Estimating Income Dynamics from Cross-Sectional Data Using Matching Techniques

Christopher Ball

WORKING PAPER 06/2016
October 2016

Working Papers in Public Finance

VICTORIA
UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA
O TE ĪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

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:
http://www.victoria.ac.nz/cpf/working-papers

Estimating income dynamics from cross-sectional data using matching techniques

Christopher Ball

Victoria University of Wellington

October 12, 2016

Contents

1 Introduction 2

2 Data 4
- 2.1 Population definition 5
- 2.2 Household survey data 6

3 Descriptive Analysis 7
- 3.1 Static income distribution measures 7
- 3.2 Measures of income mobility 8

4 Dynamic imputation methods 10
- 4.1 Nearest neighbour matching 12
- 4.2 Probabilistic matching 13

5 Distance function 14
- 5.1 Variables used in matching 14
- 5.2 Methodology 14

6 Results 17
- 6.1 Mean age-gender cohort 17
- 6.2 Nearest neighbour matching 20
- 6.3 Probabilistic matching 23

7 Conclusion 26

A Detailed weight derivation tables 28

B Sensitivity: excluding income variables from matching 35

Abstract

This paper* considers using direct matching techniques to construct synthetic panels, based on annual data from 2000 to 2015 held by Statistics New Zealand in the Integrated Data Infrastructure (IDI).[1] The IDI holds administrative tax data and household survey data on income linked together so that individuals can be tracked through the different data sets available. Thus, the IDI allows for the calculation of population level income mobility measures, and can be used for comparative analysis to validate the synthetic panel methods explored in this study.

Direct matching techniques, such as Nearest Neighbour matching, are used to construct synthetic panels for estimates of income mobility. Consideration of the variables used, and methods to construct the relative variable weights is presented. Matching techniques perform better than existing synthetic panel techniques across a range of measures. With further refinement, matching techniques may allow synthetic panels to estimate income mobility.

*The author gratefully acknowledges the support this research received from the Chair of Public Finance, Victoria University of Wellington and a grant from the Victoria University Research Fund.

[1] See Disclaimer following the references for the full IDI disclaimer.

1 Introduction

Estimates of income mobility have typically relied on longitudinal data measuring income. Prior to longitudinal data, however, the only income data available was from cross-sectional sources. While cross-sectional data are available for a much longer time period - and with a broader range of characteristics - without a way of linking units of analysis between time periods there is little cross-sectional data can say about income mobility. This study uses a recent development in New Zealand where household survey data have been linked with administrative tax data. Thus, in addition to presenting a new way of linking units of analysis across time periods, we can compare the synthetic panel results to actual income mobility measures derived from longitudinal tax data.

We follow Fields & Viollaz (2013) by attempting to answer the same question: *"are pseudo-panels a suitable substitute for true panels for estimating income mobility?"*. The main difference in methodology is the approach which we use to link observations between repeated cross-sections. This paper introduces direct matching techniques for this purpose, which link individuals in one survey to individuals in another survey through a weighting function. Compared to the averaging approaches implicit in previous work, which have yet to produce a robust technique for estimating general income dynamics from a series of cross-sectional data, this technique has the potential to preserve idiosyncrasies by matching individuals based on a set of common variables.

If pseudo-panels constructed from repeated cross-sectional data using the approach suggested in this paper can be used to reliably estimate income mobility, there are three main advantages for researchers.[2] Firstly, repeated cross-sections do not generally suffer from sample attrition. Secondly, cross-sectional data are (at least in New Zealand) more readily available than longitudinal data. Thirdly, as we use income recorded by the tax administration authority the measurement error should be dramatically reduced.[3]

Three New Zealand data sets which are linked together through the Integrated Data Infrastructure (IDI) held by Statistics New Zealand are used for this analysis.[4] The first data

[2] Heckman et al. (1997) provide an overview of why matching techniques have not typically been used by economists in the context of programme evaluation.

[3] There are sources of income that are not required to be reported to the tax authorities in New Zealand which may make household survey measures more representative of total individual/household income.

[4] Prior work has investigated both static and dynamic measures of income. See, for example, Claus et al. (2012) investigate the elasticity of taxable income using longitudinal IRD data, and Ball & Creedy (2016) provide detailed analysis of static inequality measures for the past 30 years using HES data.

set is administrative Inland Revenue Department (IRD) tax data, a longitudinal data set which holds information on all people in New Zealand who have reported income to the tax authority since 2000. The second data set is the Survey of Families, Income and Employment (SoFIE), a longitudinal household survey that followed the same families for 8 annual waves between 2002 and 2009. The third data set is the Household Economic Survey (HES), which is a repeated cross-sectional data set available from survey year 2006/07 to survey year 2014/15.[5]

Fields & Viollaz (2013) is perhaps the most comprehensive work to date which uses repeated cross-sectional data to estimate income mobility.[6] Fields & Viollaz (2013) used the household as the unit of analysis, whereas this study uses the individual as the unit of analysis as comprehensive time-varying household composition is not available. The approach taken in Fields & Viollaz (2013) was to use panel survey data with known income dynamics and apply three income mobility estimation techniques to each wave of the panel data set treated as independent cross-sectional data sets.[7] The first technique uses a means-based cohort approach outlined in Antman & McKenzie (2007), the second technique uses a dispersion based approach outlined in Bourguignon et al. (2004) and the final technique uses an alternative specification of a dispersion based technique outlined in Dang et al. (2011). The psuedo-panels created with each of the three techniques were compared to income dynamics derived from the "true" panel data. The conclusions presented by Fields & Viollaz (2013) were that pseudo-panel techniques considered do not give good results for the mobility concepts they sought to measure, and the techniques considered performed poorly on a broader range of income mobility measures.

The direct matching techniques introduced in this paper have not previously been applied to estimating income mobility. Thus, the two contributions of this paper are to construct pseudo-panels through directly matching individuals across surveys on a range of covariates, and to compare income mobility measures derived from synthetic panels with population level measures of income mobility using actual panels. Two techniques for constructing a direct match for a given individual are presented. The first method simply finds the individual in the second data set that minimises a given distance function. The second method samples

[5]HES data has only been matched into the IDI from 2006/07 onwards, although unit record data is available over a much longer time frame in an unlinked format.

[6]Other work includes Bjorklund & Jantti (1997) who estimate intergenerational mobility using cross-sectional data and regression techniques, which was subsequently re-estimated by OSterberg (2000) using administrative data. There is also specialised work by Dang & Lanjouw (2013), Navarro (2011) and Martinez Jr. et al. (2013) attempting to estimate a limited range of income mobility measures. Rubin (1986) consider statistical matching, but not quite in the same context.

[7]A general outline of each of the three estimation techniques is also provided in Fields & Viollaz (2013).

individuals in the second data set, for each individual in the first data set, based on a probability function defined by the distance function.

Data used in the study is outlined in Section 2. Given that little prior information about these data have been published previously, a range of descriptive static and dynamic income measures are presented in Section 3. Matching methodology is detailed in Section 4, with the variables selected for the matching and the construction of the distance function is outlined in Section 5. Results are presented in Section 6 and concluding remarks are made in Section 7.

2 Data

The data collection used for this study is the Integrated Data Infrastructure (IDI), an anonymised linked administrative data collection held by Statistics New Zealand. These data sets allow for individuals to be tracked across multiple collections, such as Inland Revenue Department tax records and the 2013 Census. In addition to these data, many of the household surveys such as the Household Economic Survey (HES) and the Survey of Families, Income and Employment (SoFIE) have been linked to the IDI, which allows for the use of common income measures when testing the performance of the matching methodology. More information about the IDI can be found on the Statistics New Zealand website.[8]

Statistics New Zealand (2014) make information available about the methodology used to link data in the IDI as well as progressively releasing meta data on the collections contained within the IDI.[9] There are, however, some salient points which help understanding the limitations of the present study. Firstly, the IDI is linked by creating a spine, consisting of tax information from IRD, migration information provided by Ministry of Business, Innovation and Employment, and population information from Department of Internal Affairs.[10] The spine is used to link individual records across collections, so if an individual does not appear on the spine it is not possible to find any information beyond what is available in the source data collection. Secondly, almost all collections are linked to the spine using a combination

[8] http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure.aspx

[9] http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/integrated-data-infrastructure/idi-data-dictionaries.aspx

[10] The information provided by the DIA primarily consists of Births, Deaths and Marriages registered in New Zealand. The quality of the information varies depending on the time period when it was last requested, with records prior to 1999 typically having a much lower probability of being correctly linked.

of name and date of birth.[11] The (spine) link quality is limited by the quality of the name and date of birth information, which helps explain the lower link rate seen for collections which aren't required to formally verify identity, such as household surveys. Finally, many of the data collections were not collected for research purposes and standards of data collection vary according to the collecting agency. Issues which are highlighted in the meta data have been addressed and appropriate secondary checks have been performed, but as the data are provided in a processed anonymised form there may be remaining data quality issues which IDI researchers have no practical way of identifying.

2.1 POPULATION DEFINITION

With such a rich data set we have the ability to limit the administrative data collected using information from multiple collections, such as border movements and death records. The definition of the population used for this study is:

- a member of the working-age population (between 15 and 64 inclusive throughout the entire tax year) during both study periods. This excludes children and, more importantly, public pension recipients.[12]
- linked to the IDI spine, which ensures that the individual at least has both a gender and a date of birth record.
- present in New Zealand (according to the border movements data) for at least half of the respective tax year.
- must not have a death record (through health or DIA records) before or during either tax year considered.[13]
- must have at least one IRD tax record linked to the individual for each of the tax years in question.

A comparison of the final study population size to the official estimated resident population for ages 15-64 is presented by tax year in Table 1. The estimated resident population is

[11] As at June 2016, Census 2013 is the only data set to use geographic information for matching purposes, in addition to name and date of birth.

[12] New Zealand's public pension is available almost universally (with no income or asset test) to individuals aged 65 years or more. These individuals were excluded from the study population as there dynamics are likely to be influenced by factors not considered in this study.

[13] This means that individuals who die during a tax year are excluded from that tax years dynamics.

TABLE 1: Comparison of study population to estimated resident population by tax year (15-64)

Tax Year	Estimated Resident Population	Derived Population	Percentage
2000	2,516,100	2,243,000	89.1
2001	2,531,600	2,256,000	89.1
2002	2,558,300	2,279,000	89.1
2003	2,615,400	2,328,000	89.0
2004	2,676,400	2,380,000	88.9
2005	2,721,600	2,421,000	89.0
2006	2,758,100	2,458,000	89.1
2007	2,792,900	2,485,000	89.0
2008	2,815,100	2,512,000	89.2
2009	2,836,100	2,522,000	88.9
2010	2,863,000	2,530,000	88.4
2011	2,886,100	2,546,000	88.2
2012	2,896,200	2,547,000	87.9
2013	2,899,200	2,539,000	87.6
2014	2,918,800	2,561,000	87.7
2015	2,965,400	2,544,000	85.8

Estimated resident population is the mean over the March year ending at the respective tax year, covering the 15-64 population. Available from http://www.stats.govt.nz/infoshare/default.aspx, last accessed 23rd of May, 2016.

not necessarily comparable with the study population as they are intended for different purposes, although the stability of the study population as a percentage of the estimated resident population indicates the study population is broadly comparable.[14]

2.2 HOUSEHOLD SURVEY DATA

There are two household surveys considered for this study, the Household Economic Survey (HES) and the Survey of Families, Income and Employment (SoFIE). The HES has been linked to the IDI for survey years ending in 2007 to 2015, while the SoFIE has been linked for the 8 wave duration of the survey (ending in 2009). Both of these data sets are subsets of the broader linked administrative data sets with the additional requirement that the individuals in each data set must be linked to the respective survey in the respective survey year.

[14]The final IRD data for 2015 is not available in the IDI, which is why the coverage rate dips in 2015. The currently available data is close enough to final for the purposes of this study.

The reference period for income is an additional complication with household survey data. For both household surveys the income is collected for the year immediately prior to the interview date, with interview dates ranging from the 1st of July to the 30th of June for a given survey year. As an example, HES 14/15 refers to households interviewed between the 1st of July 2014 to the 30th of June 2015. This HES year will potentially consist of income recalled from the 1st of July 2013 (if the respondent was interviewed on the 1st of July 2014) to the 30th of June 2015 (if the respondent was interviewed on this day). For this study the income considered is the administrative income recorded for the tax year which overlaps most with the interview period, so HES 14/15 respondents will have the IRD income from the 2015 tax year.

3 Descriptive Analysis

We are interested in having both similar dynamic and static measures while also getting the aggregate relative and absolute changes similar. As no-one has previously investigated this particular data set, we present both static and dynamic measures from the IRD data for the study population. Throughout this paper the unit of analysis is the individual, and the income measure is IRD income from all sources expressed in $2015. For household survey measures, the income time period is the tax year which has the most overlap with the survey period: for example, HES 2014/15 which interviewed from 1st of July, 2014 to 30th June, 2015 will use the 2015 tax year (running from 1st of April, 2014 to the 31st of March, 2015).

3.1 STATIC INCOME DISTRIBUTION MEASURES

Table 2 presents the size of the linked population, the income at each of the decile boundaries, the Gini coefficient, the mean and the standard deviation for each of the three data sets considered. Where the sample size is large enough the 95th, 99th and 99.9th percentile boundaries are presented.[15] From Table 2 the HES decile boundaries are consistently higher than the IRD boundaries, whereas the SOFIE boundaries are generally close to the IRD boundaries. It is interesting that the 2015 tax year HES sample is the closest to the IRD decile boundaries, and that this sample is roughly 50% larger than any of the other HES samples. The IRD Gini is underestimated by both the SOFIE and the HES samples, with the SOFIE data showing a broadly similar trend. The mean income broadly reflects what is seen with the decile boundaries, with the SOFIE mean income broadly comparable to the

[15] The x-th percentile is the smallest value y such that at least x percent of the sample is below y.

IRD mean and the HES mean typically above.

Looking at the static measure of income distribution suggest that the SOFIE measures are much more comparable than HES measures to the IRD population level measures. There are many possible explanations for the relatively poor performance of the linked HES sample. Three possible explanations are that the process linking HES to the IDI may induce a higher bias than SOFIE, the HES sample size is less than half the size of SOFIE in any given year and the HES sample design may have higher non-response or selection bias for income. These problems may be fixed through the HES calibration process, where representative sample weights are derived by benchmarking against known administrative totals.[16] With the systematic errors in the static income measures, the HES income dynamic measures derived need to be interpreted with caution.

3.2 MEASURES OF INCOME MOBILITY

Table 3 presents the Pearson correlation coefficient, while Table 4 presents the Spearman correlation coefficient. In both measures there is a high correlation in the subsequent time period, followed by a decay over time. There is also a trend of increasing correlation in the subsequent period over time, going from 0.8 for Pearson and 0.84 for Spearman in 2000 to 0.89 for Pearson and 0.88 for Spearman in 2015. Using this as an inverse measure of income mobility would suggest that mobility has decreased over the time period considered for the study population.

Table 5 shows the mean absolute decile change, which first calculates the number of decile moved for each member of the study population and then calculates the average over the study population. Similar to the correlation coefficients, this measure also shows steadily increasing mobility as the accounting period increases. Further, this measure also shows that income mobility has decreased between 2000 and 2015, going from 0.94 in 2000 to 0.81 in 2015.

Looking across the range of mobility measures presented suggest that time-invariant characteristics may not be enough to capture the income dynamics seen in the study population. The results from Fields & Viollaz (2013), where time-invariant characteristics were a predominant feature of the income mobility estimation procedure, have fairly weak predictive performance which further strengthens the argument for including time-varying character-

[16] Investigating alternative calibration methodologies is beyond the scope of this paper.

Table 2: Static income measures for IRD, SOFIE and HES data – 2000 to 2015

	Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
IRD	Sample Size	2,243,000	2,256,000	2,279,000	2,328,000	2,380,000	2,421,000	2,458,000	2,485,000	2,512,000	2,522,000	2,530,000	2,546,000	2,547,000	2,539,000	2,561,000	2,544,000
	Percentile																
	10%	$4,346	$4,463	$4,759	$4,790	$4,926	$5,002	$5,094	$5,245	$5,589	$5,789	$5,987	$6,337	$6,585	$6,982	$7,170	$7,298
	20%	$8,598	$8,857	$9,374	$9,628	$9,913	$10,200	$10,373	$10,848	$11,511	$11,696	$11,552	$11,845	$12,183	$12,610	$13,096	$13,418
	30%	$11,446	$11,573	$12,331	$12,747	$13,440	$14,167	$14,692	$15,397	$16,099	$16,275	$16,223	$16,551	$16,987	$17,311	$17,792	$18,331
	40%	$14,766	$15,452	$16,518	$17,209	$18,167	$19,217	$20,149	$21,146	$22,789	$23,477	$22,628	$23,011	$23,983	$25,206	$26,443	$27,316
	50%	$20,248	$21,261	$22,528	$23,359	$24,574	$25,851	$27,113	$28,325	$30,167	$31,206	$30,880	$31,518	$32,660	$34,068	$35,426	$36,445
	60%	$26,302	$27,439	$28,821	$29,784	$31,153	$32,597	$34,083	$35,438	$37,384	$38,581	$38,669	$39,612	$40,953	$42,404	$43,945	$45,000
	70%	$32,609	$33,874	$35,340	$36,375	$37,931	$39,459	$41,162	$42,756	$45,013	$46,546	$46,951	$48,104	$49,732	$51,271	$52,976	$54,175
	80%	$40,084	$41,557	$43,281	$44,425	$46,376	$48,416	$50,641	$52,638	$55,407	$57,314	$57,992	$59,698	$61,634	$63,593	$65,689	$66,972
	90%	$53,436	$55,479	$57,681	$58,732	$60,603	$62,973	$65,908	$68,519	$71,988	$74,791	$76,097	$78,239	$81,040	$83,451	$86,061	$87,432
	95%	$70,558	$70,944	$73,900	$74,434	$77,351	$80,711	$84,801	$88,092	$93,006	$96,188	$97,714	$100,637	$104,979	$108,690	$112,846	$113,987
	99%	$137,665	$131,381	$137,514	$138,342	$143,499	$149,076	$156,947	$162,495	$170,465	$174,975	$175,571	$181,796	$191,456	$199,303	$206,532	$205,758
	99.90%	$338,199	$307,131	$316,147	$318,567	$331,862	$340,098	$361,661	$376,865	$393,631	$396,798	$396,946	$414,009	$427,302	$448,889	$462,270	$449,551
	Gini	0.476	0.467	0.463	0.461	0.459	0.458	0.461	0.461	0.459	0.462	0.463	0.462	0.461	0.459	0.458	0.453
	Mean	$27,175	$27,728	$29,041	$29,635	$30,955	$32,274	$33,734	$35,064	$37,000	$38,163	$38,391	$39,544	$41,067	$42,529	$44,065	$44,822
	Standard Deviation	$32,707	$30,582	$33,384	$31,718	$33,620	$34,315	$36,158	$37,549	$40,031	$41,638	$40,978	$42,497	$43,571	$44,785	$46,290	$45,459
SOFIE	Sample Size			15,500	17,200	17,800	17,900	17,900	17,800	17,800	17,500						
	Percentile																
	10%			$5,023	$4,826	$5,118	$5,269	$5,298	$5,596	$5,836	$6,269						
	20%			$9,449	$9,634	$10,103	$10,221	$10,615	$11,198	$11,749	$12,403						
	30%			$12,834	$13,184	$13,883	$14,309	$15,051	$15,668	$16,232	$16,795						
	40%			$16,914	$17,384	$18,361	$19,326	$20,375	$21,713	$23,361	$24,421						
	50%			$22,948	$23,358	$24,773	$25,863	$26,914	$28,527	$30,469	$31,807						
	60%			$28,932	$29,718	$31,075	$32,252	$33,927	$35,602	$37,713	$38,914						
	70%			$35,638	$36,309	$37,884	$39,083	$40,879	$42,800	$45,113	$46,842						
	80%			$43,614	$44,044	$46,332	$48,212	$50,209	$52,253	$54,980	$57,662						
	90%			$57,461	$57,887	$60,046	$61,902	$65,001	$67,841	$70,786	$74,673						
	95%			$72,961	$72,852	$76,346	$78,751	$82,156	$86,408	$91,860	$94,894						
	99%			$139,078	$133,737	$140,259	$143,691	$149,233	$160,894	$169,146	$171,440						
	Gini			0.460	0.454	0.453	0.450	0.454	0.454	0.453	0.454						
	Mean			$29,475	$29,565	$31,032	$32,018	$33,485	$35,351	$37,382	$38,740						
	Standard Deviation			$34,522	$31,098	$33,317	$32,941	$36,705	$40,409	$42,733	$42,244						
HES	Sample Size								3,400	4,000	4,000	3,700	4,200	4,200	3,200	3,900	6,200
	Percentile																
	10%								$6,024	$6,236	$7,123	$6,713	$7,220	$7,524	$7,904	$7,454	$7,625
	20%								$11,648	$12,463	$13,100	$12,906	$13,258	$13,624	$14,883	$14,119	$14,406
	30%								$16,742	$17,961	$17,378	$17,501	$18,133	$18,752	$20,134	$20,152	$19,911
	40%								$24,305	$26,000	$25,992	$26,021	$26,636	$27,889	$30,267	$29,838	$29,301
	50%								$31,283	$33,201	$34,009	$34,569	$35,152	$36,446	$39,064	$38,385	$38,406
	60%								$38,228	$40,846	$41,179	$41,983	$42,826	$44,666	$47,018	$47,316	$46,671
	70%								$45,763	$48,588	$49,212	$50,515	$50,777	$53,734	$56,299	$57,219	$55,422
	80%								$55,150	$58,890	$60,964	$62,020	$63,142	$65,800	$69,138	$70,000	$67,870
	90%								$70,868	$74,946	$76,918	$79,062	$81,427	$86,827	$87,119	$89,998	$86,426
	95%								$89,993	$97,016	$102,919	$99,090	$104,903	$114,196	$111,218	$121,952	$108,836
	Gini								0.430	0.440	0.455	0.440	0.443	0.452	0.441	0.453	0.439
	Mean								$36,583	$39,586	$40,912	$41,120	$41,904	$44,542	$46,499	$47,415	$45,613
	Standard Deviation								$31,752	$37,167	$42,413	$38,638	$41,750	$45,590	$48,959	$47,305	$44,345

istics. Section 4 outlines the combination of time-invariant and time-varying variables that are used to estimate income mobility measures.

Table 3: Income correlation from IRD data

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.80	0.69	0.67	0.63	0.60	0.58	0.56	0.53	0.50	0.49	0.48	0.49	0.47	0.47	0.46
2001		0.79	0.75	0.69	0.65	0.64	0.61	0.58	0.54	0.54	0.53	0.53	0.51	0.49	0.48
2002			0.80	0.73	0.68	0.67	0.63	0.60	0.57	0.56	0.54	0.54	0.52	0.50	0.49
2003				0.82	0.76	0.72	0.69	0.64	0.61	0.60	0.58	0.58	0.56	0.54	0.53
2004					0.84	0.80	0.75	0.70	0.66	0.65	0.62	0.61	0.60	0.57	0.56
2005						0.84	0.78	0.73	0.67	0.66	0.64	0.63	0.61	0.59	0.58
2006							0.87	0.81	0.76	0.73	0.69	0.68	0.66	0.62	0.61
2007								0.86	0.79	0.77	0.73	0.71	0.69	0.65	0.64
2008									0.84	0.80	0.76	0.73	0.71	0.67	0.66
2009										0.86	0.80	0.77	0.74	0.69	0.69
2010											0.86	0.82	0.79	0.74	0.73
2011												0.87	0.82	0.78	0.74
2012													0.89	0.84	0.80
2013														0.89	0.84
2014															0.89

Table 4: Income rank correlation from IRD data

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.84	0.76	0.70	0.65	0.61	0.57	0.54	0.51	0.48	0.46	0.45	0.44	0.43	0.42	0.41
2001		0.85	0.76	0.70	0.65	0.61	0.57	0.54	0.50	0.49	0.48	0.47	0.45	0.44	0.42
2002			0.85	0.76	0.70	0.65	0.61	0.57	0.54	0.51	0.50	0.49	0.47	0.46	0.44
2003				0.85	0.77	0.71	0.66	0.62	0.58	0.55	0.53	0.52	0.50	0.49	0.47
2004					0.86	0.77	0.71	0.66	0.62	0.59	0.56	0.54	0.52	0.51	0.49
2005						0.86	0.78	0.72	0.66	0.63	0.60	0.57	0.55	0.53	0.51
2006							0.86	0.78	0.72	0.67	0.64	0.61	0.59	0.57	0.54
2007								0.86	0.78	0.73	0.69	0.65	0.63	0.60	0.57
2008									0.86	0.79	0.74	0.70	0.67	0.63	0.60
2009										0.87	0.80	0.75	0.71	0.67	0.64
2010											0.87	0.80	0.75	0.71	0.67
2011												0.88	0.81	0.76	0.71
2012													0.88	0.81	0.75
2013														0.88	0.81
2014															0.88

4 Dynamic imputation methods

Similar to the approach taken in Fields & Viollaz (2013) results from pseudo-panel income mobility estimates are compared to the same results from longitudinal data. With the range of data available three sets of comparisons are presented in the results section. First, we compare IRD information treated as a repeated cross-section to IRD information treated as longitudinal data. Second, we compare SoFIE information similarly to IRD information, with

TABLE 5: Mean absolute decile change from IRD data

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.94	1.26	1.47	1.63	1.76	1.88	1.96	2.04	2.11	2.15	2.18	2.21	2.24	2.28	2.31
2001		0.90	1.24	1.46	1.62	1.76	1.87	1.96	2.04	2.09	2.11	2.15	2.19	2.22	2.26
2002			0.91	1.23	1.45	1.62	1.75	1.85	1.95	2.01	2.04	2.08	2.12	2.16	2.21
2003				0.90	1.22	1.44	1.60	1.74	1.85	1.92	1.97	2.01	2.06	2.10	2.15
2004					0.91	1.23	1.44	1.60	1.73	1.82	1.89	1.94	1.99	2.04	2.09
2005						0.90	1.20	1.43	1.59	1.70	1.78	1.85	1.91	1.97	2.03
2006							0.88	1.19	1.41	1.55	1.65	1.74	1.81	1.89	1.95
2007								0.89	1.20	1.38	1.51	1.62	1.71	1.80	1.88
2008									0.89	1.17	1.34	1.48	1.59	1.69	1.78
2009										0.85	1.11	1.31	1.45	1.58	1.68
2010											0.83	1.11	1.31	1.46	1.59
2011												0.82	1.11	1.30	1.46
2012													0.81	1.10	1.30
2013														0.80	1.10
2014															0.81

TABLE 6: Multi-period inequality from IRD data

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.43	0.42	0.41	0.40	0.39	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
2001		0.43	0.41	0.40	0.39	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
2002			0.42	0.41	0.40	0.39	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
2003				0.42	0.41	0.40	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.38	0.38
2004					0.42	0.41	0.40	0.39	0.38	0.38	0.38	0.38	0.38	0.38	0.38
2005						0.42	0.41	0.40	0.39	0.39	0.39	0.38	0.38	0.38	0.38
2006							0.42	0.41	0.40	0.40	0.39	0.39	0.39	0.38	0.38
2007								0.42	0.41	0.40	0.40	0.39	0.39	0.39	0.38
2008									0.42	0.41	0.40	0.40	0.39	0.39	0.38
2009										0.42	0.41	0.41	0.40	0.39	0.39
2010											0.43	0.42	0.41	0.40	0.39
2011												0.43	0.42	0.41	0.40
2012													0.43	0.42	0.41
2013														0.43	0.41
2014															0.42

an additional comparison possible to the IRD longitudinal data. Finally, income measures derived from HES data are compared to income measures derived from IRD data. All of these results are presented in Section 6.

Fields & Viollaz (2013) compare three pseudo-panel generation methods to the actual panels. This study presents the mean-based approach alongside the results from the pseudo-panel techniques presented in 4. We omit the other two approaches in Fields & Viollaz (2013) as our choice of individual as the unit of analysis would require non-trivial changes to the methodology,[17] the data requirements are significant in the context of the administrative data collected, and there is little evidence to suggest the omitted methods would perform any better for New Zealand data.[18]

4.1 NEAREST NEIGHBOUR MATCHING

Assume that the data have m standardised dimensions to match on (where $X_{i,j}$ is the j-th covariate for element i), there is a known non-negative weight vector $W = (w_1, \ldots, w_m)$ and the two samples are denoted A and B. It is not necessarily the case that A and B will have equal sizes, or even have elements in common. Section 5 provides more information about the variables used and the derivation of the weights.

The first matching technique introduced is Nearest Neighbour Matching.[19] This technique goes through each data point in A and matches it to the closest data point in B. More formally each element a in A is matched to

$$\min_{b \in B} \left(d(b, a) \right) \tag{1}$$

where $d()$ is a non-negative measure of distance. For this study we use a weighted Euclidean distance measure,

$$d(b, a) = \sum_{i=1}^{m} w_i (b_i - a_i)^2. \tag{2}$$

We use the R package written by Arya et al. (2015) to find the nearest neighbours, which

[17] Investigating the methodological modifications needed to change the unit of analysis from individuals to households is the subject of a planned follow-up paper.

[18] See Section 6.1 for a summary of the mean-based approach, which similar to Fields & Viollaz (2013) has poor performance predicting income dynamics.

[19] See Knuth (1973) for more information.

implements the approach outlined in Bentley (1975) to efficiently find the k-nearest neighbours. The 2 nearest neighbours are selected, so that longitudinal data can be matched to the nearest distinct individual.[20]

4.2 PROBABILISTIC MATCHING

This study also considers an extension of the Nearest Neighbour matching to the case where individuals are selected randomly, using the respective distances to inform the probability of selection. Using the same R package as Nearest Neighbour matching (Arya et al. (2015)) we first select the nearest 100 neighbours in B for every element of A.[21] For a given element $a_j \in A$ denote $b_{[i]}$ as the i-th closest element.[22] The probability of selecting element $b_{[i]}$ as the match for a_j is

$$P\left(b_{[i]} \text{ selected as match for } a_j\right) = \frac{\frac{1}{1+d(a_j, b_{[i]})}}{\sum_{k=1}^{100} \frac{1}{1+d(a_j, b_{[k]})}} \tag{3}$$

Similar to the Nearest Neighbour matching case, for SoFIE data we choose only observations that are distinct to the individual in A we are trying to match. Computational restrictions made probabilistic matching impractical for the IRD data.

5 Distance function

This section covers the variables used to match variables across data sets, and how the relative variable weights were derived using administrative information.

[20] An extreme example will highlight why a distinct individual needs to be matched when using longitudinal data. Imagine that the weighting function always selected the correct individual to match to (a perfect match). While this would indicate that the dimensions chosen and the weighting function are searching for neighbours in the right place, the dynamics for the matched data would be exactly the same as for the panel data. This presents problems when expanding the technique to repeated cross-sectional data.

[21] Choosing the nearest 100 is noticeably more efficient than calculating the full distance matrix, while still covering most of the units that would ever be selected by random selection.

[22] Assume for the moment that ties are not possible, so that $b_{[i]}$ is unique. This can be done, for example, by adding an extra dimension following a Uniform distribution, scaled so that it is small enough that it does not affect the rankings except when ties are present.

5.1 VARIABLES USED IN MATCHING

With the linked administrative data available through the IDI we have both a comprehensive population level data set and a wide range of variables which can be used for matching. So that comparisons can be made between the accuracy of the matching techniques across the three data sets available, the same variables need to be available at the population level to be included in this specification.[23] This limits us to variables that are time-invariant, such as ethnicity and date of birth, and variables that are widely available and continuously updated, such as meshblock of residence.[24] A full list of variables used for matching is provided in Table 7.

Table 7: Matching variable definitions

Name	Definition
Age	Combination of birth month and year used for matching individuals to the IDI spine, consistently standardised so that the oldest person in the combined 2000 to 2015 data set is 0 and the youngest person is 1. Assumed to be time invariant.
Multiple Response Ethnicity	Binary indicator variables for each of the 6 high level ethnicity categories (European, Maori, Pasifika, Asian, MELAA (Middle Eastern, Latin American or African) and Other) collected by Statistics New Zealand. Assumed to be time invariant.
Meshblock	Most recently collected geocoded address information by March 31st of the tax year considered. Time varying.
Income Decile	An integer where the lowest 10% of income earning individuals in a given tax year are allocated to decile 1, the next lowest 10% are allocated to decile 2 and so on. Deciles are allocated using IRD individual income for the tax year in question.
Meshblock Income Decile	Similar to Income Decile, except assigned using the average meshblock income (weighted by the number of individuals with a tax record in the meshblock in the tax year).

[23] This, for example, eliminates educational attainment as a potential matching variable, as it is only available for people who have attained qualifications from domestic providers since 2006.

[24] Meshblocks are the smallest geographic unit of collection for official statistics in New Zealand, consisting of contiguous areas that are mutually exclusive and exhaustive of New Zealand.

5.2 METHODOLOGY

The methodology used to determine the weights considers two metrics for each variable, the probability of the subset identified for each variable containing the individual and the average size of the subset identified. The best possible variable would identify a subset for each individual of size 1, and that subset would always be the correct individual. This is the unique statistics identifier which is used to link records to the IDI spine, which we have excluded from consideration as it provide no value for cross-sectional surveys. The time invariant variables, such as age and ethnicity, will always contain the individual in question at the cost of identifying potentially large subsets on average. The time-varying variables, such as meshblock, can identify a small subset but there is no guarantee that the individual is a member of this subset. The exact metric selected for each variable is

$$\text{Variable Ratio} = \left(\frac{\#\text{ where subset contains individual}}{N}\right)\left(\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{\text{Size of identified subset for } i}{N}\right)\right) \quad (4)$$

The variable ratios for each variable and each year combination are shown in Appendix A. Using the observations from the dynamic measures presented, we group the weights by the number of years separating the data from which these variable ratios are derived, and average over all variable ratios in each group to derive the final variable ratio for each variable. Table 8 presents the final variable ratios grouped by year, standardised so that the sum across all variable weights is 1000.

TABLE 8: Relative variable weights used in matching

Variable	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
Age	282	308	324	337	346	354	361	368	374	380	385	391	397	402	404
Ethnicity 1	107	116	121	124	127	129	130	131	133	134	134	135	136	137	137
Ethnicity 2	79	86	91	95	98	100	102	105	107	109	111	113	115	117	119
Ethnicity 3	39	42	43	45	46	46	47	47	48	48	48	49	49	49	50
Ethnicity 4	48	50	50	50	50	49	48	47	46	45	44	42	41	40	38
Ethnicity 5	16	16	16	16	16	16	15	15	14	14	14	13	13	13	13
Ethnicity 6	14	15	15	16	16	17	17	18	18	19	19	20	20	20	20
Meshblock	189	171	158	149	141	134	129	122	116	112	108	103	99	97	94
Meshblock Income Decile	98	83	77	74	71	69	68	66	65	64	62	60	58	56	56
Income Decile	130	115	104	95	90	86	82	80	78	76	75	73	72	70	68

The types of variables are not consistent. Ethnicity is binary in each of the 6 categories, age is a discretised continuous variable, meshblock is geographical and the decile variables have 10 categories. The distance function used for each variable is presented in Table 9.

Table 9: Distance functions used for each variable

Name	Distance function
Age	Distance is difference in the number of months divided by the maximum difference occurring across the data sets.
Multiple Response Ethnicity	Distance is 0 if ethnicity is the same, 1 otherwise.
Meshblock	Each meshblock is defined as located at the mean of the area defined, standardised so that the Southern most point has y coordinate 0, the Western most point has x coordinate 1, the Northern most point has y coordinate 1 and the Eastern most point has x coordinate 1. Missing addresses are imputed at $(0.5, 0.5)$. The distance is calculated within this standardised coordinate system as the square deviation by x and y coordinate separately.
Income Decile	Distance is absolute # of income deciles changed divided by 9.
Meshblock Income Decile	Distance is absolute # of meshblock income deciles changed divided by 9.

6 Results

Three types of results are presented. The first type in Section 6.1 presents income dynamics based on the mean for each age-gender cohort. The second type in Section 6.2 presents income dynamics using nearest neighbour matching on the IRD, SOFIE and HES data sets. The third type in Section 6.3 presents income dynamics using the probabilistic matching technique on the SOFIE data set.

6.1 Mean age-gender cohort

The mean age-gender cohort approach is presented as a benchmark for the accuracy of regression techniques. Fields & Viollaz (2013) present a similar measure, although as mentioned previously the two measures are not comparable. In general the mean cohort method shows

noticeably more income mobility than is present in the IRD panel data. Tables 10 and 11 show correlation and rank correlation respectively, which are both substantially underestimated compared to the values calculated from IRD panel data. For example, between 2001 and 2006 Table 10 shows a correlation of 0.31 whereas the IRD panel data shows a much higher correlation of 0.64 (Table 3). Similarly, Table 11 shows a rank correlation of 0.39 over the same time period, whereas the IRD data again shows a higher rank correlation of 0.61 (Table 4).

TABLE 10: Income correlation using IRD data - Mean cohort

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.32	0.31	0.31	0.30	0.30	0.29	0.28	0.27	0.25	0.25	0.24	0.24	0.23	0.22	0.20
2001		0.33	0.33	0.33	0.32	0.31	0.30	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.23
2002			0.32	0.32	0.31	0.31	0.30	0.29	0.28	0.28	0.27	0.26	0.25	0.24	0.23
2003				0.34	0.34	0.33	0.33	0.32	0.31	0.31	0.30	0.29	0.28	0.27	0.26
2004					0.34	0.33	0.33	0.32	0.32	0.31	0.30	0.30	0.29	0.28	0.27
2005						0.35	0.35	0.34	0.34	0.33	0.33	0.32	0.31	0.30	0.29
2006							0.35	0.34	0.34	0.34	0.33	0.33	0.32	0.31	0.30
2007								0.34	0.34	0.34	0.34	0.34	0.33	0.32	0.31
2008									0.34	0.34	0.34	0.34	0.33	0.33	0.32
2009										0.33	0.33	0.33	0.33	0.33	0.32
2010											0.34	0.33	0.33	0.33	0.32
2011												0.34	0.34	0.34	0.33
2012													0.35	0.35	0.35
2013														0.35	0.35
2014															0.36

TABLE 11: Income rank correlation using IRD data - Mean cohort

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.41	0.41	0.40	0.40	0.39	0.38	0.36	0.34	0.33	0.32	0.31	0.29	0.28	0.28	0.27
2001		0.41	0.41	0.41	0.40	0.39	0.38	0.37	0.35	0.35	0.33	0.32	0.30	0.29	0.28
2002			0.42	0.42	0.41	0.41	0.40	0.39	0.37	0.37	0.36	0.34	0.32	0.31	0.29
2003				0.42	0.42	0.42	0.41	0.40	0.39	0.39	0.38	0.36	0.35	0.33	0.31
2004					0.42	0.42	0.42	0.41	0.41	0.40	0.39	0.38	0.37	0.35	0.33
2005						0.43	0.43	0.43	0.42	0.42	0.41	0.41	0.39	0.38	0.36
2006							0.43	0.43	0.43	0.42	0.42	0.42	0.41	0.39	0.38
2007								0.43	0.43	0.43	0.43	0.42	0.42	0.41	0.39
2008									0.43	0.43	0.43	0.43	0.42	0.42	0.41
2009										0.41	0.42	0.42	0.41	0.41	0.40
2010											0.41	0.41	0.41	0.40	0.40
2011												0.41	0.41	0.41	0.41
2012													0.42	0.42	0.42
2013														0.42	0.42
2014															0.43

Table 12 presents the mean absolute decile change, which shows more mobility than is present in the IRD panel data. Over the 2001 to 2006 time period the mean-age cohort method

indicates the mean absolute decile change was 2.43, whereas the IRD data in Table 5 shows less inter-decile mobility at 1.76. The multi-period Gini coefficient is in Table 13, where the coefficients again indicate more mobility than present in the IRD panel data. Continuing with the 2001 to 2006 time period, the mean-age cohort measure of multi-period Gini is 0.24, whereas the IRD data in Table 6 indicates less mobility with a multi-period Gini of 0.39.

TABLE 12: Mean absolute decile change using IRD data - Mean cohort

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	2.40	2.41	2.41	2.42	2.44	2.47	2.50	2.55	2.60	2.60	2.65	2.70	2.72	2.73	2.75
2001		2.40	2.40	2.41	2.41	2.43	2.45	2.50	2.51	2.54	2.59	2.63	2.67	2.69	2.73
2002			2.39	2.39	2.38	2.40	2.41	2.43	2.46	2.48	2.52	2.57	2.63	2.68	2.70
2003				2.39	2.38	2.39	2.41	2.42	2.44	2.45	2.47	2.50	2.57	2.61	2.67
2004					2.37	2.38	2.39	2.40	2.41	2.41	2.43	2.46	2.49	2.56	2.62
2005						2.37	2.37	2.37	2.38	2.38	2.40	2.42	2.44	2.48	2.52
2006							2.37	2.36	2.37	2.37	2.38	2.40	2.39	2.44	2.48
2007								2.37	2.38	2.36	2.37	2.39	2.38	2.41	2.44
2008									2.38	2.37	2.37	2.38	2.38	2.39	2.41
2009										2.40	2.41	2.40	2.41	2.42	2.44
2010											2.43	2.43	2.43	2.43	2.44
2011												2.41	2.41	2.41	2.41
2012													2.40	2.40	2.40
2013														2.40	2.39
2014															2.38

TABLE 13: Multi-period inequality using IRD data - Mean cohort

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.29	0.28	0.27	0.26	0.25	0.24	0.23	0.21	0.21	0.20	0.19	0.19	0.18	0.18	0.18
2001		0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.20	0.20	0.19	0.19	0.18	0.18
2002			0.28	0.27	0.26	0.25	0.24	0.23	0.22	0.21	0.20	0.20	0.19	0.19	0.18
2003				0.28	0.27	0.26	0.25	0.23	0.22	0.22	0.21	0.20	0.20	0.19	0.19
2004					0.28	0.27	0.26	0.24	0.23	0.23	0.22	0.21	0.21	0.20	0.19
2005						0.28	0.27	0.25	0.24	0.24	0.23	0.22	0.21	0.21	0.20
2006							0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22	0.21
2007								0.28	0.27	0.26	0.25	0.24	0.23	0.23	0.22
2008									0.28	0.27	0.26	0.25	0.25	0.24	0.23
2009										0.28	0.27	0.26	0.25	0.25	0.24
2010											0.28	0.27	0.26	0.25	0.25
2011												0.28	0.27	0.26	0.26
2012													0.29	0.28	0.27
2013														0.28	0.28
2014															0.29

6.2 NEAREST NEIGHBOUR MATCHING

The nearest neighbour derived correlation measures for the IRD data are presented in Table 14. The nearest neighbour correlation is lower than the actual correlation when the time

difference is less than 3 years, broadly similar with a time difference of 4 to 6 years, and higher at time differences 7 years and beyond. Using the 2001 cohort to demonstrate, at 2 years the nearest neighbour correlation is 0.13 lower - 0.63 for nearest neighbour compared to 0.76 from IRD data. After 5 years the difference is 0.03 - 0.67 for nearest neighbour compared to 0.64 from IRD data. After 8 years the nearest neighbour measure is higher by 0.09 - 0.63 for nearest neighbour compared to 0.54 from IRD data. The nearest neighbour algorithm, however, does seem to broadly match the gradually increasing correlation observed over time, increasing from 0.62 in 2000 to 0.72 in 2015 compared to 0.80 to 0.89 in the actual data over the same time period.

The nearest neighbour performance on correlation contrasts with the nearest neighbour decile change measure in Table 15, which is significantly lower than the actual decile change in Table 5. The decile change is not the only dynamic income measure using relative income that perform poorly for the IRD data; nearest neighbour rank correlation is above 0.99 in all pairs of years which is significantly higher than the 0.4 - 0.9 observed using actual dynamics.

Table 14: Income correlation using IRD data - Nearest Neighbour

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.62	0.64	0.59	0.56	0.60	0.63	0.64	0.64	0.59	0.63	0.64	0.64	0.65	0.65	0.63
2001		0.65	0.63	0.62	0.64	0.67	0.66	0.66	0.63	0.68	0.69	0.68	0.66	0.67	0.69
2002			0.63	0.61	0.63	0.64	0.64	0.64	0.61	0.65	0.65	0.65	0.64	0.65	0.66
2003				0.65	0.68	0.68	0.67	0.67	0.67	0.68	0.68	0.69	0.69	0.70	0.70
2004					0.66	0.67	0.67	0.65	0.66	0.66	0.67	0.68	0.68	0.68	0.69
2005						0.68	0.68	0.67	0.67	0.68	0.68	0.69	0.70	0.69	0.70
2006							0.68	0.67	0.67	0.68	0.67	0.68	0.69	0.69	0.70
2007								0.66	0.67	0.69	0.69	0.69	0.69	0.69	0.70
2008									0.65	0.67	0.66	0.68	0.68	0.67	0.69
2009										0.67	0.67	0.69	0.68	0.68	0.68
2010											0.68	0.70	0.70	0.69	0.70
2011												0.69	0.70	0.70	0.70
2012													0.71	0.70	0.71
2013														0.71	0.71
2014															0.72

The nearest neighbour decile change measure exhibits an atypical relationship between sample size and accuracy. Table 16 presents the SOFIE measure of decile change, where the sample sizes are near 16,000, while Table 17 presents the HES measure of decile change, where the sample sizes are near 4,000. As the sample size decreases, for this particular measure, the accuracy generally improves (although remains noticeably below actual decile change in all cases). This most likely the distance function is overly penalising relative income changes, particularly when there is a larger pool of potential 'neighbours' which can be selected. Section 6.3 addresses the issue with relative income by using information from the nearest

TABLE 15: Mean absolute decile change using IRD data - Nearest Neighbour

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.01	0.01	0.02	0.03	0.03	0.05	0.05	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.07
2001		0.01	0.01	0.01	0.02	0.03	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.06
2002			0.01	0.01	0.02	0.02	0.03	0.03	0.03	0.03	0.03	0.04	0.04	0.04	0.04
2003				0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.04
2004					0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.05
2005						0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03	0.03	0.04
2006							0.02	0.02	0.03	0.03	0.03	0.03	0.04	0.04	0.04
2007								0.01	0.01	0.02	0.02	0.02	0.03	0.03	0.03
2008									0.02	0.02	0.02	0.02	0.03	0.03	0.03
2009										0.03	0.03	0.03	0.04	0.04	0.04
2010											0.01	0.01	0.02	0.02	0.02
2011												0.01	0.01	0.02	0.02
2012													0.02	0.02	0.02
2013														0.03	0.03
2014															0.01

100 neighbours, which overestimates the decile change. Section B addresses the issue with relative income by excluding income from the distance function, which overestimates decile change by about the same magnitude as Table 15 understimates.

TABLE 16: Mean absolute decile change using SOFIE data - Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.30	0.33	0.38	0.43	0.46	0.49	0.52
2003		0.29	0.33	0.38	0.42	0.45	0.48
2004			0.29	0.33	0.38	0.42	0.46
2005				0.29	0.34	0.38	0.43
2006					0.29	0.33	0.37
2007						0.31	0.35
2008							0.30

TABLE 17: Mean absolute decile change using HES data - Nearest Neighbour

Year	2008	2009	2010	2011	2012	2013	2014	2015
2007	0.56	0.61	0.67	0.68	0.70	0.80	0.83	0.72
2008		0.56	0.64	0.64	0.66	0.80	0.83	0.70
2009			0.57	0.61	0.67	0.77	0.79	0.68
2010				0.57	0.59	0.71	0.72	0.67
2011					0.55	0.66	0.68	0.61
2012						0.59	0.63	0.57
2013							0.56	0.51
2014								0.49

Nearest neighbour multi-period inequality in Table 18 is higher than actual multi-period inequality presented in Table 6. Tables 19 and 20 present the relevant measure in the left sub-table and the absolute deviation from the administrative measure in the right sub-table.[25] They show that similar to the decile change measure, as the sample size decreases the accuracy generally improves. In most HES years, the HES measure of multi-period inequality is usually within 0.01 of the actual value. The difference in HES measures for 2015 predictably worsen, as the HES 2015 sample size is noticeably larger than any previous year.

TABLE 18: Multi-period inequality using IRD data - Nearest Neighbour

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.45	0.45	0.45	0.44	0.44	0.44	0.44	0.43	0.43	0.44	0.43	0.44	0.43	0.43	0.44
2001		0.44	0.44	0.44	0.43	0.43	0.43	0.42	0.43	0.43	0.43	0.42	0.44	0.43	0.42
2002			0.44	0.44	0.43	0.43	0.43	0.42	0.43	0.43	0.43	0.43	0.43	0.42	0.42
2003				0.44	0.43	0.43	0.43	0.42	0.42	0.43	0.43	0.43	0.42	0.42	0.42
2004					0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42	0.42	0.42
2005						0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.42	0.42	0.42
2006							0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43	0.42
2007								0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.42
2008									0.44	0.44	0.44	0.43	0.43	0.43	0.42
2009										0.44	0.44	0.44	0.43	0.43	0.43
2010											0.44	0.44	0.44	0.44	0.43
2011												0.45	0.44	0.44	0.44
2012													0.45	0.44	0.44
2013														0.45	0.44
2014															0.44

TABLE 19: Multi-period inequality using SOFIE data - Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.43	0.43	0.42	0.42	0.42	0.41	0.41
2003		0.43	0.42	0.42	0.42	0.41	0.40
2004			0.43	0.43	0.42	0.42	0.41
2005				0.43	0.43	0.42	0.41
2006					0.43	0.43	0.42
2007						0.43	0.43
2008							0.43

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.01	0.02	0.02	0.03	0.03	0.03	0.03
2003		0.01	0.02	0.02	0.03	0.03	0.02
2004			0.01	0.02	0.03	0.03	0.03
2005				0.01	0.02	0.02	0.02
2006					0.01	0.02	0.02
2007						0.01	0.02
2008							0.01

Left hand side presents the multi-period inequality measure using nearest neighbour matching, and the right hand side presents the absolute difference compared to IRD actual multi-period inequality.

This contrasts with the specification of the model in Appendix B, which excludes income as a matching variable. The income mobility derived from IRD panel data typically lies within the bounds of the two specifications, which indicates that a better set of matching variables, or a different weighting scheme, may yield more accurate estimates of income mobility from repeated cross-sectional data.

[25] This convention is used consistently in similarly structured tables throughout the text.

TABLE 20: Multi-period inequality using HES data - Nearest Neighbour

Year	2008	2009	2010	2011	2012	2013	2014	2015
2007	0.41	0.41	0.40	0.40	0.41	0.40	0.39	0.40
2008		0.42	0.42	0.41	0.42	0.41	0.40	0.41
2009			0.42	0.42	0.42	0.41	0.41	0.42
2010				0.42	0.43	0.41	0.41	0.41
2011					0.43	0.41	0.41	0.42
2012						0.42	0.43	0.43
2013							0.42	0.42
2014								0.44

Year	2008	2009	2010	2011	2012	2013	2014	2015
2007	0.01	0.00	0.00	0.00	0.01	0.01	0.01	0.02
2008		0.01	0.01	0.01	0.02	0.02	0.01	0.03
2009			0.00	0.01	0.01	0.01	0.01	0.03
2010				0.00	0.01	0.00	0.01	0.01
2011					0.00	0.01	0.00	0.02
2012						0.00	0.01	0.02
2013							0.00	0.01
2014								0.01

Left hand side presents the multi-period inequality measure calculated using nearest neighbour matching, and the right hand side presents the absolute difference compared to IRD actual multi-period inequality.

6.3 PROBABILISTIC MATCHING

Probabilistic matching is computationally challenging, both in constructing the distance matrix and calculating the 100 sets of measures for each pair of years in these data. As such, probabilistic matching was restricted to SOFIE data.

Tables 21, 22, 23 and 24 present the correlation, rank correlation, decile change and multi-period gini respectively. Compared to nearest neighbour matching correlation is less accurate as the table of differences on the right hand side are typically larger. Rank correlation measures are closer to IRD rank correlation, and the deviations are below IRD measures. As the nearest neighbour measure of rank correlation was above IRD data, this suggests by adjusting the number of units taken it may be possible to get a better measure of rank correlation using probabilistic matching. Mean absolute decile change is more accurate, and again the deviations with probabilistic matching above IRD data when nearest neighbour matching produced estiamtes below. Multi-period inequality is slightly less accurate, and again the probabilistic matching and nearest neighbour estimates sandwich the IRD value.

As probabilistic matching uses information from individuals who are further away from the nearest neighbour to inform mobility estimates, this suggests that the information about income mobility - at least within the specification defined by the distance function - deteriorates quicker than the distance function penalises the probability of selection. Within the probabilistic matching method this problem could be fixed by taking fewer neighbours, or by increasing the exponent of the penalty applied as distance from the observation increases. Either way provides evidence that taking the 100 nearest neighbours is more than enough - if not too many - for estimating income mobility.

Table 21: Correlation using SOFIE data - Probabilistic Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.42	0.41	0.39	0.28	0.34	0.32	0.29
2003		0.47	0.44	0.39	0.36	0.34	0.33
2004			0.46	0.40	0.38	0.36	0.35
2005				0.47	0.41	0.39	0.39
2006					0.40	0.39	0.41
2007						0.45	0.41
2008							0.43

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.38	0.32	0.29	0.38	0.29	0.28	0.28
2003		0.35	0.33	0.33	0.33	0.30	0.28
2004			0.38	0.40	0.38	0.34	0.31
2005				0.38	0.38	0.34	0.28
2006					0.47	0.42	0.35
2007						0.42	0.39
2008							0.41

Left hand side presents the correlation measure calculated using probabilistic matching on SOFIE data, and the right hand side presents the absolute difference compared to IRD actual correlation.

Table 22: Rank correlation using SOFIE data - Probabilistic Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.70	0.63	0.58	0.53	0.51	0.49	0.45
2003		0.70	0.63	0.59	0.54	0.51	0.48
2004			0.71	0.64	0.59	0.55	0.51
2005				0.71	0.64	0.59	0.55
2006					0.71	0.64	0.59
2007						0.71	0.63
2008							0.70

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.15	0.14	0.13	0.12	0.10	0.09	0.08
2003		0.15	0.14	0.12	0.12	0.11	0.10
2004			0.15	0.14	0.13	0.11	0.11
2005				0.15	0.14	0.12	0.12
2006					0.15	0.14	0.13
2007						0.15	0.15
2008							0.15

Left hand side presents the rank correlation measure calculated using probabilistic matching on SOFIE data, and the right hand side presents the absolute difference compared to IRD actual rank correlation.

Table 23: Mean absolute decile change using SOFIE data - Probabilistic Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	1.65	1.85	1.99	2.11	2.17	2.22	2.31
2003		1.64	1.85	1.95	2.09	2.18	2.24
2004			1.63	1.84	1.96	2.07	2.17
2005				1.63	1.84	1.95	2.07
2006					1.62	1.81	1.96
2007						1.63	1.84
2008							1.65

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.74	0.61	0.54	0.49	0.43	0.37	0.36
2003		0.74	0.63	0.51	0.49	0.44	0.39
2004			0.72	0.61	0.52	0.46	0.44
2005				0.72	0.63	0.53	0.48
2006					0.73	0.62	0.55
2007						0.74	0.64
2008							0.75

Left hand side presents the mean absolute decile change calculated using probabilistic matching on SOFIE data, and the right hand side presents the absolute difference compared to IRD actual mean absolute decile change.

Table 24: Multi-period inequality using SOFIE data - Probabilistic Nearest Neighbour

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.39	0.38	0.37	0.36	0.36	0.35	0.35
2003		0.39	0.37	0.37	0.36	0.36	0.35
2004			0.39	0.38	0.37	0.36	0.36
2005				0.39	0.38	0.37	0.36
2006					0.39	0.38	0.37
2007						0.39	0.38
2008							0.39

Year	2003	2004	2005	2006	2007	2008	2009
2002	0.03	0.03	0.03	0.03	0.03	0.03	0.03
2003		0.03	0.03	0.03	0.03	0.03	0.03
2004			0.03	0.03	0.03	0.03	0.03
2005				0.02	0.03	0.03	0.03
2006					0.03	0.03	0.03
2007						0.02	0.03
2008							0.03

Left hand side presents the multi-period inequality measure calculated using probabilistic matching on SOFIE data, and the right hand side presents the absolute difference compared to IRD actual multi-period inequality measure.

7 Conclusion

This study has presented a methodology to construct synthetic panels from cross-sectional data to estimate measures of income mobility. The methodology is based on deriving a distance function and linking individuals across data sets based on the resulting distance. The two methods, nearest neighbour matching and probabilistic matching, were applied to administrative panel data, survey panel data and repeated cross-sectional data, with the results were compared to estimates of income mobility derived from the administrative panel data.

Prior work by Fields & Viollaz (2013) found estimates of income mobility from synthetic panel techniques were poor. We find similarly poor results when the mean age cohort technique is applied to the data used in this study, which indicates that the other two techniques presented in their paper are unlikely to provide good estimates of income mobility in our case. The two direct matching techniques presented, nearest neighbour matching and probabilistic matching, have mixed performance estimating income mobility, where generally the performance depends on the (inverse) sample size, the time difference between the two years and the extent to which the measure is calculated on relative income. The differences observed with the new techniques are typically in the opposite direction to those observed using synthetic panel techniques, and the true income mobility estimates typically lie within the bounds of the two variable specifications presented, which indicates that more research on the specification could lead to a model which provides accurate measures of income mobility.

These results suggest that further work specifying the variables included in the model, and the weights assigned to them, could yield a direct matching technique which produces accurate income mobility measures. Such further work could include an alternative weighting methodology, which explicitly assigns weight based on the accuracy of the resulting dynamics from the administrative panel data. Additional variables such as education could be investigated, using the SOFIE panel data in place of the IRD data for deriving weights. There is also further work needed with the direct matching techniques to change the unit of analysis to the household or family level, or to use household or family variables as part of the matching criteria for the individual.

References

Antman, F. & McKenzie, D. J. (2007), 'Earnings Mobility and Measurement Error: A Pseudo-Panel Approach', *Economic Development and Cultural Change* **56**, 125–161. [Cited on page 3.]

Arya, S., Mount, D., Kemp, S. E. & Jefferis, G. (2015), *RANN: Fast Nearest Neighbour Search (Wraps Arya and Mount's ANN Library)*. R package version 2.5.
URL: *http://CRAN.R-project.org/package=RANN* [Cited on page 13.]

Ball, C. & Creedy, J. (2016), 'Inequality in new zealand 1983/84 to 2012/13'. Forthcoming.
URL: *http://dx.doi.org/10.1080/00779954.2015.1128963* [Cited on page 3.]

Bentley, J. L. (1975), 'Multidimensional binary search trees used for associative searching', *Commun. ACM* **18**(9), 509–517. [Cited on page 13.]

Bjorklund, A. & Jantti, M. (1997), 'Intergenerational income mobility in Sweden compared to the United States', *American Economic Review* **87**(5), 1009–18. [Cited on page 3.]

Bourguignon, F., Goh, C.-c. & Kim, D. I. (2004), Estimating individual vulnerability to poverty with pseudo-panel data, Policy Research Working Paper Series 3375, The World Bank. [Cited on page 3.]

Claus, I., Creedy, J. & Teng, J. (2012), 'The elasticity of taxable income in new zealand*', *Fiscal Studies* **33**(3), 287–303. [Cited on page 3.]

Dang, H.-A. H. & Lanjouw, P. F. (2013), Measuring poverty dynamics with synthetic panels based on cross-sections, Policy Research Working Paper Series 6504, The World Bank. [Cited on page 3.]

Dang, H.-A., Lanjouw, P., Luoto, J. & McKenzie, D. (2011), *Using Repeated Cross-Sections to Explore Movements into and Out of Poverty*, The World Bank. [Cited on page 3.]

Fields, G. & Viollaz, M. (2013), Can the limitations of panel datasets be overcome by using pseudo-panels to estimate income mobility.
URL: *http://www.ecineq.org/ecineq_bari13/FILESxBari13/CR2/p90.pdf* [Cited on pages 2, 3, 10, 12, 17, and 26.]

Heckman, J., Ichimura, H. & Todd, P. E. (1997), 'Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme', *Review of Economic Studies* **64**(4), 605–654. [Cited on page 2.]

Knuth, D. E. (1973), *The Art of Computer Programming, Volume 3: Sorting and Searching*, Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA. [Cited on page 12.]

Martinez Jr., A., Western, M., Haynes, M. & Tomaszewski, W. (2013), 'Measuring income mobility using pseudo-panel data', *Philippine Statistician* **62**(2), 71–99. [Cited on page 3.]

Navarro, A. I. (2011), 'Estimating long term earnings mobility in Argentina with pseudo-panel data*', *Revista de Analisis Economico âĂŞ Economic Analysis Review* **25**(2), 65–90. [Cited on page 3.]

OSterberg, T. (2000), 'Intergenerational income mobility in Sweden: What do tax-data show?', *Review of Income and Wealth* **46**(4), 421–436. [Cited on page 3.]

Rubin, D. B. (1986), 'Statistical matching using file concatenation with adjusted weights and multiple imputations', *Journal of Business and Economic Statistics* **4**(1), 87–94. [Cited on page 3.]

Statistics New Zealand (2014), Linking methodology used by Statistics New Zealand in the Integrated Data Infrastructure project, Technical report. Available from www.stats.govt.nz. [Cited on page 4.]

A Detailed weight derivation tables

The following tables present either the average subset size (as a proportion of the relevant population) or the predictive accuracy of the variables used in model specification. Some variables have perfect predictive accuracy, such as age, in which case they have been omitted to conserve space.

TABLE 25: Average subset size - Age

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.0018	0.0018	0.0018	0.0019	0.0019	0.0020	0.0020	0.0020	0.0021	0.0021	0.0022	0.0022	0.0023	0.0024	0.0024
2001	0.0018	0.0018	0.0018	0.0019	0.0019	0.0020	0.0020	0.0020	0.0021	0.0021	0.0022	0.0022	0.0023	0.0024	
2002	0.0018	0.0018	0.0018	0.0019	0.0019	0.0020	0.0020	0.0020	0.0021	0.0021	0.0022	0.0022	0.0023		
2003	0.0018	0.0018	0.0018	0.0019	0.0019	0.0019	0.0020	0.0020	0.0021	0.0021	0.0022	0.0022			
2004	0.0018	0.0018	0.0018	0.0019	0.0019	0.0019	0.0020	0.0020	0.0021	0.0021	0.0022				
2005	0.0018	0.0018	0.0018	0.0019	0.0019	0.0019	0.0020	0.0020	0.0021	0.0021					
2006	0.0018	0.0018	0.0018	0.0019	0.0019	0.0019	0.0020	0.0020	0.0021						
2007	0.0018	0.0018	0.0018	0.0018	0.0019	0.0019	0.0020	0.0020							
2008	0.0017	0.0018	0.0018	0.0018	0.0019	0.0019	0.0020								
2009	0.0017	0.0018	0.0018	0.0018	0.0019	0.0019									
2010	0.0017	0.0018	0.0018	0.0019	0.0019										
2011	0.0017	0.0018	0.0018	0.0019											
2012	0.0017	0.0018	0.0018												
2013	0.0017	0.0018													
2014	0.0017														
Weights	0.9982	0.9982	0.9982	0.9981	0.9981	0.9981	0.9980	0.9980	0.9979	0.9979	0.9978	0.9978	0.9977	0.9976	0.9976

TABLE 26: Average subset size - Ethnicity 1

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.67	0.67	0.67	0.67	0.67	0.66
2001	0.65	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	
2002	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.66	0.65		
2003	0.64	0.64	0.64	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65	0.65			
2004	0.63	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64				
2005	0.63	0.63	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.63					
2006	0.62	0.63	0.63	0.63	0.63	0.63	0.63	0.63	0.63						
2007	0.62	0.62	0.62	0.62	0.63	0.63	0.63	0.62							
2008	0.62	0.62	0.62	0.62	0.62	0.62	0.62								
2009	0.61	0.61	0.61	0.62	0.62	0.61									
2010	0.61	0.61	0.61	0.61	0.61										
2011	0.6	0.61	0.61	0.61											
2012	0.6	0.6	0.6												
2013	0.6	0.6													
2014	0.59														
Weights	0.378	0.373	0.371	0.367	0.364	0.362	0.358	0.356	0.353	0.35	0.346	0.34	0.34	0.335	0.34

TABLE 27: Average subset size - Ethnicity 2

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71
2001	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71	
2002	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71	0.71		
2003	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71			
2004	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71				
2005	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71					
2006	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71						
2007	0.72	0.72	0.72	0.72	0.72	0.72	0.72	0.71							
2008	0.72	0.72	0.72	0.72	0.72	0.72	0.72								
2009	0.72	0.72	0.72	0.72	0.72	0.72									
2010	0.72	0.72	0.72	0.72	0.72										
2011	0.72	0.72	0.72	0.72											
2012	0.72	0.72	0.72												
2013	0.72	0.72													
2014	0.72														
Weights	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.29	0.28	0.29	0.29

TABLE 28: Average subset size - Ethnicity 3

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88	0.88
2001	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88	0.88	0.88	
2002	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.87		
2003	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.88	0.87			
2004	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87				
2005	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87					
2006	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87						
2007	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87							
2008	0.86	0.86	0.86	0.86	0.87	0.87	0.87								
2009	0.86	0.86	0.86	0.86	0.87	0.86									
2010	0.86	0.86	0.86	0.86	0.86										
2011	0.86	0.86	0.86	0.86											
2012	0.86	0.86	0.86												
2013	0.86	0.86													
2014	0.85														
Weights	0.14	0.14	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.12	0.12	0.12	0.12	0.12

TABLE 29: Average subset size - Ethnicity 4

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.91
2001	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.90	0.90	
2002	0.88	0.88	0.88	0.88	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.89		
2003	0.86	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88	0.88	0.88	0.88			
2004	0.85	0.86	0.86	0.86	0.86	0.86	0.87	0.87	0.87	0.87	0.87				
2005	0.84	0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.86	0.86					
2006	0.84	0.84	0.84	0.85	0.85	0.85	0.85	0.85	0.85						
2007	0.83	0.83	0.84	0.84	0.84	0.84	0.84	0.84							
2008	0.82	0.83	0.83	0.83	0.83	0.83	0.83								
2009	0.81	0.82	0.82	0.82	0.82	0.82									
2010	0.81	0.81	0.81	0.82	0.82										
2011	0.80	0.81	0.81	0.81											
2012	0.79	0.80	0.80												
2013	0.79	0.79													
2014	0.78														
Weights	0.17	0.16	0.15	0.15	0.14	0.14	0.13	0.13	0.12	0.12	0.11	0.11	0.10	0.10	0.09

TABLE 30: Average subset size - Ethnicity 5

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97
2001	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97	
2002	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97		
2003	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.96			
2004	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96	0.96				
2005	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96	0.96					
2006	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.96						
2007	0.94	0.95	0.95	0.95	0.95	0.96	0.96	0.96							
2008	0.94	0.95	0.95	0.95	0.96	0.96	0.96								
2009	0.94	0.95	0.95	0.95	0.96	0.96									
2010	0.94	0.95	0.95	0.95	0.95										
2011	0.95	0.95	0.95	0.95											
2012	0.95	0.95	0.95												
2013	0.95	0.95													
2014	0.95														
Weights	0.05	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.03	0.03	0.03	0.03

TABLE 31: Average subset size - Ethnicity 6

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.94	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
2001	0.94	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	
2002	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95		
2003	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95			
2004	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95				
2005	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95					
2006	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95						
2007	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96							
2008	0.95	0.95	0.95	0.95	0.95	0.95	0.96								
2009	0.95	0.95	0.96	0.96	0.96	0.96									
2010	0.95	0.96	0.96	0.96	0.96										
2011	0.96	0.96	0.96	0.96											
2012	0.96	0.96	0.96												
2013	0.96	0.96													
2014	0.96														
Weights	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

TABLE 32: Predictive accuracy (probability) - Meshblock

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.56	0.35	0.34	0.32	0.29	0.29	0.28	0.27	0.25	0.26	0.27	0.27	0.26	0.26	0.26
2001	0.39	0.37	0.34	0.31	0.31	0.30	0.29	0.27	0.28	0.28	0.28	0.27	0.27	0.27	
2002	0.69	0.57	0.47	0.43	0.40	0.37	0.37	0.34	0.32	0.31	0.28	0.28	0.27		
2003	0.69	0.54	0.48	0.44	0.41	0.40	0.37	0.35	0.34	0.31	0.31	0.30			
2004	0.65	0.55	0.50	0.46	0.43	0.40	0.38	0.37	0.34	0.33	0.32				
2005	0.70	0.60	0.53	0.49	0.44	0.42	0.40	0.36	0.35	0.33					
2006	0.72	0.61	0.53	0.48	0.45	0.43	0.39	0.38	0.36						
2007	0.72	0.59	0.53	0.50	0.47	0.43	0.42	0.39							
2008	0.67	0.59	0.54	0.50	0.45	0.44	0.41								
2009	0.69	0.59	0.53	0.46	0.44	0.41									
2010	0.73	0.63	0.54	0.51	0.47										
2011	0.74	0.60	0.56	0.51											
2012	0.69	0.62	0.56												
2013	0.77	0.66													
2014	0.76														

TABLE 33: Average subset size - Meshblock

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.1098	0.1103	0.1101	0.1094	0.1085	0.1078	0.1067	0.1054	0.1039	0.1024	0.1009	0.0994	0.0977	0.0957	0.0915
2001	0.0993	0.0992	0.0987	0.0981	0.0975	0.0966	0.0955	0.0942	0.0928	0.0914	0.0900	0.0884	0.0866	0.0828	
2002	0.0096	0.0095	0.0094	0.0093	0.0092	0.0091	0.0089	0.0087	0.0085	0.0084	0.0082	0.0080	0.0073		
2003	0.0073	0.0072	0.0071	0.0070	0.0069	0.0067	0.0066	0.0065	0.0063	0.0062	0.0060	0.0054			
2004	0.0050	0.0050	0.0049	0.0048	0.0047	0.0046	0.0045	0.0044	0.0043	0.0042	0.0037				
2005	0.0023	0.0023	0.0022	0.0022	0.0021	0.0021	0.0020	0.0020	0.0019	0.0017					
2006	0.0016	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014	0.0014	0.0012						
2007	0.0012	0.0011	0.0011	0.0011	0.0011	0.0010	0.0010	0.0009							
2008	0.0008	0.0008	0.0007	0.0007	0.0007	0.0007	0.0006								
2009	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002									
2010	0.0002	0.0002	0.0002	0.0002	0.0001										
2011	0.0001	0.0001	0.0001	0.0001											
2012	0.0001	0.0001	0.0001												
2013	0.0000	0.0000													
2014	0.0000														
Weights	0.67	0.56	0.49	0.44	0.41	0.38	0.36	0.33	0.31	0.30	0.28	0.27	0.25	0.24	0.23

TABLE 34: Predictive accuracy (probability) - Meshblock Income Decile

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.27	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.17	0.17	0.17	0.17
2001	0.30	0.19	0.19	0.18	0.18	0.18	0.18	0.17	0.17	0.17	0.17	0.17	0.17	0.17	
2002	0.35	0.28	0.24	0.22	0.21	0.21	0.20	0.20	0.20	0.19	0.19	0.19	0.18		
2003	0.44	0.33	0.26	0.24	0.23	0.22	0.22	0.21	0.21	0.20	0.20	0.19			
2004	0.39	0.29	0.26	0.25	0.24	0.23	0.22	0.22	0.21	0.21	0.20				
2005	0.38	0.32	0.28	0.26	0.25	0.24	0.23	0.22	0.21	0.21					
2006	0.43	0.35	0.28	0.27	0.26	0.25	0.23	0.22	0.22						
2007	0.42	0.31	0.29	0.28	0.26	0.24	0.23	0.23							
2008	0.35	0.32	0.30	0.28	0.26	0.24	0.24								
2009	0.39	0.32	0.28	0.26	0.24	0.24									
2010	0.41	0.33	0.29	0.27	0.26										
2011	0.43	0.33	0.31	0.28											
2012	0.39	0.35	0.31												
2013	0.46	0.37													
2014	0.43														

TABLE 35: Average subset size - Meshblock Income Decile

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.19	0.20	0.20	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.18	0.18	0.18
2001	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.18	0.18	0.18	0.18	0.18	0.18	
2002	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11		
2003	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11			
2004	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11				
2005	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11					
2006	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11						
2007	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10							
2008	0.10	0.10	0.10	0.10	0.10	0.10	0.10								
2009	0.10	0.10	0.10	0.10	0.10	0.10									
2010	0.10	0.10	0.10	0.10	0.10										
2011	0.10	0.10	0.10	0.10											
2012	0.10	0.10	0.10												
2013	0.10	0.10													
2014	0.10														
Weights	0.35	0.27	0.24	0.22	0.21	0.20	0.19	0.18	0.17	0.17	0.16	0.15	0.15	0.14	0.14

TABLE 36: Predictive accuracy (probability) - Income Decile

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.49	0.39	0.34	0.30	0.27	0.25	0.24	0.22	0.22	0.21	0.21	0.20	0.20	0.19	0.19
2001	0.50	0.40	0.34	0.30	0.27	0.25	0.24	0.23	0.22	0.22	0.21	0.21	0.20	0.20	
2002	0.50	0.40	0.34	0.30	0.28	0.26	0.24	0.24	0.23	0.22	0.22	0.21	0.21		
2003	0.50	0.40	0.34	0.30	0.27	0.26	0.25	0.24	0.24	0.23	0.22	0.21			
2004	0.49	0.39	0.34	0.30	0.28	0.27	0.25	0.25	0.24	0.23	0.22				
2005	0.49	0.40	0.34	0.31	0.29	0.27	0.26	0.25	0.24	0.23					
2006	0.50	0.40	0.35	0.33	0.30	0.28	0.27	0.25	0.24						
2007	0.49	0.41	0.37	0.33	0.31	0.29	0.27	0.25							
2008	0.50	0.42	0.38	0.34	0.31	0.29	0.27								
2009	0.52	0.44	0.39	0.34	0.31	0.29									
2010	0.53	0.44	0.37	0.33	0.30										
2011	0.53	0.44	0.36	0.33											
2012	0.54	0.43	0.37												
2013	0.54	0.43													
2014	0.54														

Table 37: Average subset size - Income Decile

Year	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
2000	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
2001	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	
2002	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10		
2003	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10			
2004	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10				
2005	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10					
2006	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10						
2007	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10							
2008	0.10	0.10	0.10	0.10	0.10	0.10	0.10								
2009	0.10	0.10	0.10	0.10	0.10	0.10									
2010	0.10	0.10	0.10	0.10	0.10										
2011	0.10	0.10	0.10	0.10											
2012	0.10	0.10	0.10												
2013	0.10	0.10													
2014	0.10														
Weights	0.46	0.37	0.32	0.28	0.26	0.24	0.23	0.22	0.21	0.20	0.19	0.19	0.18	0.18	0.17

B Sensitivity: excluding income variables from matching

A sensitivity test is presented where income is excluded from the nearest neighbour matching. Table 38 details the weights used for the distance function in the excluding income sensitivity test. Summarising Tables 39 and 40 and 41 and 42, using only the non-income variables produces income mobility measures which are among the worst of the psuedo-panel techniques considered in this study.

TABLE 38: Relative variable weights used in matching when income is excluded

Variable	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	t+12	t+13	t+14	t+15
Age	324	348	362	372	380	387	393	400	406	411	416	422	428	432	434
Ethnicity 1	123	131	135	137	139	141	142	143	144	145	145	146	147	147	147
Ethnicity 2	90	97	101	105	107	110	112	114	116	118	120	122	124	126	128
Ethnicity 3	45	47	48	49	50	51	51	51	52	52	52	53	53	53	53
Ethnicity 4	55	56	56	56	55	54	53	52	50	49	47	46	44	43	41
Ethnicity 5	18	18	18	18	17	17	16	16	16	15	15	14	14	14	14
Ethnicity 6	16	17	17	18	18	19	19	19	20	20	21	21	21	22	22
Meshblock	217	193	177	165	154	147	141	133	126	121	116	111	106	104	101
Meshblock Income Decile	112	94	86	81	78	76	74	72	70	69	67	65	63	60	61
Income Decile	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TABLE 39: IRD data correlation - Nearest Neighbour excluding income

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.18	0.19	0.16	0.12	0.18	0.16	0.17	0.16	0.15	0.16	0.16	0.14	0.15	0.15	0.16
2001		0.18	0.14	0.17	0.17	0.18	0.15	0.16	0.15	0.15	0.16	0.16	0.15	0.15	0.16
2002			0.19	0.20	0.20	0.20	0.18	0.18	0.18	0.18	0.17	0.18	0.17	0.17	0.18
2003				0.22	0.22	0.22	0.21	0.20	0.19	0.19	0.19	0.19	0.19	0.19	0.19
2004					0.22	0.22	0.21	0.21	0.20	0.19	0.19	0.20	0.19	0.19	0.19
2005						0.24	0.23	0.22	0.21	0.21	0.21	0.21	0.21	0.20	0.20
2006							0.24	0.23	0.21	0.22	0.22	0.21	0.22	0.21	0.21
2007								0.23	0.23	0.22	0.22	0.22	0.22	0.22	0.21
2008									0.22	0.23	0.22	0.23	0.22	0.22	0.22
2009										0.22	0.23	0.22	0.22	0.22	0.22
2010											0.23	0.24	0.23	0.23	0.23
2011												0.24	0.24	0.24	0.23
2012													0.25	0.25	0.25
2013														0.26	0.25
2014															0.26

TABLE 40: IRD data rank correlation - Nearest Neighbour excluding income

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.25	0.25	0.24	0.23	0.23	0.20	0.20	0.17	0.18	0.17	0.17	0.17	0.16	0.16	0.18
2001		0.26	0.24	0.24	0.22	0.20	0.21	0.18	0.17	0.17	0.18	0.17	0.16	0.15	0.18
2002			0.28	0.26	0.25	0.23	0.22	0.21	0.19	0.18	0.18	0.18	0.18	0.18	0.18
2003				0.28	0.26	0.24	0.23	0.22	0.20	0.19	0.19	0.18	0.18	0.18	0.17
2004					0.28	0.26	0.24	0.23	0.22	0.20	0.20	0.19	0.19	0.19	0.18
2005						0.29	0.27	0.25	0.24	0.22	0.22	0.21	0.20	0.20	0.19
2006							0.29	0.27	0.25	0.24	0.23	0.22	0.21	0.21	0.20
2007								0.29	0.27	0.25	0.24	0.24	0.22	0.22	0.21
2008									0.29	0.27	0.26	0.25	0.24	0.23	0.22
2009										0.28	0.27	0.26	0.24	0.24	0.22
2010											0.28	0.27	0.26	0.25	0.23
2011												0.29	0.27	0.26	0.25
2012													0.30	0.28	0.27
2013														0.30	0.29
2014															0.31

TABLE 41: IRD data mean absolute decile change - Nearest Neighbour excluding income

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	2.73	2.73	2.76	2.79	2.79	2.87	2.87	2.91	2.91	2.93	2.93	2.94	2.95	2.96	2.93
2001		2.69	2.76	2.78	2.80	2.85	2.85	2.91	2.92	2.94	2.91	2.93	2.95	2.96	2.92
2002			2.67	2.71	2.75	2.79	2.82	2.85	2.89	2.90	2.90	2.91	2.92	2.92	2.92
2003				2.66	2.71	2.76	2.80	2.82	2.86	2.89	2.90	2.90	2.91	2.91	2.92
2004					2.65	2.71	2.76	2.78	2.82	2.86	2.87	2.88	2.89	2.90	2.91
2005						2.64	2.70	2.74	2.78	2.82	2.83	2.85	2.87	2.88	2.89
2006							2.64	2.69	2.74	2.77	2.80	2.82	2.84	2.85	2.87
2007								2.64	2.70	2.74	2.76	2.79	2.81	2.83	2.86
2008									2.65	2.69	2.72	2.74	2.78	2.80	2.83
2009										2.67	2.70	2.73	2.77	2.79	2.82
2010											2.67	2.70	2.73	2.76	2.79
2011												2.65	2.69	2.73	2.76
2012													2.64	2.67	2.72
2013														2.63	2.67
2014															2.61

TABLE 42: IRD data multi-period Gini - Nearest Neighbour excluding income

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
2000	0.37	0.38	0.38	0.38	0.38	0.37	0.36	0.37	0.37	0.37	0.37	0.37	0.36	0.36	0.37
2001		0.37	0.38	0.36	0.35	0.35	0.35	0.34	0.35	0.35	0.35	0.35	0.36	0.35	0.35
2002			0.37	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.36	0.35	0.35
2003				0.37	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
2004					0.37	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35	0.35
2005						0.37	0.36	0.36	0.35	0.35	0.35	0.35	0.35	0.35	0.35
2006							0.37	0.36	0.36	0.36	0.36	0.36	0.35	0.35	0.35
2007								0.37	0.36	0.36	0.36	0.36	0.36	0.35	0.35
2008									0.37	0.37	0.36	0.36	0.36	0.36	0.35
2009										0.37	0.37	0.36	0.36	0.36	0.35
2010											0.37	0.37	0.36	0.36	0.36
2011												0.38	0.37	0.37	0.36
2012													0.38	0.37	0.37
2013														0.38	0.37
2014															0.37

About the Authors

Christopher Ball is a Senior Analyst at the New Zealand Treasury.
Email: Christopher.Ball@treasury.govt.nz

VICTORIA UNIVERSITY OF WELLINGTON
TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI

Chair in Public Finance
Victoria Business School

Working Papers in Public Finance