the proportion of probabilistic links for the Asian ethnic grouping is much lower

than that for other groupings of ethnicities. However, the proportion of the Asian population that was not theoretically linkable (the residual population) was higher than for other ethnic groupings because of migration histories – a larger proportion of people of Asian ethnicities than other ethnicities had been usually resident overseas at the previous census, and a larger proportion of New Zealand-born Asian population were still under five years of age.

One of the characteristics of probabilistic linking is that the quality of these links may be lower and may generate additional inherent bias in the data. Bias of this nature is likely to be especially problematic for relatively rare events. We can see one indication of this in the apparent differences in levels of ethnic mobility between deterministically linked data and probabilistically linked data. For example, table 9 shows that while we probabilistically linked 4.0 percent of Māori records in 2006 to 2001, the percentage of the probabilistically linked people that appear to have changed ethnic identification is significantly higher. However, overall, the proportion of links made deterministically was high, and it remains for us to investigate whether this additional bias may be largely restricted to particular age groups, especially in the younger adult working ages.

**Table 9**

**Percentage of links linked probabilistically**

By ethnic grouping and ethnic mobility  
2001–2006 Censuses

Ethnic grouping	Total in ethnic grouping 2006	Same ethnic grouping 2001 and 2006	In ethnic grouping 2006, but not 2001 (gain)	In ethnic grouping 2001, but not 2006 (loss)	Total in ethnic grouping 2006	Same ethnic grouping 2001 and 2006	In ethnic grouping 2006, but not 2001 (gain)	In ethnic grouping 2001, but not 2006 (loss)
Number of links					Percentage probabilistically linked			
Māori	280,623	247,032	33,597	26,397	4.0	3.7	6.0	6.9
Pacific	121,533	112,005	9,528	8,748	4.8	3.8	15.7	16.9
Asian	128,382	121,143	7,239	8,422	2.6	1.9	14.3	13.0

**Note:** Counts have been independently rounded to base 3.

**Source:** Statistics New Zealand

Table 10 provides a summary by ethnic grouping across the whole period currently covered by the NZLC. It covers the relationship between the theoretical populations and the number of achieved links either to the previous census or to the subsequent census. The counts in this table are consistent with the groupings as defined by the [Statistical Standard for Ethnicity 2005](#).

**Table 10****Link counts for theoretical populations**By level 1 ethnicity grouping<sup>1</sup>


1981–2006 Censuses

	European	Māori	Pacific	Asian	MELAA <sup>2</sup>	Other	Not stated
1981 Census							
Linked to previous census	..	..	..	..	..	..	..
Not linked to previous census	..	..	..	..	..	..	..
Linked to subsequent census	1,926,585	211,335	59,733	23,544	4,767	1,020	17,853
Total theoretical population	..	..	..	..	..	..	..
1986 Census							
Linked to previous census	1,850,796	197,502	51,657	23,475	1,416	375	12,468
Not linked to previous census	612,663	147,576	41,022	12,378	690	258	20,220
Linked to subsequent census	1,957,257	223,701	66,483	32,427	1,878	423	12,810
Total theoretical population	2,662,563	377,931	106,773	44,382	2,637	717	34,248
1991 Census							
Linked to previous census	1,968,348	222,726	69,873	31,113	2,151	138	5,232
Not linked to previous census	514,617	143,880	54,645	15,546	981	57	8,532
Linked to subsequent census	1,888,146	228,447	84,111	54,372	3,171	156	5,040
Total theoretical population	2,678,448	402,261	143,139	71,676	4,470	225	14,400
1996 Census							
Linked to previous census	1,911,447	265,668	92,328	57,993	4,410	567	15,585
Not linked to previous census	632,955	176,463	65,133	30,861	2,403	291	26,043
Linked to subsequent census	1,874,538	269,268	102,213	86,811	6,951	627	15,489
Total theoretical population	2,742,843	482,091	176,463	122,535	9,897	990	43,791
2001 Census							
Linked to previous census	1,864,965	252,795	99,651	84,393	6,228	405	13,881
Not linked to previous census	690,939	197,817	78,753	47,565	4,554	249	23,214
Linked to subsequent census	1,937,316	273,429	120,750	129,564	11,484	477	15,672
Total theoretical population	2,753,931	488,802	202,845	179,457	16,401	780	39,105
2006 Census							
Linked to previous census	1,675,362	280,623	121,533	128,382	10,482	295,062	11,748
Not linked to previous census	594,750	206,964	90,321	76,224	8,124	95,205	17,604
Linked to subsequent census	..	..	..	..	..	..	..
Total theoretical population	2,270,112	487,587	211,854	204,609	18,606	390,270	29,352

1. Statistical Standard for Ethnicity 2005

2. Middle Eastern / Latin American / African

**Note:** Counts have been independently rounded to base 3.**Symbol:**..figure not applicable**Source:** Statistics New Zealand



## 5 Variability in link rate for core population characteristics

This section covers the effect of overall link rates on core population characteristics. This will help explain some of the differences in linking of different subpopulations and point to areas requiring further research into biases.

### Summary of main findings

Reasons for differences observed in link rates include factors such as age and sex. There is a clear difference in link rates by age. We are less likely to be able to link younger adults than people in the mid-adult years, between 40 and 65 years of age.

Younger adults are much more mobile than other sections of the population and are more likely to incorrectly recall their address at the previous census, or to have provided incomplete or erroneous information for other key variables. Younger adults are also more likely to have been usually resident overseas at the previous census, or to have been in New Zealand but not counted in the previous census. In the former case, they will not be included in the theoretical population; in the latter they may be included in the theoretical population, but we may not have linked them.

Males are harder to link than females because they are more likely to be missed by the census and more likely to have missing or different data between one census and the next.

### Age, sex, and ethnicity

Age, sex, and ethnicity are the three most frequently used variables for population analysis. Variation of the link rates for any of these variables has implications for researchers. Any bias in the distribution of linking for different age groups impacts on longitudinal analysis of population groups as they age through a population. Differences in link rates by sex have consequences for analysis of gendered processes. Linking differences between ethnic groupings may also impact on inter-ethnic comparability.

Figure 3 shows the relationship between the age and sex profile of the theoretical population against the population of achieved links. The graph compares age by sex distributions of the theoretical population with the population of people who specified in the 2006 Census that they were of Māori ethnicity. We can clearly see the much younger profile of the Māori population. The graph shows that the linking pattern is very similar for both populations across ages.

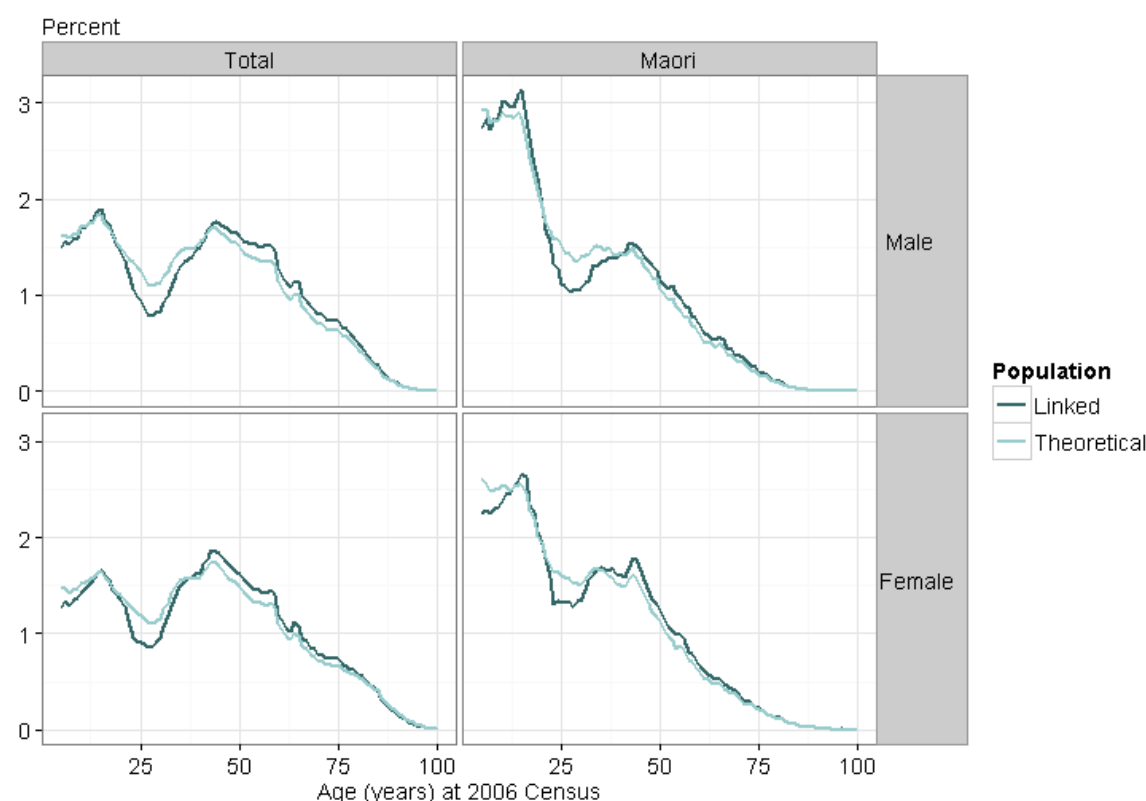
For both males and females, younger adults are under-represented in the linked populations. Male Māori children are better represented, relatively, than other children. This graph demonstrates in broad terms that the lower overall link rates for Māori is because of their younger average age and higher mobility. It is not only because of something intrinsic to people's ethnic identification.



**Figure 3****Age distribution by sex of linked and theoretical populations**

Māori and total populations

2006 Census



Source: Statistics New Zealand

**Link rates by age and sex**

Link rates primarily follow a pattern that is related to age rather than cohort, as can be seen in figure 4. The age profile of link rates follow very similar patterns for each of the linked census pairs. This is not to say that there are no cohort effects, rather the age effects dominate in this aspect of the data.

The changing link rates and biases between the census pairs reflects New Zealand's social and economic history, with the recent period since 1996 characterised by considerable international migration, as a result of larger numbers of people both arriving and leaving. Populations have also become much more mobile within New Zealand.

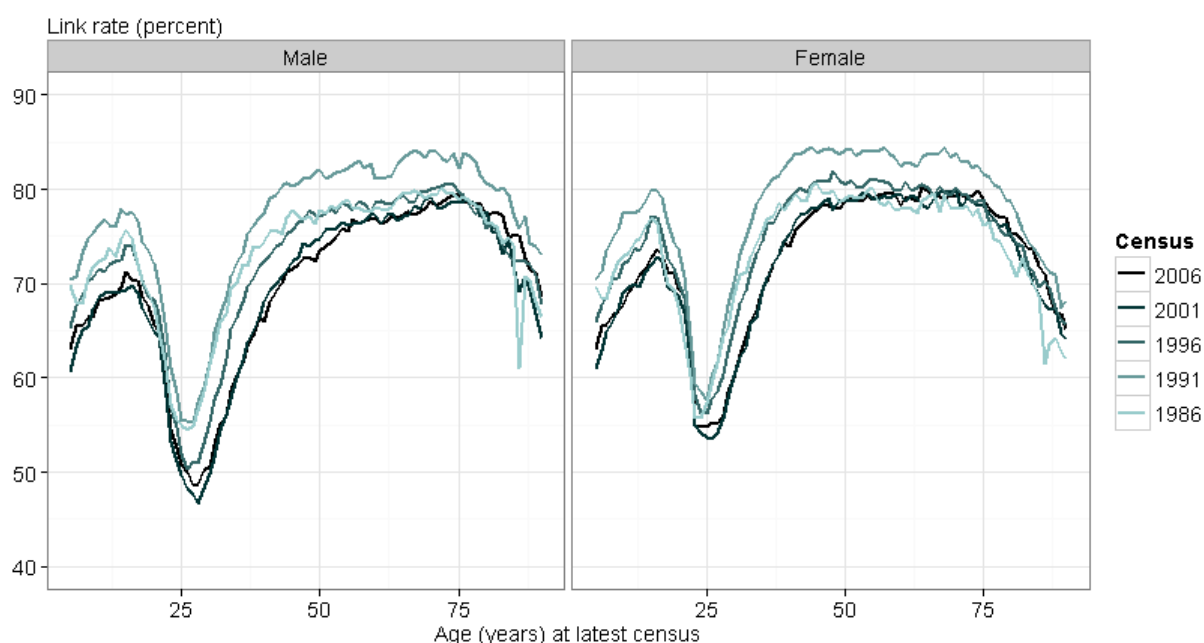
Simultaneously, public compliance in form completion has decreased and precision in the recall of details has diminished. The 1986–1991 census pair has the highest overall link rates, but reflects a period of economic difficulties and lower migration. This census pair straddled two periods of social uncertainties and transitions, but in each case the relatively low levels of population mobility constrained the effect.

A comparison of the five linked pairs shows a general progression over time of declining link rates. The 1986–1991 pair is slightly aberrant in that it has a higher link rate at almost all ages. The progressive decline in link rate is stronger among males than females. Although females exhibit the same pattern as males, the decline is more marked for males.

**Figure 4****Percentage of theoretical population at census t linked to census t-1**

By age and sex

1986–2006 Censuses



**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

**Source:** Statistics New Zealand

The link rates for the 2006 theoretical population, linking to 2001, range from almost 80 percent for adults in the 50–75-year age group, to approximately 50 percent for the 25–29-year age group. As we see in figure 4, this is a pattern typical of all census pairs. Not only are male link rates lower than females' at all ages, both the age and the cohort effects are more pronounced among males.

The behaviours that lead to these linking differences relate to age at the previous census, so that the low rates for the 25–29-year age group relate to behaviours inherent to people aged 20–24 years. The latter age group is a highly transitional one – typically they move between education and work, are involved in much more short-term international travel, and move residence often.

Two associated factors give rise to the cohort effect. Not only are the same age groups hard to link across time, but over time the cohorts are prone to carry these behaviours forward. This stems in part from changing life patterns over time; for example, people leave home at later ages than they used to, or remain more highly mobile for longer. Hence, while the graphs all show that 25-year-old females and 27-year-old males (those who were 20 and 22 years, respectively, at the previous census) are the hardest to link; both sexes have become increasingly more difficult to link over time at older ages. For females, we see this especially up to the age of 50 years, and for males up to 75 years. The reasons for this remain to be analysed.

**Link rates by age, sex, and ethnicity**

These characteristics impinge on population groups in different ways. We have noted age and cohort effects in the linking patterns above. If we plot the same data by ethnicity, as well as age and sex, we see important differences. For example, in figure 5 we see quite distinctive differences between the link rates by age and sex for people in different groupings of ethnicities. However, the general pattern is similar for all groupings, with

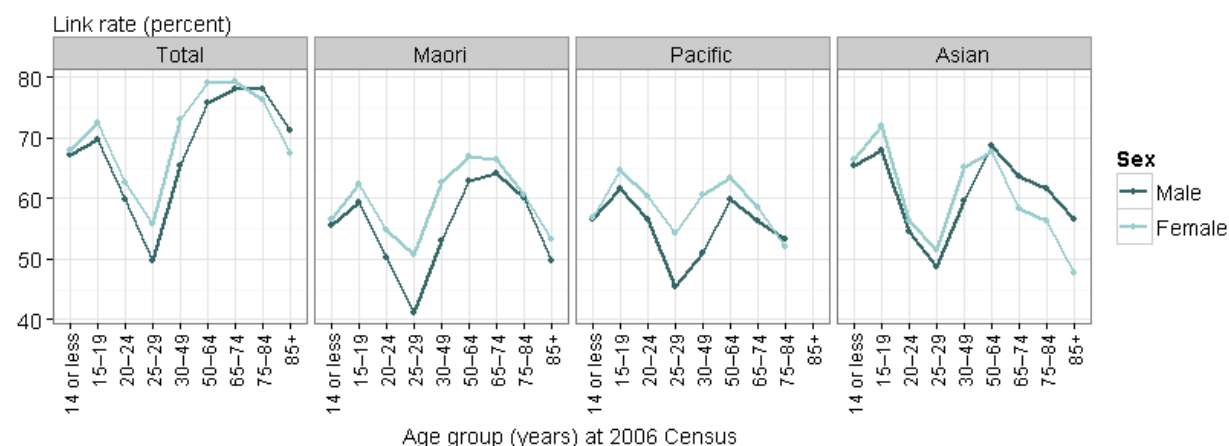
higher link rates for children and middle-aged adults, lowest rates for younger adults, and declining rates for older people.

**Figure 5**

**Percentage of theoretical populations linked to census t-1**

By age group, sex, and ethnic group

2001–2006 Censuses



**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

**Source:** Statistics New Zealand

Figure 5 illustrates that children with at least one Asian ethnicity in 2006 have high link rates. The majority of these children are New Zealand-born and have good education and language skills. There is little difference between males and females. Conversely, older people with at least one Asian ethnicity have lower link rates, with quite markedly lower link rates for females than males. This reflects several aspects of their migration histories.

People with Māori and Pacific ethnicities show lower link rates overall, with males showing significantly lower rates than females. With the exception of younger adults, where link rates for males of Māori ethnicity drop to 40 percent, Māori are more successfully linked overall than people of Pacific ethnicities. This difference is due in part to birth-place, education levels, mobility, proportion with multiple ethnicities, and residential location.

## Other population characteristics

The 2001–2006 linked population is used above to illustrate the effect of specific characteristics. The immediately striking feature is that there is very little difference with respect to these characteristics between males and females. The exception is partnership status, where older not-partnered females tend to be more successfully linked than older not-partnered males (figure 6).

### Employment, country of birth, and usual residence

The selected characteristics in figure 6 show a consistent pattern. These variables represent attributes associated with people who are more settled, more highly educated, and less mobile. For example, employed people are more likely to provide satisfactory information than unemployed people. The striking aspect of this characteristic is that the linked proportion of the theoretical population is almost independent of either age or sex. The proportion is also largely independent of age and sex for the binary New Zealand-born / overseas-born, and for people who had moved versus people who had not moved over the previous five years.

### **Qualifications**

Qualifications – whether a person has specified a formal qualification or specified they had no qualification on their census form – is dependent on age. Younger adults are more likely to be not linkable, or non-linked, if they have no qualification than if they do.

### **Home ownership and partnership status**

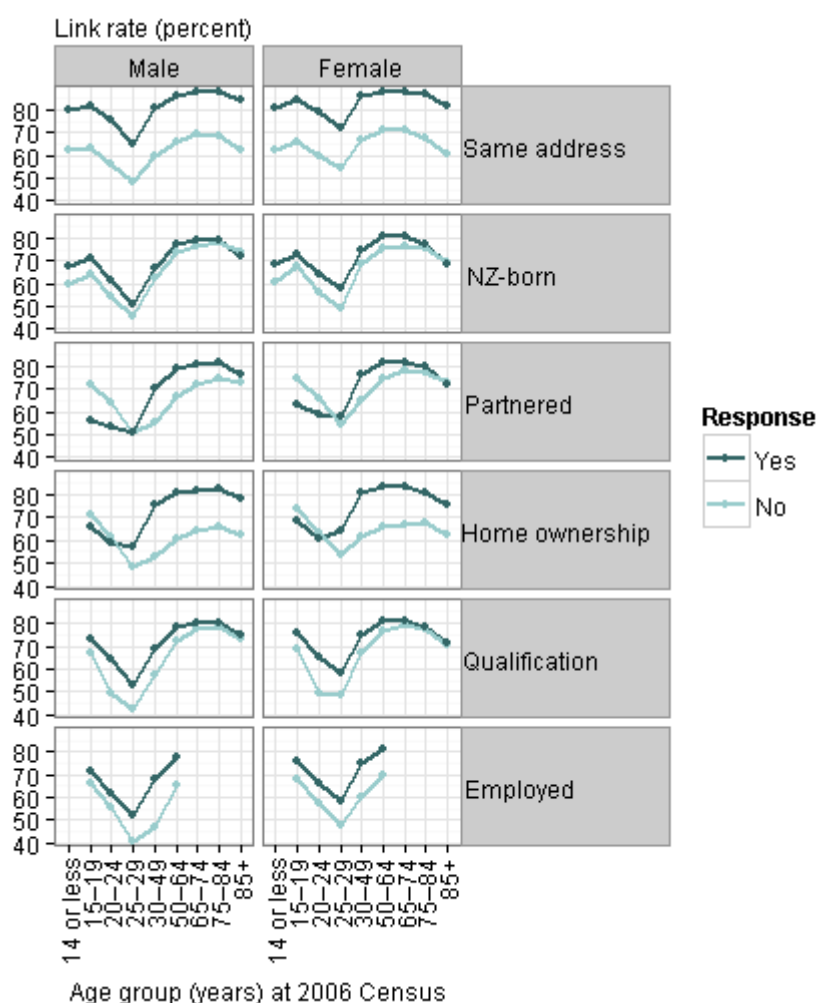
Home ownership and partnership status are different from the other characteristics compared here. They are however similar to each other. Atypically, for younger adults, both males and females are less likely to be linked if they do own their own home or are partnered. This is opposite to the expected pattern, which is exhibited by older people.

For home ownership, the differences are minor for the very young and this is more likely to be a feature of the quality of home ownership data, with few people under 25 years of age owning their own home.

For partnered people, the difference is much larger than for those who own their own home. The explanation is straightforward, however. Younger adults who have formed partnerships and/or purchased their own home, almost all changed their address in the process, and were most-likely to have done so in the previous five years. This reflects the relatively high mobility of this group, rather than a feature of the characteristic itself. Home ownership is associated with lower mobility, and we can see this in the similarity between the pattern for older home owners and the movers/non-movers in the same age groups. Conversely, older couples are more likely to have been in the same relationship for more than five years and to have moved address together, while older non-partnered people may also own their own homes and not have moved in the previous five years.

**Figure 6**

**Link rate by age group and sex for selected characteristics of the 2006 theoretical population**  
2001–2006 Censuses



**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

**Source:** Statistics New Zealand

## Occupation

The binary characteristics above show some of the typical features of the NZLC data. However, there is considerable diversity within each of the groupings at the level shown above. For example, the link rates for the employed, relative to those not employed (ie unemployed or not in the labour force), are largely consistent across age and sex. However, there are differences in the link rates within the employed population when we disaggregate it by occupation group.

Figure 7 shows link rates by occupation. We have grouped the stated occupations by the first digit of the NZSCO99 classification. Group 9 in this classification includes both labourers and employed people who did not specify their occupation. Although this is, therefore, a diverse group, it tends to include both people who have high employment churn and hard-to-link people who have either provided information that is hard to code, or no information for many fields. The high level of mobility is likely to be the key factor. Agriculture and fisheries workers are also very mobile, frequently in itinerant employment, and they show a similar pattern to the younger adult ages.

For both males and females, the first four categories (managerial, professional, and other skilled workers) have higher link rates than other categories. Those in the 20–24-year age group in these categories have a much higher link rate than this age group as a whole. The lower rates for the 25–29-year age group result from their higher mobility: they were 20–24 years old at the time of the previous census, and many will have moved from tertiary education during the intercensal period and relocated for employment.

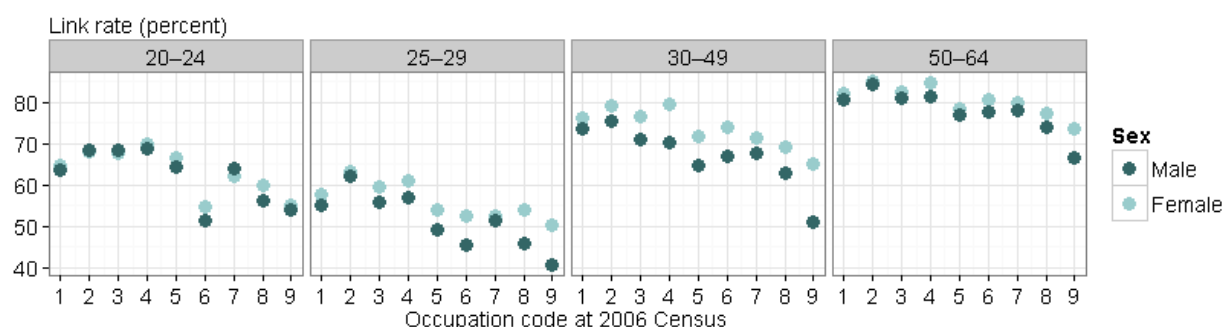
People who were in short-term tenancies, as many students are, at the time of the previous census, have greater difficulty accurately recalling where they were living when, or indeed if, they completed the previous census questionnaire.

Service and trade workers (occupation codes 5 and 7; see figure 7) cover a very diverse range of skills and experience, and the two categories have similar link rates. We are much more likely to link women than men employed in these categories. This contrasts with the first four categories, where there are smaller differences by sex.

One factor that contributes to the better link rates for women in younger age groups is that these are the prime child-rearing years, when many women work part time and are consequently less mobile.

**Figure 7**

**Link rate by age group, sex, and occupation code stated at the 2006 Census**  
2001–2006 census pair



**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census. The NZSCO99 occupation codes are

- 1 Legislators, administrators, and managers
- 2 Professionals
- 3 Technicians and associate professionals
- 4 Clerks
- 5 Service and sales workers
- 6 Agriculture and fishery workers
- 7 Trades workers
- 8 Plant and machine operators and assemblers
- 9 Elementary occupations (including residuals)

**Source:** Statistics New Zealand

The underlying causes of these variations are complex and variation in one variable is only very rarely independent of link biases in other variables.

## Region of residence

Subnational data analysis is an extremely important aspect of New Zealand demography. New Zealand is a geographically and socially diverse country, with important local characteristics. Any national-level analysis also needs to be aware of subnational

diversity that impinges on national trends. We therefore need to consider biases that may derive from any strong geographic differences in link rates.

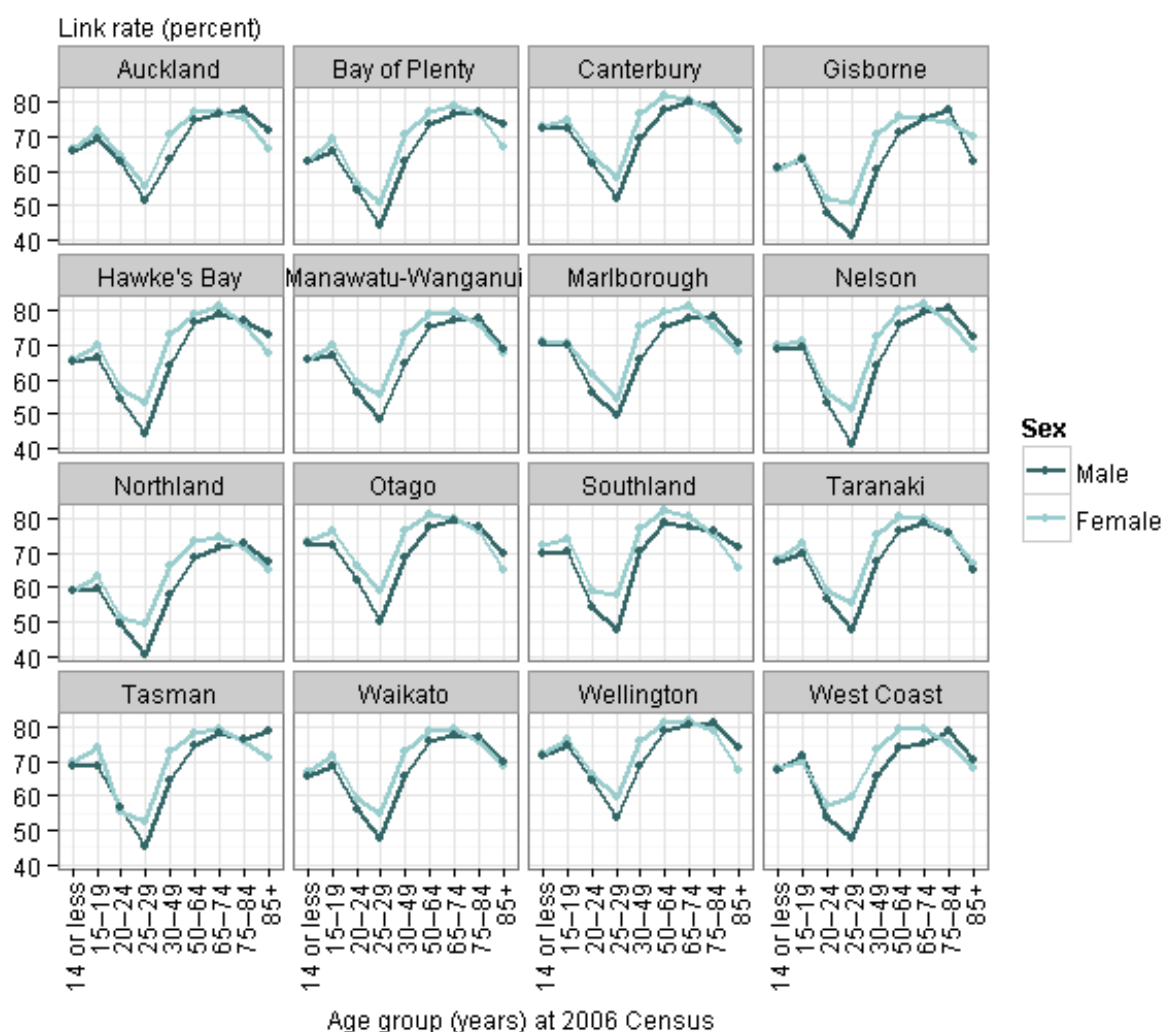
Figure 8 shows the link rates by age, sex, and region of residence at the 2006 Census. The highly urbanised regions of Canterbury, Wellington, and Auckland have almost identical link rate patterns. Regions such as Hawke's Bay, Nelson, the West Coast, and Gisborne – each with large seasonal and semi-skilled or unskilled labour forces and above-average unemployment – are characterised by lower link rates among younger adults. Seasonally employed people tend to have lower link rates because of their high level of mobility, and greater propensity to be missed by the census.

These differences tend to be quite strongly differentiated by sex. Males are more likely to not be linked. In some regions such as West Coast, the difference between male and female link rates is greater than in other regions. The West Coast region is also notable for having higher sex ratios and an older age profile than most other regions. While this is partly related to the relatively high proportion of the region's population that is rural, other socio-economic factors contribute to the pattern. The combination of high sex ratios, older median age, and low link rates suggest that linkage bias for this region will be different from other regions.

Data users should take considerations such as these into account in any comparison between regions or area types.

**Figure 8**

**Link rate by age group, sex, and region of residence in 2006**  
2001–2006 Censuses



**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

**Source:** Statistics New Zealand

## New Zealand deprivation index

The relative levels of areal deprivation contribute to the diversity in link rates between regions, seen above. We can demonstrate this by plotting link rates against a geographic classification that incorporates areal typologies. The [New Zealand socio-economic deprivation index](#) (and score) provides a suitable vehicle for this purpose. The New Zealand Deprivation Index (NZDep) groups areas into deciles according to levels of deprivation. The least-deprived areas are in decile 1 and the most-deprived areas are in decile 10.

Figures 9a and 9b show a direct correlation between deprivation, as measured by NZDep, and the link rate. The rates are independent of age and sex, with the link rates for age and sex in each decile following a very similar pattern. As levels of deprivation decrease, the link rate rises uniformly except, in a minor way, at the oldest ages for the less-deprived areas (figures 9a and 9b).

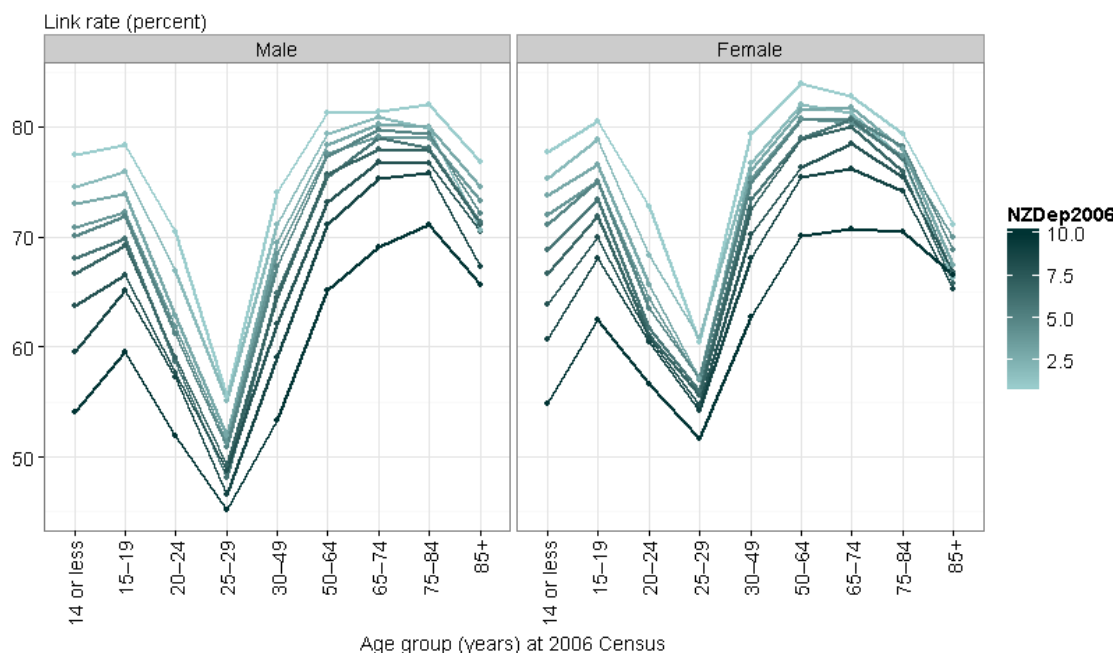
The relationship between NZDep and link rates explains a substantial part of the differences between regions, reflecting the distribution of the deprivation deciles across regions. At least in part, this derives from the close correlation between the reasons for link failure and the variables used to derive NZDep. This does not, of course, explain regional variation entirely, and residual differences may help identify other regional processes contributing to linkage bias.

**Figure 9a**

### Link rate by age group, sex, and NZDep2006<sup>1</sup>

NZDep2006 = decile 1–10

2001–2006 census pair



1. NZDep2006 is the NZ socio-economic deprivation index defined at the 2006 Census.

**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

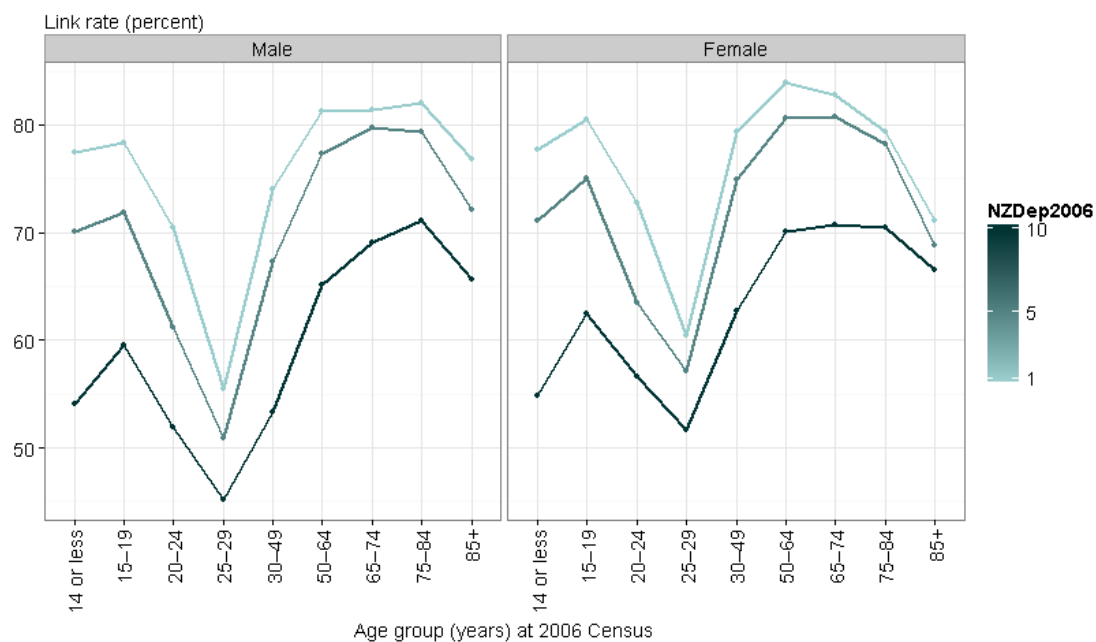
**Source:** Statistics New Zealand



**Figure 9b****Link rate by age group, sex, and NZDep2006<sup>1</sup>**

NZDep2006 = decile 1, 5, and 10

2001–2006 census pair



1. NZDep2006 is the NZ socio-economic deprivation index defined at the 2006 Census.

**Note:** The link rate is the percentage of the theoretical population at the latest census linked to the previous census.

NZDep2006 = decile 1, 5, and 10

**Source:** Statistics New Zealand



## 6 Preparing for future developments

In this section we outline preparatory work done towards the future development of the linked census as an enduring resource for population analysis and monitoring. Included in this discussion is the preliminary consideration of methodological challenges when adding 2013 Census data, and the investigation into potentially adding data from the 1976 Census.

We also discuss linking other datasets in order to make the resource a powerful tool for analysing demographic processes, and as statistical support in health and education research.

### Adding 2013 Census data to the NZLC

Creating the 2006–2013 census pair is the highest priority for the future development of the NZLC. The 2013 Census is not only the most recent census, it is also the set of census data that the majority of users need to relate their research to. To facilitate this, we consider adding 2013 Census data as urgent, in order to ensure that the NZLC is up to date and that we are prepared for the later addition of future censuses.

Developing the 2006–2013 census pair will present methodological challenges specific to this pair. Of immediate concern is the effect of having a seven-year gap between the two censuses, but also having a question that asked for respondent address five years ago (ie in 2008). We will have to accord greater weight to the question on number of years at current address, with additional work to assess the probability of having moved address two years ago or less.

This will entail developing a uniquely defined linking methodology for the 2006–2013 census pair, including further investigation of the use of family and household information to achieve and validate links. The investigation and review of this methodology and construction of the census pair is part of the NZLC phase 3 development. We plan to include the 2006–2013 census pair in Statistics NZ's microdata access environment, with the existing NZLC datasets, during phase 3.

A potential benefit of developing additional techniques to build the 2006–2013 census pair is being able to use these techniques to investigate improvements to link rates in earlier pairs. Potential options include the use of the 'years at current usual address' census question. If successful across a seven-year period, we may apply this option to ten-year periods to enable some 'leap-frog' linking across non-adjacent censuses.

### Investigating potential for including the 1976 Census

A linked 1976–1981 census pair would complete the historical set of linked pairs of censuses for which we have available electronic datasets. We put considerable effort into our investigation, because a series start date of 1976 would add the potential to investigate the demographic history of an important transitional period in New Zealand's recent past. While the ultimate outcome and assessment of viability was negative, the process that led to this outcome is described in some detail to fully illustrate some of the challenges in working with historical census data.

In order to build a 1976–1981 census pair, we required concordances that would allow us to compare the key matching variables. Unfortunately, concordances for 1976 to 1981 had been largely lost. A preliminary task was to recover and verify these concordances. An additional concern was identifying which classification sets were used during the

census coding, as some apparently contemporary classifications differed for many variables. These difficulties formed the initial focus of our analysis.

## Key tasks

In order to demonstrate the viability of developing a 1976–1981 pair, there were four principal tasks, all of which have value and relevance beyond the present project:

1. Establish a concordance of country of birth between the 1976 Census (as coded) and the 1981 Census (as coded).
2. Identify area unit codes used, identify locations of area units, and establish the area unit concordance.
3. Establish other concordances for key variables to ensure that the resulting data can be used effectively.
4. Assess the viability of creating a 1976–1981 pair.

We completed tasks 1 and 2. Task 3 is on-going as time permits, and forms part of the metadata development.

## Recovering country of birth and area unit codes

The area unit classification (AU76) and concordance of this classification to the 1981 'area unit five years ago' classification (AU5YR81) provides the mechanism for matching 1981 data back to 1976 data, based on the 1981 area unit of usual residence five years ago.

The initial pass over the data involved comparing area units of usual residence five years ago for overseas-born residents in 1981, with area units of usual residence in 1976. At least one overseas-born usual resident was in most area units in 1976, and these people were distributed by age and sex in such a way that we could confidently link at the area-unit level in the majority of cases.

The primary focus was on people who reported in 1981 that they lived at the same address in 1976, on the assumption that non-movers in 1981 with specified birthplace had a greater chance of being able to be used to identify the relevant area unit in 1976. Movers were also available, and we used them to support results. However, we considered that coding was likely to be of slightly lower quality relative to non-movers, because coders in 1981 would have needed to manually search for the origin area unit number and, hence, introduce further potential error. We also used movers to verify concordant area units in some cases.

However, before we could do this, we had to develop a concordance of countries of birth. We then used the country of birth, age, and sex to build the area unit concordance.

### *Resources available*

Resources available included:

- the electronic master files for the 1976 and 1981 Censuses
- a partial set of urban area-unit maps for 1976, 1981, and 1986
- surviving incomplete hard-copy country of birth classifications and codefiles for 1975 (which differed from the 1976 codes used in the census)
- 1981 country of birth classification and a 1981 area unit codefile list with descriptors
- historical geographic reference material such as published tables, Wise's Directories, atlases.

## Task 1: Building the country of birth concordance

The process for building the country of birth concordance was to start with overseas-born non-movers, as this group was most likely to have the most reliable data. By comparing

data at the level of single year of age and sex it was possible to identify, for larger birthplace populations, the country of birth codes for 1981 and 1976. Single year of age was used because date of birth was not collected in 1976.

Fortunately some codes and descriptors were available for both 1976 and 1981. However, it was discovered that we needed to check every code, as in some cases the codes used in the census differed from those in the classifications available, and no complete code lists were extant. We easily checked the larger countries from published country of birth tables, but data were not available for all individual countries from published tables.

We encountered some difficulties with country of birth data. There were some trivial changes such as change of country name; for example, Rhodesia to Zimbabwe. In some more complex cases, countries had split or merged. Examples include the split of the Gilbert and Ellice Islands, and the reunification of Germany.

In recovering the country code concordance, it was important to ensure that the codes concurred were those actually used. For example, New Zealand was coded to 0 for operational purposes in 1976, rather than the classification code 609. Where possible, we verified the labels associated with each code against published counts.

The purpose of the exercise at this stage was to provide an instrument to recover the area unit concordance. This meant that it was not critical to ensure that we identified all country codes, though in the event we were able to recover a near-complete codefile.

## **Task 2: Building the area unit concordance**

Having established a working draft of the country of birth concordance, data was extracted for the people born overseas in the 1981 Census, who gave an area unit of usual residence five years before. This included an indicator to distinguish between movers and non-movers. Non-movers would have been substantially easier to process in 1981, since the coders would have simply needed to code the area unit from the area unit of the current address, whereas movers required coders to search again for the relevant area unit code. This suggests that the non-movers would have a higher quality of coding.

A set of urban maps from 1976, including the 1977 revisions that applied to the 1976 Census and the equivalent urban maps for 1981 gave a good starting point for building a concordance. We were not able to find maps for rural and smaller urban areas. A paper typescript of 1975 area unit codes and names helped add labels to codes when we were sure they were equivalent to the 1977 revisions. The 1981 area unit classification was available, which supplied labels to all the 1981 New Zealand area unit codes. In some cases it was found that names had been re-used for different areas.

We extracted two sets of data. From 1981, we used year of birth, country of birth, sex, area unit five years ago, and area unit of usual residence. Other geographic scales, such as local authority and statistical division, were also included to assist with identifying broad locations. We then derived age at 1976 from the 1981 data, because year of birth is not available for 1976 Census data.

The area unit classification was radically re-ordered geographically between 1976 and 1981, with the result that previously adjacent area units in 1976 (in terms of code numbers) were no longer adjacent in 1981. Moreover, area unit naming conventions, especially in the case of splits, created difficulties. Frequently a split was labelled merely as, for example, northern part and southern part with no indication of location or original name. In a few cases the numbering convention was also inconsistent.

We have identified most of the area units satisfactorily. Exceptions are those in the block of codes 1071 to 1097 – these were not used as area units of usual residence in the 1976 Census. We have not verified a further 17 area units.

From tasks 1 and 2 we now have:

- a concordance of census codes used for country of birth in 1976 to country of birth in 1981.
- a concordance of area unit in 1976 to area unit five years ago in 1981.

### **Task 3: Ongoing work for the 1976–1981 census pair**

- Concordances of other key variables as time permits, as part of the metadata development for the longitudinal census. This will provide an additional resource for users of 1976 Census data independently of the linked census.

### **Task 4: Assessing the viability of building a 1976–1981 linked pair**

A number of features of the data emerged during this part of the exercise, which will be significant for building a linked 1976–1981 census pair.

Country of birth can vary for important migrant groups; for example, a person may be coded to 'England' in one census but 'United Kingdom' in the other. All combinations of the countries comprising the 'United Kingdom' exist in the data. We also observed variations with other groups of countries also occur. This is not unique to 1976–1981.

Because of respondent approximation and age heaping, age can vary by one or two years if both age in 1976 and year of birth in 1981 were recognised correctly. But common digit misrecognitions or errors, similar to those observed in more recent censuses, were identified, with 3s and 9s, or 2s, 1s, and 7s, for example, being easily mistaken. Errors in other than the last digit could result in ages being different by decades in some cases.

This type of error also affects other recorded fields such as country of birth, sex, and area unit. Overall however, the coding was of good quality. While recognition and coding errors are common across all censuses, the special difficulty for the construction of a 1976–1981 census pair is that age, and not date of birth, was collected in 1976. Having date of birth on one side (1981) but only age on the other (1976) will affect the quality of the linking and will also affect how well we can link records.

An exploratory exercise assessed the number of records we could deterministically link from 1981 back to 1976. This involved attempting to link the 1981 theoretical population of 2,793,021 records back to 1976. We acknowledge that, for this test, the 1981 theoretical population was not as rigorously defined as would normally be required when building a census pair. However, for the purpose of this exploratory exercise, and given the limitations in the data noted above, this is satisfactory. This test returned only 5,471 unique exact links; representing a link rate of only 0.2 percent.

Few other viable linking variables emerged. Country of birth is useful for those born overseas, but this would not add more than a few thousand links at best. We considered using Māori ethnicity, since this was available from both censuses, but the social and political events of the period contributed to an unusual level of inconsistency in ethnic reporting between Māori and Pacific ethnicities. Moreover, this would, at best, only marginally improve linked rates for Māori.

Using other variables such as the year of arrival in New Zealand would add little to what is already possible with country of birth, though years at usual residence could provide a more general discriminator to supplement age, sex, and area unit information.

Neither 1976 nor 1981 Census databases contain age and sex imputation indicators. Substitute records were not created in either census.

The unsatisfactory achieved link rate for step 1 of the deterministic linking stage is a result of:

- the lack of date of birth variables in 1976 and concerns about the quality of the age data in 1976
- the limited options for additional choices of linking/blocking variables
- not being able to identify imputed variable responses
- the quality of usual residence coding to area unit level being unknown. While we were able to recreate a near complete set of geographic concordances at area-unit level, we also identified concerns about the quality of the area unit coding.

### *Final assessment*

In summary, an assessment of the 1976 data suggests that building a 1976–1981 census pair, using the methodology available to us, is not viable at this stage. In order to reverse this decision, we would need to undertake a substantial amount of extra research. However, our assessment shows that even then the results would possibly be unsatisfactory. Given the very limited available resources and the far greater urgency to add other data (notably the 2013 Census, and birth and death registrations), we see no strategic advantage in pursuing the 1976–1981 census pair at this time.

## **Viability of linking international travel and migration data**

One of the objectives of the longitudinal census is to create a demographic tool that would enable users to analyse basic demographic processes and flows. Including migration data in the longitudinal census would, along with birth and death information, enable analysis of all types of flows into and out of the population. Linking migration data would add substantial analytical power to the longitudinal census tool, as well as contribute directly to the resource as a continuous demographic accounting tool. Linking migration data is a task for the future development of the longitudinal census. The objective here is to assess the viability of undertaking this.

### **Challenges**

Linking migration data presents some major challenges. People arrive in, and depart from, New Zealand for short periods, extended periods, and permanently. The number of border crossings is high. We are limited in our capacity to link migration data to census records by availability of reliable variables. Any information given at the time of arrival or departure about a person's actual or intended location within New Zealand may not be the same location at which they were counted in either the previous or the subsequent census.

### **Benefits**

The benefit for the analysis of both population and migration is high. Census data gives a good basis for the interpretation of mobility of people within New Zealand – at least relative to their locations at the five-yearly census points. However, the data does not provide information on intermediate international trips, on people who move out of New Zealand following a census or arrive prior to a census. This information would add considerably to our information on the exchange of people between countries and to our understanding of diasporic behaviours.

Linking migration data to the census would also contribute information to our migration knowledge. It would add information on ethnicity, educational qualifications, occupations, and work histories to the analysis of migration. This, in turn, would contribute to the production of population ethnic estimates and projections, and other processes. There is, therefore, significant benefit in linking migration events to the NZLC.

## Linking test

A sample of 2,000 migration records were selected from June 2001 (N=999) and January 2006 (N=1001). This allowed an examination of data taken from soon after the 2001 Census and shortly before the 2006 Census. We selected data for a period when both birthplace and date of birth were available, and excluded overseas visitors as we assumed we would not be able to match these to the census, even if they were in New Zealand at the time of a census and actually counted. However, we included both permanent and long-term (PLT) migrants and short-term New Zealand resident travellers, as they are of interest to the resident population of New Zealand. New Zealand resident travellers made up most of the sample, and allowed a good test of whether linking was likely to be successful.

We linked departures to the 2001 Census, and arrivals to the 2006 Census. We also tested links where people may have arrived and departed during the intercensal period, and so would have appeared in both censuses. We need to further investigate whether we can establish links that may allow leap-frog links – if a person was away during a particular census but had been present for a previous or subsequent census.

**Table 10**

### Number of migration records in test sample, by passenger type June 2001 and January 2006

Passenger type	June 2001	January 2006
Short-term NZ resident travellers	961	957
Long-term overseas visitors	0	11
Long-term NZ resident migrants	0	11
Permanent and long-term migrants	40	20

**Source:** Statistics New Zealand

## Results

An initial assessment suggests the available address information within the migration data is insufficient for effective linking. The quality of location information in migration data has varied greatly over time, and has never been sound at anything close to the granularity required for linking. The quality of location information requires substantially more research, and the limited number of alternative linking variables renders any linking of historical migration data difficult.

However, for recent and on-going border crossings there is the possibility of using additional information from both migration and census data, which suggests that future linking may be a real possibility. Without this additional information, linking is likely to be very sparse. Linking of historical data is likely to not be possible given that for the earlier years all international travel and migration data was sampled.

## Recommendations

We propose postponing further investigation of migration to census linking until we can fully evaluate available information. Linking the 2013 Census to the 2006 Census may have additional information available to enable high-quality linking. This will greatly contribute to our knowledge of how to link international travel and migration data with census data.

The initial longitudinal datasets will be limited to the 1981–2013 Censuses, along with births and deaths for at least the later part of this period. We propose to postpone linking migration data until we have linked census and vitals data. We expect that we will only be able at best to link migration data from May 2013 at the earliest, and that the potential to

link back to the 2013 Census will be limited to those who were in New Zealand on 5 March 2013, and left after the end of April 2013.

## **Linking families and dwellings**

### **Linking families**

During development of the linked census pairs, we observed the potential to find additional links, both within pairs and between pairs, if we also linked dwellings and families. This would, moreover, enable the analysis of households and of families.

To facilitate the analysis of families a series of family indicators have been derived. The Data Users' Guide describes these more fully. The family indicators are attached to the individual records. They are designed to identify whether the person belongs to a family for which all or some other family members are also linked in the later census in each census pair.

For those linked families, we can derive longitudinal changes to family membership and family composition. This process will enable a large range of research topics into family structures, family well-being, and many aspects of social environmental demography.

These indicators will also provide a potential pathway for improving the data linking over time. Additional individual links may be possible if not all members of a family are linked. It may enable a missing member to be linked through family connections with members who are linked. Link failures can arise if relevant information is missing, someone is away from home on census night, or has been recorded only by way of a substitute form. Additional links may also be possible across more than one pair using this information.

### **Linking dwellings**

A linked set of dwellings would provide a valuable component of the longitudinal census for researchers dealing with areal statistics. Users can establish the majority of links between private dwellings from the data in the existing census pairs. We have yet to investigate the potential of the residual occupied private dwellings in which no links have been established, together with links to identified unoccupied dwellings that were occupied at a neighbouring census date.

We could also link non-private dwellings, as this would also enable information on dwelling-type changes, and transitions for some dwellings between private and non-private dwellings. These changes can occur because of real-world changes in the use of the dwelling, as well as by changes in coding practices between censuses. Linked dwelling information is another pathway towards improving the link rates in the longitudinal census and for adding value to the data for use in:

- dwelling estimates
- household and family estimates and projections
- housing dynamics
- maintaining an address register
- analysing longitudinal crowding.

This additional data resource will form a part of the future development of the longitudinal census.

## **Linking birth and death registration records to the NZLC**

The potential for the NZLC to facilitate research on socio-economic characteristics of fertility and mortality outcomes, as well as health and educational related differences, is high.

Initial investigation into adding birth and death registration confirms the value of adding this information to the NZLC. The process entails linking deaths to census records in the



immediately prior census, and births to the immediately following census. The [New Zealand Census Mortality Study \(NZCMS\)](#) has linked death registrations in the years immediately following a census to the census records, with cause of death information drawn from [New Zealand Health Information Service](#) mortality data. The NZLC project has, independently of the NZCMS, investigated the feasibility of linking birth and death registrations to the nearest census, followed by linking to the relevant census pair.

For the NZLC, eligible death records include registrations of New Zealand residents occurring in the five-year period following a census, and who were eligible to be counted at the last census. Similarly, eligible birth records include registrations in the five-year period before a census to New Zealand resident mothers.

### **Eligible census records**

Eligible census records for deterministic linking to birth occurrences in the five-year period before a census are the people who:

- stated they were not born five years ago
- stated New Zealand as their country of birth
- provided perfect information on the key linking variables (day, month, and year of birth; sex; and usual meshblock code).

Census records eligible for linking to deaths occurring in the five-year period following a census are the people who:

- provided perfect information on the key linking variables (day, month, and year of birth; sex; and usual meshblock code).

### **Deterministic linking process**

We applied a simple step-wise deterministic linking methodology by blocking on the date-of-birth variables, sex, and geographic area of residence (in birth records, the mother's address is used as a proxy for the child's address). The approach was step-wise based on the geographic location recorded, progressing up the geographic hierarchy at three different levels of the classification (meshblock, area unit, territorial authority). Each step identified unique links. Residual records that were not linked but still had potential to be linked, were passed on to the following step that used a broader geography.

The preliminary investigation into linking births and deaths to the nearest census was exploratory and requires further refinement of the methodology followed by a review of the resulting link rates. For the three censuses, 1996, 2001, and 2006, however, good results were obtained from the initial deterministic linking process. Registrations of events before 1991 produced significantly lower link rates because of limitations in the address coding. For this reason we have excluded initial results of linking births to the 1986–1991 Censuses and deaths to the 1981–1986 Censuses from this report.

### **Results**

Table 11 uses the 2006 Census as an example. The table summarises the steps outlined above and presents the results of linking births and deaths records to the census. For birth registrations, 56 percent of the births were linked to the census deterministically, of which 70 percent were linked at the first step. For death registrations, 75 percent were linked, of which 82 percent were linked at the first step.

Linking of births required two extra steps that were not relevant for deaths. Multiple births present a difficulty if it is not possible to identify which baby belongs to which census record. Step 4 looked for links that could be made in these situations. The final step made use of information about the mother and looked in the household or family of which she was a member. Each of these steps contributed less than 2 percentage points to the links.

**Table 11****Number of births and deaths records linked to the 2006 Census, and link rates for deterministic linking**

For births occurrences 2001–2006 and deaths occurrences 2006–2011

Deterministic linking step (blocking on date of birth, sex, address)	Example: Linking of births (2001–2006) to 2006 Census		Example: Linking of deaths (2006–2011) to 2006 Census	
	Unique links at step (000)	Cumulative link rate (percent of eligible events)	Unique links at step (000)	Cumulative link rate (percent of eligible events)
Step 1: Meshblock code of address in eligible records	112.4	39.1	83.8	58.7
Step 2: Area unit code of address in residual records from step 1	24.5	47.6	6.9	63.5
Step 3: Territorial authority code of address in residual records from step 2	18.6	54.1	16.5	75.1
Step 4 (births): Same-sex multiple births in residual records from step 3	3.9	55.4	..	..
Step 5 (births): Mother's date of birth and household identifier in census	1.8	56.0	..	..
<b>Total</b>	<b>161.2</b>	<b>56.0</b>	<b>102.7</b>	<b>75.1</b>

**Symbol:**

.. figure not applicable

**Source:** Statistics New Zealand

Further investigation of these data will consider ways of refining the linking methodology. Other variables included in births and deaths registrations are also available in census data, which are potential linking variables. The additional shared information is country of birth (of the mother in births data), ethnicities (of both mother and child in births data), Māori descent, marital status, occupation, years in New Zealand (deaths data), and occupation. However, we have not considered these variables in the initial deterministic linking stage due to varying, and often unknown, quality of this information in the registration data. Most of these variables may also, independently of quality concerns, validly change between collections. We will explore a probabilistic linking stage using combinations of blocking and linking variables at a later NZLC development phase.

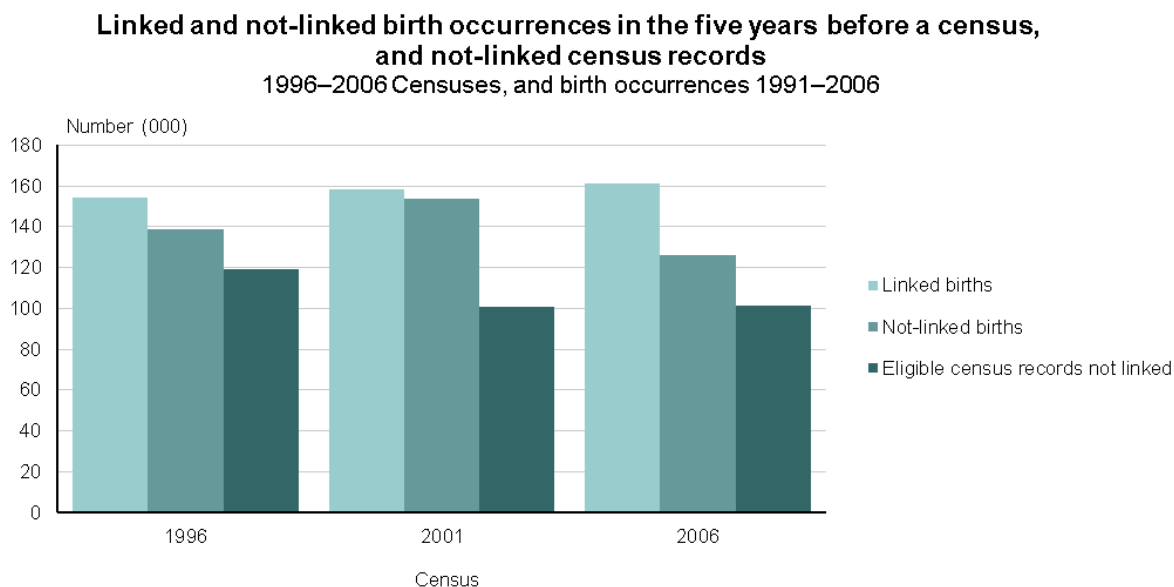
**Linked and non-linked records***Births*

Results from linking five-year periods of birth occurrences preceding each of the 1996–2006 Censuses, respectively, showed a consistent proportion of just over one-half of the birth records linked to the nearest census (52.6 percent in 1996, 50.5 in 2001, and 56.0 in 2006).

Figure 10 illustrates the relationship between the linked and non-linked birth registrations, and the residual eligible census records. Noticeably, the number of non-linked birth

records is consistently higher than the number of non-linked eligible census records. Reasons for this difference include the emigration of younger adults, who were leaving New Zealand with children born during the five-year period before a census. The child may have died, though the number of infant deaths is relatively small. More significant, in terms of the effect on the numbers of unlinked births, is the census undercount – parents may not have returned a census form for the youngest members of their family.

**Figure 10**



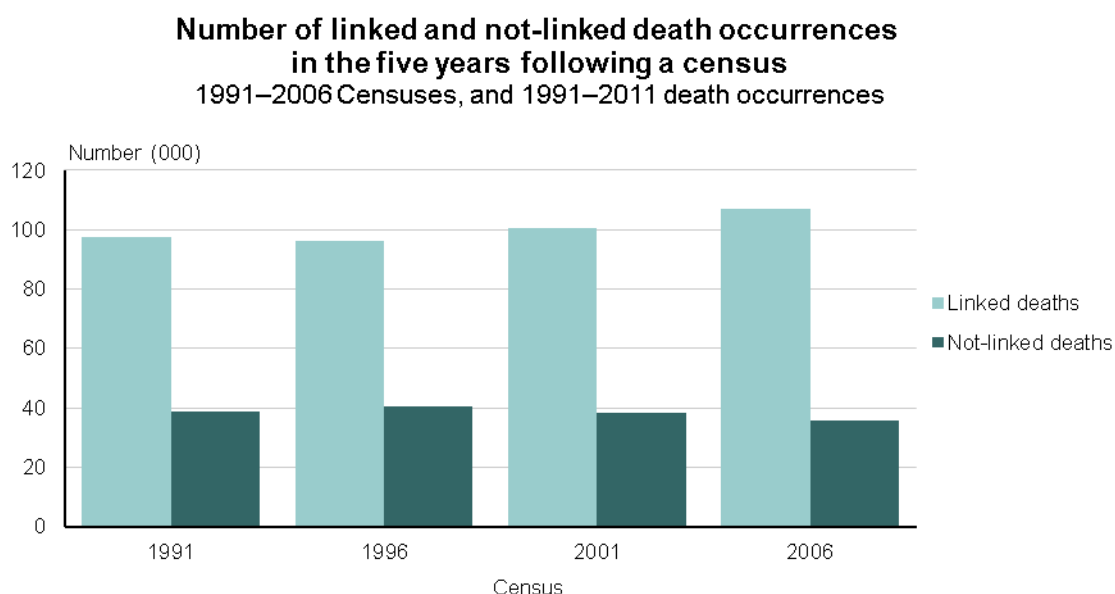
Source: Statistics New Zealand

#### *Impact of migration data on link rates for births*

International travel and migration has a direct but complex impact on the potential link rates. For example, in the 2001–2006 intercensal period, around 14,800 people who would have been aged under five years at the time of the 2006 Census, had permanently arrived from overseas on a PLT basis. Among these were more than 1,500 people who had been born in New Zealand for whom there should be both an eligible birth record and a valid census record. During the same intercensal period, around 11,600 PLT migrants, who were aged under five years in 2006, left New Zealand. The vast majority (88 percent) of the departures were New Zealand-born, for which there would have been an eligible but unlinked birth record. These departures account for around 10 percent of the unlinked births.

#### *Deaths*

Figure 11 shows the number of deaths able to be linked by the deterministic linking process to each of the 1991–2006 Censuses. The proportion of linked death records to the number of eligible death records, or the link rates, was between 70 percent in 1996 and 75 percent in 2006.

**Figure 11**

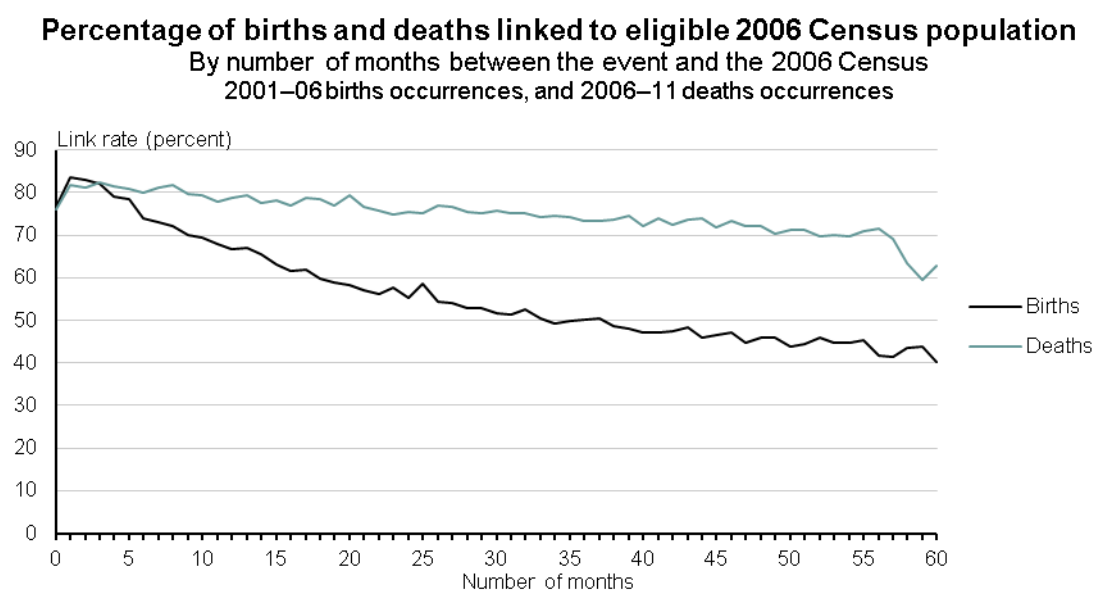
### Link rate attrition for births and deaths

Figure 12 shows the attrition in the link rates for both births and deaths. The ability to link deteriorates as the temporal distance between the events increases. For example, when linking between censuses, both non-response and failure to accurately recall affect respondents' answers to the address five years ago question.

In the case of births and deaths registrations, no question asks where the deceased lived at the last census, nor where a newborn is likely to be living at the time of the next census. Therefore, the stated residence of the deceased at the time of death, and usual address of the newborn's mother are used to link to the nearest census.

Figure 12 also shows the attrition rate when linking births to the 2006 Census as time before the census increases. Link rates are high for the six months before the census, but drop off as time before the census increases. This pattern suggests that location is stable for babies up to six months of age – when high link rates of around 80 percent are achieved. The link rate drops off fairly sharply to around 50 percent for children entering the education system around three years of age, before settling to a steady decline to around 45 percent by the age of five years.

In the case of deaths, most deaths occur at older ages among less-mobile populations. The main linking difficulty is where people have moved into residential care after the census. Nevertheless, this usually involves no more than one move, and figure 12 shows that the attrition of the link rates as the time from census increases is relatively small. Link rates drop from 80 percent for deaths close to the census, to around 70 percent even five years after the census.

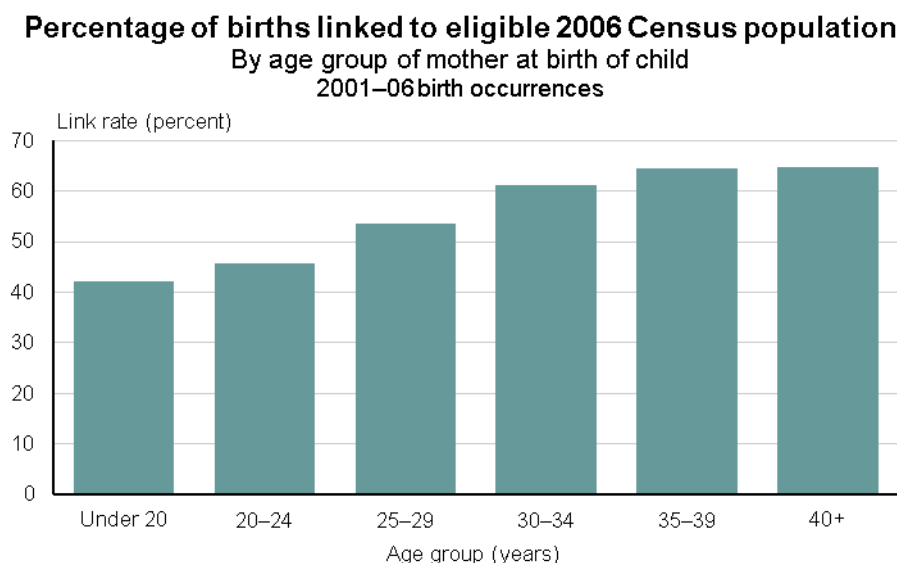
**Figure 12**

### Characteristics of births and deaths linkage to the census

The relationship between deaths and the census differ from that between births and the census. While there are similarities in linking, each has additional characteristics that affect linkability. Many of these differences are related to the characteristics such as age at death or, in the case of births, the age of the mother.

#### *Births*

Births, for example, typically occur to mothers in the highly mobile younger adult years. Young mothers are usually in the process of forming partnerships, possibly moving for their child's education, or transitioning to home ownership. They may record a temporary address on the birth registration. The similarity of the pattern of link rates by age of mother suggests that at least part of the attrition rate is driven by link rates in the census pairs for mothers under the age of 25 years at the time of their child's birth, as illustrated in figure 13.

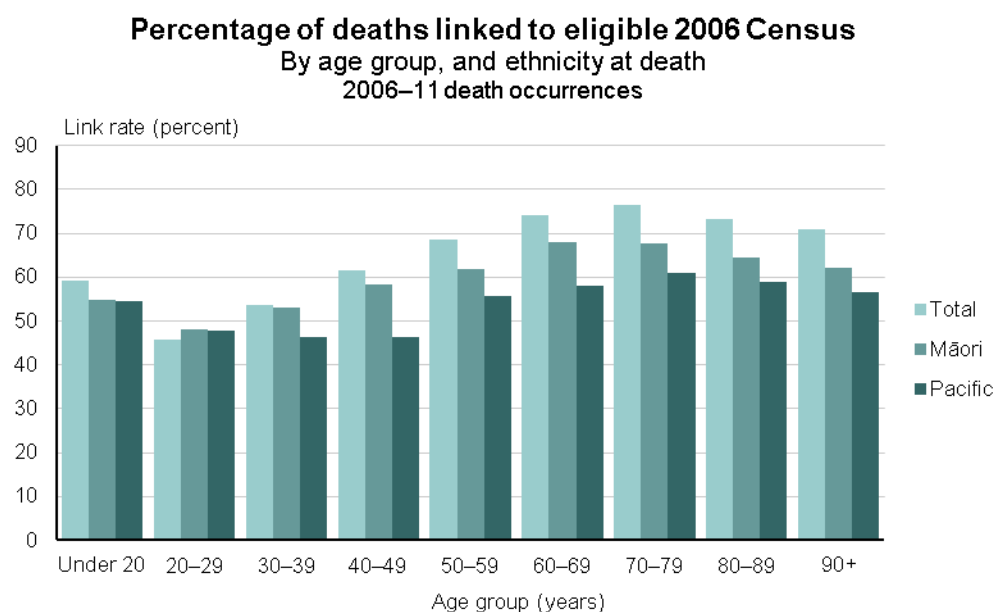
**Figure 13**

### *Deaths*

There is a correlation between link rates and age. Because New Zealand has relatively low infant mortality, most deaths occur at older ages. Link rates in the census pairs vary sharply by age and we would expect that, for similar reasons, linking deaths back to a census record would also be subject to an age bias (as shown in figure 14). Typically, deaths for 20–29-year-olds are the most difficult to link back to a census record. Part of the reason is that this age group is a highly mobile sector of the population, so the recorded address for the deceased is more likely to differ from their recorded census address, and the census record is more likely to be missing the key information.

There is also a correlation between ethnicity and link rates. In the case of deaths, death registrations record ethnicity by proxy. Unexpectedly, the link rate for both Māori and Pacific deaths in the 20–29-year age group is marginally better than for total deaths, implying that deaths among other ethnicities are likely to be more difficult to link. Further research is required to explain the reason for this, but users should note that relatively few deaths occur in this age group.

Nevertheless, the link rates for both Pacific and Māori deaths are lower at all other ages. Because the quality of the blocking variables in the death registration data is known to be higher than that of the census, with extremely small levels of missing data, we would expect that the ethnic and age patterns of linking are very similar to those seen in the 2001–2006 linked census pair.

**Figure 14**

### Future development of births and deaths linkage to the NZLC

We will investigate other aspects of the inherent biases in linking births and deaths in the next stage of the NZLC development. The phase 3 development of the NZLC will refine the current deterministic linking methodology followed by a probabilistic linking step. In phase 3, the final linked records of intercensal birth and death registrations to the nearest census will be subsequently linked to the relevant census pair, and included in the NZLC database.

The addition of other administrative data sources will be investigated as part of future developments of the NZLC. Of considerable value would be the inclusion of cause-of-death data. This would provide a rich data source for analysing the relationship between life course and work histories in relation to mortality, for the major causes of death. Linking this data to hospitalisation data also has potential for examining the relationship between life histories, morbidity, and mortality, in ways currently not possible.

Similarly, linking births data provides important information, such as birth weights and gestation periods. But there is additional potential to link to childhood health data and education data to eventually provide, as the time scale of the longitudinal census expands, a rich source of information to study well-being from birth to death.

The potential for education-data analysis is also high. The primary interest is between educational performance and outcomes. While the qualifications collected by the census do provide a useful frame, there is little information on when and where respondents gained their qualifications. Of policy interest is the efficacy of educational development strategies, and the relationship between life course outcomes for people at different educational levels.

### Microdata access development

Currently, the NZLC database is only available in the Statistics NZ Data Lab environment. The key reasons for this are that the data source is still being developed, and that Statistics NZ is legally required to ensure the security and confidentiality of the data. Statistics NZ is investigating future developments, including extending access to include secure remote access to the NZLC for approved researchers, and the derivation of a set

of weights to compensate for link bias in longitudinal datasets. At a later stage, and as resources permit, there is the potential to disseminate longitudinal tabular summaries on socio-economic topics.

Prerequisites for data users include full metadata, clear instructions on data release, constraints related to confidentiality, and symmetrical datasets where possible. The census pairs currently have a unique reference ID for each of the 7.3 million records included in at least one census theoretical population (linked or not linked to a previous census), which enables linked records to be joined across pairs and across specific subsets of data. We are continuing with work on optimising these datasets and ensuring they meet user requirements.





## 7 Summary of recommended future developments

The recommended future developments for the NZLC include:

1. Creation of the 2006–2013 census pair

This will include refining methodologies as discussed in chapter 6, and linking the 2013 Census data to the 2006 Census data. We will then include the 2006–2013 census pair in the NZLC microdata access product.

2. Confidentiality rules

Confidentiality rules will be established to enable dissemination of longitudinal tabular outputs, and analytical and research outputs in a form that will prevent disclosure of confidential individual information. These will include rules for both weighted and unweighted counts. A set of interim working rules has been prepared and is currently in use. These interim rules make use of the methodological standards used in the census and in other social collections.

3. Finalising linking of birth and death registrations

This will involve linking birth registrations to the nearest later census, and death registrations to the nearest previous census. Because of data quality limitations, we initially intend to cover the period 1991 to 2006. These links will be included in the NZLC microdata access product. While most deaths will be linked to records in the theoretical population of the previous census, some deaths will be linked to the residual population of the previous census. Similarly, births linked to the nearest census will appear in the census pair only if that record has been linked to the subsequent census. All other birth links will be to records that have not been linked. We will investigate consequential linking bias.

4. Further investigate into linking international travel and migration data

This would include investigating the time depth of the data eligible for linking. We will also investigate future developments in both migration and census information that may enable good-quality linking between censuses and international travel and migration events.

5. Investigate linking to other administrative data sources

We will investigate administrative data sources as potential inclusions. For example, adding cause of death data to linked death registrations would greatly enhance research into mortality and morbidity topics. External agencies currently hold these data. We propose here to investigate the processes and protocols of adding this information to the NZLC.

6. Refine family- and household-level analytical capability

Currently, the NZLC includes indicators that identify completely linked and partially linked families. We will review and refine these indicators to enable sophisticated family- and household-level analysis. A number of current research proposals will, if they proceed, add considerably to our understanding of what is needed, and what needs to be included, to improve analysis at this level.

7. Statistical analysis of bias, and development of weighting methodology

Weighted data will be required for many purposes. Weighting of longitudinal data, as part of the NZLC, will require substantial in-depth statistical analysis of bias and coverage. We propose investigating the possibility of creating sets of weights that may be applied in various research scenarios.



---

## References

Brown, P and Gray, A. (2009). Inter-ethnic mobility between the 2001 and 2006 Censuses: The statistical impact of the 'New Zealander' response. In Statistics New Zealand, [Final report of a review of the official ethnicity statistical standard 2009](#) [sic: sc. 2005] (pp27–36). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Didham, R (2005). [Understanding and working with ethnicity data](#). Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Errington, C, Cotterell, G, von Randow, M, & Milligan, S (2008). [A guide to using data from the New Zealand census: 1981–2006](#). A University of Auckland Family and Whānau Wellbeing Project paper. Available from [www.stats.govt.nz](http://www.stats.govt.nz).

Statistics New Zealand (2013). [Developing a historical longitudinal census dataset in New Zealand: A feasibility study](#). Available from [www.stats.govt.nz](http://www.stats.govt.nz).