A Longitudinal Database for the Analysis of Family Incomes in New Zealand

Nazila Alinaghi, John Creedy and Norman Gemmell

WORKING PAPER 06/2020
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

A Longitudinal Database for the Analysis of Family Incomes in New Zealand[*]

By

Nazila Alinaghi, John Creedy and Norman Gemmell[†]

Abstract

This paper describes the construction of a unique longitudinal family-level dataset that allows the dynamics of family incomes in New Zealand to be examined over the period, 2000 to 2017. The data are obtained from the New Zealand Integrated Data Infrastructure, requiring a complex linking exercise to be carried out. The dataset provides a basic resource for economic analyses of income inequality in which substantial attention is paid to the accounting period over which income is measured, and the nature of income changes over calendar time for different date-of-birth cohorts.

[*] 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 are grateful to Megan Gath of Statistics New Zealand for sharing the codes used for research undertaken on the feasibility study of census transformation programme; Inny Kang of Statistics NZ for assistance with 2001 and 2006 stand-alone censuses; Catherine Layne of Statistics NZ for assistance with IDI data; Sarah Crichton and Robert Templeton of the NZ Treasury for assistance in linking the 2001 Census to the IDI. We have also benefited from discussions with Robert Templeton, Sarah Crichton, and Christopher Ball regarding the data.

[†] Wellington School of Business and Government, Victoria University of 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 longitudinal dataset that allows the dynamics of family incomes in New Zealand to be examined. While it has long been recognised that income inequality and poverty depend on the changing size and composition of families, and on the consequent dynamic characteristics of family incomes, the vast majority of studies of inequality in New Zealand rely on cross-sectional comparisons, using an annual accounting period.[1]

Information about income dynamics of families or households is extremely rare in New Zealand. The most extensive survey with a longitudinal nature in New Zealand is the Survey of Family, Income and Employment (SoFIE). It covers nationally representative households, but this survey was conducted only for a limited duration. It contains eight annual 'waves', from October 2002 to September 2010, and covers around 11,000 households (consisting of about 22,000 individuals).[2]

It may be thought that national population censuses, which contain a wealth of demographic information about individuals and their families, can be linked over time. However, censuses often lack detailed information about incomes. Total personal income is recorded as the total before-tax income received by an individual in the 12 months preceding census day, and is collected using income bands instead of exact dollar amounts. Furthermore, censuses are too infrequent to provide sufficient information about the time spent in particular states, and provide no data on life-course events such as having children, or family formation and/or dissolution. The Census in New Zealand is normally conducted every five years.[3]

The present research has been made possible by access to the New Zealand Integrated Data Infrastructure (IDI), which contains a wide range of administrative and survey data sources from various government agencies and non-government organisations (NGOs).[4] Unlike the Nordic countries, for example, New Zealand does not have unique identifiers which can be used across government departments with which individuals come into contact, and there are

[1] The small number of studies investigating individual (as opposed to family or household) income dynamics in New Zealand are discussed briefly by Creedy *et al.* (2019) and Alinaghi *et al*. (2020).
[2] Carter and Gunasekara (2012) and Carter *et al*. (2014) provide graphical descriptions of income mobility in New Zealand using SoFIE data.
[3] However, following the 2009 and 2010 Canterbury earthquakes, the 2011 census was postponed to 2013.
[4] At the time of the present exercise, the only census year linked to the IDI was the 2013 census. Due to the data quality issues, the release of 2018 Census data has been postponed; for further details on the issues see 2018 Census External Data Quality Panel (2020). At the time of writing, only two tables of the 2018 Census were available within the IDI, namely individual and dwelling. However, the individual level data provided lack a family identification number and, therefore, cannot be used for family construction purposes.

strict rules (within the Privacy Act) about data-sharing among agencies. Hence, linking datasets involved an extensive deterministic and probabilistic matching exercise.[5] The structure of the IDI can be described as a 'central spine' through which a range of administrative and survey datasets are linked at the individual level. This is briefly described in Section 2.

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. A huge advantage of this linking is that, in further analyses of inequality and poverty dynamics, there is no need to rely on relatively small samples: the full relevant population data are available for each component data source. However, it is necessary to restrict the IDI spine population to the subset of interest. Since this research is primarily built on an individual database compiled and described in Alinaghi *et al*. (2020), the subset includes New Zealand residents who have received (at least) one form of income over the period of study, 2000 – 2017. A substantial challenge here is to add information about the relationships between individuals within families or households to the database for individuals where the only comprehensive source of family or household membership within the IDI is the 2013 census. Alternatively, there are multiple administrative data sources where these relationships, including parents-children and partners, are observed at one or several points in time. However, the study of income dynamics at the family or household level requires linked relationships across time. Earlier attempts to construct families and/or households from linked administrative data are briefly discussed in Section 3.

In constructing and using longitudinal data, a basic decision must be made regarding the unit of analysis. In cross-sectional studies of income (or consumption) inequality and poverty, different unit types have been used depending on the precise nature of the research question. The choice is among individuals, 'families' and households. No information is usually available about the precise nature of income sharing among family and household members, and an 'equal-sharing' assumption is often used. In the present context, the fundamental unit must be the individual, in view of the fact that families and households experience changes in size and composition over time. Yet, welfare measures, or metrics, often require information - at any point in time - on the total resources available to the individual. Hence, it is necessary to be able to trace connections with other individuals and their characteristics over time. The

[5] For a detailed description of the IDI linkage; see Statistics NZ (2014).

choice of unit is discussed briefly in Section 4, where it is explained that the present exercise involves connections with other close 'family' members.

Section 5 explains how the different sources are linked to generate data on income and various demographic characteristics for families. Section 6 describes how the data from various sources are combined, and the basic form of the dataset. Brief conclusions are in Section 7.

2. The Integrated Data Infrastructure

Statistics NZ's Integrated Data Infrastructure (IDI) is an anonymised research collection of national and regional data sources systematically and securely linked. The IDI contains a wide range of administrative data sources from government agencies, Census, 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. 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. 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.[6]

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
1. Ministry of Social Development (MSD): social benefits[7]

[6] Short-term visitors such as tourists are excluded from the spine.
[7] 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.

2. Census 2013: ethnicity and qualification
3. Ministry of Education (MoE): education and training attainments
4. Ministry of Health (MoH): ethnicity
5. Ministry of Business, Innovation and Employment (MBIE): border movements and visa decisions
6. Department of Internal Affairs (DIA): births, deaths, marriages and civil unions
7. Statistics NZ Derived tables (summarised across the available data sources): date of birth (age), gender, ethnicity, person overseas spell, geographic information (from several data 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, for example, 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 of 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. The main characteristics of a local identifier is that it remains 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).

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 longitudinal phenomena. However, the size and scope of the IDI makes its use equally challenging. The IDI contains more than 1000 tables sourced from 14

entities, each of which is structured differently.[8] Data dictionaries provided for most of these tables contain detailed information about the variables used. However, the quality and the level of details provided vary substantially depending on the source agency.

3. Earlier Use of Administrative Records to Link Families/Households

In the absence of a high-quality panel dataset which provides a nationally representative picture of New Zealand families or households and their income and labour dynamics over time, families/households can be constructed using administrative data sources. The key challenge is that administrative datasets observe parent-child and couple partnership relations at only one point (or at best several points) in time. However, the study of income dynamics at the family or household level requires linked relationships across time.

In New Zealand, there have been few attempts to construct family and/or household units from administrative data sources. This section briefly surveys these earlier attempts.

3.1. New Zealand Census Transformation Programme

As a result of the Census transformation strategy agreed by the New Zealand government in 2012, a series of studies investigating the feasibility of providing census-type information from administrative data sources has been conducted by Statistics NZ.[9] The potential for such a transformation has previously been explored by the Nordic countries where, in Denmark (since 1981), Finland (since 1990), and Norway and Sweden (since 2011), the census has been entirely substituted by administrative data.[10]

The most extensive in a series of studies in New Zealand, exploring the potential for administrative sources to provide census-type information on households and/or families, was undertaken by Gath and Bycroft (2018). They constructed datasets at two levels, households and families, based on usually-resident individuals at census day.[11] Since the current administrative datasets do not collect the information required to construct families or households in accordance with the Statistics NZ's standard definitions, they use address

[8] For example, the number of tables available from Ministry of Education is 16. There exist another 18 tables which provide classifications of variables when required.
[9] Publications on this can be found in https://www.stats.govt.nz/methods-and-standards/census-transformation-programme/.
[10] The fact that each individual who resides in any of the Nordic countries has a unique identifying code facilitates such transformation. In New Zealand, on the other hand, the Privacy Act reflects a strong concern for individual privacy and prevents sharing of data across government agencies.
[11] As mentioned earlier, the IDI spine aims to include all people who have ever been a 'resident' in New Zealand. This allows a resident population to be selected at a given point in time (the corresponding date in Gath and Bycroft (2018) is 5th March 2013, the census day).

information to construct proxy households. People in these households are then grouped into families using the family information known to the government agencies. Due to the quality issues with address information and the limitations of geocoded address in the administrative datasets, it is likely that some individuals are incorrectly placed into households.

The aim of the Gath and Bycroft (2018) study is to evaluate the extent to which census household and family information can be obtained from existing administrative records. As a result, the information derived from administrative sources is benchmarked against the 2013 Census. To construct households, individuals who share address at a certain point in time are grouped into one household. For this, address information available in several sources within the IDI is used. However, for families the most recent relationships, relative to 5th March 2013, including parent-child and partnership information, are first collected. The next step undertaken is to use this information in conjunction with address data to create family nuclei within households. Gath and Bycroft (2018) conclude that despite the limitations observed, there is potential for providing household information on an aggregate level. However, the potential is severely limited with respect to families, largely due to the lack of family coverage in the administrative data sources.

3.2. Other Studies

The latest attempt to provide better identification of households through improving quality of address information is made by the Social Wellbeing Agency (SWA, 2020). In this study a series of detailed cleaning rules is applied to improve the accuracy of address information provided by different sources. This information is used to construct households. These households are then benchmarked against the 2013 Census and against several other surveys available within the IDI; namely the General Social Survey (GSS), Household Economic Survey (HES) and Household Labour Force Survey (HLFS). The SWA (2020) conclusion is that household-level matching rates are variable over time and across different validation sources used. This is attributed to the observed differences in the accuracy rates of different sources and the sample size differences between the census and the surveys.

In recognition of the limitations associated with address information used in the earlier household exercises, the focus of the current study is on the construction of a family-level database, as in the second level (families) of data construction undertaken by Gath and Bycroft (2018). However, an important objective of the current exercise is to capture changes in family formation over time. Therefore, instead of collecting the most-recent recorded

relationships relative to the 5th March 2013 census date used in Gath and Bycroft (2018), relationships within each tax year (from 1st April each year to 31st March the following year) are collected. Further, in this study an age limit of under 18 years old for dependent children is added. Finally, the constructed families in Gath and Bycroft (2018) are benchmarked against the 2013 census, whereas family information used here includes the 2013 census along with the two other stand-alone censuses in 2001 and 2006.

4. The Unit of Analysis

The question of the appropriate unit of analysis in longitudinal surveys is well captured in the following quote from Buck *et al.* (1995), in their report prepared 25 years ago for the New Zealand Treasury:

> The 'unit of analysis' in virtually all longitudinal surveys is an individual person, not the family or household. (This contrasts with cross-sectional social surveys which, depending on their purposes, may use any one of these different units of analysis as their focus.) The reason for the focus on individuals is very simple: it is impossible to define a longitudinal family or household in any rigorous way which would enable the unit to be followed over time. New families and households are continually being created, and existing ones have ever-changing memberships (and may cease to exist). By contrast with this flux, the concept of an 'individual' is stable in a longitudinal context. This does not mean that longitudinal surveys cannot tell us about families and households and their dynamics – quite the opposite. But the necessary information is derived from individuals who are related to their family or household context (which changes over time). Buck *et al.* (1995, p. 2).

The starting point of any attempt to construct a longitudinal database in which families can be identified is thus the formation of a database for individuals, such as that described in Alinaghi *et al*. (2020).

The question then remains of the extent to which each individual, as the fundamental unit, can be linked at any time to other individuals who may be part of the same household or family, and who may be considered to share their resources to some extent. Linking individuals to all household members at any time, within the IDI, presents a severe challenge, largely associated with difficulties of identifying precise residential addresses; see Gath and Bycroft (2018) and SWA (2020), among others.

The present study is therefore limited to linking individuals to 'family' members, where the family is defined as: either a single adult living alone, single adult living with children, or a couple with or without dependent children. It must be acknowledged that this definition

excludes certain types of family relationship, such as where multiple generations of adults are living in the same household as a close family unit.

5. Constructing Families

This section explains the construction of a new family-level dataset using several administrative data sources within the IDI. As mentioned above, this dataset is constructed from an initially compiled individual-level database described in Alinaghi *et al*. (2020). In order to collect family relationship information, three steps were undertaken. In the first step, the unique identities of those individuals for whom at least one income record in the tax register data exists over the period of study (2000 – 2017) were collected. In the second stage, parents and children were linked. The third stage involves establishing relationships between adult partners with and without children. These last two stages are described in subsections 5.1 and 5.2.

5.1. Parent-Child Relationships

Relationship information linking parents and children has been extracted from several data sources. These include New Zealand registration of births from the Department of Internal Affairs (DIA), social welfare benefits information from the Ministry of Social Development (MSD), Working for Families (WfF) tax credit payments jointly from the Inland Revenue Department (IRD) and MSD, and visa information from the Ministry of Business, Innovation and Employment (MBIE).

The DIA birth registrations data contain birth information dating back to 1848, although the introduction of digital storage of records began in 1998.[12] The data include unique identifiers for both parents (where they are recorded) and the child. Since parent-child links derived from this data source encompass individuals with a wide range of ages (who are not necessarily the focus of this study), further refinement is required. In this study, a child in the definition of the 'family' refers to a dependent child and, therefore, those aged 18 years and over are excluded from the final dataset at any given year. In this definition of family, a person who is 18 years or older is dealt with as an adult but it is possible that these young adults are living with their parents and, therefore, are part of the economic family unit. However, unless sufficiently high-quality address data are readily available, this is hard to distinguish.

[12] Correspondence with the Department of Internal Affair confirmed that, since a September 2019 archive refresh, the majority of birth records between 1990 and 1998 have been fully digitised.

To capture any change in family formation over time, a 'reference date' is set annually. The reference date is a time period over which the family relationship information is collected. It can be a calendar year, a fiscal year, or a specific date. For the reference date to be consistent with the income data, a reference year is based on the tax year: this runs from 1st April each year to 31st March the following year.

The target population of DIA-Births dataset is everyone born in New Zealand (and children born overseas but adopted in New Zealand) and, therefore, those who are not born in New Zealand cannot be linked to their parents. To deal with this concern, other complementary data sources such as MSD, WfF, and MBIE are used to improve the coverage. It is important to stress that family information of individuals who migrated to New Zealand prior to 1997 or those who do not require a visa to live and work in New Zealand, such as Australian migrants, are not available in the MBIE data sources (and might not be available in the DIA data if none of their children is born in New Zealand). However, if they are in receipt of any benefits, family information might be available in either the MSD or WfF sources.

The second source from which parental relationships can be extracted is MSD benefit records for children. This provides the unique identifiers of parents (the main beneficiary recipient) and dependent children receiving benefits at any time from 1993. A beneficiary is a person or family in receipt of main benefits. These include Unemployment Benefit (UB), Domestic Purposes Benefit (DPB), Widow's Benefit (WB), Emergency Maintenance Allowance (EMA), Independent Youth, Orphan's and Unsupported Child, Sickness Benefit (SB) and Invalid's Benefit (IB). Since this source stores the dates of inclusion in receipt of benefit for children, the reference date is again set to capture those parental relationships that exist in the relevant tax year.

The third source of parental relationship data relates to WfF, a package designed to provide targeted social assistance to middle- and low-income families with children. This package consists of three categories, namely WfF tax credits, and two additional welfare benefits: Accommodation Supplement, and Childcare Assistance. The information is administered jointly by MSD and IRD. The dataset for children provides detailed information on receiving families along with the relationship period (start and end dates) since 1999. As with the MSD data source, the reference date is set to capture parental relationships recorded in each tax year.

The next data source is for immigration data provided by the MBIE. This contains information on all migrants who have ever lodged applications with Immigration New Zealand from 1997. Since migrants who have applied for and been granted a residence visa are the population of interest, migrants with temporary visas such as visitor visa (and those whose application was declined) are excluded.

A complication with this data source is that a single application number can be associated with several clients. While this allows family members to be identified and linked, it may cause an incorrect classification of some individuals as forming a family when they do not actually belong to the same family. For example, individuals for whom their agents, sponsors or employers make a single visa application on their behalf are all assigned the same application number. To avoid such misclassification, individuals are divided into two categories: adults and children according to their age (18 and above, or less than 18 years, respectively). Applications in which the number of adults is more than two, or applications without children, are excluded. Second, to infer a parental relationship, an age difference of at least 14 years between the oldest child and the oldest adult is required. Third, the tax year restriction is applied to the decision date based on when the final decision on the visa application is made.

Parent-child relationships extracted from these data sources are then combined using the IDI unique person identifiers. Data sources are prioritized according to their coverage of the provided parental links and are checked in sequence. The most comprehensive data source is the DIA-Births. Therefore, if a parental link for a given individual is available in the DIA-Births, it is recorded as final. Otherwise, availability of links in the WfF, migration data from MBIE, and MSD is subsequently checked. This process is repeated for both parents (where available) and for all tax years from 2000 to 2017. The corresponding identification numbers for parents (parent 1 and parent 2) may be represented in the reverse order when different data sources are used. To link the family members correctly, each parent and his/her relevant children are collected separately. This information is then combined using partnership relations. In cases where more than one relationship is available, only the latest one is retained.

While it is possible to infer a partnership between the two parents from the data sources described earlier, partnership status may change over time. Furthermore, the steps undertaken can only capture partners with children and those without children are not represented in this

part of the final dataset. The following section outlines the further steps undertaken to address these shortcomings.

5.2. Partnerships

The next part of relationship information is between adult partners. Formal relationships such as marriages and civil unions are fully captured by administrative data sources, but informal partnerships can only be partially captured.[13] That is because registration of marriages and civil unions is compulsory, but information on a partnership relationship is required only for birth certificates, and social assistance where benefit entitlement depends on partnership status.

The records of marriages and civil unions registered in New Zealand can be found in the DIA datasets. Records of marriages that occurred overseas but were attended by a NZ Representative are also included in this dataset. The DIA marriage and civil union registrations provide detailed information about each partner, date of marriage or civil union, and dissolutions where applicable. The Civil Union Act 2004 came into effect from 26 April 2005. Since then, the data have been available in the DIA dataset. Under this act, both same-sex and opposite-sex civil unions can be registered. Same-sex marriages in New Zealand have been legally recognized since August 2013.

To construct partnership links for this study, two individuals are linked as partners if they are married at any point before the end of the relevant tax year, with no records of dissolution in that year. For example, 31 March 2000 is considered as the end point of tax year 2000. Alternatively, if there is any record of dissolution, it should refer to a date when the corresponding tax year is over. Otherwise, these two individuals are not linked as partners in that particular year. This process is repeated for each tax year from 2000 to 2017.

As explained in the previous subsection, children and parent links (up to two parents) are collected from the DIA birth registrations. These data can be used to infer partnerships between two parents at the time of the child's birth, if both parents are listed in the birth registration. To be considered as partners in a given tax year, the child should be born during the relevant tax year.

[13] Around one in five New Zealanders who are living in a relationship have chosen not to marry (336,591 people identified themselves as having a partner but not legally married in Census 2001). For the full report, see https://www.beehive.govt.nz/release/questions-and-answers-civil-union-and-relationships-statutory-references-bills.

The next data source from which partnership information can be collected is the Benefit Dynamics Datasets (BDD) provided by the MSD. These datasets include information on all individuals who are in receipt of a working-age social welfare benefit since 1993. However, partnership information for superannuitants is not currently available in the MSD datasets. The partnership dataset, in particular, provides benefit histories of partners in receipt of benefits. Given that the time-spell information is provided for each record, partners' links are retained if they are together at any point during a given tax year.

The third source of partnership information consists of WfF datasets. The Working for Families package (including WfF tax credits, Accommodation Supplement, and Childcare Assistant) provides details of benefit recipients and their partners. As with other datasets, partnership information is collected for those individuals in a relationship over a given tax year.

Previous administrative data sources used to obtain partnership information are mainly focused on local residents. However, the NZ population consists of many individuals who migrated from overseas at some point. In particular, the administrative data sources examined earlier may be representative of the overall national population, but this does not include those individuals (and their partners, where applicable) who recently migrated to New Zealand. That is partly because those individuals (and their families) are less likely to interact with government agencies. It is also possible that no corresponding record can be found in the latest census conducted. To deal with this problem, immigration data from MBIE are used. An important caveat, when these data are used to infer partnership information, is that two applicants with the same application number might not necessarily be partners. To avoid any incorrect inference, partnerships are inferred only for those partners who have children together.

Partnerships information collected from all these data sources are then combined using the IDI unique person identifiers. An individual may appear several times in the combined dataset, in which case, only the latest indicated partner from any source in each year is retained. This process is repeated for all the tax years from 2000 to 2017.

5.3. Additional Relationship Information from Census Data

As mentioned above, family information provided by administrative data sources tends to be relatively accurate, as confirmed by Gath and Bycroft (2018) when the constructed 'family

nucleus' are benchmarked against the families available in the 2013 Census. However, the population covered by administrative data is restricted to those individuals who have interacted with certain government agencies, and where the details are required and collected. By contrast, census datasets target complete coverage of New Zealand's population, but they provide family information only at particular points in time. Linking census datasets provides an opportunity to add a longitudinal aspect to a series of census snapshots. However, such an exercise is not feasible within the IDI, because the only census data available at the time of the present data construction exercise is the 2013 census.[14]

Since income data within the IDI is available from 1999, any census data prior to this date would add little value to the existing datasets. Therefore, the other censuses that can potentially be used to create linked datasets are 2001 and 2006, none of which is available within the IDI environment.[15] Access to these two stand-alone datasets has been made available to the project by Statistics NZ. However, unless the proper linking between the individual records in the stand-alone censuses and the records in the spine is formed, the additional family information available in the 2001 and 2006 censuses cannot be used. To link the records from each of these censuses to the spine manually, linking variables are required. Ideally, the first and last names can be used as linking variables. However, in accordance with the Privacy Act 1993, personal identifiers are removed before these datasets are made available to the project. As a result, linking variables are limited to date of birth (itself limited to the year and month of birth), gender, and usual residence. The latter is classified according to the 'meshblock', defined as the smallest geographical areas in the NZ geographical classification, representing roughly 30 to 60 dwellings and/or 60 to 120 residents. Among the three proposed linking variables, only gender is readily available. The following section elaborates the issues related to use of the other two linking variables, namely date of birth and usual residence.

The main difficulty in forming a manual linkage relates to the date of birth variable. This is because an age variable instead of date of birth is provided. While age is derived from date of birth, it is not specific enough to be used as one of the key linking variables. To overcome this difficulty, two shortened versions of these datasets, including the date of birth variable, were subsequently requested and provided by Statistics NZ to the project. However, to

[14] At the time of writing, only two tables of the 2018 Census are available within the IDI, namely individual and dwelling. However, the available individual level data lack any family identification number and, therefore, cannot be used for family construction purposes.
[15] Access to these two stand-alone datasets can be provided by Statistics NZ Microdata team upon request.

protect privacy, the precise date of birth was not provided, only the year and month of birth were available. Also, when these shortened versions were sorted by dwelling level identifier (**batch_nbr**), it was revealed that the records in these versions lacked a unique person-level identifier within the dwelling. To increase linking precision, 17 additional variables were then requested and provided to the project. These variables include gender, ethnicity, family role, legal and social marital status, qualification, income and occupation, among others. To have a stand-alone census with date of birth instead of age, the latest versions of these datasets (the shortened versions) were first linked to the longer versions initially provided. However, due to unavailability of person-level identifiers and the precise birth date in the datasets provided, further cleaning was required. For this, records with an exact match on all these variables were identified as duplicate records, which were then excluded from the final datasets. That is because the values for most of these variables were reported missing. Doing so, the dates of birth available in the shortened versions were added to the existing stand-alone census datasets. Records without a date of birth or those for which a date of birth was not identifiable were excluded. The next step was to attach dwelling information including meshblock codes to the individual-level records.[16]

The last linking variable is usual residence. Information about where people live is gathered by various government agencies. The recorded information is updated by organisations when a change of address is notified. This information is then provided to Statistics NZ, and are coded such that addresses in text form are converted to standard geographic locations.

Two central geographic (or address notification) tables are derived by Statistics NZ and available within the IDI. Ten sources (across seven agencies) currently contribute to these tables which 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 Autority (Motor Vehicle Regsiter addresses and Driver License Registration addresses).

The data recorded in these tables contain a range of geographical information such as meshblock, area units, Territorial Authorities (TA), District Health Board (DHB) areas, and

[16] Statistics NZ provided dwelling files separately. To append dwelling information to the rest of variables, common identifiers (batch-nbr) were used. Records with insufficient information for key linking variables were excluded.

regions. 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 coded 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 with the higher quality are prioritised.[17] As a result, the number of notification records in the prioritised table is almost one quarter of those in the full table. Meshblocks of each individual's address are determined using the full table in the IDI (the IDI refresh 20200120 is used).[18] In order to be able to compare the area classification over time (with earlier census), a meshblock concordance table is used for mapping.[19]

The next step was to add the two key linking variables, namely date of birth and gender, to the residential address. Finally, these stand-alone census datasets are linked to the IDI spine using core linking variables. To do so, records with the same linking variables from the IDI are required. The records of this dataset are then compared with records from each of the stand-alone censuses using linking variables. The values of linking variables for each pair of records are checked to see the level of agreement between them. If the core linking variables are in agreement, the link is successfully created and a global unique identifier used within the IDI and across different datasets (**snz_uid**) is added to each record. Observations linked to more than one record in the IDI spine are excluded from the final datasets.[20] Since partnership relations particularly in the form of *de facto* relationships are not extensively covered by administrative data sources, these two censuses are mainly used to collect the relevant information. Therefore, the final step is to identify partners who are both linked to the IDI spine. For this, a variable which contains family role information is used. As a result of the steps undertaken, the final datasets contain the global unique identifiers for both partners.

[17] Administrative address sources are classified by Statistics NZ 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 agency indicates that a residential address is both required and obtained from clients and, therefore, have a higher quality.

[18] For 2001 and 2006 censuses, addresses recorded prior to 1st January 2006 and 2007 are derived, accordingly. These choices were based on trial-and-error to achieve the largest possible matching.

[19] The Meshblock concordance table contains a concordance of annual meshblock patterns. The current meshblock (2020) was converted to the corresponding meshbocks in previous years, namely 2001 (MB2001) and 2006 (MB2006).

[20] The existence of the name and day in the date of birth could potentially improve the linking substantially but these are not provided for confidentiality reasons.

The 2013 census provides complete coverage of partners in the NZ population at the time of the census. That is mainly because this census is matched at the individual level to the IDI with a reasonably high match rate. The overall linkage rate of the 2013 census usual-resident population to the IDI was 92.4 percent. The main linking variables were full name, date of birth, gender, meshblock of usual residence, and country of birth (Gibb *et al*., 2016). Given the fact that the relationship information derived from the earlier censuses (2001 and 2006) are limited to those partners where both are linked to the IDI spine, the relationship information derived from the 2013 census is divided into two files. One of these files follows the other censuses and is limited to partner relationships, while the other file contains parent-child relationships. For the sake of consistency, parent-child links derived from the 2013 census are restricted to dependent children, defined as those aged under 18 at the time of census. This means that young adults, those aged 18 and over, who choose to live with their parents are excluded from the final population.

The process of compiling family-level information over the period of study is explained in the following section.

6. Collecting Relationship Information from All Sources

As discussed earlier, the relationship information collected from administrative data sources is on a yearly basis. However, during the period of study, three censuses, namely 2001, 2006, and 2013, were conducted. While the first two censuses provide additional partnership information, the latter provides details on partners' relations as well as parent-child links. Therefore, for those years in which more than one data source is available, the records from both data sources, administrative and census, need to be combined. This information is then used to construct a family-level dataset.

Since the fundamental unit of analysis used here is the individual, family information is examined for all the records available in the individual-level database, described in Alinaghi *et al*. (2020). This means that a total of 5,393,874 individuals are checked for any family information available in the administrative/census data sources. Inevitably, this information is not available for every record in the individual-level database.

Due to the data limitations explained earlier, a narrow definition of a family is applied here: A family is defined as either a single adult living alone, single adult living with one or more children, or a couple with or without dependent children. To be able to examine the role of

changes in family composition and the consequences for income mobility and poverty, the relevant information on partners and dependent children (where available) are added to each individual record. Table 1 illustrates the structure of the family-level dataset by showing a number of hypothetical entries. Each individual can fall into one of the four possible family type categories mentioned earlier. Since the number of children can vary from a minimum of zero to a maximum of 24 in some years, a variable which records the total number of children for each family across each tax year is created.

Table 1: Structure of the Family-level Dataset: Simple Hypothetical Example

Year	Snz_uid	Partner_snz_uid	Child#1_snz_uid	…	Child#24_snz_uid	NO_kids	Fam_type
2000	105185	.	.	.	.	0	Single
2001	105185	.	.	.	.	0	Single
.	.					.	.
.	.					.	.
.	.					.	.
2016	105185	.	.	.	.	0	Single
2017	105185	.	.	.	.	0	Single
2000	30836	.	21620736	.	.	1	Single&Dep
2001	30836	.	21620736	.	.	1	Single&Dep
.	.					.	.
.	.					.	.
.	.					.	.
2016	30836	.	21620736	.	.	1	Single&Dep
2017	30836	.	21620736	.	.	1	Single&Dep
2000	598352	58082025	.	.	.	0	Couple
2001	598352	58082025	.	.	.	0	Couple
.	.					.	.
.	.					.	.
.	.					.	.
2016	598352	58082025	.	.	.	0	Couple
2017	598352	58082025	.	.	.	0	Couple
2000	39682	8774120	39714454	.	.	1	Couple&Dep
2001	39682	8774120	39714454	.	.	1	Couple&Dep
.	.					.	.
.	.					.	.
.	.					.	.
2016	39682	8774120	39714454	.	.	1	Couple&Dep
2017	39682	8774120	39714454	.	.	1	Couple&Dep

As illustrated in Table 1, each segment refers to one form of family type. In the first segment, an individual with the person identifier **snz_uid** = 105185 is a single adult, living alone over the period of study. The next segment reports the family information of a single adult

(**snz_uid** = 30836) who is living with a child with **child#1_snz_uid** = 21620736, over the whole period. The next two segments refer to couples without, and with, a child. As mentioned earlier, the possible number of children ranges from 0 to 24 and, therefore, there are 24 columns to record the maximum number of children for a given family over time. For simplicity, this table shows cases where the family structure remains the same over the period of study, 2000 – 2017.

As a result of partnership formation and dissolution, and the arrival or departure of children, the family structure quickly becomes complicated. An example of a more complex structure is illustrated in Table 2. As before, each segment reports the relationship between the main individual (person with **snz_uid**) and other family members where the family consists of more than one member. In contrast to the previous table, a family type in this table may change from one period to the next depending on changes in the family composition. The first segment of Table 2 refers to an individual with the person-identifier of 104181. This single adult forms a partnership relation in 2016 and one year later, a child is born to this family. Thus, the family type remains as 'single' until 2016, after which it becomes a couple, and then a couple with a dependent child, in the next two years. The next segment refers to a single adult living with a dependent child until 2016, when a new adult member, a partner of the main adult member, is added to the family (**Partner_snz_uid** = 69810).

Table 2: Structure of the Family-level Dataset: Complex Hypothetical Examples

Year	Snz_uid	Partner_snz_uid	Child#1_snz_uid	Child#2_snz_uid	.	NO_kids	Fam_type
2000	104181	.	.	.	.	0	Single
2001	104181	.	.	.	.	0	Single
.	.					.	.
.	.					.	.
.	.					.	.
2016	104181	2036781	.	.	.	0	Couple
2017	104181	2036781	104181	.	.	1	Couple&Dep
2000	63803	.	20736	.	.	1	Single&Dep
2001	63803	.	20736	.	.	1	Single&Dep
.	.					.	.
.	.					.	.
.	.					.	.
2016	63803	69810	20736	.	.	1	Couple&Dep
2017	63803	69810	20736	.	.	1	Couple&Dep
2000	352598	8082025	.	.	.	0	Couple
2001	352598	8082025	.	.	.	0	Couple
.	.					.	.
.	.					.	.
.	.					.	.
2016	352598	.	.	.	.	0	Single
2017	352598	769871	.	.	.	0	Couple
2000	9682	74120	4454	.	.	1	Couple&Dep
2001	9682	.	4454	.	.	1	Single&Dep
.	.					.	.
.	.					.	.
.	.					.	.
2016	9682	885544	4454	99010	.	2	Couple&Dep
2017	9682	885544	.	99010	.	1	Couple&Dep

The third segment of Table 2 illustrates the case of a couple family that experiences a dissolution in 2016, and the main individual adult forms a new partnership relation in 2017. The last segment is of a couple family with a dependent child, where the partnership relation is dissolved in 2001. Therefore, the main adult lives with a dependent child for several years. In 2016, the main adult member (**Snz_uid** = 9682) forms a partnership relation with a person with identifier 885544. They become a couple family with two dependent children. One year later, in 2017, one of the children departs the family and forms a new family.

The next step is to add income information to these yearly family-level datasets. To do so, individual level income data are linked to the family-level datasets using the individual-level identifiers (**Snz_uid**). Table 3 shows the structure of the family-level dataset when income information for adult members of the families is added. For the sake of simplicity, only

relevant information corresponding to the tax year 2000 is shown here. As an example, the first row of Table 3 refers to an individual (with identifier = 105185) whose gender is male, and who was born in March 1980.[21] This individual receives an income of Y_{2000} in the year 2000. In this year, he is considered as a 'Self-employed' individual (SE) who spent 17 days overseas during the tax year 2000 (from 1st April 1999 to 31st March 2000).

The third row refers to a female individual born in August 1960 (**Snz_uid** = 598352). This person receives wage and salary income (W&S) of YYY_{2000}. Her partner in the year 2000 is a person with **Snz_uid** = 58082025, born in September 1957, who is a wage earner with an income of XXX_{2000}.

The year-ended variables such as 'Taxinc_2000' are created for the whole period, 2000 to 2017. This means that taxable income of a given individual is available for the following years (e.g., Taxinc_2001,…, and Taxinc_2017). However, a missing value, shown as '.', is assigned to variables for which the relevant information is not available/applicable. For example, in cases where the family type is identified as 'single', the individual does not have either a partner or child, and, therefore, all the corresponding variables (cells) take the missing values; see, for example, first row of Table 3.

Finally, a unique family-level identifier is constructed. The creation of a unique identifier at the family level allows any changes in family composition to be followed over time. In cases where the adult member or members of the family remain constant over time, the same identifier is assigned to a family throughout. To create a unique family-level identifier with such characteristics, the identifiers for adult members are combined. These are positioned sequentially, with the smaller identifier first followed by the other identifier. In cases where there is only one adult member in the family (single and single with dependent children), the individual-level identifier is assigned as a family-level identifier, and therefore the two identifiers are identical.

Examples are illustrated in Tables 4 and 5. The former refers to a simple hypothetical example, while the latter represents more complex situations. As discussed earlier, the fundamental unit of the family dataset is the individual and in cases where a family consists of couples (with or without children), the two adult members of the family appear in this dataset each once. However, these two records refer to the same family, and therefore the

[21] As explained earlier, to reduce risk of spontaneous recognition, the exact date of birth is not reported within the IDI (day is removed for confidentiality purposes) and therefore, day 15th of month of birth is assigned as a day of birth.

family identifier is set to be identical. To retain one record per family in a given year, the older partner is considered as the 'reference person'. In cases where the ages of both partners are equal, the reference person is the male. In the rare case of same-sex partners who have the same age, the partner with the smaller individual-level identifier is considered as the reference person. The family income is then calculated by pooling the income across adult members of the family.

In Table 4, the first row refers to a single adult who lives alone in year 2000. The family-level identifier assigned to this single-person family is identical to the individual-level id (**Snz_uid** = **Fam_id** = 105185). The taxable income for the family is also identical to the income of the single adult member of this family. The last row, on the other hand, shows a couple family with a dependent child. Since this family consists of two adult members, the individual-level identifiers are first sorted for the adult members, so that the family-level identifier starts with the smaller number. In this case, the smaller identifier number belongs to the main individual (**Snz_uid** = 39682), and therefore, to create a family-level identifier, this number is positioned first (39682 < 8774120). It is then followed by the individual-level identifier for the partner (**Partner_snz_uid** = 8774120). Doing so, the family-level identifier is **Fam_id** = 396828774120. By pooling income across adult members of this family, taxable income for the family in year 2000 is calculated as $YYYY_{2000}$ + $XXXX_{2000}$ (taxable income corresponding to the main individual and the partner are $YYYY_{2000}$ and $XXXX_{2000}$ respectively).

The construction of the family-level identifier for more complex cases is shown in Table 5. This table corresponds to Table 2, where a change in family compositions is observed. As shown in Table 2, the person with **Snz_uid** = 104181 is identified as a single-adult family. The family composition corresponding to this individual changes in 2016 when a couple-type family is formed. As described earlier, the family id assigned to this family for years 2000 to 2015 is 104181 (**Fam_id** = **Snz_uid**). However, since another adult member is added to this family in 2016, this family is no longer recognised as the same family, and therefore a new family id including individual-level identifiers for both adult members is assigned to this family. The last three rows of Table 5 show a change in family composition and therefore change in **Fam_id**, corresponding to an individual with **Snz_uid** = 9682.

Table 3: Family-level Dataset with Income Information: Simple Hypothetical Example (corresponding to records in Table 1)

Snz_uid	Sex	Dob	Taxinc_2000	SE_2000	No_Days_ Overseas_2000	Partner_ Snz_uid_2000	Partner _Sex_2000	Partner _Dob_2000	Partner _Taxinc_2000	Partner _SE_2000	NO _Kids_2000	Fam_ Type_2000
105185	male	15mar1980	Y_{2000}	SE	17	.	.	.	.	.	0	Single
30836	male	15apr1967	YY_{2000}	W&S	.	.	.	.	.	.	1	Single&Dep
598352	female	15aug1960	YYY_{2000}	W&S	24	58082025	male	15sep1957	XXX_{2000}	W&S	0	Couple
39682	female	15oct1950	$YYYY_{2000}$	SE	.	8774120	male	15may1950	$XXXX_{2000}$	SE	1	Couple&Dep

Table 4: Final Family-level Dataset (with Family Identification Number and Family Income): Corresponding to the Simple Hypothetical Example

Snz_uid	Taxinc_2000	Partner_Snz_uid_2000	Partner_Taxinc_2000	Taxinc_Fam_2000	Fam_id	Fam_Type_2000
105185	Y_{2000}	.	.	Y_{2000}	105185	Single
30836	YY_{2000}	.	.	YY_{2000}	30836	Single&Dep
598352	YYY_{2000}	58082025	XXX_{2000}	$YYY_{2000} + XXX_{2000}$	59835258082025	Couple
39682	$YYYY_{2000}$	8774120	$XXXX_{2000}$	$YYYY_{2000} + XXXX_{2000}$	396828774120	Couple&Dep

Table 5: Final Family-level Dataset (with Family Identification Number and Family Income): Corresponding to the Complex Hypothetical Example

Snz_uid	Partner_ Snz_uid_2000	Partner_ Snz_uid_2001	…	Partner_ Snz_uid_2016	Partner_ Snz_uid_2017	Fam_id
104181	.	.	.	.	.	104181
104181	.	.	.	2036781	2036781	1041812036781
63803	.	.	.	.	.	63803
63803	.	.	.	69810	69810	6380369810
352598	8082025	8082025	.	.	.	3525988082025
352598	.	.	.	.	.	352598
352598	.	.	.	.	769871	352598769871
9682	74120	.	.	.	.	968274120
9682	.	.	.	.	.	9682
9682	.		.	885544	885544	9682885544

7. Conclusions

This paper has described the construction of a unique longitudinal family-level dataset, built on an individual-level database containing over five million individual taxpayers in New Zealand over the period, 2000 to 2017. A family is defined as either: a single adult living alone, a single adult living with children, or a couple with or without dependent children.

To construct a database at the family level, a range of information on family members, including children and partners, has been compiled and added to the individual-level database. The construction of unique identifiers at the family level allows changes in family composition to be followed over time. The main components of the family-level identifier are individual-level identifiers for adult members of a family. Therefore, an individual may appear several times, in association with different families, in the final database which covers almost seven million families.

The present data construction exercise allows a more extensive analysis of income inequality and mobility than has previously been possible for New Zealand. The next research exercise using this dataset will involve the important stage of comparing various summary measures - using the cross-sectional information for each year - with corresponding measures obtained using alternative data sources, such as the Household Economic Survey. This process will provide valuable information about whether the various conditions used to provide the extensive linking needed in the present data-construction exercise have inadvertently introduced data selection biases.

References

Alinaghi, N., Creedy, J. and Gemmell, N. (2020) Constructing a longitudinal database for the analysis of individual incomes in New Zealand. *Working Papers in Public Finance*, 05/2020, Wellington School of Business and Government, Victoria University of Wellington.

Buck, N., Ermisch, J.F. and Jenkins, S.P. (1995) *Choosing a Longitudinal Survey Design: The Issues*. University of Essex, ESRC Research Centre on Micro-Social Change (Report commissioned by New Zealand Treasury).

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: University of Otago Department of Public health.

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

Census External Data Quality Panel (2020) *Final report of the 2018 Census External Data Quality Panel*. Wellington: Statistics New Zealand. Retrieved from www.stats.govt.nz.

Creedy, J., Gemmell, N. and Laws, A. (2019) Relative income dynamics of individuals in New Zealand. *New Zealand Economic Papers*. Available online at: https://www.tandfonline.com/doi/abs/10.1080/00779954.2019.1665574.

Gath, M. and Bycroft, C. (2018) *The Potential for Linked Administrative Data to Provide Household and Family Information*. Wellington: Statistics New Zealand. Retrieved from www.stats.govt.nz.

Gibb, S., Bycroft, C. and Matheson-Dunning, N. (2016). *Identifying the New Zealand Resident Population in the Integrated Data Infrastructure (IDI)*. Wellington: Statistics New Zealand. Retrieved from www.stats.govt.nz.

Social Wellbeing Agency (2020) *Constructing Households from Linked Administrative Data: An Attempt to Improve Address Information in the IDI*. Wellington: Social Wellbeing Agency.

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