THE UNIVERSITY OF AUCKLAND
FACULTY OF SCIENCE
Department of Statistics

Department of Statistics, University of Auckland

Summer Scholarship 2014-2015

Construction of life-course variables for the New Zealand Longitudinal Census (NZLC)

Name: Chris Liu

Supervisor: Roy Lay-Yee (COMPASS), Alan Lee (Stats)

Degree: Master of Science

Host Unit: COMPASS

Summary

The New Zealand Longitudinal Census has linked individuals across the 1981-2006 New Zealand Censuses. To make best use of this dataset requries that life-course variables are derived by combining data across the six censuses. However, the questionnaires, items, and formats in each census are not necessarily the same. The aim of this project is to harmonise data variables across the censuses from 1981 to 2006 to allow the assessment of change over time; some life-course variables may need to be derived from a combination of existing variables.

The results from this project will enable future studies based on the New Zealand Longitudinal Census, for example, the influence of life-course variables on the risk of mortality.

Abstract

The aim of this project is to use data across 6 censuses (1981-2006) to create ‘life-course variables’ for each person, e.g. socio-economic status, health, education, work, family ties, and cultural identity.

Raw variables in each census were identified and harmonised, so they are consistent from 1981-2006 to allow the assessment of change over time.

The following variables were created for each census: Age, Ethnicity, Iwi, NZSEI/E&I, NZ Deprivation and Household income quintiles, Unemployment, Education, Welfare receipt, Living alone, Partnership, Household size, Overcrowding index, Housing tenure, Moved in last year/last 5 years, Born in NZ, Number of years in NZ, Smoking, Language, Ethnic density, Religion, Long-term health problem or disability, Access to phone, internet and car, Voluntary work or caring.

In addition, the following variables were created to characterise the overall census data, indicating the number of individuals who were: In NZSEI/E&I classes 5&6, In lowest NZ Deprivation quintile, In lowest household income quintile, Unemployed, On welfare, Living alone, In overcrowded residence, In rented accommodation, Moved in last year, Moved in last 5 years, Smoker, Religious.

Furthermore, education level changes were tracked across censuses.

SAS programmes used to create the new data sets as well as a data dictionary documenting the life-course variables and their characteristics were produced.

The results of this project will provide an infrastructure for further studies based on census data to describe trends or for longitudinal analysis.

Table of contents

List of figures

1.0 Introduction

The New Zealand Longitudinal Census has linked individuals across the 1981-2006 New Zealand Censuses. Previously each Censuswas only a snapshot at a single time point, but now individuals can be tracked through their life course. This enables the assessment of associations of various factors with various outcomes through the life course. In each census, a proportion of individuals were identified as eligible for linking to the previous census; this eligible population was restricted to those people who:

- were old enough to have been alive at the last census (older than 5 years old)
- were in New Zealand at the previous census
- had completed a census form at the previous census (Didham et al, 2014)

Figure 1: Population 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	4027947	3737277	3618303	3373926	3263283
Theoretical population available for linking to census t-1	3285978	3122175	3018918	2925849	2884221
Percentage of usually resident population available for linking	81.6	83.5	83.4	86.7	88.4

Source: New Zealand longitudinal census 1981–2006

A problem with the census data is that census questions in different years have different formats since 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." (Didham et al, 2014). Therefore data variables are not necessarily consistent over censuses.

The aim of this project was to use data across 6 censuses (1981-2006) to create 'life-course variables' for each person, e.g. socio-economic status, health, education, work, family ties, and cultural identity. Data variables needed to be harmonised (made consistent) from 1981-2006 to allow the assessment of change over time, and new variables may be needed to be derived from a combination of existing variables. This

would make the New Zealand Longitudinal Census more usable and enable further studies based on these data.

Tasks of this project included (1) identify the raw variables in each census, (2) harmonise variables across censuses, (3) derive variables according to specification, (4) track changes in variable characteristics across censuses, and (5) derive cross-census characteristics.

This report will describe the methods and processes of harmonising variables, results, as well as concerns, and possible future work. It will conclude on how this project can be useful for future studies and analysis.

2.0 Methods

There are two main steps involved.

2.1 Identify the raw variables in each census

At the beginning, there were 29 data sets including individual census set, family set, geography set, ethnicity set, and spine set (containing general information for all records from 1981 to 2006). There were also metadata files including database design, data concordances and simple data dictionaries for each type of data set.

The Data Design file contains all the variables in each data set, I checked these variables names with data dictionaries and questionaires for each year to identify what these variables represented and what each variable category represented.

There were some problems at the beginning.

- 1981, 1986 and 1991 data dictionaries were not available in the project data folder.
- All dwelling data sets were not available.
- Dwelling numbers in 1996 data set were incorrect, they were actually 1991 dwelling numbers.

We requested these data from Statistics New Zealand and eventually obtained the correct data.

There were additional problems.

- New Zealand Socio-economical Index and Elley & Irving Index were not available. Eventually I had to create them using a previous report.
- Religion classification was not consistent over the years, especially 1986 had no general classification data. Eventually I went through all religion categories and decided which religions were for example Christian.
- 1981 Amenities variables contained household information about Television, Telephone and so on. They were named A,B,C,D and E, however we did not know what they actually represented, I compared the counts in our data set with the 1981 census result, and determined which variable represented Telephone.
- Mesh block numbers used in 1981 and 1986 used the 2001 classification, other years used the 2006 classification, I converted them to the 2006 classification using the data concordances file provided by Statistics New Zealand.
- The 1986 Usual residence indicator had missing values. I modified it as: if family code is 9 (guest/visitor) and family number is missing then usual residence is 2 (No), otherwise it is 1 (Yes).

2.2 Harmonise and Derive variables according to specification

Once I determined the categories in each variable using data dictionaries for each census, the next step was to harmonise them and create new variables.

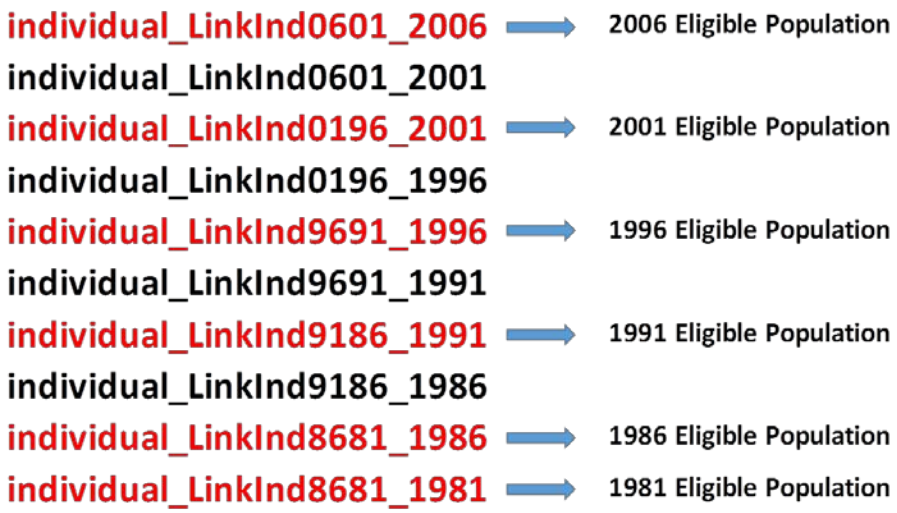
The project provided a list of variables we needed to create as well as their categories. I classified all variable categories into more general categories which were consistent over all censuses. This involved reading through data dictionaries for each census. I also used two reports produce by COMPASS and Statistics New Zealand: “A Guide To Using Data From The New Zealand Census: 1981-2006” (Errington et al 2008), and “Family Wellbeing Indicators” (Milligan et al 2006). They covered information about variables in each census, their comparability, their categories in each census and some general classifications (for example, religion variables had around 200 categories; the report classified them into general categories such as No Religion,Buddhist, Christian, Hindu, Islam/Muslim, Judaism, Maori Christian, Spiritualism/ New Age Religions and Other). Most of the variables we required were covered by these reports; they provided us a framework for this project.

Once the categories were classified, the next step was to create new variables using SAS Enterprise Guide in the Statistics New Zealand Datalab.

Instead of six data sets for six censuses, there were 10 data sets containing records for the theoretical population in each census, since censuses were paired with adjacent ones.

I checked each data set and determined the data sets that contained all records we needed, as shown in Figure 2.

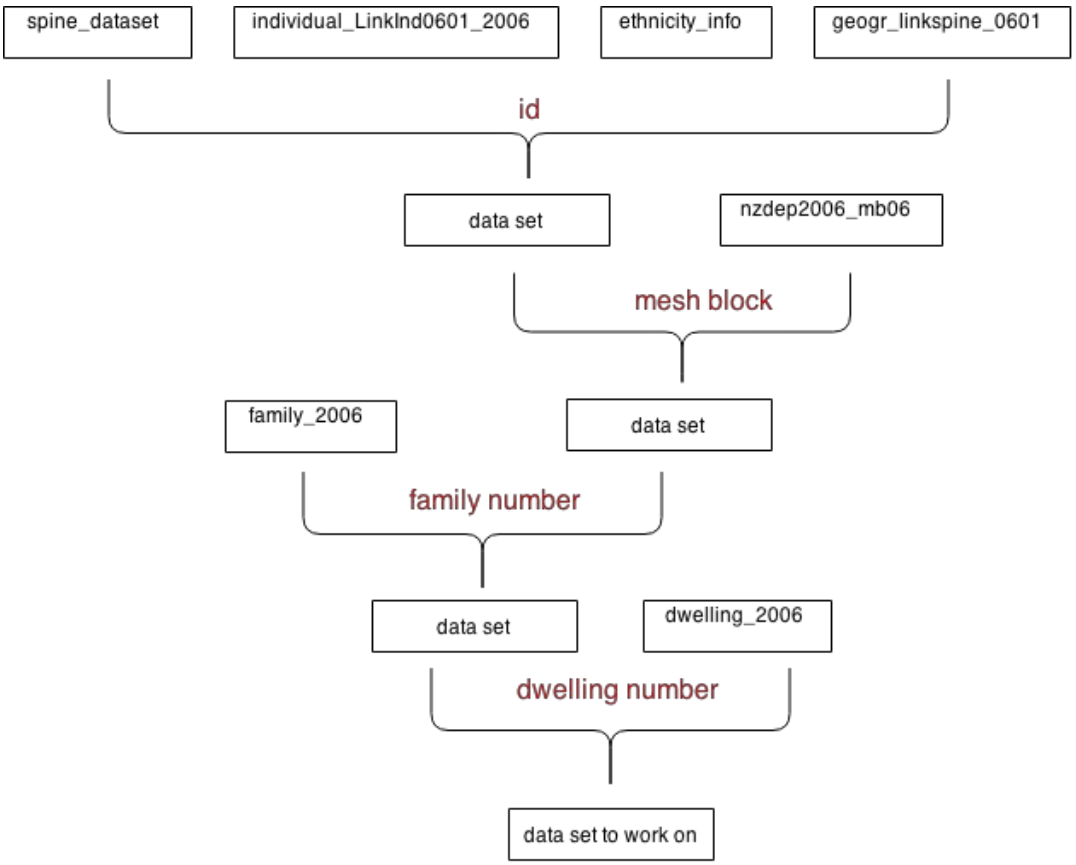
Figure 2: Eligible Data Sets



After that, I merged the data sets relevant to each census to come up with a data set containing all variables required, so we could work on it. Spine data set, individual data set, ethnicity data set, and geography data set were merged together first by ID, then the New Zealand Deprivation data set were merged by mesh block number, then the family data set were merged by family number. And eventually we merged the dwelling data set by dwelling number. However, 1981 and 1986 dwellling data sets had no dwelling number; instead we merged them by mesh block number and dwelling number within the mesh block.

Figure 3 shows 2006 census merging process as an example.

Figure 3: 2006 Census merging process



Then I wrote SAS code for each census to create new harmonised variables according to specification. Below is an example of how I harmonised the 2001 Partnership Status variable.

Figure 4: 2001 Partnership Status Harmonisation

```
*Create partnership status;
marital_status_legal=input(legal_marital_status_code_01,f8.0);
marital_status_social=input(social_marital_status_code_01,f8.0);
partner_01=.;

if marital_status_legal in (11) OR marital_status_social in (200, 211) then
    partner_01=0;
else if marital_status_legal in (33) OR marital_status_social in (223) then
    partner_01=1;
else if marital_status_legal in (31,32) OR marital_status_social in (221, 222) then
    partner_01=2;
else if marital_status_legal in (21) OR marital_status_social in (100, 111, 121) then
    partner_01=3;
else if marital_status_legal in (77,99) AND marital_status_social =999 then
    partner_01=9;
```

Finally cross-census variables covering the period 1981 to 2006 such as the number of times unemployed, and the number of times situated in the lowest household income quintile were created. The cross-census data set merged all censuses by ID number and determined which censuses a respondent had participated in and the number of vaild (non-unstated) answers given.

2.3 Notes on variables

2.3.1 Individual Census Life-Course Variables

Age

age5year: Age in 5 year band

Age information was collected in all six censuses, we only needed to convert single year to 5 year bands.

Ethnicity

EurOther: European or Other

Mao: Maori

Pac: Pacific

Asian: Asian

EthNS: Ethnicity Not Selected

Above variables are all binaries. Information was extracted from the “Ethinicity_info” data set. “EurOther” includes European, MELAA and Other. In the original data set, missing values represent “No”; I modified them to “0” which makes more sense in our case.

ethnicity: Prioritised Ethnicity variable. A prioritisation method was used: if the respondent belonged to multiple ethnicities, we only assigned them to one ethnicity according to an order. The order is Maori, Pacific, Asian, then European and others.

ethnic_density: Ethnic density/fractionalisation - mesh block level. This is a number representing the percentage of the population having the same ethnicity as the respondent within the same mesh block.

Calculation of ethnic density requires mesh block number. The 1981 and 1996 censuses used 2001 mesh block numbers while other censuses used 2006 mesh block numbers. In order to make variables consistent over censuses, 1981 and 1986 data sets were converted to 2006 mesh blocks. However, as some 2001 mesh blocks can correspond to more than one 2006 mesh block, there could be some comparability issues.

New Zealand Deprivation Quintile

Nzdep: New Zealand Deprivation Quintile. I converted the original New Zealand Deprivation Index which has 10 categories to quintiles. It was not available in 1981 and 1986 since the New Zealand Deprivation Index only existed after 1991.

Unemployment

unemp_lf: Unemployment indicator (labour force only). This variable only considered those who were in the labour force. If the respondent was not in the labour force then the unemployment indicator has value “0” since the respondent is technically not unemployed.

unemp_nonlf: Unemployment indicator (all). This variable includes everyone. If the respondent was not working, then the unemployment indicator has value “1”, regardless of labour force status.

labrforce: Labour force indicator. This variable has values ‘1’ or ‘0’ depending on being in the labour force.

Education

education: Education Level. The 1981, 1986 and 1991 censuses had separate variables for school qualification and post-school including tertiary qualification, so I combined all variables.

Welfare

Welfare: Welfare Recipient Indicator (0/1). This includes sickness benefits, invalids benefits, student allowance and other government benefits.

ACC: ACC Payment Indicator (0/1). ACC information was not collected in 1981 and 1986.

super: Superannuation, pension and annuities indicator (0/1).

Living Alone

*live_alone**:*** Living alone indicator (0/1).

In each census, there was a proportion of respondents who answered that they were living alone, but if we checked the dwelling records, they appeared to be living with others, so theoretically speaking they were not living alone. However, people living in the same dwelling might not necessarily have been interacting with each other, so even if they were living together in the same dwelling physically, they could be still living alone in a psychosocial sense. Therefore, we will still use the respondents’ answers.

In the 1986 census data set, there was only one ‘living arrangements’ variable, unlike other years. It had missing values in it which represented people under 15 years old. After checking with other data sets about the treatment of “Children” in this type of question, we can assume that all people under 15 years old were not living alone.

In 2001 and 2006, there was a category”7777” which according to the data dictionary meant “Response Unidentifiable”, but it was also used in the “Living with Other” binary variable. My assumption was that if the respondent had given an answer which was not

partner, child, parent, sibling or flatmate, it was considered to be “Response Unidentifiable”. However, we cannot treat it as a residual category (like with other variables) since we know the respondent was at least living with someone. Therefore, I treated all “7777” answers as not living alone.

Partnership Status

Partner: Partnership Status. There were legal marital status, social marital status and de facto status variables available in some censuses. I combined all of them to get the information required.

In 1996, 2001 and 2006 there was a category called “Non-Partnered, not further defined” in the social marital status variable. By default I treated this category as “Never Married” as there was simply no extra information I could find.

Moving

moved_in_lastyear: Indicates if respondent moved residence in last year.

moved_in_5years: Indicates if respondent moved residence in the last 5 years.

In the 2006 data set, more than 30,000 respondents answered that they had been living at their usual residence for less than 5 years, but they also answered the ‘usual residence 5 years ago question’ in the positive. Basically, they reported that their usual residence 5 years ago was the same place they have been living less than 5 years. Our first thought was that these respondents were not counting the years rigorously, since the usual residence 5 years ago question was asking for a specific date, so they might have treated 4 years and 200 days as 5 years. However, after some detective work, we found out that of these 30,000 respondents, more than 10,000 answered they had been staying at this address for less than 1 year. Clearly this was not some counting error; we could only conclude and assumed that these respondents were actually living at the same address 5 years ago, but moved away and came back in the last 5 years, and it was quite possible, for example, for university students if they moved to another city for study. This applied to all other data sets.

For those who answered “not stated” or had an “unidentifiable answer”, and where some of them answered the ‘usual residence 5 years ago’ question, we could use this information to derive the ‘moved in last 5 years’ indicator. If a respondent was living somewhere else in New Zealand 5 years ago, then the respondent definitely moved within the last 5 years. On the other hand, if a respondent was living in the same address as now, we cannot simply conclude that the respondent hadn’t moved in the last 5 years. For example, university students might have lived at their family home 5

years ago, then moved to another city for study, then after 3 or 4 years, they might have moved back to home again; the census question cannot track such changes.

The 1991 ‘years at usual residence’ question separated “Not stated” into 2 groups: “not stated” and “not stated but at least 5 years”. As we mentioned above, we could not determine if a person had moved in the last 5 years purely based on ‘address 5 years ago’ since they could have moved away and come back, but with this new category “not stated but at least 5 years”, we can now separate those who have moved and those who stayed for at least 5 years. This reduced the number of “not stated” respondents significantly compared to other censuses.

New Zealand Born

nz_born: Indicated if respondent was born in NZ.

Number of Years in New Zealand

yrs_in_nz_max97: Number of years in New Zealand since arrival for long term residence, upper limit 97 years or more. Not available in 1981 since 1981 variable has an upper limit of 50 years or more.

yrs_in_nz_max50: Number of years in New Zealand since arrival for long term residence, upper limit 50 years or more.

‘Years in New Zealand’ question was not asked in 1991.

I assumed that once people came to live in NZ permanently there were no long-term gaps when they stayed overseas between the date of arrival and current date.

People who were born in NZ often gave unspecified answers, so I needed to identify them and assign them with their ages. Again I assumed there were no long term gaps when they stayed overseas between their birth and current date.

One interesting fact about the data sets: the 1996 data set (perhaps others as well) had processed raw data already, for example, ‘years in new Zealand’ was derived from variables ‘month arrival’ and ‘year arrival’ to New Zealand. If one had no arrival date then the ‘years in New Zealand’ variable would have an ‘unstated’ value as well. However, there were records that had arrival date but ‘year in New Zealand’ variable still had an ‘unstated’ value. With further investigation, we realised that these were unrealistic answers, for example, one says he arrived in Dec 1912 but at the time of census (1996) he was only 32 years old, so this arrival date was simply not possible.

Smoker

Smoke: Indicated if respondent is a regular smoker.

Smoking information was only collected in 1981, 1996 and 2006.

Language

Language: Indicated the language(s) used by respondent.

Language information was only collected in 1996, 2001 and 2006.

The 1996 official language variable had fewer categories than 2001 and 2006 variables; there was no category about other languages. Therefore, I combined official language and detailed language variables together to derive “language”.

Disability, FWWP mention

Health: Long term health problem indicator

Disability: Long term disability indicator

overall_dsb: Overall long term health or disability indicator; if either ‘long term health problem’ indicator or ’long term disability’ indicator has value “1” then “overall_dsb” is “1”.

This information was only collected in 1996, 2001 and 2006. (Milligan et al, 2006)

For “*overall_dsb”*, I only assigned it with ‘9’ when both health problem and long term disability variables had an ‘unstated’ value, but if only one of them was ‘unstated’, I followed the other variable. This was to make sure all censuses were consistent with the 1996 variables.

Voluntary work (outside the household)

Unpaid: Voluntary work or caring (outside the household) indicator.

No information on voluntary work outside of the household in the 1981 census could be retrieved as there were no relevant questions.

The 1986 census asked about household duties as part of employment activities (Q16), and there was also a separate question for number of hours spent on voluntary work (Q15). The example given in the question also suggested that it was collecting information on voluntary work done outside of the household.,Thus if a respondent ticked any option other than “Nil hours”, then we assumed that it was a “Yes” for voluntary work (outside of the household). (Errington et al, 2008)

The 1996, 2001 and 2006 data were fine, but 1996 variables need careful handling.

Activity_count_96 counted the number of specified activities unpaid for the people in the SAME household (Q37).

Unpaid_count_96 counted the number of specified activities unpaid for people in a DIFFERENT household (Q38).

In addition, the 1986 census asked how many hours the respondent normally spent on voluntary work on a weekly basis. The 1991 census asked if the respondent had done any voluntary work in the last week. The 1996, 2001 and 2006 censuses asked if the respondent had done any voluntary work in the last 4 weeks.

Private Dwelling Indicator

Private_dwell: Indicator in the dwelling if private or non-private type.

This variable was not of primary interest but will be helpful when deriving other dwelling variables.

Visitor

Visitor: Indicated if one was visitor/guest in the household based on usual residence variable in “spine_dataset”.

This variable was not of primary interest but will be helpful when deriving other dwelling variables.

Housing Tenure

Tenure: Housing tenure. Basically a dwelling should either be owned or rented by the resident. There was an extra category “Not Applicable”, this applies to all visitors in the dwelling since they were simply guests.

Household Size

Household_size: Number of usual residents or number of occupants of the dwelling on the census night.

The 2001 and 2006 data sets had information about usual residents in a dwelling. All other data sets only had information about the number of occupants of the dwelling on census night. This could affect the comparability between census years.

If the dwelling is a non-private or visitor-only (only available in 2001 and 2006) dwelling, then there was no household, hence household size was 0. Also maximum household size was limited to 20 in order to make data consistent.

Access to Motor Vehicle

Motor: Access to motor vehicles indicator.

The 1991 variable has unknown categories 6, 7, and 8 which only occurred in non-private dwellings; we decided to treat them as “5 or more” vehicles”, which was the same as category 5.

I have set up several categories instead of just ‘yes’ or ‘no’, since access to a motor variable was given in the dwelling data set. Whether a person has access to a motor vehicle was hard to determine, for example: children under age 15 cannot have a legal license to drive the vehicle, a visitor to the dwelling does not necessarily have access to the dwelling’s vehicle, and also non-private dwellings such as hospitals and military camps all have vehicles but not all residents can access them.

Access to Internet

Internet: Access to Internet indicator.

This variable was only relevant to the 2001 and 2006 censuses.

Similar to ‘access to motor vehicles’ indicator, non-private dwelling and visitors have separate categories.

Access to Telephone

*Phone***:** Access to Telephone or Mobile Phone indicator.

This information was not collected in 1986 and 1991.

Similar to ‘Access to Motor Vehicles’ indicator, non-private dwelling and visitors have separate categories.

The 2006 Census started to collect both Telephone and Mobile Phone information.

In the 1981 data set, the ‘household amenities’ question was asked. There were 7 amenities: Telephone, Electric clothes drier, Auto Washing machine, Non-auto Washing machine, Colour TV, Black & white TV and Deep Freeze. However, there were only 4 amenities variables, a, b, c and d; there was no explanation about what they represented in the data dictionary. I went to the library and obtained a copy of the 1981 Census results and compared the counts with our data set. Eventually I found out that Amenities_a represented Telephone and Electric clothes drier, Amenities_b represented Auto Washing machine and Non-auto Washing machine, Amenities_c represented Colour TV and Black & white TV and Amenities_d represented Deep Freeze.

There were levels 0, 1, 2, 3, and 9 for each of these amenity variables.

Using amenities_a as an example, ‘0’ meant this household had neither Telephone nor electric clothes drier, ‘1’ meant it had only Telephone, ‘2’ meant it had both Telephone and Electric clothes drier, ‘3’ meant had only Electric clothes drier, and ‘9’ meant ‘not stated’.

Iwi Affiliation

iwi_01=Te Tai Tokerau/Tamaki-makaurau (Northland/AKL)

iwi_02=Hauraki (Coromandel)

iwi_03=Waikato/Te Rohe Potae (Waikato/King Country)

iwi_04=Te Arawa/Taupo (Rotorua/Taupo)

iwi_05=Tauranga Moana/Mataatua (Bay of Plenty)

iwi_06=Te Tai Rawhiti (East Coast)

iwi_07=Te Matau-a-Maui/Wairarapa (Hawke's Bay/Wairarapa)

iwi_08=Taranaki

iwi_09=Whanganui/Rangitikei (Wanganui/Rangitikei)

iwi_10=Manuwatu/Horowhenua/Te Whanganui-a-Tara (Manuwatu/Horowhenua/Wellington)

iwi_11=Te Waipounamu/Wharekauri (South Island/Chatham Islands

iwi_21=Iwi Not Named, but Waka or Iwi Confederation Known

iwi_22=Iwi Named but Region Unspecified

iwi_23=Hapū Affiliated to More Than One Iwi

iwi_99=Don't Know/Refused to Answer/Response Unidentifiable/Response outside Scope/Not Stated

The above variables were all binaries. Each of them represented an Iwi region or category.

iwi_count=Indicated how many iwis the respondent related to; the count increased by 1 if any of the iwi variables above had a value of 1 (except 99)

iwi_nonspecified=Indicated if all iwi variables were unspecified

Iwi questions were not asked in 1981 and 1986. Also the 1991 census collected up to 3 iwis, while other censuses collected up to 5 iwis.

Also, the 1991 census did not have “Iwi Not Named, but Waka or Iwi Confederation Known” category.

iwi_nonspecified variable could be useful since it might identify those who were not Maori and had no affiliation with iwis at all. In our data set, we still have records who claimed to be not of Maori descent or not of Maori ethnicity but still had and affiliation with iwis.

In addition, there was a change of region in the 2006 census. Ngati Tama ki Te Upoko o Te Ika (Te Whanganui-a-Tara/Wellington) was previously coded to Ngati Tama (Taranaki). After some inspection, I determined that the numbers of records affected were very few, and ignorable.

Religion

religion_none = No Religion

religion_other = Other Religion

religion_christian = Christian Religion

religion_residual = Object to answer/Uncertain/Don't know/Not Stated/Not Applicable/Out of scope/Unidentifiable

The 1981, 1986, 1991 and 1996 censuses only allowed respondents to give 1 religion while the 2001 and 2006 censuses allowed respondents to give up to 4 religions. This has caused some trouble in this project. There were people in 2001 and 2006 who answered they believed in Christian, Hindu, Islam and Tao - it could have been true, but we cannot ever find out. Therefore, binaries were used.

The 2001 and 2006 data were more comparable since they used the same format of codes, whereas 1986 was the most difficult one with 500 different categories and many of them only appearing in 1986, making the 1986 religion data least comparable with any other years.

In the data sets, we were not given broad religion classifications such as Christian, Catholic or Hindu. Instead we have very specific religion categories such as “Protestant” and “Divine Light Mission”. Using the report “A Guide to Using Data from the New Zealand Census” (Errington et al, 2008) and Census Data Dictionaries, we were able to classify these religions to 4 categories: None, Christian, Other and Residual category. However, there was no information on 1986 religions, so I had to classify them using the data dictionaries of other censuses and information on the web, since coding was based on my own discretion and help from my supervisor, the results could be biased.

In addition, it seemed that Statistics New Zealand also changed its classification of religions over the years. Some religions were classified as ‘Christian’ in 1981 but became ‘other’ religion in 2006. Therefore, in order to make data consistent, I tried to apply the 2006 classification to other census years where possible, though there might have been religions not taken into account.

Household Income Quintile

hhld_income_quintile: Population was divided into 5 groups based on their household income. The bottom represented the 20 percent (theoretically) of the population with the lowest household income, while the top quintile represented the 20 percent of the population who receive the highest income.

Since household income in the census was in the form of grouped data which had 13 or 14 categories, we did not know the actual amount. Dividing population will less likely to result perfect 5 groups each contains 20% of the population (please clarify this sentence).

For 2001 and 2006 census data, we also used family income to rank the population in order to balance the quintile groups.

Crowding Index

Index: Crowding Index, calculated using Figure 5 formula (Statistics New Zealand, n.d.)

Figure 5: Crowding Index Formula

$$\text{Crowding Index} = \frac{0.5 * \text{Children Under 10} + \text{Couples} + \text{Others aged 10 or above}}{\text{Number of Bedrooms}}$$

Given the information available to us, it was a problem to calculate the crowding index for usual residence since the census allocates dwelling numbers according to respondents' census night addresses. For those who were not staying at the same residence as their usual residence on census night, there was not enough information to find out which dwelling they usually lived in. Therefore, the crowding index we could calculate was based on the census night living arrangements.

Furthermore, since the data sets only contained "eligible population" who were in NZ 5 years ago and were older than 5 years old, we omitted a significant proportion of the population (there could be entire families not showing up in our data sets). This was somewhat problematic as the number of children under 10 years and number of couples' information would not so accurate, flowing on to the crowding index. More importantly, if we did not have the full population in our data sets, we could not even accurately estimate the number of people living in the dwelling. In the future, it would be a good idea to attach the non-theoretical population data sets to obtain a more accurate crowding index.

The 1991, 1996, 2001 and 2006 data sets used both usual household composition and family type information to estimate the number of couples.

The 1981 and 1986 data sets only had usual household composition available. Dwellings containing three or more families could have inaccurate crowding indices since there was no couple information available for them. Therefore we could not

determine the number of couples for dwellings that had no clear usual household composition, and the crowding index would be overestimated for these dwellings.

The 2001 and 2006 data sets had the ‘number of usual residents’ variable but other censuses only had the ‘number of occupants on census night’ variable. This could cause problems when calculating the crowding index, since, for example, a household could have 8 people usually but there was only 1 people staying there on census night. As the number of occupants could be inconsistent with usual household composition information, to avoid the problem of having negative counts, we set the minimum to 0.

Also, there was an upper limit for the ‘number of bedrooms’ variable which was different in each census; this could cause some problems. For example, it was possible that a dwelling had 40 bedrooms and 40 usual residents, but the ‘number of bedrooms’ variable would only have an upper limit of 20 bedrooms, so it would have been calculated as overcrowded while it was in fact not.

New Zealand Socio-Economical Index / Elley & Irving Index

NZSEI: New Zealand Socio-Economical Index / Elley & Irving Index

The 1981, 1986 and 2001 censuses used Elley & Irving index while the 1991, 1996 and 2006 censuses used New Zealand Socio-Economic Index

The Elley & Irving index used a slightly different index of occupation which did not match perfectly with the occupation list used in the census. Thus the 1981, 1986 and 2001 index could contain errors as I manually modified the occupation list to be consistent.

2.3.2 Cross-Census Variables

These were variables, derived in the course of this project that were designed to span across all the census years, and intended to characterise the life course of a respondent over this period (1981 to 2006).

Participation: Number of times the respondent participated in the census from 1981 to 2006.

Census Year: Which censuses the respondent participated in.

low_inc_quintile: Number of times a respondent was in NZSEI/E&I classes 5&6 (lowest quintile).

low_nzdep: Number of times a respondent was in lowest New Zealand Deprivation quintile. (Most deprived quintile, 5)

low_nzsei: Number of times a respondent was in lowest household income quintile.

unemploy: Number of times a respondent was Unemployed (only considered those in labour force).

labf: Number of times a respondent was in labour force.

fare: Number of times a respondent was on welfare.

alone: Number of times a respondent was living alone.

crowded: Number of times a respondent was living in an overcrowded residence (If crowding index was greater than 1 then it was considered an overcrowded dwelling).

rented: Number of times a respondent was living in rented accommodation.

movedlastyr: Number of times a respondent moved in last year.

moved5yrs: Number of times a respondent moved in last 5 years.

smoker: Number of times a respondent was a smoker.

religious: Number of times a respondent was religious (reported a religion).

In addition, each of the above variables also had two extra components:

Number of Appearances: The number of times this question appeared in the census from 1981 to 2006.

Number of Valid Answers: The number of times the respondent provided a valid answer which was not ‘unstated’, ‘unidentifiable’, ‘out of scope’ or any other residual category.

Notes

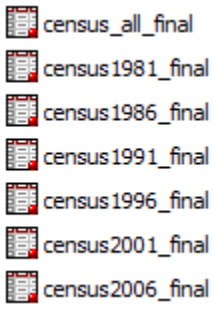
- Detailed SAS code and Data Dictionaries can be found in the Appendix (as separate files).
- ‘Readme’ files – documenting locations and descriptions of all project outputs (data sets, data dictionaries, SAS programs, presentation, report) - can be found in the Appendix

3.0 Results

There are 3 types of results produced by this project.

3.1 Data sets

Figure 6: Data sets created



As shown in Figure 5, there were six final census data sets containing life-course variables, and one final data set containing cross-census variables. A list of life-course variables created for each census can be found in Appendix A and the list of cross-census variables can be found in Appendix B.

The 2006 Census has 3,285,978 records, the 2001 Census has 3,122,175 records, the 1996 Census has 3,018,918 records, the 1991 Census has 2,925,849 records, the 1986 Census has 2,884,221 and the 1981 Census has 2,078,427 records. (Numbers have been randomly rounded to multiple of 3, as required by Statistics NZ)

1981 has significantly less records compared to the other censuses because it only contains the population that could be linked to 1986, while other censuses contain records that could be linked to a previous census.

The “Census_all_final” data set contains cross-census variables; there are 7,399,134 records, i.e. the total theoretical population from 1981 to 2006.

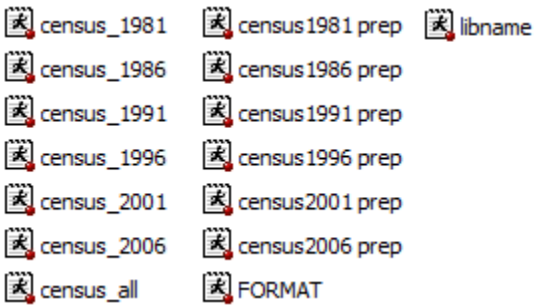
3.2 Data Dictionary

The Data Dictionary contains information about Life-course variables as well as cross-census variables created, including availability in each census, values and categories, and variables used to derive them in each census.

The Data Dictionary can be found in a separate Appendix file.

3.3 SAS Code

Figure 7: SAS Code Files



Files ending with “prep” are programmes used to merge data sets. The ‘Libname’ file is used to set up libraries to use in the SAS environment. The ‘FORMAT’ file contains format data which can be used with the SAS ‘format’ option in many commands. ‘Census_all’ contains the code used to generate cross-census variables. Other files starting with “census_” are programmes used to generate life-course variables for each census.

These files can be found in a separate Appendix.

3.4 Challenges

For the ‘education’ variable, a new residual category, ‘unidentifiable’, was introduced in 2001.Ssome variables showed a significant drop in category 0 answers, while the residual category increased markedly. It is possible that before 2001, if the respondent gave an ‘unidentifiable’ answer, that it was classified as 0 (No or None) rather than assigned to a residual category. If it is true then category 9 could have been overstated before 2001, and would lead to some problems when describing trends in category 0.

Figure 8: Changes in Education variable

education_96			education_01	
0	34%		0	26%
1	31%		1	35%
2	25%		2	18%
3	7%		3	10%
9	2%		9	11%

3.5 Further work

Similar projects, such as the Family Whanau and Wellbeing Project, have been undertaken to which we could compare our results in order to validate our outputs, such as the distributions of the NZ Social economic Index/ Elley & Irving Index, and the Crowding Index. Also it might be a good idea to use complete data sets by merging the individual data sets and non-theoretical population data sets to create some composite variables, such as the crowding index. However, it might be better if we could request these variables from Statistic NZ directly (if they were available) rather than creating them ourselves, as we cannot be sure if we would have produced the correct values.

In addition, more work is required to derive more life-course variables that span across years, such as Changes in Occupation, Age partnered, length of partnership, Number of children, and Age retired. These variables will provide some more interesting information but will be much more complicated to create.

4.0 Conclusions

The key outcomes of this project are: (1) SAS final datasets that are ready for analysis; (2) SAS programmes used to generate life-course and cross-census variables; and (3) a data dictionary which will be useful for understanding the variables.

Imagine if a researcher wants to analyse the changes in New Zealand society over the period 1981 to 2006; this project can provide a lot of information on life-course variables such as education, work, cultural and socio-economic status. The longitudinal component opens up a new dimension for research that can now be done to track people through the life course while the harmonised life-course variables make assessment of changes over time much easier. As a result of this project, it now takes much less effort to write code to describe trends and to perform statistical analysis.

Of course there is much more work to be done in the future to improve the New Zealand Longitudinal Census and this project was just the beginning.

5.0 References

Didham, R, Nissen, K and Dobson, W (2014). *Linking censuses: New Zealand longitudinal census 1981–2006.* Available from 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*. Wellington, N.Z.: Statistics New Zealand.

Milligan, S.; A. Fabian; P. Coope and C. Errington (2006). *Family Wellbeing Indicators*, Statistics New Zealand, Wellington.

Statistics New Zealand, (n.d.). *Number of Rooms/Bedrooms*. Retrieved from http://www.stats.govt.nz/methods/classifications-and-standards/classification-related-stats-standards/number-of-rooms-bedrooms/definition.aspx.

Appendix A: Individual Census Variables

Census variables	**category 99999 means such question or information was not collected**					
harmonisation	**_xx represent the census year**					
Variable	description	values/categories		Notes	Availability	
age5year_xx	Age in 5 year band	0	0-4 years old	Converted from single years into 5 year bands	2006	✓
		1	5-9 years old		2001	✓
		2	10-14 years old		1996	✓
		...	...		1991	✓
		26	125-129 years old		1986	✓
					1981	✓
EurOther_xx	European or Other	0	No	Available in ethnicity_info dataset.	2006	✓
Mao_xx	Maori	1	Yes	Binary dummy variables are used instead of multi-levels	2001	✓
Pac_xx	Pacific Islander			variables	1996	✓
Asian_xx	Asian			EurOther_xx includes European, MELAA and Other	1991	✓
EthNS_xx	Ethnicity Not selected				1986	✓
					1981	✓
nzdep_06	New Zealand Deprivation Quintile	1	Least Deprived	Original classes were from 1 to 10, converted to scale of 1 to 5	2006	✓
		2			2001	✓
		3			1996	✓
		4			1991	✓
		5	Most Deprived		1986	✗
		.	Missing value		1981	✗
unemp_lf_xx	Unemployment indicator (labour force only)	0	No	Does not includes those not in the labour force	2006	✓
unemp_nonlf_xx	Unemployment indicator (all)	1	Yes	Includes those not in the labour force	2001	✓
labrforce_xx	Labour force indicator	.	Missing value (Age under 15)		1996	✓
					1991	✓
					1986	✓
					1981	✓
education_xx	Education Level	0	No Qualification	1991 have separate variables for school qualification and	2006	✓
		1	School Qualification	tertiary qualification, so we need to combine them	2001	✓
		2	Post-school Qualification		1996	✓
		3	Tertiary Qualification		1991	✓
		9	Not Stated/Unidentified/Out of Scope		1986	✓
		.	Missing value (Age under 15)		1981	✓

welfare_xx		0	No	sickness benefits, invalids benefits, student allowance and	2006	✓
		1	yes	other government benefits are counted as welfare	2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
		.	Missing value (Age under 15)	if all income sources unstated, then give code 9	1991	✓
					1986	✓
					1981	✓
ACC_xx	regular acc payment receipt indicator	0	no	Dervied from income source variables	2006	✓
		1	yes		2001	✓
		9	not stated		1996	✓
		.	Missing value (Age under 15)		1991	✓
					1986	✗
					1981	✗
super_xx	Superannuation, pension and annuities indicator	0	no	Dervied from income source variables	2006	✓
		1	yes		2001	✓
		9	not stated		1996	✓
		.	Missing value (Age under 15)		1991	✓
					1986	✓
					1981	✓
live_alone_xx	Living alone indicator	0	no	Derived from living arrangement variables	2006	✓
		1	yes	7777 means unidentifiable, but it is not classified as "9"	2001	✓
		9	Not Stated/Unidentified/Out of Scope	since it means the respondent is not living alone	1996	✓
					1991	✓
					1986	✓
					1981	✗
partner_xx	Partnership Indicator	0	Never Married	2006, 2001 and 1996 have both legal and social marital status	2006	✓
		1	Widowed or Bereaved	1986 and 1981 have both legal and de facto status	2001	✓
		2	Separated-Divorced	Social Marital Status have a category "non partnered nfd"	1996	✓
		3	Partnered-Married	It is bit ambiguous, since it could mean both single and divorced	1991	✓
		9	Not Stated/Unidentified/Out of Scope	but we assumed it is single	1986	✓
		.	Missing value (Age under 15)		1981	✓
moved_in_lastyear_xx	Indicate if respondent moved in the last year	0	no	Derived from years_at_addr_code variable	2006	✓
		1	yes		2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
					1991	✓
					1986	✓
					1981	✗

moved_in_5years_xx	Indicate if respondent moved in the last 5 years	0	no	Derived from years_at_addr_code variable and	2006	✓
		1	yes	addr_5years_ago_code	2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
					1991	✓
					1986	✓
					1981	✗
nz_born_xx	Indicate if respondent was born in NZ	0	no	If birth_country_code_06 is 1201 then NZ_born is 1	2006	✓
		1	yes		2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
					1991	✓
					1986	✓
					1981	✓
yrs_in_nz_max97_xx	Number of years in New Zealand since arrival for	0	Less than 1 year	Respondents who were born in NZ usually have unspecified	2006	✓
	long term residence	1		answers, so need to do extra data manipulation to find them	2001	✓
		to	In single years	and assign them with their ages as years in NZ	1996	✓
		96			1991	✗
		97	97 years or more		1986	✓
		99	Not Stated/Unidentified/Out of Scope		1981	✗
yrs_in_nz_max50_xx	Number of years in New Zealand since arrival for	0	Less than 1 year	Respondents who were born in NZ usually have unspecified	2006	✓
	long term residence	1		answers, so need to do extra data manipulation to find them	2001	✓
		to	In single years	and assign them with their ages as years in NZ	1996	✓
		49			1991	✗
		50	50 years or more		1986	✓
		99	Not Stated/Unidentified/Out of Scope		1981	✓
smoke_xx	Indicate if respondent is a regular smoker	0	no		2006	✓
		1	yes		2001	✗
		9	Not Stated/Unidentified/Out of Scope		1996	✓
		.	Missing value (Age under 15)		1991	✗
					1986	✗
					1981	✓

language_xx	Indicate the language(s) used by respondent	0	Maori Only		2006	✓
		1	English Only		2001	✓
		2	English and Maori (No Other)		1996	✓
		3	Maori and Other (No English)		1991	✗
		4	English and Other (No Maori)		1986	✗
		5	English, Maori and Other		1981	✗
		6	Other Languages Only			
		8	No Lanugage/Not Applicable			
		9	Not Stated/Unidentified/Out of Scope			
health_xx	Long term health problem indicator	0	No		2006	✓
		1	Yes		2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
					1991	✗
					1986	✗
					1981	✗
disability_xx	Long term disability indicator	0	No		2006	✓
		1	Yes		2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
					1991	✗
					1986	✗
					1981	✗
overall_dsb_xx	Overall long term health or disability indicator	0	No	Combines information form previous 2 variables, if any of them	2006	✓
		1	Yes	is 1, then overall_dsb indicator is 1	2001	✓
		9	Not Stated/Unidentified/Out of Scope		1996	✓
				only assign overall_dsb with 9 if both variables have unstated	1991	✗
				value	1986	✗
					1981	✗
unpaid_xx	Voluntary work or caring (outside the household)	0	No	1991 has an indicator of involvement in all voluntary work	2006	✓
		1	Yes	1986 only has has a variable indicating hours spent on voluntary	2001	✓
		9	Not Stated/Unidentified/Out of Scope	work	1996	✓
		.	Missing value (Age under 15)		1991	✓
					1986	✓
					1981	✗

private_dwell_xx	Private Dwelling Indicator	0	No	Useful for identifying household, since non-private dwellings	2006	✓
		1	Yes	have no usual residents	2001	✓
					1996	✓
					1991	✓
					1986	✓
					1981	✓
visitor_xx	Indicate if one is visitor/guest in the household,	0	No	Usual resident variable info is partitally availabe for 1986 data,	2006	✓
	based on usual resident var	1	Yes	so some imputation methods were used for 1986	2001	✓
		.	Missing	if family_code =9 and family_nbr is missing, then	1996	✓
				usual residence shoulbe be 2 (no), visitor=1	1991	✓
					1986	✓
					1981	✓
tenure_xx	Housing Tenure	1	Owned or Held in a family trust		2006	✓
		2	Rented		2001	✓
		8	Not Applicable (Non-Private or visitor only)		1996	✓
		9	Unspecified		1991	✓
					1986	✓
					1981	✓
household_size_xx	Household size	0	No household (non-private/visitor only)	2001 and 2006 have usual resident information available	2006	✓
		1-19	Represent number of residents	Other years only have number of occupants at the dwelling	2001	✓
		20	20 or more residents	on census night, there could be minor comparable issues	1996	✓
				if there are visitors or absentees at the dwelling	1991	✓
					1986	✓
				2001 and 2006 also indicate visitor only dwellings, they have	1981	✓
				no usual residents neither		
				max is limited to 20 to make data consistent		
motor_xx	Access to motor vehicles	0	No	1991 has unknow category 6,7,8, we treated them as 5 and more	2006	✓
		1	Yes		2001	✓
		2	Visitor (Private Dwelling)		1996	✓
		7	Under age person (Under 15 year old)		1991	✓
		8	Non-Private Dwelling		1986	✓
		9	Not Stated/Unidentified/Out of Scope		1981	✓

internet_xx	Access to Internet	0	No		2006	✓
		1	Yes		2001	✓
		2	Visitor (Private Dwelling)		1996	✗
		8	Non-Private Dwelling		1991	✗
		9	Not Stated/Unidentified/Out of Scope		1986	✗
					1981	✗
phone_xx	Access to Telephone or Mobile Phone	0	No	1996 has unknown category 3, we consider it as unidentifiable	2006	✓
		1	Yes		2001	✓
		2	Visitor (Private Dwelling)		1996	✓
		8	Non-Private Dwelling		1991	✗
		9	Not Stated/Unidentified/Out of Scope		1986	✗
					1981	✓
iwi_yy_xx	01=Te Tai Tokerau/Tamaki-makaurau (Northland/AKL)	0	No	Binaries are created instead of variable with multiple levels	2006	✓
	02=Hauraki (Coromandel)	1	Yes	iwi_01_xx represent Northland/Auckland	2001	✓
	03=Waikato/Te Rohe Potae (Waikato/King Country)			iwi_08_xx represent Taranaki	1996	✓
	04=Te Arawa/Taupo (Rotorua/Taupo)			and etc	1991	✓
	05=Tauranga Moana/Mataatua (Bay of Plenty)				1986	✗
	06=Te Tai Rawhiti (East Coast)				1981	✗
	07=Te Matau-a-Maui/Wairarapa (Hawke's Bay/Wairarapa)					
	08=Taranaki					
	09=Whanganui/Rangitikei (Wanganui/Rangitikei)					
	10=Manuwatu/Horowhenua/Te Whanganui-a-Tara (Manuwatu/Horowhenua/Wellington)					
	11=Te Waipounamu/Wharekauri (South Island/Chatham Islands					
	21=Iwi Not Named, but Waka or Iwi Confederation Known					
	22=Iwi Named but Region Unspecified					
	23=Hapū Affiliated to More Than One Iwi					
	99=Don't Know/Refused to Answer/Response Unidentifiable/Response Outside Scope/Not Stated					
iwi_count_xx	Indicate how many iwis does the respondent	1-5		count+1 if any of iwi variables above have value 1, except 99		
	relate to.	9	Unspecified			
iwi_nonspecified_xx	Indicate if all iwi variables are unspecified	0	No	this variable could be useful since it might identify those who		
		1	Yes	are not maori and have no affilation with iwis at all		
				In our data set, we still have records who are not maori descedent		
				, not maori ethnicity but still have affilation with iwis		

religion_none_xx	religion binary variables	0	No		2006	✓
		1	Yes		2001	✓
religion_other_xx		.	Missing		1996	✓
					1991	✓
religion_christian_xx					1986	✓
					1981	✓
religion_residual_xx	Object to answer/Uncertain/Don't know/Not Stated/Not Applicable/					
	Out of scope/Unidentifiable					
hhld_income_quintile_xx	Divide Population into 5 groups based on their	1	Lower household income group	We do not have actual amount of income but grouped data,	2006	✓
	household income	2		it is unlikely to have perfect quintiles each contains 20% records	2001	✓
		3			1996	✓
		4		2001 and 2006 also used family income to balance groups	1991	✓
		5	Higher Household income group		1986	✓
					1981	✓
		888	Respondent is a visitor			
		999	Not Stated			
		.	Missing value (non-private/visitor only)			
ethnic_density_xx	Ethnic density/fractionalisation - meshblock level		Decimal numbers	We excluded those who have not stated their ethnicity	2006	✓
	a demical number represent the proportion of			All meshblocks were converted to 2006 classification	2001	✓
	respondents' ethnicity in the meshblock				1996	✓
					1991	✓
ethnicity_xx		.	Not stated	Prioritisation method was used, Maori then pacific then asian	1986	✓
		1	Maori	then european and others	1981	✓
		2	Pacific			
		3	Asian			
		4	European and others			
index_xx	Crowding index	Other	The higher the more crowded	2006 used number of usual residents	2006	✓
		888	Respondent is a visitor	Other years used number of occupant on census night	2001	✓
		999	Not Stated		1996	✓
		.	Missing value (non-private/visitor only)	bedroom upper limits are different in each census	1991	✓
					1986	✓
	(0.5*numbe of child under 10+number of couple+others)/number of bedrooms			1981 and 1986 data set do not have family variables	1981	✓
				So we cannot determine couple information for some hhlds		

nzsei_xx	Divide population into 6 groups by their nz social	1	Highest	1981, 1986 and 2001 used elley-irving index	2006	✓
	economical index or elley irving index	2		1991, 1996 and 2006 used NZ socio-economic index	2001	✓
		3			1996	✓
		4		Elley-Irving index used a slightly different index of occupation	1991	✓
		5		which do not match perfectly with the occupation list used in	1986	✓
		6	Lowest	census, so 1981, 1986 and 2001 index could contain errors	1981	✓
		999	Not Stated			
		.	Missing (Under 15 YO)			

Details about variables categories in each census are available in separate file

Appendix B: Cross-Census Variables			
FOR OVERALL CENSUS DATA SET			
Variable	description	values/categories	
Participation	Number of times participated in Census	0	No
Census1981	Participation in 1981 Census	1	Yes
Census1986	Participation in 1986 Census		
Census1991	Participation in 1991 Census		
Census1996	Participation in 1996 Census		
Census2001	Participation in 2001 Census		
Census2006	Participation in 2006 Census		
low_inc_quintile_appear	Number of times Household Income Quintile Appeared Overall	lowest Household Income Quintile	
low_inc_quintile_valid	Total Valid Household Income Quintile Answers Given		
low_inc_quintile_numtimes	Number of times in lowest Household Income Quintile		
low_nzdep_appear	Number of times NZDEP Appeared Overall		
low_nzdep_valid	Total Valid NZDEP Answers Given		
low_nzdep_numtimes	Number of times in lowest NZDEP		
low_nzsei_appear	Number of time3 NZSEI/E&I Appeared Overall)		
low_nzsei_valid	Total Valid NZSEI/E&I Answers Given		
low_nzsei_numtimes	Number of times in NZSEI/E&I 5 and 6		
unemploy_appear	Number of times Unemployment Appeared Overall		
unemploy_valid	Total Valid Unemployment Answers Given		
unemploy_numtimes	Number of times being unemployed		Only considered those in labou force
labf_appear	Number of times Labour Force Appeared Overall		
labf_valid	Total Valid Labour Force Answers Given		
labf_numtimes	Number of times not in labour force		
fare_appear	Number of times Welfare Receipt Appeared Overall		
fare_valid	Total Valid Welfare Receipt Answers Given		
fare_numtimes	Number of times being a welfare receipt		
alone_appear	Number of times Living Alone Appeared Overall		
alone_valid	Total Valid Living Alone Answers Given		
alone_numtimes	Number of times living alone		
crowded_appear	Number of times Crowding Index Appeared Overall		
crowded_valid	Total Valid Crowding Index Answers Given		
crowded_numtimes	Number of times living in a crowded dwelling		Crowding index >1 is crowded
rented_appear	Number of times Rented Appeared Overall		
rented_valid	Total Valid Rented Answers Given		
rented_numtimes	Number of times living in a rented dwelling		
movedlastyr_appear	Number of times Moved in last year Appeared Overall		
movedlastyr_valid	Total Valid Moved in last year Answers Given		
movedlastyr_numtimes	Number of times Moved in last year		
moved5yrs_appear	Number of times Moved in last 5 years Appeared Overall		
moved5yrs_valid	Total Valid Moved in last 5 years Answers Given		
moved5yrs_numtimes	Number of times Moved in last 5 years		
Smoker_appear	Number of times Smoking Appeared Overall		
Smoker_valid	Total Valid Smoking Answers Given		
Smoker_numtimes	Number of times being a Smoker		
Religious_valid	Total Valid Religion Answers Given		
Religious_numtimes	Number of times Religious		
Religious_appear	Number of times Religion Appeared Overall		

Appendix C: SAS Code

Available as separate files

Appendix D: Presentation

Available as separate files

Appendix E: Readme.txt on S: Drive

There are 5 Folders.

Final Work: Contains all files relevant to use the life-course variables and data sets

1. Files name ended with "prep" are SAS codes used to merge data sets for each census
2. Files name started with "census_" are SAS codes used to generate life-course variables and data sets
3. FORMAT.SAS: format file can be used with certain SAS command
4. Libname.SAS: can be used to set up library names for SAS environment.
5. Data Dictionary.xlsx: Data dictionary
6. How to Use.txt: Instruction of how to use these SAS codes to create data sets.

Metadata: some references files I used to created certain variables

1. mb01tomb06: conversation file of mesh block 2001 to mesh block 2006.
2. NZSEI files: look–up table of occupation and nz socio–economical index and elley & Irving index
3. religion classification NZLCS v1: religion variable category classification.

Presentation and report: Contains presentation and reports

1. Presentation.pptx: Presentation about this project
2. Appendices.zip: Contains all appendices for the report, basically a wrap up of what is in the "Final Work" folder.
3. Report.docx: A report of this project

References: All the files I used as references while working on the project

1. Forms: 6 Individual and 6 Dwelling Census forms for each census
2. Data Dictionaries: 6 Data Dictionaries for each census
3. a-guide-to-using-data-from-the-nz-census.pdf: a report produced by COMPASS earlier, contains notes on many variables for census 1981-2006, also some general classifications.
4. family-wellbeing-report-2006.pdf: Another report produced by COMPASS.
5. linking-censuses-nzlc-1981-2006.pdf: Report of the New Zealand Longitudinal Census.

Stats NZ rules: contains regulations of releasing data from Datalab computers

1. microdata-output-guide-2014.pdf: General guidelines
2. NZLC Confidentiality Rules v1.2.docx: Rules relevant to NZLC

Work in Progress: Other files I produced when working on this project

1. Construction of life-course variables for NZLC.docx: Description and Outline of this project.
2. Notes.docx: Notes I made when I was working on the project, contains thoughts and findings about data sets and variables
3. Project Schedule.docx: Initial Time Schedule of this project, never updated
4. To Dos.docx: contains a list to-do tasks when I was working on the project, also some problems I found and their solutions.
6: Variables Request.xlsx: list of variables that were relevant to the project but was not given at the first place

Appendix F: Readme.txt on Datalab server

There are five folders.

Census: All the census data, except nzdep data sets which is in \\wprdfs08\RO-MAA2013-18 Linkage Bias Longitudinal Census\Updated data dec 2013'

1. Dwelling and household data sets: contains dwelling and household variables
2. ethnicity_info: Ethnicity data for all people from 1981 to 2006.
3. family data sets: Family variables, there are no 1981 and 1986 data sets.
4. geography data sets: with names like "geogr_linkspine", they contain geography variables like Area Unit, Mesh Block for each person in each census.
5. Individual data sets: 10 data sets as censuses are paired adjacently, it ended with a year represent which census' question and variables this data set contains.
6. Non-theoretical populations: Contains all non-theoretical population for each census.
7. Spine_dataset: Contains general information such as gender, usual residence indicator, family numbers, dwelling number each census for all people from 1981 to 2006.

Created data sets: The final data sets created for this project contains life-course variables

1. Individual Census data sets: contains life course variables for each census.
2. Census_all_final: contains cross census variables, such as number of times being unemployed and number of times being religious.
3. Census_across_final: contains variables that track a person's life—course, currently only 1 variable is created, highest education 1981-2006

Data Dict: some Data dictionary files from Stats NZ

1. 1996-census—classification-counts—people: I used it to determine some variables categories in 1996.
2. Data concordances NZLCS v1: Contains concordances data, I used it to determine iwi and mesh block across years.
3. Data Dictionary Construction of Life—course variables: Data dictionary for this project.
4. Data dictionary NZLCS V1: Data Dictionary for NZLC, not complete.
5. Database design NZLCS: Data Design of all the data sets.
6. Metadata for ethnicity_info: Data Dictionary for ethnicity_info data set.
7. Metadata for individual_linkind0601_2006: Simple Data Dictionary for 2006 census, not complete.
8. Metadata for Spine dataset: Data Dictionary for spine data set.

Metadata: some references files I used to created certain variables

1. mb01tomb06: conversation file of mesh block 2001 to mesh block 2006.
2. NZSEI files: look—up table of occupation and nz socio—economical index and elley & Irving index
3. religion classification NZLCS v1: religion variable category classification.

NZSEI: NZ socio—economical index reports

1. 1991, 1996 and 2006 NZSEI reports, I used these to create NSZEI look-up table.

Other files in the Home directory:

1. "How to use file": Instructions of creating life-course variables and data sets using SAS codes provided.
2. SAS codes