Constructing a Longitudinal Database for the Analysis of Individual Incomes in New Zealand

Nazila Alinaghi, John Creedy and Norman Gemmell

WORKING PAPER 05/2020
November 2020

Working Papers in Public Finance

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

Chair in Public Finance
Wellington School of Business and Government

The Working Papers in Public Finance series is published by the Victoria Business School to disseminate initial research on public finance topics, from economists, accountants, finance, law and tax specialists, to a wider audience. Any opinions and views expressed in these papers are those of the author(s). They should not be attributed to Victoria University of Wellington or the sponsors of the Chair in Public Finance.

Further enquiries to:
The Administrator
Chair in Public Finance
Victoria University of Wellington
PO Box 600
Wellington 6041
New Zealand

Phone: +64-4-463-9656
Email: cpf-info@vuw.ac.nz

Papers in the series can be downloaded from the following website:
https://www.wgtn.ac.nz/sacl/centres-and-chairs/cpf/publications/working-papers

Constructing a Longitudinal Database for the Analysis of Individual Incomes in New Zealand[*]

By

Nazila Alinaghi, John Creedy and Norman Gemmell[†]

Abstract

This paper describes the construction of a unique longitudinal individual-level dataset that allows the dynamics of individual incomes in New Zealand to be examined. The data are obtained from the New Zealand Integrated Data Infrastructure, and cover approximately 5,393,874 taxpayers, for whom a range of information including, but not limited to, taxable income, gender, ethnicity, education level and location have been compiled. The availability of suitable data has previously been a constraint on income dynamics research. The present data construction exercise allows a more extensive analysis of individual income inequality and mobility than has previously been possible.

[*] This paper is part of a project on 'Measuring Income Inequality, Poverty and Mobility in New Zealand', funded by an Endeavour Research Grant from the Ministry of Business, Innovation and Employment (MBIE) and awarded to the Chair in Public Finance at Victoria University of Wellington. We have benefited from discussions with Sarah Crichton, Robert Templeton, and Christopher Ball regarding the data.

[†] Wellington School of Business and Government, Victoria University of Wellington, Wellington, New Zealand.

Disclaimer

The results presented in this study are the work of the authors, not Statistics New Zealand (Statistics NZ); they are not official statistics. They have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Statistics NZ. The opinions, findings, recommendations, and conclusions expressed in this paper are those of the authors, not Statistics NZ, or Inland Revenue.

Access to the anonymised data used in this study was provided by Statistics NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this paper have been confidentialised to protect these groups from identification and to keep their data safe. Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further details can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The matching of different data sources on the IDI spine is done by Statistics NZ. These datasets are anonymised thereafter and made available to researchers. The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. The tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.

1. Introduction

This paper describes the construction of a unique dataset that allows the dynamics of individual incomes in New Zealand to be examined. The dataset provides a basic resource for economic analyses of income inequality where the accounting period over which income is measured is important.

The vast majority of empirical studies investigating income inequality and poverty provide snapshot information based on cross-sectional data. These studies, while informative, tend to overstate the degree of income inequality and poverty by necessarily ignoring transitory and life-cycle variations. Longitudinal data are needed to provide a more accurate picture of long-run income mobility and poverty persistence. However, despite the recognition of this point, access to such data has been highly restricted in New Zealand: previous sources are discussed briefly in Section 2.

Until recently, access to unit-record data was restricted to New Zealand government departments for bona fide research or statistical purposes. Unlike a number of other countries, access to the micro-data in the form of 'Confidentialised Unit Record Files' (CURFs) has generally been no less stringent. However, there has been something of a 'revolution' regarding access by university and other researchers to micro-data collected by Statistics New Zealand (henceforth, Statistics NZ), in recent years. Since 2012, it has been possible for approved researchers to examine data in Statistics NZ datalabs. The individual records are anonymised, and strict conditions apply to the nature of summary statistics reported or shared with other researchers.

A second important development is the establishment by Statistics NZ of an Integrated Data Infrastructure (IDI), which is made available to researchers in the datalabs. The IDI collects a range of administrative and survey datasets, linked through a central spine. Currently, there are more than nine million uniquely identified individuals on the spine, many of whom are former New Zealand residents who have since left or died. An advantage is that various datasets available within the IDI can cover a wide range of subject areas for an 'ever resident' New Zealand population. However, for this study it is necessary to restrict the IDI spine population to the subset of individuals who are not only usual residents but have received income at some stage during the period of study and therefore interacted with the Inland Revenue Department (IRD). This paper describes how a dataset relating to individuals was constructed from the IDI.

The IDI is described briefly in Section 3. Section 4 describes the use of the IRD data within the IDI to construct longitudinal income information. A substantial challenge is to use other administrative data sources in the IDI to obtain a range of demographic characteristics of the individuals for whom longitudinal profiles are obtained. Section 4 explains how a range of demographic variables can be linked.

2. Earlier Longitudinal Income Data in New Zealand

Limited information about income dynamics and mobility in New Zealand has previously been available. Longitudinal income information, if only for a few years, has been scarce. The earliest analysis of longitudinal data in New Zealand appears to be the unpublished study by Smith and Templeton (1990), using the Personal Income Survey for eight consecutive years (1980 – 1987).

A survey with a longitudinal nature is the Survey of Family, Income and Employment (SoFIE). This survey contains only eight annual 'waves', from October 2002 to September 2010, and covers around 11,000 nationally representative households (about 22,000 individuals). For income mobility studies using SoFIE data see Carter and Gunasekara (2012) and Carter et al. (2014).

In the absence of a large-scale longitudinal survey, the most effective method of obtaining information about a sample of individuals over a number of years is to use administrative data.[1] Such a dataset for individuals was compiled in the mid-1990s by the New Zealand's Inland Revenue Department, although only three consecutive years (1991, 1992 and 1993) of incomes were available for a limited number of age cohorts. Creedy (1996) uses these tax return data to examine the dynamics of earnings over the life cycle for males and females in New Zealand. A further, larger, dataset of around 30,000 individuals was constructed by the Inland Revenue Department in 2014. This sample contains longitudinal information covering a 19-year period from 1994 to 2012. It represents a 2 per cent sample of individual taxpayers with PAYE earnings or those individuals for whom a tax return was filed: the selection was based on the last two digits of taxpayers' IRD numbers. The characteristics of income dynamics using these data are explored by Creedy *et al*. (2019). A major advantage of using

[1] Some early studies used surveys containing retrospective histories, though these are subject to obvious recall bias. An alternative, though little-used approach is to attempt to construct a synthetic longitudinal dataset by using matching techniques with a time series of cross-sectional surveys. In New Zealand, the only attempt to produce such a dataset is by Ball (2016).

this type of tax return data is that it provides the most accurate reported income (and its component sources) with the minimum measurement error.

One difficulty with the IRD data is that only limited information on demographic characteristics such as date of birth and gender was collected. Gender had to be derived from the nominated title given by individuals (Mr, Miss, Mrs, Ms), but those without gender-identifying titles (such as Dr) were given a probability-weighted random allocation. Furthermore, the data could not be made publicly available to researchers.

3. The Integrated Data Infrastructure

Statistics NZ's Integrated Data Infrastructure (IDI) is an anonymised collection of national and regional data sources systematically and securely linked. The IDI contains a wide range of administrative data sources from government agencies, Censuses, Statistics NZ surveys, and non-government organizations linked at the individual level. These data collections in the IDI are linked through a 'central spine'. The IDI spine aims to include all people who have ever been resident in New Zealand.[2] It is constructed by probabilistically linking a set of three key collections. These are:

1. Inland Revenue Department: tax administrative data
2. Ministry of Business, Innovation and Employment (MBIE): visa information
3. Department of Internal Affairs (DIA): population information.

By combining tax records from 1999 onwards, New Zealand birth records since 1920, and long-term visas since 1997, the IDI spine covers the target population, an 'ever resident' New Zealand population.[3]

A number of other administrative data collections are linked directly to the spine. The collections that are of relevance to the present study are listed below where, as above, the relevant ministry is first given, followed by the nature of the data:

1. Inland Revenue (IRD): income and tax
2. Ministry of Social Development (MSD): social benefits[4]
3. Census 2013: ethnicity and qualifications

[2] This includes individuals who were born in New Zealand, permanent residents, people with a visa that allows them to reside, work or study in New Zealand, and those who can live and work in New Zealand without requiring a formal visa.
[3] Short-term visitors such as tourists are excluded from the spine.
[4] The Working for Families (WfF), a package designed to provide targeted social assistance to middle- and low-income families with children, is administered jointly by MSD and IRD.

4. Ministry of Education (MoE): education and training attainments
5. Ministry of Health (MoH): ethnicity
6. Ministry of Business, Innovation and Employment (MBIE): border movements and visa decisions
7. Department of Internal Affairs (DIA): births, deaths, marriages and civil unions
8. Statistics NZ Derived tables (summarised across the available data sources): date of birth (age), gender, ethnicity, person overseas spell, geographic information (from several sources)

These other data sources, such as education and health, are each linked to the spine separately and, therefore, are not directly linked to each other. This means that links between education and health records are available only for those individuals existing on the spine.

Datasets within the IDI are deterministically linked where common unique identifiers are available. Otherwise, personal variables such as full name, date of birth, and address are used for probabilistic matching; see Statistics New Zealand (2014), for details on the linking methodology. Personal identifiers such as names, addresses, and exact dates of birth are then removed to protect privacy and confidentiality. Instead encrypted, unique identifiers are assigned by Statistics NZ (snz_uid). These linkages are re-constructed every time a new refresh is added. The IDI is updated ('refreshed') up to four times a year to include new datasets and to update the existing data.

Individuals can be linked across different datasets using these unique identifiers, which are changed and reassigned in each 'refresh': the refresh archive used for the present exercise is IDI_Clean_20200120. Another unique identifier is a local one (source dependent identifier) derived by Statistics NZ. These remain constant for an identity across different refresh archives. An example is snz_ird_uid, a local identifier in Inland Revenue, constructed from the IR unique identifier (the IRD number).

It is apparent that by providing an appropriate infrastructure, the IDI allows 'big data' techniques to be applied to several national and regional data sources, so that various datasets can be systematically and securely linked at the individual level. This provides a unique opportunity for researchers and policy makers to study inherently longitudinal phenomena such as inequality and poverty at the national level. However, the size and the scope of the IDI makes the use equally challenging. The IDI contains more than 550 tables sourced from

14 entities, each of which is structured differently.[5] Data dictionaries are provided for most of these tables but the quality and the level of details vary substantially depending on the source agency.

4. Individual Income Data

The most extensive source of income information in the IDI is from Inland Revenue (IR), the New Zealand government's tax revenue collection agency. This information is derived from four main IR income tax forms, namely EMS, IR3, IR4S and IR20. Each of these is explained below.

In New Zealand, all businesses with paid employees are obliged to deduct and withhold income tax at source under the Pay-As-You-Earn (PAYE) system. This information is transferred to the Inland Revenue by filing the Employer Monthly Schedule (EMS), a mandatory monthly reporting requirement. Individuals who earn income other than salary and wages, interests, dividends and/or taxable Māori authority contributions are required to declare such incomes by filing an IR3 (an individual income tax return).

Company shareholder details are filed through the IR4S. This includes details of all shareholders, directors and relatives of shareholders who receive remuneration (with no PAYE deducted). The filing is required for all active and New Zealand resident companies. Partnership and Look-Through Companies (LTCs) are required to file an IR7 (labelled IR20 in the IDI dataset). An LTC is a form of company structure with limited liability where income and expenditure can be transferred to shareholders directly. The IR3 tax return is the most comprehensive form, including income information otherwise reported in the IR4S and IR20.

In addition, an IR-produced Personal Tax Summary (PTS) records income and tax deductions such as interest, dividends and tax credits for salary and wage earners. Under certain circumstances, this summary is sent automatically to taxpayers by IR or individuals can request a PTS to confirm whether that are paying the right amount of tax.

Income information for a given individual is sometimes available from more than one source: for example, IR3 and EMS. Thus, to construct a dataset covering the taxpayer population over the period of study (2000 to 2017), the following steps are undertaken. Firstly, data

[5] For example, the number of tables available from Ministry of Education is 16. Another 18 tables provide classification of variables when required.

sources are prioritized according to their comprehensiveness and checked in sequence. The most comprehensive source is IR3, followed by the PTS and then EMS. Therefore, if income information for a given individual is available from the IR3, it is recorded as final. Otherwise, availability in the PTS, then EMS, is checked. This procedure continues until all available taxpayers' income information is collected.

Following the steps described earlier, taxable incomes for individuals aged from 15 to 100, in any given year over the period of study (2000 to 2017), were collected. As the IDI has been anonymised, the day of birth is not available within the IDI. As a result, each individual's age is calculated based on year and month of birth.[6]

To examine the number of individuals for which taxable income is available for the whole period of study, a variable which records the number of missing incomes is generated. As expected, the number of missing incomes can range from a minimum of zero to a maximum of 18. The first row of Table 1 shows that there are about 1.6 million records with available taxable income for the full period of study.

Table 1: Number of Missing Income Observations

Years of Missing Incomes	Number of Observations	Percentage
0	1,605,192	29.76
1	236,457	4.38
2	188,733	3.50
3	171,009	3.17
4	167,892	3.11
5	165,306	3.06
6	165,273	3.06
7	166,098	3.08
8	175,134	3.25
9	172,545	3.20
10	168,717	3.13
11	188,901	3.50
12	183,270	3.40
13	187,641	3.48
14	209,061	3.88
15	250,020	4.64
16	461,340	8.55
17	531,285	9.85
Total Number of Observations	5,393,874	100

[6] There are about 100 observations which are linked to the spine but their sex is recorded as 'Null'. These are excluded from the final dataset.

Table 2 reports the number of individuals who can be tracked over time. The left-hand side refers to the number of observations derived from forward tracking, while the right-hand side shows the number of individuals who can be linked from backward tracking.

Table 2: Number of Individuals Who Can be Tracked over Time

Forward Time Period	Number of Observations	Backward Time Period	Number of Observations
2000 – 2001	2,706,264	2017 – 2016	3,374,817
2000 – 2002	2,596,416	2017 – 2015	3,167,538
2000 – 2003	2,507,487	2017 – 2014	3,001,617
2000 – 2004	2,427,453	2017 – 2013	2,853,174
2000 – 2005	2,351,232	2017 – 2012	2,724,786
2000 – 2006	2,276,466	2017 – 2011	2,601,636
2000 – 2007	2,179,797	2017 – 2010	2,488,128
2000 – 2008	2,106,831	2017 – 2009	2,386,293
2000 – 2009	2,033,793	2017 – 2008	2,285,229
2000 – 2010	1,965,630	2017 – 2007	2,182,179
2000 – 2011	1,909,920	2017 – 2006	2,084,046
2000 – 2012	1,852,917	2017 – 2005	1,990,845
2000 – 2013	1,795,470	2017 – 2004	1,903,752
2000 – 2014	1,742,280	2017 – 2003	1,820,487
2000 – 2015	1,695,045	2017 – 2002	1,742,265
2000 – 2016	1,649,100	2017 – 2001	1,670,214
2000 – 2017	1,605,195	2017 – 2000	1,605,195

Tables 3 and 4 present summary statistics (mean, median, and number of observations) of taxable income for the population aged 15 to 100 years old, over the 18-year period, 2000 to 2017. Table 3 includes only positive incomes, while Table 4 includes zero incomes.[7] For inflation adjustment, the base used is the first quarter of the year 2017. The CPI (Consumer Price Index) is obtained from the RBNZ website.[8]

[7] There are several obvious errors (infeasible values) in the reported taxable incomes (mainly under the IR3 table) for the following years, 2005, 2010, 2011, and 2012 within the IDI. These errors were fixed manually.
[8] This is reported under 'Economic Indicators': https://www.rbnz.govt.nz/statistics/.

Table 3: Summary Statistics for Individual Taxable Incomes (Excluding Zero Values)

			Inflation Adjusted		
Year	**Mean**	**Median**	**Mean**	**Median**	**Number**
2000	26,199	15,558	38,113	22,632	2,762,877
2001	25,667	16,374	36,230	23,113	2,788,077
2002	26,829	17,433	36,915	23,987	2,828,949
2003	27,445	18,164	36,834	24,378	2,883,957
2004	28,689	19,197	37,918	25,372	2,946,687
2005	29,812	20,185	38,335	25,956	3,012,834
2006	31,054	21,194	38,652	26,379	3,061,671
2007	32,296	22,205	39,203	26,953	3,108,504
2008	34,040	23,817	39,974	27,969	3,149,688
2009	35,208	24,583	40,154	28,037	3,169,503
2010	35,389	23,894	39,551	26,704	3,179,970
2011	36,122	24,230	38,643	25,922	3,218,658
2012	37,891	25,165	39,910	26,505	3,253,146
2013	39,459	26,264	41,207	27,427	3,273,705
2014	40,398	27,369	41,550	28,150	3,326,118
2015	41,590	28,533	42,669	29,273	3,400,860
2016	43,244	29,550	44,181	30,190	3,478,767
2017	44,095	30,537	44,095	30,537	3,568,758

* All the Mean and Median values are in NZ dollars ($).

Since the time spent overseas by an individual may partially explain the observed fluctuation of incomes over years, this information is also added to the income data. As stated earlier, income data are on an annual basis and therefore, to be consistent, the time spent overseas is calculated during the tax year (1st of April to 31st of March each year). For this, the 'person overseas spell' table, sourced from MBIE migration data, is used.[11] In this table, for each overseas spell, the start and end date, along with the length (in days) are reported.

[11] The MBIE migration tables keep the record of all border movements in and out of New Zealand. These are collected from passports.

Table 4: Summary Statistics for Individual Table Incomes (Including Zero Values)

			Inflation Adjusted		
Year	**Mean**	**Median**	**Mean**	**Median**	**Number**
2000	25,536	14,858	37,148	21,614	2,834,616
2001	24,980	15,549	35,261	21,948	2,864,706
2002	26,106	16,555	35,921	22,779	2,907,303
2003	26,674	17,218	35,800	23,109	2,967,267
2004	27,882	18,191	36,850	24,043	3,032,061
2005	28,957	19,109	37,237	24,572	3,101,727
2006	30,081	19,949	37,439	24,829	3,160,797
2007	31,531	21,220	38,275	25,758	3,183,876
2008	33,161	22,676	38,942	26,629	3,233,121
2009	34,280	23,362	39,095	26,644	3,255,306
2010	34,451	22,569	38,502	25,223	3,266,577
2011	35,201	22,971	37,658	24,574	3,302,823
2012	36,966	23,945	38,935	25,220	3,334,614
2013	38,502	25,001	40,208	26,108	3,355,032
2014	39,458	26,111	40,584	26,855	3,405,285
2015	40,703	27,320	41,759	28,029	3,475,005
2016	42,391	28,416	43,309	29,031	3,548,775
2017	43,310	29,530	43,310	29,530	3,633,420

* All the Mean and Median values are in NZ dollars ($).

5. Demographic Characteristics

Basic demographic characteristics such as gender, date of birth, and ethnicity are available in several data sources within the IDI. Gender and date of birth (age) are collected from the personal details dataset, which records Statistics NZ's best estimate of demographic information derived from multiple collections in the IDI using a set of specific rules. For ethnicity, the relevant data are collected from multiple sources. This is explained in further detail below. Table 5 reports the proportion of males and females in the final dataset.

Table 5: Summary Statistics for the New Zealand Taxpayers Gender

Gender	**Number of Observations**	**Percentage**
Male	2,731,068	50.63
Female	2,662,806	49.37
Total Number of Observations	5,393,874	100

5.1 Ethnicity Information

One of the main demographic variables describing the population is ethnicity. It is a measure of cultural identity and relates to ethnic group(s) with which people identify or feel they belong. In the New Zealand source datasets, ethnicity is a self-identified concept and individuals may specify multiple ethnicities. According to the statistical standard, all official statistics measuring ethnicity should have the capacity to collect a minimum of three and a maximum of six ethnic group responses per individual. The six include: European, Māori, Pacific Peoples, Asian, MELAA (Middle Eastern, Latin American and African), and other ethnicities.

Information on ethnicity can be found in several data sources within the IDI. This includes the 2013 Census along with administrative data sources such as DIA (births), ACC, MoE, MoH and MSD. The level of detail and quality of the data provided can vary from one source to another.

In collecting ethnicity information, two points are important. Firstly, the recorded information for an individual may differ across different data sources. This is partly because individuals may provide different responses to ethnicity question asked by different agencies. Different data sources can also record ethnicity information differently. Secondly, ethnic identity may change over time.

The ethnicity information for this study is collected preferentially from the 2013 Census which has the highest quality and coverage, followed by MoH and Statistics NZ 'personal details' datasets.[12] In addition to collating information from these data sources, a prioritised ethnicity variable is constructed. This means that individuals are classified into one ethnic group in a prioritised order of Māori, Pacific Peoples, Asian, European, MELAA (Middle Eastern, Latin American, and African), and Other. To construct this variable, an individual is classified as Māori, if a person's ethnic code in one of the three data sources is Māori. This process is repeated for other ethnic groups in order. Note that individuals with Asian ethnicity are further classified into three sub-groups including Indian, Chinese, and other Asians.

The summary of prioritised ethnicity is reported in Table 6. The values corresponding to the Asian ethnicity group are reported in parentheses, and reflects the summation of the three sub-categories of Indian, Chinese, and other Asians.

[12] The 'Personal details' dataset records Statistics NZ's best estimate of demographic information. This is derived from several data sources in the IDI.

Table 6: Summary Statistics for the New Zealand Taxpayers Ethnicity (Prioritised Ethnicity)

Prioritised Ethnicity	Number of Observations	Percentage
Māori	594,954	11.03
Pacific	289,104	5.36
Asian:	(647,685)	(12.01)
Indian	206,241	3.82
Chinese	194,739	3.61
Other Asians	246,705	4.57
European	3,105,621	57.58
All Others	756,510	14.03
Total Number of Observations	5,393,874	100

Table 7 reports the number of observations for each ethnic group without prioritisation. As stated above, more than one ethnicity can be recorded for a given individual and therefore, the total value adds up to more than the population size. For example, the number of observations which fall under the 'Pacific' group is 289,104 in the Table 6, this becomes 321,213 in Table 7. This is because there are 4,737 observations who identify themselves as both Indian and Pacific and there are 27,372 observations with both Māori and Pacific ethnicities.

Table 7: Summary Statistics for the New Zealand Taxpayers Ethnicity (Non-Prioritised Ethnicity)

Non-Prioritised Ethnicity	Number of Observations	Percentage of population
Māori	594,954	11.03
Pacific	321,213	5.96
Asian	(656,943)	(12.18)
Indian	208,119	3.86
Chinese	201,327	3.73
Other Asian	247,497	4.59
European	3,361,911	62.33
MELAA	73,926	1.37
Others	79,233	1.47
Unknown*	620,556	11.50
Total Number of Observations	5,708,736	106
Population Size	5,393,874	

*Note: 'Unknown' category in this table is reported under the 'All Others' sub-category in the Table 6.

5.2. Life Event Data

The Department of Internal Affairs (DIA) dataset within the IDI provides life events information such as births, deaths, marriages and civil unions registered in New Zealand. The notification of births and deaths was made compulsory in 1848. However, the recorded information may not be complete until later, when the digital storage of paper records occurs.[13]

Information about births for those individuals with a minimum of one income record is collected from the DIA_births. Table 8 reports the number of individuals in each birth cohort's category.

Table 8: Summary Statistics for the New Zealand Taxpayer Birth Cohorts

Birth Cohorts	Number of Observations	Percentage
1899 – 1903	2,043	0.04
1904 – 1908	9,771	0.18
1909 – 1913	27,960	0.52
1914 – 1918	53,619	0.99
1919 – 1923	86,595	1.61
1924 – 1928	117,282	2.17
1929 – 1933	133,383	2.47
1934 – 1938	149,682	2.78
1939 – 1943	188,925	3.50
1944 – 1948	243,000	4.51
1949 – 1953	280,452	5.20
1954 – 1958	326,139	6.05
1959 – 1963	387,501	7.18
1964 – 1968	402,420	7.46
1969 – 1973	444,312	8.24
1974 – 1978	459,300	8.52
1979 – 1983	535,146	9.92
1984 – 1988	569,142	10.55
1989 – 1993	524,529	9.72
1994 – 1998	366,954	6.80
1999 – 2001	85,713	1.59
Total Observations	5,393,874	100

Note: The total does not add precisely because of the Statistics NZ confidentiality rule requiring random rounding to base 3 (RR3).

Information about deaths for those individuals with a minimum of one income record is collected from the DIA_deaths source data. A summary of this information is given in the

[13] The birth registration records in the DIA is digitised since 1998. However, in correspondence with Department of Internal Affair, it is confirmed that since September 2019 refresh archive, the majority of birth records between 1990 and 1998 have been fully digitised.

Table 9. Since the first tax year is 2000, only death records which occurred after 1st April 1999 are collected.

Table 9: Summary Statistics for the New Zealand Taxpayers Death Records

Death Year	Number of Observations	Percentage
1999	9,777	2.58
2000	12,996	3.42
2001	13,989	3.69
2002	14,964	3.94
2003	15,522	4.09
2004	16,521	4.35
2005	16,851	4.44
2006	18,132	4.78
2007	18,945	4.99
2008	20,151	5.31
2009	20,832	5.49
2010	21,213	5.59
2011	23,148	6.10
2012	23,844	6.28
2013	23,964	6.31
2014	25,779	6.79
2015	26,898	7.09
2016	26,919	7.09
2017	29,106	7.67
Total	379,551	100

5.3 Qualification Data

There are several administrative data sources from which education and training information can be collected. These include education and qualification data from the Ministry of Education along with the highest qualification reported in the 2013 Census.

5.3.1. Ministry of Education Data

The main source of education and training information in the IDI is Ministry of Education data (MoE). It contains data on primary and secondary education, tertiary education, and industry and targeted training.

The aim is to collect the highest qualification an individual obtains each year. To do so, several MoE tables have been used. The first is the 'student qualification' table which keeps the full records of the highest secondary school qualification achieved by individuals aged 15 and over. In New Zealand, students must attend school until the age of 16. A nationally recognized qualification for senior secondary school students is NCEA (National Certificate of Educational Achievement) administered by NZQA (New Zealand Qualification

Authority). The three-level certificate of NCEA (level 1-3) can be achieved by students in Year 11 through Year 13. The student qualification table contains detailed information on provider code, qualification level code, exam results, and completion year among others.[14] Quality of work for a given level can be distinguished by Excellence or Merit. Table 10 summarizes the business rules used to define NCEA attainment.

Table 10: Business Rules to Define NCEA Attainment

Level of Qualification	Qualification Type	Qualification Code	Exam Results	NCEA Qual
NCEA 3	Certificates	1039	Achieved with Excellence(E)	39
			Achieved with Merit (M)	38
			Achieved with No Endorsement (ZZ)	37
			Not Achieved (N)	36
NCEA 2	Certificates	0973	Achieved with Excellence(E)	29
			Achieved with Merit (M)	28
			Achieved with No Endorsement (ZZ)	27
			Not Achieved (N)	26
NCEA1	Certificates	0928	Achieved with Excellence(E)	19
			Achieved with Merit (M)	18
			Achieved with No Endorsement (ZZ)	17
			Not Achieved (N)	16

New Zealand's tertiary education system comprises 10 levels. Level 1 to 3 are comparable to NCEA qualifications, level 4 to 6 cover trades, technical and business qualifications, and levels 7 to 10 covers degrees qualifications including graduate and postgraduate qualifications. Data on tertiary level qualifications can be found in the MoE's completion table. The earliest data available in this table is for 1994.

All quality-assured qualifications in New Zealand, including senior secondary school, tertiary and industry-based standards and qualifications, are covered in the New Zealand

[14] This does not include qualifications obtained overseas. Also, the records prior to year 2003 have been excluded because the National Student Index (NSI) was not available before then and hence the matching rate to the IDI was poor.

Qualifications Framework (NZQF). Table 11 summarizes the classification used to record the level of qualification in New Zealand.

Table 11: NZQF Structure: Level and Qualification Type

Level of Qualification	Qualification Types
10	Doctoral Degree
9	Master's Degree
8	Postgraduate Diplomas and Certificates, Bachelor Honours Degree
7	Bachelor's Degree, Graduate Diplomas and Certificates
6	Diplomas
5	
4	Certificates
3	
2	
1	

Source: New Zealand Qualifications Framework

To classify qualifications according to the NZQF, the Qualification Award Category Codes (QACC) have been used, as shown in Table 12.

Table 12: Qualification Levels

Level	Qualification Types	QACC Codes
10	Doctoral Degree	1 & 10
9	Master's Degree	11
8	Postgraduate Diplomas and Certificates, Bachelor Honours Degree	12 – 14
7	Bachelor's Degree Graduate Diplomas and Certificates	20 21
6 5	Diplomas	25 – 33
4 3 2 1	Certificates	34-60

In New Zealand, workplace learning operates through an industry training system arranged by Industry Training Organizations (ITOs). This workplace-based training can lead to a formal qualification (at levels 1 to 6) listed on the NZQF. The detailed information on type and level of program, field of study, duration of training (start and end date) and the level of credits achieved are provided in the 'industry training' table.

The education data collected from different MoE tables, including upper secondary school, tertiary and industry training are then combined. Given that individuals can complete multiple qualifications within a year, only the highest level is retained.

5.3.2 The 2013 Census

The Census also collects information on the qualification achieved by individuals at a certain point in time. This information is derived from the answers provided to the relevant questions on training and education asked of individuals.[15] In the Census, only individuals aged 15 and over were asked about their training and education. Also, the qualification information recorded in the Census rely on the respondent's interpretation of the highest qualification which in some cases might be different from the registered record derived from the MoE tables.

The Census question on the highest qualification requires a 'written-in' response. Therefore, in addition to non-response cases, a specific qualification level could not be assigned in the case of ambiguous responses. Moreover, a different level of qualification could be assigned if generic responses such as 'Diploma' or 'Certificate' were provided. For example, either of level 5 or 6 qualification could be assigned where the response is 'Diploma'.

While administrative MoE sources may provide high-quality information on education and training, it does not capture the whole New Zealand population. This is particularly pronounced in cases where Census data can be a reasonable alternative. These include the following:

i. The earliest date for which qualification completions data (in tertiary level qualifications) can be found is 1994; this means that there is no record for those individuals who achieved their qualifications before then.
ii. There are no records in administrative data for those individuals who did not complete a formal qualification.
iii. Qualifications obtained overseas might not be captured in administrative data.

For these and other reasons, the two data sources are complementary and are therefore combined.

[15] These are questions 26 to 28 of the 2013 Census individual form.

The next step is to combine the highest qualification information derived from the register data (MoE tables) with those recorded in the 2013 Census. The main variables collected are the highest level of qualification along with the corresponding completion date. This involves a number of challenges.

To keep track of how the educational profile of an individual changes over time, the qualification completion is an important variable. However, this is not specified where the qualification information is derived from the 2013 census. In such cases, the census date, 5th March 2013 is recorded as a proxy for completion date.

It is possible that some individuals completed more than one qualification at any given year. This may result in several qualification records per individual in a given year.[16] When this occurs, and qualifications are in the same level, the latest date is prioritised. Otherwise, when qualifications are at different levels the one that corresponds to the highest qualification gained is selected.

Another challenge occurs when an individual with a high qualification level obtains a lower level qualification later in life. For example, a person with a PhD degree (level 10 qualification) could subsequently obtain a graduate diploma (level 7 qualification) in another field. Since this is considered as an extra qualification, when a higher level of qualification is followed by a lower one, this is replaced by the highest prior level achieved.

Table 13 reports the summary of the highest qualification achieved by individuals over the period of study. These levels of qualifications are categorised into four broader groups, details are reported in Table 14.

Table 15 reports the summary of the highest qualification achieved by individuals corresponding to the four broader categories.

[16] The exact date when these qualifications were achieved might be different.

Table 13: Summary Statistics for the Highest Qualification Achieved by Individuals

Level of Qualification	Number of Observations	Percentage
No Qualification	434,028	8.05
Level 1 Certificate	342,081	6.34
Level 2 Certificate	423,750	7.86
Level 3 Certificate	452,829	8.40
Level 4 Certificate	483,861	8.97
Level 5 Diploma	201,051	3.73
Level 6 Diploma	174,675	3.24
Bachelor Degree	555,990	10.31
Post-Graduate and Honours Degrees	151,245	2.80
Master Degrees	107,301	1.99
Doctorate Degree	28,743	53
Overseas Secondary School Qualification	131,808	2.44
Response Unidentifiable	85,692	1.59
Not Specified	42,675	79
Missing	1,778,145	32.97
Total	5,393,874	100

Table 14: Highest Qualifications Broader Categories

Level of Qualification	Categories included
No Qualification	No Qualification
School	Level 1-3 Certificate and Overseas Secondary School
Post-School	Level 4 Certificate and Levels 5-6 Diploma
University Degrees	Bachelor and Post-graduate and Honours & Masters & Doctorate Degrees
Missing	Response Unidentifiable and Not Specified & Missing

Table 15: Summary Statistics for the Highest Qualification Achieved by Individuals

Level of Qualification	Number of Observations	Percentage
No Qualification	434,028	8.05
School	1,350,468	25.04
Post-School	859,587	15.94
University Degrees	843,279	15.63
Missing	1,906,512	35.25
Total	5,393,874	100

5.4 Geographical Information

Information about where people live is collected by various government agencies. The recorded information is updated by organisations when a change of address is notified. The information provided to Statistics NZ is then geocoded; that is, text addresses are converted to standard geographic locations.

Two central geographic (or address notification) tables are derived within the IDI. Ten sources (across seven agencies) are used. These include: ACC client addresses; 2013 Census; Inland Revenue (IR) tax registration addresses; Ministry of Education secondary school records; Ministry of Social Development (residential and postal addresses); Ministry of Health (Primary Health Organisation registers (PHO) and National Health Index records (NHO)); and New Zealand Transport Authority (Motor Vehicle Register addresses and Driver License Registration addresses).

The data recorded in these tables include a range of geographical information such as 'meshblock', area units, territorial authorities (TA), District Health Board (DHB) areas, and regions.[17] Despite the similarities observed, there is an important difference between the two tables. While one of these tables, 'address notification – full', provides a full list of every geocoded address by collating all address change notifications, the other, 'address notification', uses a simple set of business rules to limit the full address table to a best-guess list of residential addresses. In other words, the second table is a prioritised version of the former where addresses from sources regarded as higher quality are prioritised.

For the exercise described here, administrative address sources are classified into two 'Quality Tiers' (Tier 1 and Tier 2) where quality is defined based on characteristics of the source data (whether a residential address is required and obtained by the agency or not). Accordingly, the Tier 1 sources include addresses where an agency indicates that a residential address is both required and obtained from clients and therefore has a higher quality. As a result, the number of notification records in the prioritised table is almost one quarter of those in the full table. Table 16 shows the Quality Tiers and ranking in the corresponding Tier.

[17] A meshblock is the smallest geographical unit used by Statistics NZ. The median size of a meshblock was approximately 87 people in the 2006 Census.

Tables 16: Prioritised Address Source Code

Address Source Code		Tier	Rank in Tier
Census (here 2013)	CEN	1	1
Ministry of Social Development – Residential address	MSDR	1	2
Ministry of Health – Primary Health Organisation registers	MOHP	1	3
Ministry of Health – National Health Index records	MOHN	1	4
New Zealand Transport Autority – Motor Vehicle Regsiter address	NZTM	1	5
ACC client addresses	ACC	2	1
Inland Revenue – tax registration addresses	IRD	2	2
Ministry of Social Development – Postal address	MSDP	2	3
Ministry of Education secondary school records	MOES	2	4
New Zealand Transport Authority – Driver License Register address	NZTD	2	5

Source: Derived from Metadata Address Notification Source within IDI

It is important to recognize that neither of these tables provide full coverage of residential movements in New Zealand. The coverage is particularly low for those individuals with less contact with government agencies where addresses are obtained. Furthermore, most address records available within the IDI are notifications of address updates and, therefore, the associated date refers to the time when the information is provided to an agency not the actual date when the residential move occurred.

For this study, the prioritised address notification table is used. Note that the number of records per individual varies extensively. Over the period examined (2000-2017), it ranges from 1 to 104 depending on the number of times an update is provided. This would be problematic when several distinct addresses in a given year are recorded. To deal with this issue, a variable which represents the length of stay (based on notification date and replacement date) is created. In a given year, the record with the longest length is assumed to be the main location and all the other records in that year are dropped. Doing so, the maximum number of records for an individual decreases to 40.

6. Conclusions

This paper has described the construction of a unique individual-level dataset from administrative and survey data sources available within the IDI. The focus of this study is the New Zealand taxpayer population (aged 15 to 100) where at least one income record in the tax register data over the period of study (2000 – 2017) exists. The final population covers approximately 5,393,874 taxpayers for whom, a range of information including, but not limited to, taxable income, gender, ethnicity, education level and location have been compiled.

The availability of suitable data has been a constraint on income dynamics research but a more general increase in the availability of good-quality data on the full population provides new opportunities. The present data construction exercise allows a more extensive analysis of individual income inequality and mobility than has previously been possible.

References

Ball, C. (2016) Estimating income dynamics from cross-sectional data using matching techniques. *Working Papers in Public Finance*, 06/2016. Victoria University of Wellington, Wellington, New Zealand.

Carter, K., and Gunasekara, F.I. (2012) Dynamics of income and deprivation in New Zealand, 2002–2009. A descriptive analysis of the Survey of Family, Income and Employment (SoFIE). Wellington: Department of Public health, University of Otago.

Carter, K., Mok, P., and Le, T. (2014) Income mobility in New Zealand: A descriptive analysis. *New Zealand Treasury Working Paper*, no. 14/15.

Creedy, J. (1996) Earnings Dynamics over the life cycle: new evidence for New Zealand. *New Zealand Economic Papers*, 30, 131-153.

Creedy, J., Gemmell, N. and Laws, A. (2019) Relative income dynamics of individuals in New Zealand. *New Zealand Economic Papers*, 1-18.

Smith, H. and Templeton, R. (1990) A Longitudinal Study of Income. *Statistics New Zealand.*

Statistics New Zealand (2014) Linking Methodology used by Statistics New Zealand in the Integrated Data Infrastructure Project, Technical Report. Wellington: Statistics New Zealand. Available from www.stats.govt.nz.

About the Authors

Nazila Alinaghi is a Research Fellow in Public Finance at Victoria Business School, Victoria University of Wellington, New Zealand.
Email: nazila.alinaghi@vuw.ac.nz

John Creedy is Professor of Public Finance at Victoria Business School, Victoria University of Wellington, New Zealand.
Email: john.creedy@vuw.ac.nz

Norman Gemmell is Professor of Public Finance at Victoria Business School, Victoria University of Wellington, New Zealand.
Email: norman.gemmell@vuw.ac.nz

VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA
Chair in Public Finance
Wellington School of Business and Government

Working Papers in Public Finance