Australian National University

Crawford School of Public Policy

TTPI

Tax and Transfer Policy Institute

Using census, social security and tax data from the Multi-Agency Data Integration Project (MADIP) to impute the complete Australian income distribution

TTPI - Working Paper 8/2021
April 2021

Nicholas Biddle
Centre for Social Research and Methods, & Tax and Transfer Policy Institute,
Australian National University

Dinith Marasinghe
Centre for Social Research and Methods, Australian National University

Abstract

The distribution of income, income dynamics and how observable characteristics predict an individual's position on the income distribution are all core aspects of economics and social science research, and of keen interest to policy makers. Researchers approach these topics using a combination of cross-sectional surveys, panel studies, and administrative datasets. In Australia, all three types of datasets have been used historically to help answer such questions, without any one individual dataset being without limitations in terms of sample size, sample representation, quality of income data, or longitudinal availability. A relatively new dataset – the Multi-Agency Data Integration Partnership (MADIP) Basic Longitudinal Extra (BLE) – has the potential to extend our knowledge of income in Australia by combining income-related data from a targeted survey (the 2011 or 2016 Censuses of Population and Housing), income tax records at the individual level, and information on access to social security. As individual datasets, there are limits of each. However, one way to overcome the limitations of the individual datasets on the MADIP BLE is to combine them to create a synthetic income measure for each individual. For 2011, this is a relatively straightforward exercise, as there are three sources of information for each individual. For the other years though, there are only two sources of information – PIT and SSRI. To overcome this limitation, we borrow information from the first wave of data (2011) to help estimate income for the remaining years (2012-16). After testing nine machine-learning approaches using a training and test dataset from the MADIP BLE 2011, we were able to generate a synthetic income measure that performed far better than either tax or census data alone in matching the HILDA income distribution, and was also able to capture income dynamics reasonably well, albeit with some understating of income dynamics. This new synthetic income data is available for further analysis for over 15 million individuals, compared to only around 17,000 for HILDA and even less for other sample surveys.

JEL Codes: C52, D31
Keywords: MADIP data, machine learning, model validation and selection, imputation, income distributions, income dynamics

∗*The authors would like to thank the Australian Bureau of Statistics for helping make data available for this project, Professor Robert Breunig for discussion on an earlier version of the paper, and an anonymous reviewer for very helpful comments.*

Tax and Transfer Policy Institute

Crawford School of Public Policy

College of Asia and the Pacific

+61 2 6125 9318

tax.policy@anu.edu.au

The Australian National University

Canberra ACT 0200 Australia

www.anu.edu.au

The Tax and Transfer Policy Institute (TTPI) is an independent policy institute that was established in 2013 with seed funding from the federal government. It is supported by the Crawford School of Public Policy of the Australian National University.

TTPI contributes to public policy by improving understanding, building the evidence base, and promoting the study, discussion and debate of the economic and social impacts of the tax and transfer system.

The Crawford School of Public Policy is the Australian National University's public policy school, serving and influencing Australia, Asia and the Pacific through advanced policy research, graduate and executive education, and policy impact.

1 Introduction and overview

Economists, other social scientists, and policymakers have a keen interest in the distribution of income, its dynamics and how observable characteristics predict an individual's position on the income distribution. These topics are answered in different countries using a combination of cross-sectional surveys, panel studies, and administrative datasets (Auten and Splinter 2019; Jauch and Watzka 2015; Bhardwaj 2018; Adler and Schmid 2012).

In Australia, to tackle these questions, researchers have been well served by sample surveys on the income distribution, but these are limited for relatively small population groups or precise points on the distribution. Furthermore, Australian researchers have made limited use of administrative data, at least historically. This is not because the administrative data doesn't exist, but because of privacy and practical challenges with linking individuals and making that data available to external researchers. As a consequence, there is much researchers do not know about the trends, distribution and dynamics of income in Australia.

The four main sources of income data used in Australia for research on income and income dynamics are, Personal income tax (PIT) records; the Household, Income, and Labour Dynamics in Australia (HILDA) survey; Census of Population and Housing; and Survey of Income and Housing (SIH). Combined, we can generate a reasonable picture of income dynamics in Australia (see the recent Productivity Commission (2018) report for a broad overview). In isolation though all these data sources have limitations when studying income distributions and income dynamics and without them being linked, there are gaps in our overall picture. These limitations are discussed below.

Repeated cross-sectional Personal income tax (PIT) records are administrative data covering all taxpayers. The PIT data contains a 1% of sample records for 2004-2011 and a 2% for sample records for 2011-2016. These samples are chosen pseudo-randomly. PIT records provide limited information about wealth, tax-exempt income, and socio-demographic variables. Furthermore, the PIT data as it is made available to the public are both bottom and top coded (the bottom and top 1%) to preserve confidentiality. Finally, and perhaps most importantly, PIT data only contain information on individuals who have completed a tax return in a given financial year. As a result, PIT records do not include information on approximately 40% of adults who are generally at the bottom of the income distribution (Atkinson and Leigh 2007). Therefore, PIT data can be useful for understanding trends at the middle and the top of the income distribution, but this exclusion results in a top-biased income distribution that is not reflective of the entire population in Australia (Tran and Zakariyya 2019).

Household, Income and Labour Dynamics in Australia (HILDA) Survey is a nationally representative household-based panel study that commenced in 2001, with 9742 households surveyed in Wave 18, the most recent wave available (Summerfield et al. 2019). Data from 2019 has been collected but not released, whereas data for Wave 20 was being collected at the time of writing this report. HILDA Survey contains a rich set of information on income. Namely, household income, wealth, consumption and socio-demographic variables. While it

provides valuable scope for cross-sectional research on income, it's most notable attribute is its longitudinal design which enables the study of income dynamics (Wilkins 2015; Sila and Dugain 2019; Productivity Commission 2018). The main limitation of HILDA though is its relatively small sample size (at least compared to administrative records and the Census). As a result it may fail to capture income dynamics and income distribution information of households/individuals at the top of the income distribution. Furthermore, because it is a longitudinal survey, HILDA is also affected by non-random attrition and consequently, it has become less representative through time (Watson and Wooden 2004).

The Australian Census Population and Housing is conducted every five years. It collects gross income information of each household member in broad income bands and is both bottom and top coded. This censoring leads to a much higher level of information loss relative to PIT, HILDA and SIH. Furthermore, the censoring mechanism makes the data unsuitable for in-depth analysis of inequality such as producing Gini coefficients.[1] Another limitation of the Census is the inconsistency of income bands used in different waves of data collection. This inconsistency makes it difficult to differentiate true income distributional changes from changes that occur due to income band changes. However, due to its large sample size and the fact that it collects information on those inside and outside of the tax system, the Census is still widely used in many studies that examine income for small geographic areas or subpopulations. (Hunter and Gregory 1996; Athanasopoulos and Vahid 2003; Markham and Biddle 2018; Biddle and Montaigne 2017). Income dynamics can also be studied for a randomly selected sample of the population (about 5 per cent of records) using the Australian Census Longitudinal Dataset (ACLD)

Finally, The Survey of Income and Housing (SIH) is a cross-sectional household survey which collects extensive information on income, sources of income, household wealth, household and individual socio-demographics. The most recent SIH provides a relatively large sample of 14,060 households for July 2017 to June 2018. Over the years, due to its large and representative sample size, SIH has been used in numerous studies to examine topics in income (Bray 2014; Athanasopoulos and Vahid 2003; Productivity Commission 2018). However, despite the attractiveness of SIH, it has a few shortcomings when examining income. Firstly, the SIH is a cross-sectional survey and therefore the dynamics of income across years is out of scope. Moreover, SIH has undergone periodical changes to survey methods. These changes are likely to affect the income information captured in repeated cross-sections. Consequently, this makes it difficult to distinguish true changes in income trends from changes in income that emerged from methodological changes.

A dataset that has the potential to fill the data gaps discussed above is the Multi-Agency Data Integration Partnership (MADIP) Basic Longitudinal Extract 2011 (BLE2011) dataset. The Basic Longitudinal Extract (2011) relates to the Australian population in 2011 and includes individual income information, socio-demographic information, and social security payment information

[1] Generalization techniques can be used to approximate Gini Coefficients when presented with Ordinal data. (Peñaloza 2019)

for the period 2011-2016. The sample size is approximately 22.5 million and contains 122 variables out of which 74 variables contain information for multiple years which allows for longitudinal studies of income.

In order to develop the MADIP BLE 2011, four data sources were linked at an individual level. Namely, Medicare Enrolments Database (MEDB) and Medicare Benefits Schedule (MBS) data; 2011 Census of Population and Housing (Census) data; Personal Income Tax (PIT) datasets; and Social Security and Related Information (SSRI) data. Furthermore, 12 derived (demographic) variables were also included. The four datasets were linked using multiple statistical linkage methods without the use of a unique identifier and a final linkage rate of 66.5% was achieved across the four datasets.[2] [3]

In terms of income data, MADIP BLE 2011 has income information from 3 sources for the period 2011-2016. Namely, 2011 Census, PIT and SSRI. In terms of 2011 Census income, the dataset provides cross-sectional Total Personal Income (weekly) in 12 categories. PIT dataset provides a variety of income variables for the financial years 2010-11 to 2015-16. These include wages and salaries, government allowances, pensions and payments, total income and taxable income. In this study, since we are interested in the total income distribution of individuals, we will focus only on the total income variable in PIT. Furthermore, for the rest of the paper, the total income in PIT will be referred to as PIT income. The final measure that relates to income, SSRI payment information, is provided as 28 binary variables and represent whether an individual received a particular benefit in a given year. Table 1 provides an overview of these income sources.

Table 1: Income sources in MADIP BLE 2011

Income source	Type	No. of categories	Availability
2011 Census income	Categorical	12	2011
PIT income	Categorical	253	2010-11 to 2015-16 financial years
SSRI income	Binary	28	2011 to 2016

Despite MADIP BLE 2011 overcoming some of the limitations of other datasets discussed above, that is, despite having a large sample with longitudinal income data and extensive socio-demographic information for each individual, MADIP BLE 2011 still faces a few shortcomings that diminishes its initial effectiveness when examining the income distribution. The main

[2] The linkage of PIT 2010-11 to MEDB records achieved a linkage rate of 93.4%. The linkage of SSRI 2011 to MEDIB records achieved a linkage rate of 94.6%. Finally, the linkage rate of Census 2011 to MEDB was 66.5%. Furthermore, given the low linkage rate between Census 2011 and MEDB, a weight was included for Census 2011 records to account of this low linkage which allows the researchers to weight results to reflect the Census 2011 population.

[3] Refer to Biddle *et al* (2019) for an extensive introduction to MADIP BLE 2011 and income measures in MADIP BLE 2011.

limitation is that when each individual dataset in MADIP BLE 2011 is used in isolation to study income, the limitations of these individual datasets will resurface. For example, if PIT income is used in isolation, the resulting income distribution will be biased towards top income earners and will not be representative of the entire population. This is primarily because PIT income only includes individuals who have lodged a tax return in a given financial year. As a result, PIT excludes individuals with low income that have no incentive to lodge a tax return. Contrastingly, if 2011 Census income is used in isolation, the resulting income distribution will not be reflective of the true population because Census income fails to capture the top income distribution. This is mainly due to top-level censoring that occurs at $104,000. Furthermore, the 2011 Census uses wide dynamic income bands which results in a significant amount of information loss, and is only available at one point in time.

One way to overcome the limitations of the individual datasets on the MADIP BLE is to combine them to create a synthetic income measure for each individual. For 2011, this is a relatively straightforward exercise, as there are three sources of information for each individual. For the other years though, there are only two sources of information – PIT and SSRI. One option with this dataset is to combine the SSRI in a deterministic way with the data from PIT. That is, if an individual received a particular payment at a point in time, then the rules for that payment could theoretically be used to ascribe a social security income to that individual. This could then be combined with their tax data to obtain an estimate of their total income.

There are two challenges though with using this approach. First, there is variation in the level of payments someone might receive from the different programs with some receiving the full payment and some receiving a part payment. The income received from a given payment can vary based on demographic, household, other income, and asset characteristics. Only a limited amount of that information is available on the dataset. The second limitation is that it is unclear from the data available on MADIP as to whether the income that would be estimated from SSRI has or has not been included in the person's tax record. Once again, the dataset does not have the information available to be able to estimate that interaction accurately.

To overcome the limitations of the SSRI and PIT data alone, we borrow information from the Census and other records in the first wave of data (2011) to help estimate income for the remaining years (2012-16). With that motivation in mind, the goal of this paper is to outline and test a methodology for the creation of a continuous longitudinal measure of income, that is representative of the entire population (aged 15 or more), for each of the years in the 2011 MADIP BLE. The remainder of the paper is structured as follows. In the next section, we define our measure of synthetic income and outline the proposed estimation methodology. The results are presented in Section 3 (cross-sectional results) and Section 4 (longitudinal results). Section 5 summarises and provides some concluding comments.

2 Methodology

2.1 Defining synthetic Income

The underlying methodology for this project is to use social security payment information (SSRI) and Census data to predict/impute income data for those individuals who are not available on the tax system for the first year of data (2011), and then use the estimated relationship between these datasets to impute income for subsequent years when Census data is not available.

The income measure that is the subject of this paper is total income, as defined by the so-called 'Canberra Group' (UNECE 2011). This is defined as a combination of Current Transfers and Primary Income, with the latter made up of Property Income and Income from Production, with the latter part of this sub-component made up of Income from Employment and Income from Household Production of Services for Own Consumption. According to the Canberra Group, Total Income can further be adjusted to create Disposable Income (Total Income minus Current Transfers Paid). While this may be a better measure of a person's standard of living, transfers paid is not able to be estimated accurately from Census data and cannot therefore be estimated with the training dataset.

Definition 1. For individual i in period t, synthetic income $y_{i,t}^{s}$ is defined as:

$$y_{i,t}^{s} = \begin{cases} PIT_{i,t} & \text{if } PITy_{i,t} = \mathbb{R} \\ \widehat{cy}_{i,t} & \text{if } PITy_{i,t} \neq \mathbb{R} \end{cases}$$

Where $PIT_{i,t}$ represents PIT total income for individual i at time t obtained from personal income tax records, $\widehat{cy}_{i,t}$ represents the predicted income for individual i at time t and is a function of census data and SSRI f(census data, SSRI).

This definition of synthetic income allows us to overcome the limitations of MADIP BLE 2011 income data discussed in Section 1. More specifically, by allowing the synthetic income distribution $y_{i,t}^{s}$ to take the values of $PIT_{i,t}$ when $PIT_{i,t} = \mathbb{R}$ (i.e. when $PIT_{i,t}$ is a real number) captures the income of individuals who have lodged a tax return in period t. This will include individuals whose income in a particular year was greater than the tax-free threshold, as well as those who otherwise need to complete a tax return. It should be noted that $PIT_{i,t}$ is not just the taxable income of those who are in the tax system, but rather the total income. This will include some (but not necessarily all) income received even if it is exempt from taxation.[4]

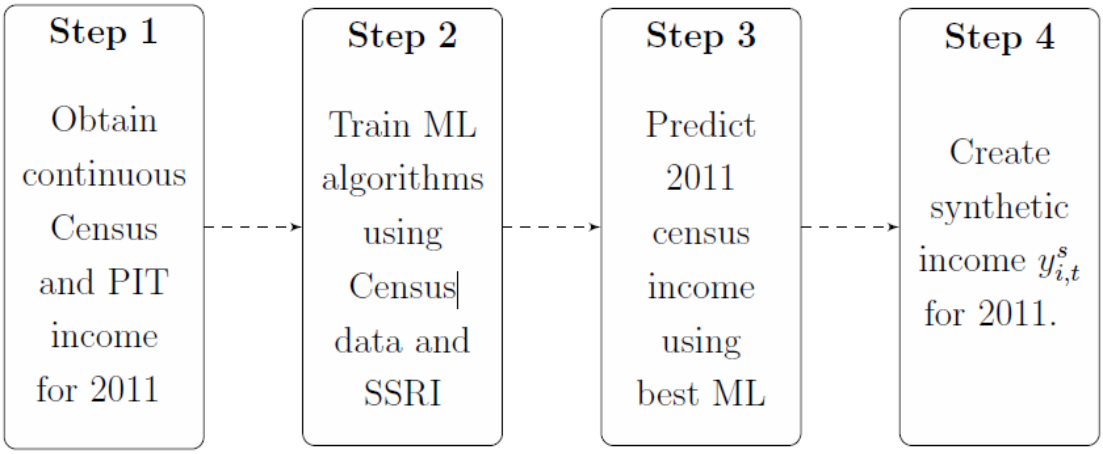
By allowing the synthetic income distribution $y_{i,t}^{s}$ to take up the values $\widehat{cy}_{i,t}$ when $PIT_{i,t} \neq \mathbb{R}$ (i.e. when $PIT_{i,t}$ is missing) we are able to obtain an estimate of income for those individuals who had no incentive to lodge a tax return in period t. This will primarily include individuals whose income in a particular financial year was less than the tax-free threshold, but will also

[4] We test whether it is more accurate to use $\widehat{cy}_{i,t}$ for the entire sample, rather than just those outside the tax system. Estimates using this approach are less accurate in aggregate relative to our comparison dataset.

include those whose income is exempt from taxation. Therefore, by combining $PIT_{i,t}$ and $\widehat{cy}_{i,t}$ we are able to obtain a continuous measure of income that is representative of the entire population.[5]

The methodological approach to create a synthetic income distribution for 2011 ($y^{s}_{i,2011}$) entails four main steps. Figure 1 outlines these steps. Sections 2.2 to 2.5 provide an in-depth discussion of each step outlined in Figure 1. Section 2.5.1 discusses the methodology used to create a synthetic income distribution for the years 2012 to 2015.

Figure 1: Methodological approach



2.2 Step 1: Obtain continuous Census and PIT income for 2011

In this first step, the goal was to approximate a continuous measure of income for both 2011 Census income and 2011 PIT income. This was achieved in two steps. First, we regressed an Interval regression (Generalized censored) model on the data and secondly, using the coefficients of the interval regression model, a continuous measure of income was estimated. This procedure is outlined below.

Interval regression for 2011 Census income and 2011 PIT income can be presented as: [6]

$$y^{*} = \beta \mathbf{x} + \epsilon \qquad (1)$$

where y^{*} is the true (unobserved) income. $\mathbf{x}$ is a vector of explanatory variables and $\epsilon \sim (0, \sigma^{2})$. The censoring mechanism used to create income categories is given as:

$$y = 1 \quad \text{if} \quad y^{*} < A_{1}$$

[5] The method of converting categorical income to continuous variable is discussed in section 2.2

[6] On ABS DataLab, an Interval regression model can be estimated using the *Interval Regression* package on STATA, *IntReg* package on R or manually coded-up on Python. GitHub link for Python code to will be made available soon.

$$\vdots$$

$$y = j \quad \text{if} \quad A_{j-1} < y^* < A_j$$

$$\vdots$$

$$y = \text{J} \quad \text{if} \quad A_J < y^*$$

Where A_{j-1} and A_j are income categories and $j \in [1,12]$ and $j \in [1,253]$ for 2011 Census income and 2011 PIT income, respectively.

This model then maximizes a likelihood function similar to an Ordered Probit model except that it uses the boundaries that were explicitly specified.[7]

$$\ell = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\Phi\left(\frac{A_j - \beta \mathbf{x} + \epsilon}{\sigma} \right) - \Phi\left(\frac{A_{j-1} - \beta \mathbf{x} + \epsilon}{\sigma} \right) \right]^{d_{ij}}$$

A consistent and unbiased estimate of a continuous income for an observed grouped income observation will be:[8]

$$E[y_i | A_{j-1} < y^* < A_j] = \hat{\beta}\mathbf{x} + \hat{\sigma} \left[\frac{\phi\left(\frac{A_{j-1} - \hat{\beta}\mathbf{x} + \epsilon}{\hat{\sigma}} \right) - \phi\left(\frac{A_j - \hat{\beta}\mathbf{x} + \epsilon}{\hat{\sigma}} \right)}{\Phi\left(\frac{A_j - \hat{\beta}\mathbf{x} + \epsilon}{\hat{\sigma}} \right) - \Phi\left(\frac{A_j - \hat{\beta}\mathbf{x} + \epsilon}{\hat{\sigma}} \right)} \right] \qquad (3)$$

where ϕ and Φ are the density function and the cumulative distribution of the standard normal respectively.

Theoretically, Equation (3) is identical to estimating the conditional expectation of a truncated normal distribution. Therefore, in our case, this would be identical to estimating a continuous income value for an individual given this individual belongs to an income category $A_{j-1} < y^* < A_j$.

The consistency and the unbiasedness of the Interval Regression model coefficients and Equation (3) are strongly based on the assumption of normality. Given the usual Gamma distribution of income, it was important to check whether our data satisfied the normality assumption, and if not, to transform our data into a log-normal normal distribution. We conducted several normality tests to assess the normality assumption.

[7] Results of the interval regression given in Appendix 7.1

[8] Stewart (1983): On Least Squares Estimation when the Dependent Variable is Grouped

First, we tested the assumption of normality using a standard Jarque-Bera (JB) test. This was an essential test to decide whether to transform our data to a log-normal distribution. According to the JB test for 2011 Census income, there was insufficient evidence to conclude that the normal distribution assumption is unreasonable at a 10% level of significance. The results from the JB test implied that our original 2011 Census income approximated a normal distribution. The main justification for this observed normal distribution (and not the usual gamma distribution common in income distributions) was the wide dynamic income bands and top-level censoring that occurred at $104,000, thus eliminating the long-tail observed in standard income distributions. Furthermore, as a second normality test, we constructed a standardized normal probability plot (Normal P-P plot) for both 2011 Census income distribution and log 2011 Census income[9]. According to the figures, it was evident that for the 2011 Census income, the deviations from the identity line (45-degree line) were minimal. This indicated that the 2011 Census income approximated a normal distribution. However, for log 2011 census income, the deviations were significant which indicated that a log transformation was not needed to transform the data into a normal distribution.[10]

For 2011 PIT income, a log transformation was needed because the normality assumption was violated. This was confirmed by a JB test as well as a Normal P-P plot.[11]

Table 1 in the Appendix provides the results of the Interval regression model. According to the model the variables included are all independently and jointly statistically significant at a 5% significance level. We then proceeded to approximate a continuous value of income for both 2011 Census income and log 2011 PIT income using equation (3). We then transformed the continuous log-PIT income to nominal values using the equation:

$$\hat{y}_{nominal} = \exp(\hat{\beta}_1 + \hat{\beta}_2 + \ldots + \hat{\beta}_n + \hat{\sigma}^2/2).^{12}$$

2.3 Step 2: Train ML algorithms using Census data and SSRI

After obtaining the continuous values of income for both 2011 Census income and 2011 PIT income, we then proceeded to train our ML algorithms using other Census data (described below) and Social Security payment information (SSRI). In this step, we treated our continuous income variables obtained from *step 1* as our independent variable (target) and other census data and SSRI as our dependent variables (features).

In terms of ML algorithms, we used 9 regression-based algorithms from SciKit-Learn in Python. Namely:

1. Linear regression – Benchmark model

[9] Graphs provided in Appendix 7.3

[10] Model fit was also used to test the suitability of interval regression. This was done by comparing the model log-likelihood of an ordered Probit model (a highly robust model that does not depend on normality) and interval regression log-likelihood. Results are given in Appendix 7.2

[11] Graphs are given in Appendix 7.4

[12] Given the large sample size, this predictor, on average, provides a closer prediction to the actual value relative to using the exponential function by itself.

2. Ridge regression
3. Bayesian Ridge regression
4. Decision Tree regression
5. Random Forest regression
6. Extra Trees regression
7. Gradient Boosting regression
8. Histogram-based Gradient Boosting regression
9. Multi-layer Perceptron (MLP) regression

We then compared the performance of each of these algorithms using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Adjusted R^2. Given the large number of outliers present in data, we used MAE as our primary criterion for model selection.[13]

2.3.1 Data pre-processing and feature selection

The first step in the data pre-processing stage was to subset our dataset such that individuals with a 2011 PIT income record were isolated from individuals without a 2011 PIT income record. This significantly reduced the variance in income in our dataset and thereby allowed the ML algorithms to be trained on a dataset with less variation. As a result, predictions of all models improved by nearly two-fold. The intuition behind this step was to minimize the variation in income by grouping individuals with similar observed and unobserved characteristics based on their level of income. After achieving this subset, the sample size was reduced to $n = 5{,}846{,}780$.[14]

In terms of feature selection, we included features to capture both 'time-invariant' and 'time-variant' variation in individual income. More specifically, we sourced data from 2011 Census to capture the 'time-invariant' variation/characteristics in income and 2011 Social Security information (SSRI) to capture 'time-variant' variation in income. These features are discussed below.

2.3.1.1 Time-invariant features

Given that the Census data is only available for 2011 and not for subsequent years (i.e. 2012 to 2016) in MADIP BLE 2011, it was imperative to only include features that had a fixed or a time-invariant effect on income. More specifically, we ignored variables that had an effect on income but had a high probability of changing every year. For example, we excluded variables such as 'Full-time/Part-time student status', 'Rent', and 'Relationship in the household' because these variables may have a high probability of taking different values each year and this data was not available on MADIP BLE 2011.

Given the data limitation, we included the following variables to capture the 'time-invariant' characteristics in income. *Sex, Indigenous status, Core activity need for assistance, Industry of employment, Occupation, and Highest level of education*. These variables are not completely immutable, with Indigenous status (Campbell et al. 2018) and disability both changing

[13] Distribution of income with respect to each feature used is provided in Appendix 7.5
[14] Individuals aged 15 or less were also removed in this step.

through time. Furthermore, according to (Wilkins and Lass 2018), the national average tenure in a given job in Australia is 3.3 years. Furthermore, those who do change jobs over a five year period tend to stay in the same industry or occupation. Therefore, we make the assumption that the Industry of employment and Occupation are 'sticky' and do not change every year. All the variables included as 'time-invariant' variables are categorical. Table 2 highlights the data preprocessing/transformation used for each of these variables and the box plots of each of these variables are given in Appendix 7.5.

Table 2: Data pre-processing used for time-invariant variables

Variable	Type	No. of categories	Pre-processing used
Sex	Numerical	3	One-hot encoding
Indigenous status	Numerical	3	One-hot encoding
Core activity need for assistance	Numerical	3	One-hot encoding
Highest level of education	Numerical	13	One-hot encoding
Industry of employment	Mixed	253	Categories collapsed to 106 then used One-hot encoding[15]
Occupation	Mixed	51	Categories collapsed to 8 then used One-hot encoding[16]

2.3.1.2 Time-variant features

Since the Census data features only captured the 'fixed' variation in income, we included a 28 Social Security payment information (SSRI) to capture the 'time-variant' differences in income. Our assumption is that the main determinant of variation in income for those who are not in the tax system is their social security payments. MADIP BLE 2011 dataset contains SSRI data for each individual from 2011 to 2016. The complete list of SSRI variables included is given in the appendix. These variables are dichotomous and represent whether an individual received a particular benefit in the corresponding year. Unfortunately, the level of income received as part of that payment is not available on the MADIP BLE 2011 (or able to be determined using other information), and it is therefore necessary to estimate rather than calculate income support.

The rationale for including SSRI as 'time-variant' features is demonstrated in Figure 2. Figure 2[17] shows the proportion of individuals receiving Social security across the income distribution. It is evident from the figure that approximately 80% of households receive social security benefits if they are below the 20th taxable, household equivalised income percentile. Moreover, it is evident that even at the 100th household equivalised income percentile, a small

[15] Collapsing categories to 106 provided a better model with a lower error.
[16] Collapsing categories to 8 provided a better model with a lower error.
[17] Biddle, N., Breunig, R., Markham, F. and Wokker, C., 2019. Introducing the Longitudinal Multi-Agency Data Integration Project and Its Role in Understanding Income Dynamics in Australia. Australian Economic Review, 52(4), pp.476-495

proportion of households (less than 2%) are receiving social security benefits. Therefore, this result presents us with a logical justification to include SSRI variables as features to capture 'time-variant' heterogeneity in individual income, especially at the bottom of the income distribution.

Furthermore, we also considered correlation matrices for both 'time-invariant' and 'time-variant' features in order to prevent including features that are not correlated 2011 Census income. This allowed us to mitigate the bias-variance tradeoff in our ML algorithms. It is evident from the figures that the included variables are correlated with the target variable and thereby improved ML predictions. These figures are presented in Appendix 7.

Figure 2: Social security across the income distribution

2.3.1.3 Missing data

Figure 3 shows the sparsity of the 2011 snapshot of the data set (before creating the subset mentioned in Section 2.3.1). The X-axis of the figure shows the features (both time-invariant and time-variant) and the Y-axis indicates each individual. In the figure, each 'white' space indicates missing values for each individual for a particular variable. Furthermore, Table 3 shows the missing size and proportion of each variable. Since SciKit-Learn ML algorithms do not handle missing data, we followed three separate techniques to handle the missing data. The 'first' technique was to exclude all missing values from both the target (y) variable and features (X variables) and then train our ML algorithms. We obtained a sample size of 1,832,467 from utilizing this technique. The 'second' technique was to one-hot encode each missing record of X variables. We obtained a sample size of 2,753,435 from using this technique. We proceeded with the latter because it significantly minimized the information loss relative to the 'first' technique, which in turn minimized the risk of potential bias from omitting data. Furthermore, the 'second' approach also improved the model predictions relative to the first technique. In addition to the aforementioned techniques, we also used

iterative imputation techniques to impute missing features. However, this technique resulted in worse predictions relative to others.[18]

After preprocessing our data and handling for missing data, we then divided our dataset into a' train set' (70%) and a 'test set' (30%). In total, 1,954,897 observations were allocated to the 'train set' and 837,813 observations were allocated to the 'test set'. The next step we undertook was to set up our ML algorithms to be trained. When training our ML algorithms, we utilized 5-fold cross-validation on our train set and the hyperparameters of our ML algorithms were tuned using both a Randomized search and a Grid search.[19] For example, a Randomized search was used to isolate a set of locally optimal hyperparameters for each ML algorithm and then a Grid search was used around those hyperparameters to further tune our models. In both searches, 5-fold cross-validation was used. The 5-fold cross-validation technique is outlined in Figure 4. In each iteration, $\frac{1}{5}$th of the training set was held out (the highlighted block), such that the held-out subset was used as the validation set and the other $\frac{4}{5}$th subsets were put together to form a training set. The error estimation was then averaged over all 5 iterations to approximate the total performance of our ML algorithms. Finally, our ML algorithms were tested on our 'test set' to evaluate their performance on *out-of-sample* observations. Table 4 provides the set of hyperparameters used to tune each model and Table 5 provides the model performance evaluated on the 'test set'.

[18]To compare these techniques, we trained and tested our ML algorithms using each of these techniques. The technique that resulted in the best model performance was selected. Model results obtained using the 'first' technique are presented in Appendix 6.5

[19] Given the computational constraints, an exhaustive Grid Search was infeasible. All models were trained, validated and tested on the ABS DataLab which runs on ten E5-2650 processors @2.0GHz (each with 8 cores) and 200 GB of physical memory. This computing power is shared among other DataLab users. Furthermore, ABS DataLab servers are restarted weekly. Therefore, this constraints the model training process to a maximum of 7 days.

Figure 3: Missing data

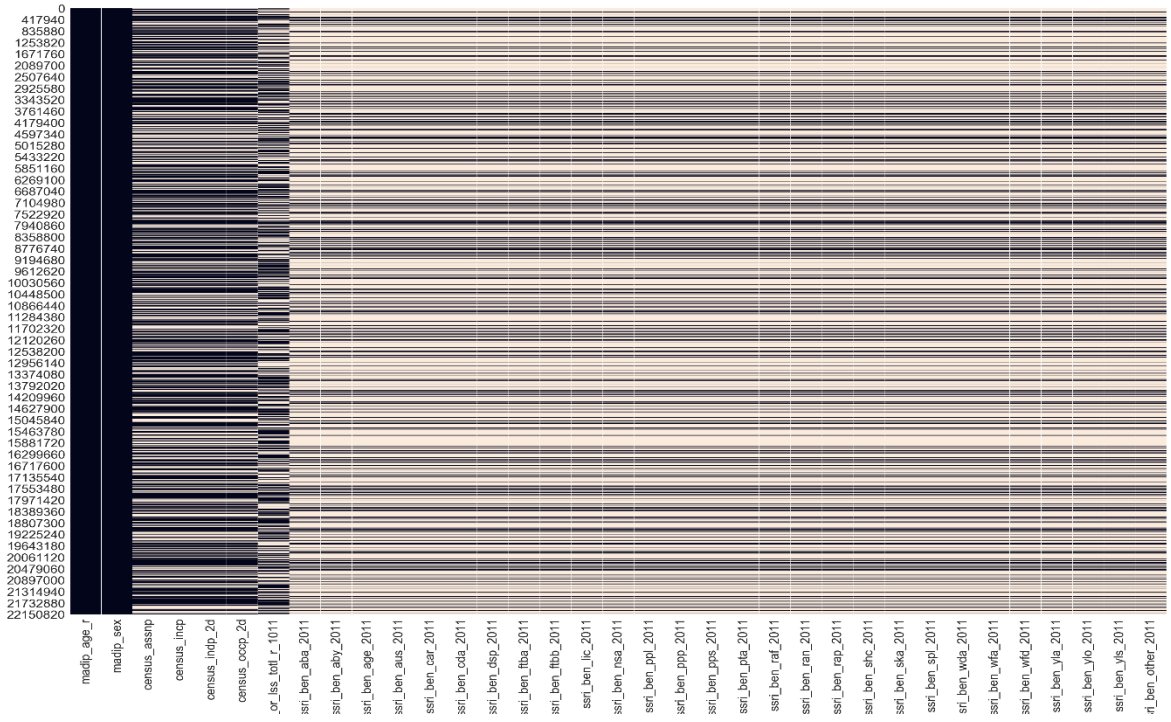


Table 3: Proportion of missing data in each feature prior to handling missing data

Variable	**Missing size**	**Missing proportion**[20]
Sex	0	0.00%
Age	0	0.00%
Indigenous Status	3,349,696	44.70%
Core activity need for assistance	2,537,723	43.40%
Industry of Employment	2,556,356	43.70%
Occupation	2,550,887	43.62%
Highest level of Education	4,608,477	61.40%
28 SSRI variables	2,544,274	43.51%

[20] Sample size n = 5,846,780

Figure 4: 5-fold cross-validation

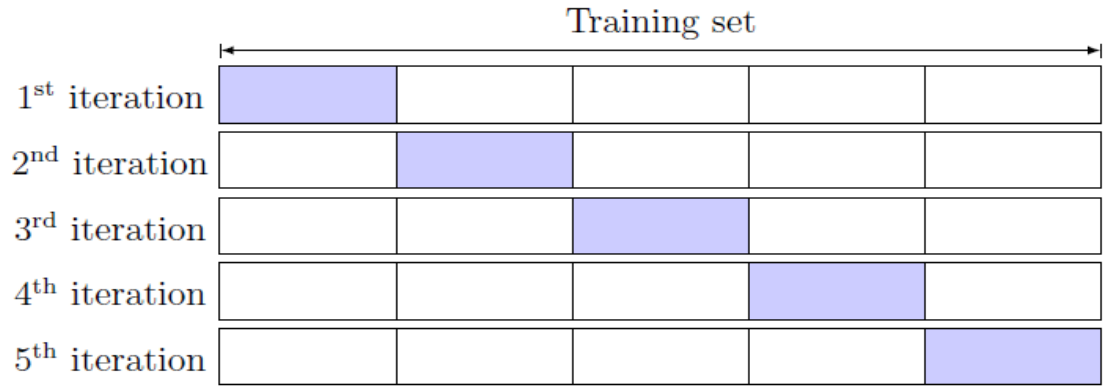


Table 4: Set of hyperparameters used to tune ML algorithms

ML algorithm	Hyperparameters
Linear Regression	Intercept ={True,False} Normalize = {True, False}
Ridge Regression	Intercept = {True,False} Normalize = {True, False} Regularization $\alpha = \{x \in \mathbb{R}_+ \mid 1 \leq x \leq 100\}$
Bayesian Ridge Regression	Intercept ={True,False} Normalize = {True, False}
Decision Tree Regression	Criterion = {MSE, MAE} Splitter = {Best, Random} Max depth = $\{x \in \mathbb{Z}_+ \mid 5 \leq x \leq 200\}$ Min samples split = {2,5,10,15} Min samples leaf = {[1,2,4,6,10]}
Random Forest Regression	Estimators = $\{x \in \mathbb{Z}_+ \mid 50 \leq x \leq 2000\}$ Criterion = {MSE, MAE} Splitter = {Best, Random} Max depth = $\{x \in \mathbb{Z}_+ \mid 3 \leq x \leq 200\}$ Min samples split = {2,5,10,15} Min samples leaf = {1,2,4,6,10} Max features = {Auto, Sqrt} Bootstrap = {True, False}
Extra Trees Regression	Identical to Random Forest
Gradient Boosting Regression	Loss = {LS, LAD, Huber} Estimators = $\{x \in \mathbb{Z}_+ \mid 50 \leq x \leq 2000\}$ Learning rate = {0.01,0.05,0.1,0.2} Criterion = {MSE, MAE, Friedman MSE} Max depth = $\{x \in \mathbb{Z}_+ \mid 3 \leq x \leq 200\}$ Min samples split = {2,5,10,15} Min samples leaf = {1,2,4,6,10} Max features = {Auto, Sqrt} Huber $\alpha = \{x \in \mathbb{R}_+ \mid 0 \leq x \leq 1\}$
MLP Regression	Given in Appendix

Table 5: ML Algorithm Results

Model	MAE	MSE	RMSE	R^2	Time (m)[21]
Linear models					
Linear Regression	155	63,500	251	0.4	0.23
Ridge Regression	154	63,240	251	0.41	0.11
Bayesian Ridge Regression	159	65,536	256	0.38	0.37
Tree-based model					
Decision Tree Regression	128	56,130	236	0.47	1.1
Ensemble models					
Random Forest Regression	122	46,816	216	0.58	91
Extra Trees Regression	125	51,692	227	0.55	110.7
Gradient Boosting**	105	43,275	208	0.6	145.4
Histogram-based Gradient Boosting Regression	124	47,674	218	0.56	31
Neural Network					
MLP Regression	116	45,183	212	0.58	313.5

Results from Table 5 can be distilled into 4 sections. Namely, results of Linear models, Tree-based model, Ensemble models and the Neural Network. Focusing on Linear models, it is evident that the Ridge Regression model outperformed both the Linear Regression (benchmark model) and the Bayesian Ridge Regression model in terms of model errors (MAE and RMSE) and the model fit (R^2). This can be attributed to the regularization (α) parameter used in the Ridge Regression model. The regularization parameter generally reduces the magnitude of the model's coefficients and reduces the model complexity, this in turn, improved the generalization of the model to *out-of-sample* data. However, relative to Ensemble models, all linear models performed poorly in terms of model predictions.

When comparing the results between our Tree-based algorithm and Ensemble models, it is evident that the Ensemble methods outperformed the Decision Tree Regressor in terms of their predictive capability. For example, the ensemble methods were on average 10% more accurate (in terms of MAE) than the Decision Tree Regressor. This edge over the Decision Tree Regressor can be attributed to multiple trees used in Ensemble methods which significantly reduced the probability of a model overfit which in turn improved their predictive power. When comparing the Ensemble methods to our Linear algorithms, the Ensemble methods were on average 30% more accurate (in terms of MAE). When comparing the results of Ensemble methods, the Gradient Boosting Regressor outperformed both the Random Forest

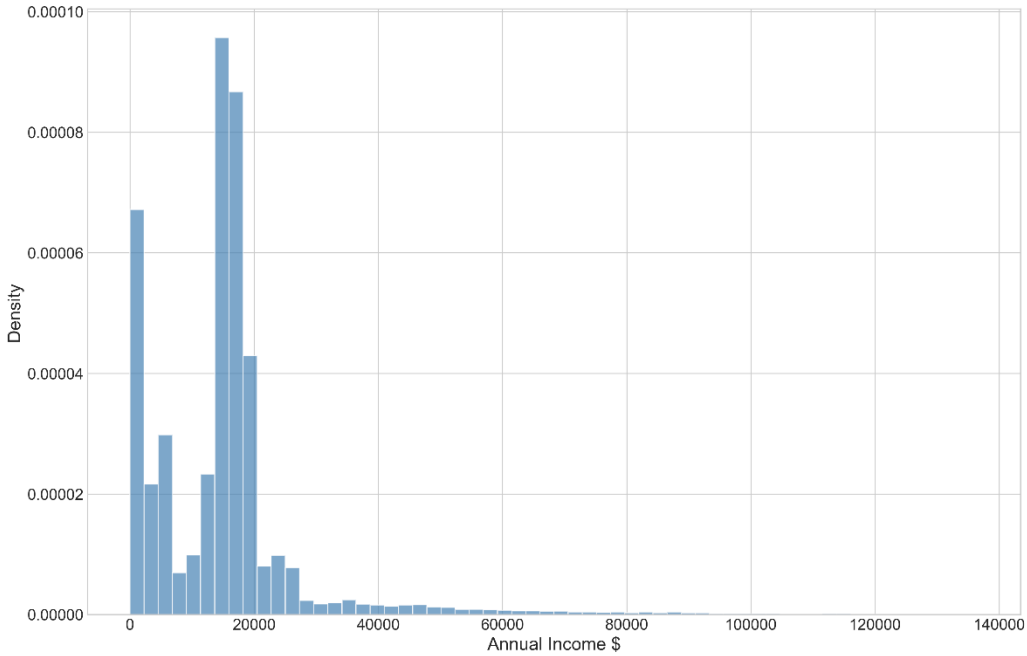
[21] Time taken to run one iteration of the default model. Parallel processing was utilized when cross-validating the ML algorithms.

Regressor and Extra trees Regressor in terms of MAE, RMSE and R^2. The superiority of the Gradient Boosting Regressor model, in this case, can be attributed to the robust loss function that was used, i.e, the Huber loss function.[22]

2.4 Predict 2011 Census income

The next step in our methodology was to use the best performing ML algorithm (Gradient Boosting Regressor), to predict the 2011 Census income ($\widehat{cy}_{i,2011}$) of individuals who do not have a PIT income ($PIT_{i,2011} \neq \mathbb{R}$). The predictions yielded approximately 3 million observations. The predicted values and the histogram of errors are presented in Figure 5 and Figure 6, respectively.

Figure 5: Predicted income values



[22] The Huber loss function combines the properties of the *least absolute deviation (LAD) loss function* and the *least squares (LS) loss function.* Therefore, the Huber function is more robust to outliers than the LS loss function and also more precise close to the minima than the LAD loss function. The latter is due to its differentiable properties around its minimum of 0 (Huber 1964).

Figure 6 Histogram of errors

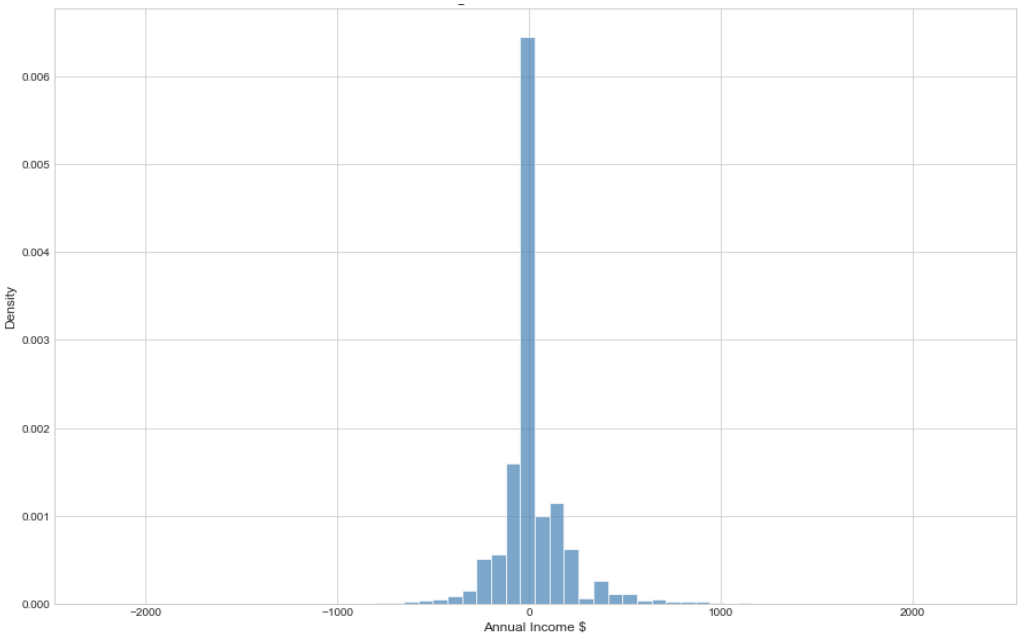


Figure 5 suggests the imputed/predicted income have a mean concentrated approximately around $15,000. [23] This result was highly desirable because it indicated that the model predicted income values primarily for individuals close to or below the tax-free threshold (i.e. individuals not expected to be present in 2011 PIT). Furthermore, Figure 6 highlights that the errors from the ML algorithm are normally distributed, thus, indicating the unbiasedness in the model's predictions.

2.5 Step 4: Create synthetic income $\boldsymbol{y_{i,t}^s}$ for 2011.

The final step for our cross-sectional income estimate was to develop a synthetic income distribution using Definition 1 outlined in Section 2. That is,

$$y_{i,2011}^{s} = \begin{cases} PITy_{i,2011}, & \text{if } PITy_{i,2011} = \mathbb{R} \\ \widehat{cy}_{i,2011} & \text{if } PITy_{i,2011} \neq \mathbb{R} \end{cases}$$

2.6 Step 5: Synthetic income for 2012 to 2015

To create a synthetic income distribution for years 2012 to 2015, we utilized the ML model from 'Step 2' to predict income in a specific year by using that specific years' *time-variant* features and base year's features. For example, the 2012 Census income ($\widehat{cy}_{i,2012}$) was predicted using 2012 *time-variant* features and 2011 *time-invariant* features.

2.7 Step 6: Constructing comparison dataset

[23] Complete descriptive statistics are provided in Appendix 7.9

To compare the accuracy of the synthetic income distribution, *Gross regular income* from HILDA for each corresponding year was used as our benchmark. To improve the comparability between HILDA income and the synthetic income distribution $y_{i,t}^{s}$, HILDA income was truncated to a minimum income of $0 and a maximum value of $250,500. Furthermore, all income values were adjusted for inflation using the 2016 Consumer Price Index (CPI) values. Finally, appropriate HILDA weights were used ensure the HILDA results reflected the Australian population, and jackknife replicate weights (which take in to account the complex sample design of HILDA) were used to create all standard errors and confidence intervals.

In addition to HILDA, synthetic income distribution was compared to other income data available on MADIP BLE 2011 as well. Namely, Census income and PIT income. Furthermore, Census weights were used to ensure Census income reflected the true Australian population.[24] The results of these distributions were compared both cross-sectionally (Section 3) and longitudinally (Section 4). When comparing cross-sectionally, we primarily focused on the accuracy of descriptive statistics, income percentiles and the Gini Coefficient of each year. In terms of longitudinal comparisons, income dynamics of individuals were compared.

3 Results – Cross-sectional validation

Table 6 provides a cross-sectional comparison of income data currently available in MADIP BLE 2011 (2011 Census income and 2011 PIT income), 2011 synthetic income and 2011 HILDA gross regular income. Figure 7 and 8 provide the 2011 HILDA income distribution and synthetic income distribution income, respectively.

To compare and discuss each income distribution presented in Table 6 we will first compare the income sources that are currently available on MADIP (i.e. Census income and PIT income) to HILDA income. In this comparison, we will isolate the key differences between these income sources and their shortcomings relative to HILDA. In the second step, we will compare the synthetic income distribution to HILDA income distribution. More specifically, we will explore their key similarities/differences, and most importantly, highlight how the synthetic income distribution overcomes the shortcomings exhibited by the Census income and PIT income.

[24] Census weights were used to account for the low linkage rate between MEDB and Census variables in MADIP BLE 2011. Furthermore, if a predicted synthetic income value for an individual was based on census income, then appropriate census weights used to ensure the predicted synthetic income value reflected the true population.

Table 6: 2011 MADIP BLE income data, synthetic income and HILDA income

	HILDA income	Census Income	PIT income	Synthetic income
Descriptive statistics				
Sample size	17,612	9,970,192	12,415,603	15,208,313
Mean	47,991	42,356	56,688	46,778*
Median	35,032	28,773	45,579	33,199
Standard deviation	46,759	35,258	49,605	46,672
Income percentiles				
1st percentile	0	0	0	0
5th percentile	123	0**	1,688	273**
10th percentile	4,035	5,754	8,440	4,398**
25th percentile	15,655	14,254	23,070	16,318*
50th percentile	35,032	28,773	45,579	33,199
75th percentile	66,634	63,750	75,965	64,711*
90th percentile	104,041	97,103	111,978	100,724**
95th percentile	134,655	125,520	149,117	133,361**
99th percentile	241,394	128,836	281,915	258,281**
Income statistic				
Gini coefficient	0.488	0.460	0.439	0.480

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

Figure 7: 2011 HILDA Income Distribution

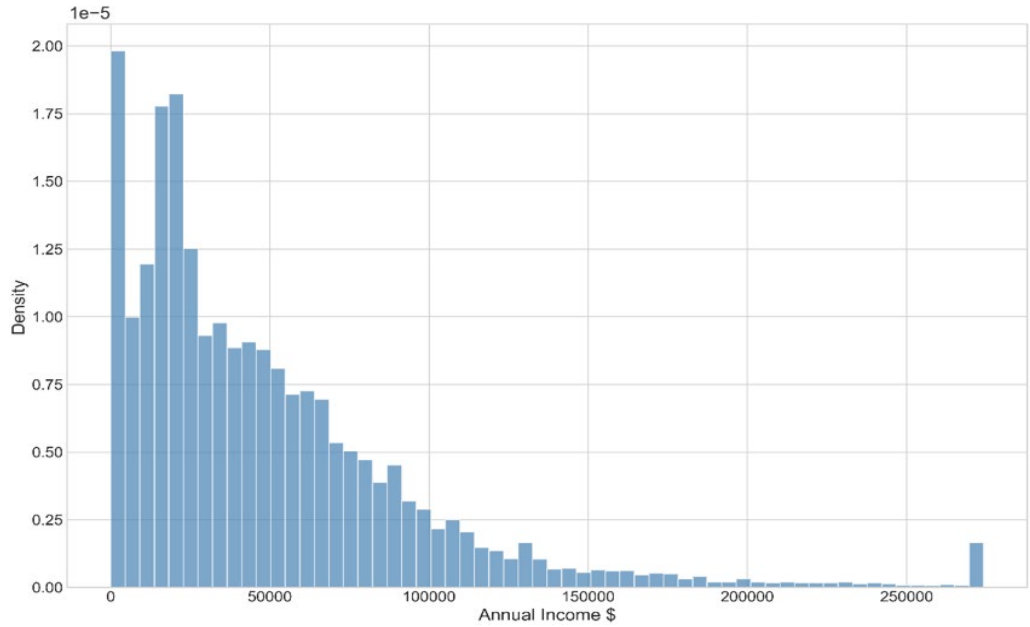
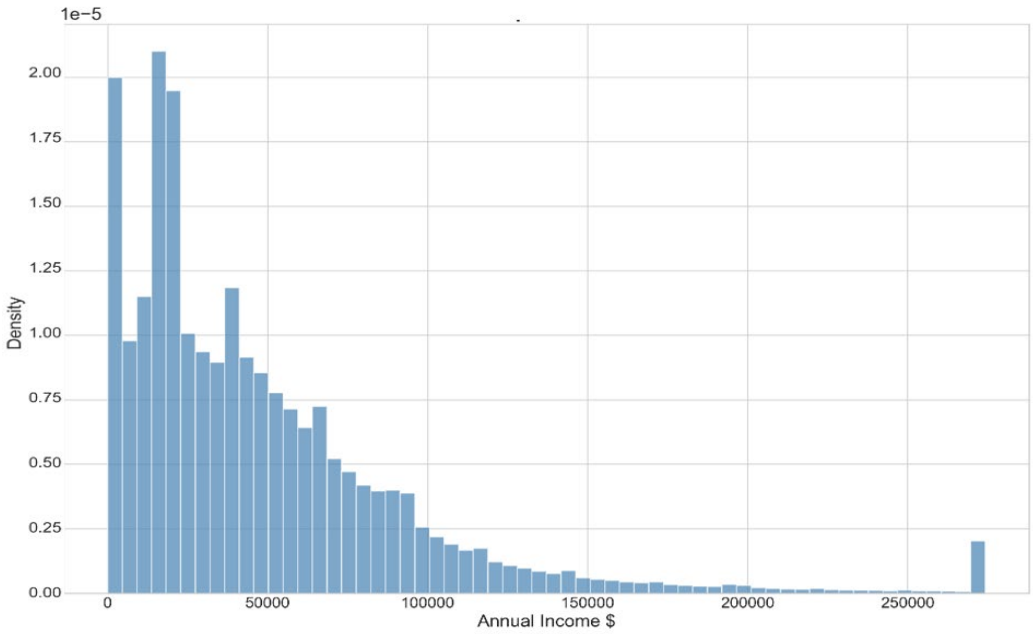


Figure 8: 2011 Synthetic Income Distribution



3.1 Comparison between Census income, PIT income and HILDA income

In terms of descriptive statistics, we immediately observe that the mean and median of Census income and PIT income are significantly different from each other, and also to HILDA. We observe that the mean and median of Census income distribution are significantly lower relative to HILDA (41,560 vs 49,086 and 28,643 vs 36,670, respectively). Whereas in the case of PIT income, these statistics are considerably overestimated relative to HILDA. Furthermore, we also observe that none of the descriptive statistics of Census income and PIT income falls within the 99% confidence interval of HILDA descriptive statistics.

This phenomenon is consistent with the shortcomings of Census and PIT income discussed in section 1. That is, Census income is biased towards low-income earners because it fails to capture the top of the income distribution (due to top-level censoring at $104,000) and PIT income is biased towards high-income earners because it fails to capture a proportion of individuals who are below the tax-free threshold. Based on these descriptive statistics, it is evident that the Census income distribution and PIT income do not mimic the HILDA income distribution. More specifically, this suggests that the Census income and PIT income fail to capture the "true" income distribution of the Australian population.

To further extend our comparison, we shift our focus to the income percentiles and Gini coefficients of these income distributions. We discern a similar phenomenon as discussed above. That is, we observe that Census income tends to understate income percentiles and PIT income overstates income percentiles relative to HILDA. Furthermore, except for the 5th income percentile of Census, no other income percentiles are within the 99% confidence interval of HILDA.[25] Lastly, in terms of the Gini coefficient, the Gini derived from PIT income is considerably understated relative to HILDA. This result is also consistent with the shortcoming of PIT income highlighted in section 1. More specifically, since PIT income fails to capture a proportion of individuals who are below the tax-free threshold (i.e. individuals in the bottom of the income distribution), PIT income portrays a higher level of equality in the income distribution which in turn results in a lower Gini coefficient relative to other distributions. In the case of Census income, the Gini coefficient is relatively closer to the HILDA income than PIT. However, given the inaccurate income percentiles and descriptive statistics of the Census income distribution, the Gini of Census should not be taken at face value and distributional income analyses should be conducted with caution. Furthermore, since Census income in MADIP BLE 2011 is only available for 2011, distributional income analyses using Census income is constrained to 2011.

3.2 Comparison between synthetic income and HILDA income.

In terms of descriptive statistics, we observe that the mean and median of the synthetic income distribution are significantly closer to HILDA. More specifically, the mean of the synthetic

[25] The 5th percentile of HILDA income is not statistically significant at a 10% significance level. This percentile has a Jackknife standard error of 82.10. Furthermore, the 99% confidence interval ranges from [-37,284]. Furthermore, in 2014, 2015, and 2016, the 5th percentile was omitted by STATA.

income distribution is within the 99% confidence of HILDA, despite not using HILDA at all in training or implementing the model. In terms of income percentiles, we observe that all income percentiles expect for the median fall within either 95th or 99th centile confidence of HILDA income. This suggests that the synthetic income distribution not only mimics the HILDA income distribution closely but outperforms both the Census and PIT income in terms of accuracy. More precisely, this suggests that the synthetic income is closer to the "true" income distribution than both Census and PIT income.

Focusing on the Gini coefficient of the synthetic income distribution, it is evident that the Gini is notably closer to the Gini of HILDA than both Census income and PIT income. This result further suggests that the synthetic income distribution captures the representativeness of the income distribution better than both Census and PIT income. Finally, another aspect in which synthetic income surpasses Census income and PIT income is the relatively larger sample size. In 2011, the sample size of synthetic income was approximately 18% larger than PIT income and 34% larger than Census income. As a result, the synthetic income distribution may facilitate a more representative analysis of income relative other income sources in MADIP BLE 2011.

Cross-sectional PIT income, synthetic income and HILDA income results for years 2012 to 2016 are given below. The results demonstrate an identical pattern to 2011. For all years, we observe that that PIT income overstates income percentiles relative to HILDA income. Furthermore, except for the 99th percentile from 2012 to 2016, all other incomes percentiles of PIT are not within the 99% confidence intervals of HILDA income. In terms of Gini coefficients of PIT income, similar to 2011, they are consistently underestimated relative to HILDA. This result implicitly suggests that PIT income fails to capture the Australian income distribution.

In contrast, we observe that for all years the synthetic income distribution mimis the HILDA income distribution similar to 2011. Focusing on individual years, in 2012 (Table 7), except for the median, all income percentiles and the mean were within 95% or 99% confidence intervals of HILDA income. In 2013 (Table 8), all income percentiles except for the 5th and the 99th percentile fell within the 95% or 99% confidence interval. In 2014 and 2015 (Table 9 and Table 10, respectively), all statistical measures fell within the 95% or 99% confidence interval of HILDA. Finally, in 2016 (Table 11), all income percentiles except for the median were within the 95% or 99% confidence interval. In terms of the Gini coefficient, the Gini of synthetic income is significantly closer to HILDA than the Gini coefficient of PIT income.

Table 7: 2012 PIT income, synthetic income and HILDA income comparison

	HILDA income	PIT income	Synthetic income
Descriptive statistics			
Sample size	17,475	11,797,101	14,931,293
Mean	47,991	56,688	46,778*
Median	35,032	45,579	33,199
Standard deviation	47,959	49,605	46,672
Income percentiles			
1st percentile	0	0	0
5th percentile	216	3,857	431*
10th percentile	4,867	11,573	4,964**
25th percentile	16,923	27,004	17,082**
50th percentile	37,155	49,049	34,754
75th percentile	69,227	79,911	66,750*
90th percentile	108,708	118,489	104,262
95th percentile	140,952	158,169	138,465**
99th percentile	255,599	276,107*	265,345**
Income statistic			
Gini coefficient	0.482	0.420	0.472

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

Table 8: 2013 PIT income, synthetic income and HILDA income comparison

	HILDA income	PIT income	Synthetic income
Descriptive statistics			
Sample size	17,501	11,369,247	14,745,429
Mean	50,057	62,118	48,761**
Median	35,032	50,007	34,951*
Standard deviation	48,069	51,011	48,103
Income percentiles			
1st percentile	0	0	0
5th percentile	126	4,839	431
10th percentile	4,329	12,367	4,839**
25th percentile	16,902	27,423	17,082**
50th percentile	36,604	50,007	34,951*
75th percentile	69,375	81,195	67,214*
90th percentile	109,867	122,062	105,930*
95th percentile	142,194	164,004	141,419**
99th percentile	259,879	269,396*	269,396
Income statistic			
Gini coefficient	0.484	0.421	0.473

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

Table 9: 2014 PIT income, synthetic income and HILDA income comparison

	HILDA income	PIT income	Synthetic income
Descriptive statistics			
Sample size	17,512	11,014,669	14,587,119
Mean	49,900	62,685	48,719**
Median	36,528	50,731	35,041*
Standard deviation	47,642	50,929	47,755
Income percentiles			
1st percentile	0	0	0
5th percentile	(omitted)	4,707	503
10th percentile	4,659	13,075	4,707**
25th percentile	17,147	27,719	17,259**
50th percentile	36,528	50,929	35,041*
75th percentile	68,913	82,111	66,457**
90th percentile	107,677	123,951	105,123**
95th percentile	143,569	165,791	141,733**
99th percentile	256,886	262,024**	262,024**
Income statistic			
Gini coefficient	0.482	0.414	0.472

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

Table 10: 2015 PIT income, synthetic income and HILDA income comparison

	HILDA income	PIT income	Synthetic income
Descriptive statistics			
Sample size	17,606	10,519,657	14,363,406
Mean	50,033	63,413	48,666**
Median	36368	50,723	35,318*
Standard deviation	47,483	50,723	47,222
Income percentiles			
1st percentile	0	0	0
5th percentile	(omitted)	4,618	711
10th percentile	4,218	13,856	4,483**
25th percentile	17,173	29,252	17,322**
50th percentile	36,368	50,723	35,318*
75th percentile	69,902	82,624	67,228*
90th percentile	108,590	124,707	105,205**
95th percentile	140,624	167,815	140,103**
99th percentile	253,063	257,112**	257,112**
Income statistic			
Gini coefficient	0.481	0.410	0.469

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

Table 11: 2016 PIT income, synthetic income and HILDA income comparison

	HILDA income	PIT income	Synthetic income
Descriptive statistics			
Sample size	17,694	10,304,098	14,270,757
Mean	50,345	64339	48,950**
Median	36368	53287	35,017
Standard deviation	47,020	51057	47,371
Income percentiles			
1st percentile	0	0	0
5th percentile	(omitted)	4567	575
10th percentile	4095	13702	3,812**
25th percentile	17600	28927	17,294**
50th percentile	37394	53287	35,017
75th percentile	70200	83737	67,497*
90th percentile	109440	126367	106,067**
95th percentile	140000	168997	140,577**
99th percentile	250500	254257**	254,257**
Income statistic			
Gini coefficient	0.477	0.410	0.470

** 95% confidence interval of HILDA
* 99% confidence interval of HILDA

4 Results – Longitudinal validation

Many of the most pressing income-related research questions are focused on dynamics, rather than static distributions. In order to compare the longitudinal performance of the synthetic income distribution, we constructed one-year income movements in the income distribution from 2011 to 2016, by the initial quintile for HILDA income, PIT income, and synthetic income. Results are provided in Table 12, 13 and 14, respectively.

This measure shows for each quintile (20%) of the income distribution, the proportion of individuals moving to a lower quintile, remaining in the same quintile and moving to a higher income quintile over a one-year time frame. As an example of the interpretation of the results, the 2nd row of Table 12 (i.e. Second quintile) shows that, of those in the second quintile in any year between 2011 and 2016, on average, 16% moved to a lower income quintile, 61% remained in the second quintile and 23% moved to a higher income quintile.

When comparing the income dynamics of HILDA and the income dynamics of the synthetic income distribution (Table 12 and Table 14, respectively), it is evident that the synthetic income distribution performs considerable well longitudinally. More precisely, we observe a substantial number of individuals moving between income quintiles similar to HILDA. However, it is also evident that the proportion of individuals remaining in the same quintile is always higher for the synthetic income relative to HILDA. For example, in the synthetic income distribution, over a year, the proportion remaining the bottom quintile is approximately 78%, whereas, in HILDA, the proportion remaining in the bottom quintile is 68%. This greater 'stickiness' relative to HILDA (i.e. proportion remaining in the same income quintile) is evident in all income quintiles. The 'stickiness' is highest in the 1st quintile (10% higher relative to HILDA). Furthermore, unsurprisingly, this higher 'stickiness' results in subdued movement among other income quintiles as well (movements up and down) relative to HILDA.

The higher 'stickiness' observed in the synthetic income distribution relative to HILDA stems from two main sources. Namely, the lack of 'time-variant' variables available in MADIP BLE 2011 and the low level of movement in PIT income. Focusing on the latter, Table 13, shows the one-year movements in the PIT income distribution by initial quintile. We observe that the proportion of individuals that remained in the same quintile in PIT income is significantly higher relative to HILDA. For example, in PIT income, of those in the bottom quintile in any year between 2011 and 2016, on average, 74% remained in the same quintile. This is significantly higher relative to HILDA (68%). Moreover, this higher level of 'stickiness' in PIT is persistent in all income quintiles except for the second quintile. As a consequence, since the synthetic income distribution incorporates income observations from PIT income, we expect this 'stickiness' in PIT income to 'spillover' to synthetic income. Furthermore, it should be noted that given the significantly different income percentiles in PIT relative to synthetic income, the 'spillover' of stickiness from PIT to synthetic income may not be uniform. For example, we can expect the 'stickiness' observed in the bottom quintile of PIT to be spilt over to the 'bottom' and 'second' quintile of the synthetic income distribution.

As highlighted previously, the other source that contributes to the higher level of 'stickiness' of the synthetic income is the lack of 'time-variant' variables available in MADIP BLE 2011. As outlined in Section 2, a strong assumption was placed on the invariability of individual characteristics through time when training our ML algorithms. This assumption was made as a workaround to account for the lack of 'time-variant' features available in MADIP BLE 2011. One implication of this assumption (on longitudinal results) is that unless a substantial change in SSRI features was observed on a yearly basis for an individual, there is a high probability they will remain in the same initial income quintile due to lack of other 'time-variant' features. There are two *possible* workarounds to minimize the 'stickiness' that stems from the lack of 'time-variant' features.

The first workaround is to incorporate more 'time-variant' features into our ML training stage. However, at present, this approach is not possible due to data limitations of MADIP BLE 2011. The second workaround is to utilize the MADIP *Modular Product 2011-2016* data to train the ML algorithms. One distinct difference between the MADIP BLE 2011 and the MADIP Modular Product 2011-2016, is the availability of ordinal categorical SSRI data. More precisely, in MADIP Modular Product, in addition to the dichotomous SSRI variables which indicate whether an individual received a particular benefit in a given year, it also contains the dollar amount each individual received. The amount received is presented as an ordered categorical variable.[26]

We have applied the methodology discussed in this paper to a MADIP Modular product with continuous SSRI information to create a synthetic income distribution. In terms of cross-sectional results, the results were identical to the results presented in this paper. That is, the descriptive statistics and income percentiles of the synthetic income distribution (created using continuous SSRI) were also within the 95% or 99% confidence interval of HILDA. However, in terms of longitudinal results, there was a remarkable improvement. The 'stickiness' observed in each quintile was reduced significantly. For example, the stickiness in the bottom quintile dropped to 72% (from 78% obtained using binary SSRI), the stickiness of the second quintile dropped from 69% from 62%, the stickiness of the third quintile dropped from 63% to 60%, and finally stickiness of fourth and fifth quintile remained unchanged. As a result of this reduction in stickiness, the longitudinal results of the synthetic income converged towards the HILDA results. These results demonstrate the methodology introduced in this paper is consistent, robust, and most importantly, outperforms the income variables that are currently available on MADIP BLE 2011, in terms of capturing the 'true' income distribution.

[26] For each benefit in a given year, the dollar amount received is given in 253 ordinal categories. Each category represents a 100 dollar increment. However, depending on the project, the SSRI variables can be obtained as continuous variables.

Table 12: Income mobility – HILDA income

One-year movements in the income distribution, by initial quintile (%)			
	Moved down	**No change**	**Moved up**
Bottom quintile	0%	68%	31%
Second quintile	16%	61%	23%
Middle quintile	22%	54%	24%
Fourth quintile	25%	62%	12%
Top quintile	19%	81%	0%

Table 13: Income mobility – PIT income

One-year movements in the income distribution, by initial quintile (%)			
	Moved down	**No change**	**Moved up**
Bottom quintile	0%	74%	26%
Second quintile	19%	58%	23%
Middle quintile	22%	60%	18%
Fourth quintile	22%	66%	12%
Top quintile	19%	82%	0%

Table 14: Income mobility – Synthetic income

One-year movements in the income distribution, by initial quintile (%)			
	Moved down	**No change**	**Moved up**
Bottom quintile	0%	78%	22%
Second quintile	12%	69%	18%
Middle quintile	18%	63%	19%
Fourth quintile	21%	67%	12%
Top quintile	17%	83%	0%

5 Concluding comments and future work

The level and distribution of income within a country, and within sub-populations of that country, are one of the key measures of wellbeing for a society. Furthermore, income dynamics (the extent to which a person's income in a given year predicts their income in a subsequent year), as well as the predictors of income dynamics are both vital for understanding the extent to which very high incomes or very low incomes are entrenched in a society. A person or household can more easily manage one year of low income by drawing down on savings or making use of government support. Many years of low income can lead to far more negative outcomes than just one year of low income. Similarly, a person with one year at the very top of the income distribution can save some of that income for future uncertainty and is likely to have higher subjective wellbeing for that particular point in time. But it is less likely to give them long-lasting political or social power.

Australian researchers are relatively well supplied with survey data that captures static and dynamic income distributions for the population as a whole, with the Australian government investing heavily in the (longitudinal) Household, Income and Labour Dynamics in Australia (HILDA) survey, as well as the (cross-sectional) Survey of Income and Housing (SIH). However, this survey data is less useful for small population groups, or for very precise points on the income distribution (the top 1% or 2%, or those just above or below the poverty line). Increasingly, administrative data is being used to supplement this survey data, but they have their own limitations. By definition, tax data only has information on those in the tax system, and social security data only has information on those who are in the social security system. Census data, while solving the low sample problem and representativeness biases, only has income in ranges and for one particular point in time.

A relatively new dataset, the Multi-Agency Data Integration Partnership (MADIP) Basic Longitudinal Extra (BLE) 2011, provides an alternative source of information by linking Census data, tax data, and social security data at the individual level, with some income information from all three constituent datasets, but with a far larger sample size than any of the surveys discussed above. The aim of this paper was to propose a methodology for constructing a synthetic measure of total income for five waves of data (linked longitudinally) that draws power from each of the constituent datasets, but overcomes the limits of each. This income measure was then validated cross-sectionally and longitudinally against a high-quality sample survey (HILDA).

After testing nine machine-learning approaches using a training and test dataset from the BLE 2011, we were able to generate a synthetic income measure that performed far better than either tax or census data alone in matching the HILDA income distribution, and was also able to capture income dynamics reasonably well, albeit with some understating of income dynamics. The big difference though is that we have synthetic income data for over 15 million individuals, compared to only around 17,000 for HILDA.

There are limitations to our income measure though, some of which may be overcome by utilising more information from the MADIP data environment. First, we are reliant on grouped Census and tax data to train and estimate our models. While there is no prospect of continuous Census data being available, there is continuous tax data that may improve the accuracy of predictions. Secondly, we have limited time-varying characteristics on our dataset (the potential reason for our 'stickier' data) and are reliant on social security data to capture income dynamics. Any measures of education, occupation, or industry that vary through time from outside of the tax system may improve our model. Thirdly, we have used 2011 data to project income dynamics data forward, but there may be research questions that are better suited to project 2016 data backwards. We are in the process of replicating our methods and estimates using the 2016 BLE. Fourth, we derive a synthetic measure of total income, rather than taxable or disposable income. In many circumstances, it is the latter two income types that are of greater interest for researchers and policy makers. The methodology utilised in this paper, with some adjustments, could be tested against these types of income in the future.

A final limitation is that our income estimates are at the individual level. Given a significant amount of intra-household sharing, income analysis often utilises household or family-level income, suitably equivalised to take into account household sharing. We can use our synthetic income in 2011 to create a household equivalised measure using the household identifier on the Census, but this is not available for non-Census years. The BLE 2011 would benefit from a more dynamic household identifier, if it were possible to construct using some of the non-Census datasets.

Despite these limitations, some of which are in the process of being overcome and others requiring additional data linkage/construction, the synthetic income measure developed in this project provides a robust measure of income levels and dynamics for a very large sample of Australian adults, and if analysed carefully can help support our understanding of access to economic resources in Australia and how it varies through time, across small population groups, and within individuals.

6 References

Adler, M. and Schmid, K., 2012. Factor Shares and Income Inequality — Empirical Evidence from Germany 2002-2008. *No 82, IAW Discussion Papers, Institut für Angewandte Wirtschaftsforschung (IAW)*

Australian Bureau of Statistics*, Multi-Agency Data Integration Project Basic Longitudinal Extract,2011-2016 (2011 Cohort)*, Confidentialised Unit Record File (CURF), DataLab. Findings based on the use of ABS Microdata.

Athanasopoulos, G. and Vahid, F., 2003. Statistical Inference and Changes in Income Inequality in Australia. *Economic Record*, 79(247), pp.412-424.

Atkinson, A. and Leigh, A., 2007. The Distribution of Top Incomes in Australia. *Economic Record*, 83(262), pp.247-261.

Auten, G. and Splinter, D., 2019. Top 1 Percent Income Shares: Comparing Estimates Using Tax Data. *AEA Papers and Proceedings*, 109, pp.307-311.

Bhardwaj, R., 2018. Explaining Income Inequalities: A Cross Sectional-Study of OECD Countries. *SSRN Electronic Journal*.

Bhat, C., 1994. Imputing a continuous income variable from grouped and missing income observations. *Economics Letters*, 46(4), pp.311-319.

Biddle, N. and Montaigne, M., 2017. Income Inequality in Australia - Decomposing by City and Suburb. *Economic Papers: A journal of applied economics and policy*, 36(4), pp.367-379.

Biddle, N., Breunig, R., Markham, F. and Wokker, C., 2019. Introducing the Longitudinal Multi-Agency Data Integration Project and Its Role in Understanding Income Dynamics in Australia. *Australian Economic Review*, 52(4), pp.476-495.

Bray, R., 2014. Changes in Inequality in Australia and the Redistributional Impacts of Taxes and Government Benefits. Measuring and Promoting Wellbeing: How Important is Economic Growth?

Cameron, A. and Trivedi, Pravin, (2005), Microeconometrics, Cambridge University Press.

Campbell, P., Biddle, N. and Paradies, Y., 2018. Indigenous identification and transitions in Australia: exploring new findings from a linked micro-dataset. Population, 73(4), pp.771-796.

Greene, William & Hensher, David. (2009). Modeling Ordered Choices: A Primer. 10.1017/CBO9780511845062.

Huber, P., 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), pp.73-101.

Hunter, B. and Gregory, R., 1996. An Exploration Of The Relationship Between Changing Inequality Of Individual, Household And Regional Inequality In Australian Cities. *Urban Policy and Research*, 14(3), pp.171-182.

Jauch, S. and Watzka, S., 2015. Financial development and income inequality: a panel data approach. *Empirical Economics*, 51(1), pp.291-314.

Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), pp.2825–2830.

Peñaloza, Rodrigo. (2016). Gini coefficient for ordinal categorical data. 10.13140/RG.2.2.34095.94886.

Productivity Commission 2018, *Rising inequality? A stocktake of the evidence*, Commission Research Paper, Canberra.

Stewart, M., 1983. On Least Squares Estimation when the Dependent Variable is Grouped. *The Review of Economic Studies*, 50(4), p.737.

Summerfield, M., Bright, S., Hahn, M., La, N., Macalalad, N., Watson, N., Wilkins, R. and Wooden, M. (2019). *HILDA User Manual – Release 18*. Melbourne Institute: Applied Economic and Social Research, University of Melbourne.

Tran, C, Zakariyya, N, March 2019, Tax progressivity in Australia: Facts, measurements and estimates. *Tax and Transfer Policy Institute – Working paper 5/2019*

United Nations Economic Commission for Europe (UNECE) 2011 Canberra Group Handbook on Household Income Statistics

Wilkins, R., 2014. Evaluating the Evidence on Income Inequality in Australia in the 2000s. *Economic Record*, 90(288), pp.63-89.

Urban Sila & Valéry Dugain, 2019. "Income, wealth and earnings inequality in Australia: Evidence from the HILDA survey," OECD Economics Department Working Papers 1538, OECD Publishing

Wilkins, R., 2015. Measuring Income Inequality in Australia. *Australian Economic Review*, 48(1), pp.93-102.

Wilkins, Roger and Inga Lass (2018) *The Household, Income and Labour Dynamics in Australia Survey: Selected Findings from Waves 1 to 16*. Melbourne Institute: Applied Economic & Social Research, University of Melbourne

Wooden, Mark & Watson, Nicole. (2004). Sample attrition in the HILDA survey. *Australian Journal of Labour Economics (AJLE).* 7. 293-308.

7 Appendix

7.1 Interval regression results

Number of observations = 9,895,864

Uncensored = 0
Left-censored = 49,343 LR chi2(35) = 3086484.70
Right censored = 640,919 Prob > chi2 = 0.00
Interval-censored
9,205,602 Log likelihood = -22671343

	Coefficient	P>|z|
Female	-1.11	0.000
Indigenous	-0.27	0.000
Disability	-0.52	0.000
Postgrad	1.76	0.000
Certificate	-0.025	0.000
HSC	-1.08	0.000
NoEducation	-1.1	0.000
SSRI_1	-4.97	0.000
SSRI_2	-1.54	0.000
SSRI_3	-2.4	0.000
SSRI_4	-1.94	0.000
SSRI_5	-1.79	0.000
SSRI_6	-0.18	0.000
SSRI_7	-2.37	0.000
SSRI_8	-0.0007	0.112
SSRI_9	-0.92	0.000
SSRI_10	-1.76	0.000
SSRI_11	-2.64	0.000
SSRI_12	0.23	0.000
SSRI_13	-1.41	0.000
SSRI_14	0.092	0.000
SSRI_15	-2.66	0.000
SSRI_16	0.459	0.000
SSRI_17	0.204	0.000
SSRI_18	0.234	0.000
SSRI_19	-1.19	0.000
SSRI_20	-1.58	0.000
SSRI_21	-3.55	0.000
SSRI_22	-2.47	0.000
SSRI_23	-0.21	0.000

SSRI_24	-1.663	0.000
SSRI_25	-1.44	0.000
SSRI_26	-3.28	0.000
SSRI_27	-2.19	0.000
SSRI_28	-0.874	0.000
Constant	7.84	0.000

7.2 Model fit comparison using log likelihood

In this section, we conducted an informal test to assess whether the fit of an Interval regression model was satisfactory relative to a standard Ordered Probit model. Ordered Probit models are highly robust models for ordered categorical variables. Furthermore, the Ordered Probit models are not based on the assumption of normality. Therefore, by comparing the log-likelihood (fit) of these of the Interval Regression model to the log-likelihood of an Ordered Probit model, we can assess the suitability of the Interval regression model. A similar or "close" log-likelihood value between the models indicate that there were no violations of the normality assumption. In the case of the 2011 Census income, the log-likelihood of the Interval Regression model showed a slight deviation from the log-likelihood value from the Ordered Probit model. This indicated the normality assumption was not violated and interval regression was satisfactory. However, for the 2011 PIT income, large deviations were noted. This indicated that the interval regression model was not satisfactory without a log-transformation.

Model fit comparison – 2011 Census income

- Interval regression – log-likelihood = -22671343
- Ordered probit – log-likelihood = -22463640

Model fit comparison – 2011 PIT Income

- Interval regression – log-likelihood = -36482325
- Ordered probit – log-likelihood = -34087697

In addition, this test was also conducted to see the improvements applying log transformation to 2011 Census income and 2011 PIT income.

Model fit comparison – 2011 Census income vs log 2011 Census income

- Interval regression – log-likelihood = -22671343 (2011 Census income)
- Interval regression – log-likelihood = -23003625 (log 2011 Census income)

A deterioration in model fit was noted when using log Census income relative to Census income.

Model fit comparison – 2011 PIT income vs log 2011 PIT income

- Interval regression – log likelihood = -36482325 (2011 PIT income)
- Interval regression – log likelihood = -35009711 (log 2011 PIT income)

An improvement in the model was evident when using the log transformation.

7.3 Probability plot of 2011 Census and log Census income distribution

Figure 8: Probability plot of log 2011 Census

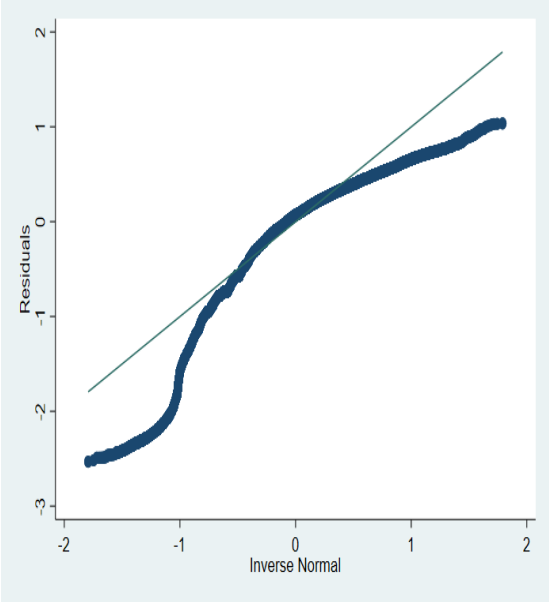
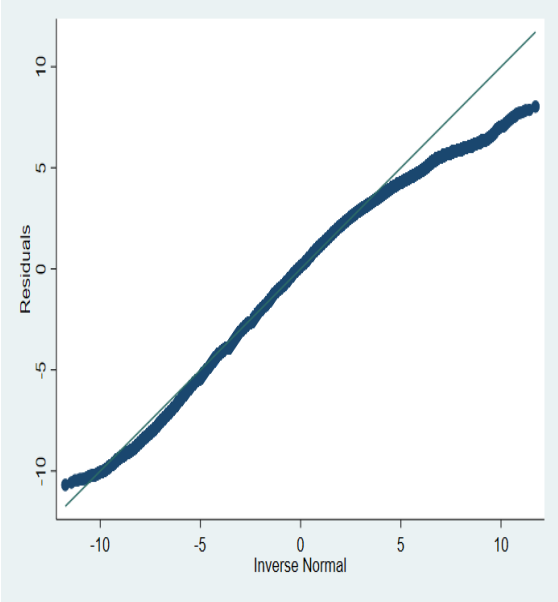


Figure 7: Probability plot of 2011 Census income



7.4 Probability plot of 2011 PIT and log PIT income distribution

Figure 10: Probability plot of log 2011 PIT income

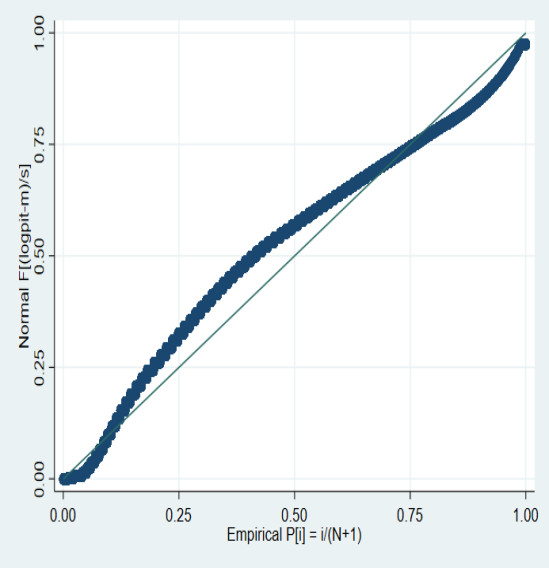
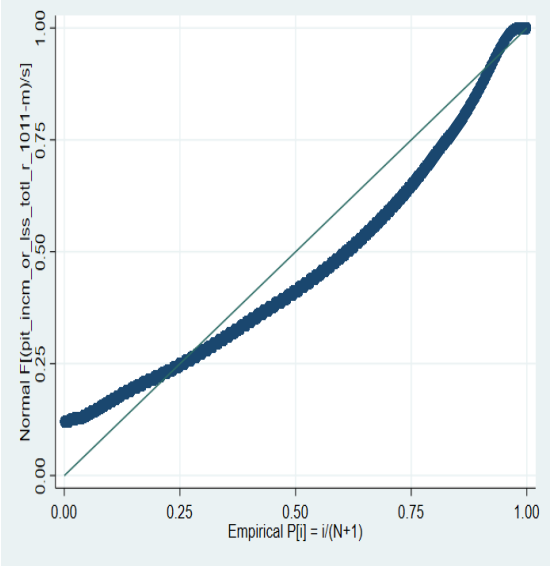


Figure 9: Probability plot of 2011 PIT income



7.5 Boxplots of features

Figure 11: Census income by disability status

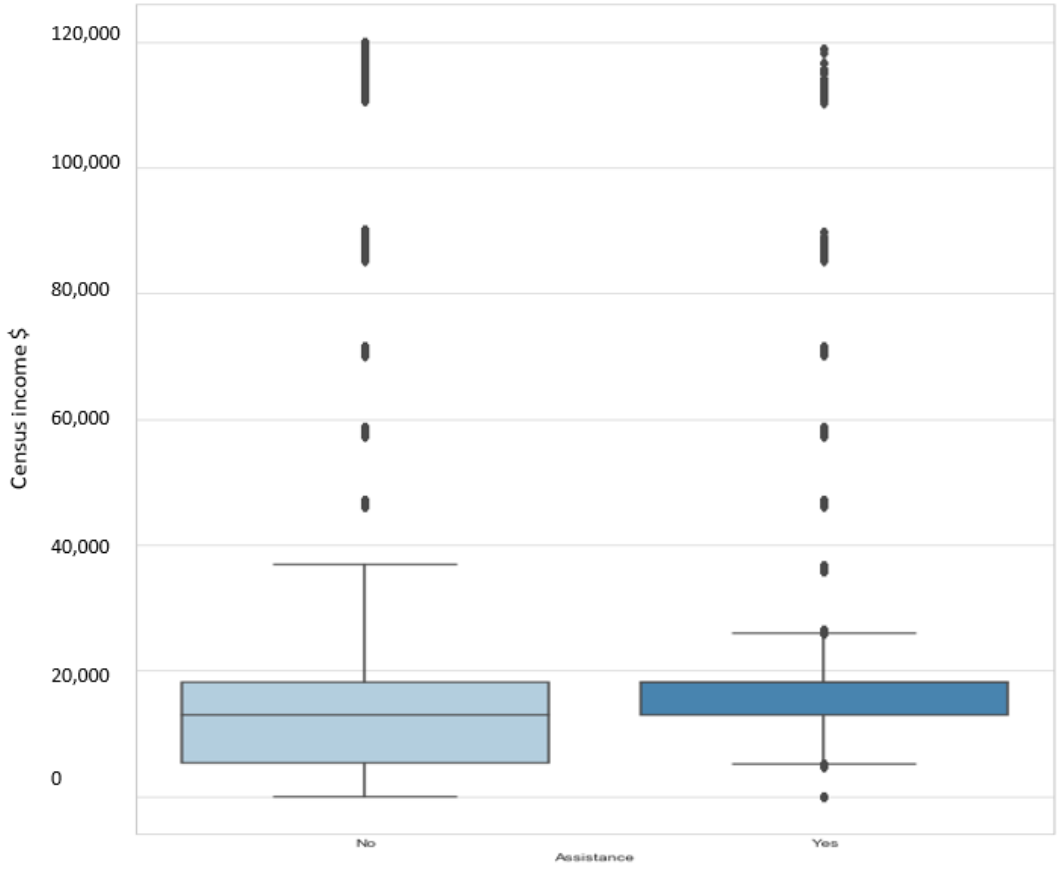


Figure 12: Census income by Age

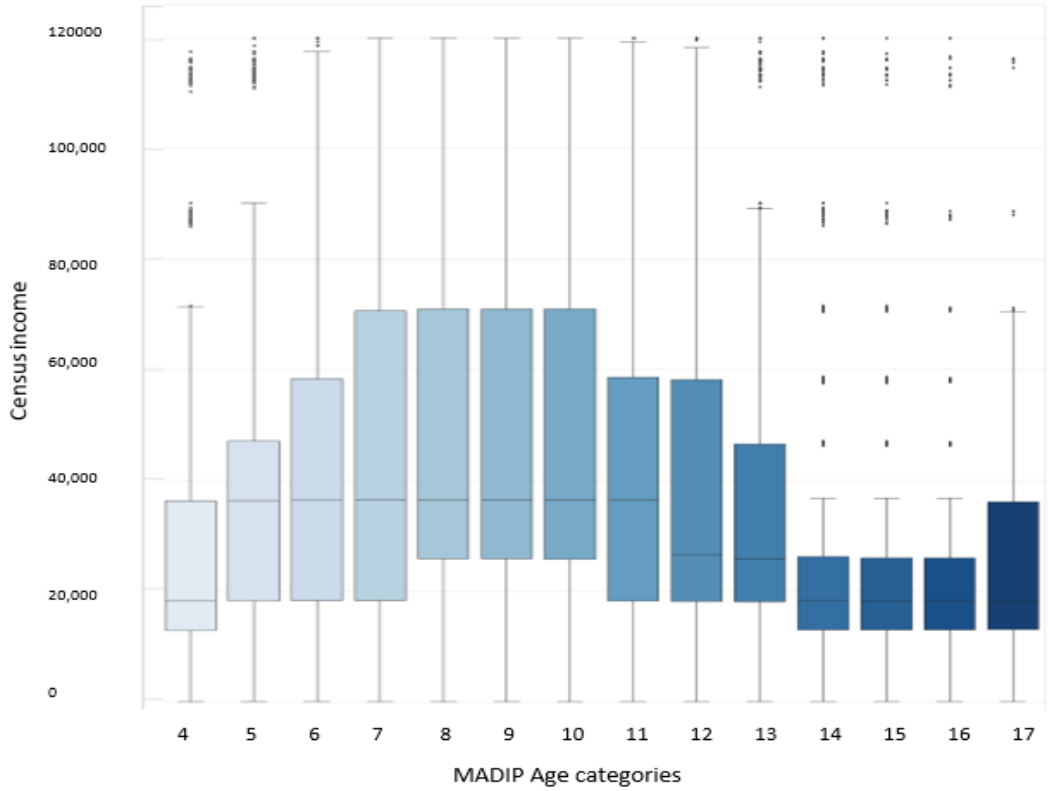


Figure 13: Census income by Education

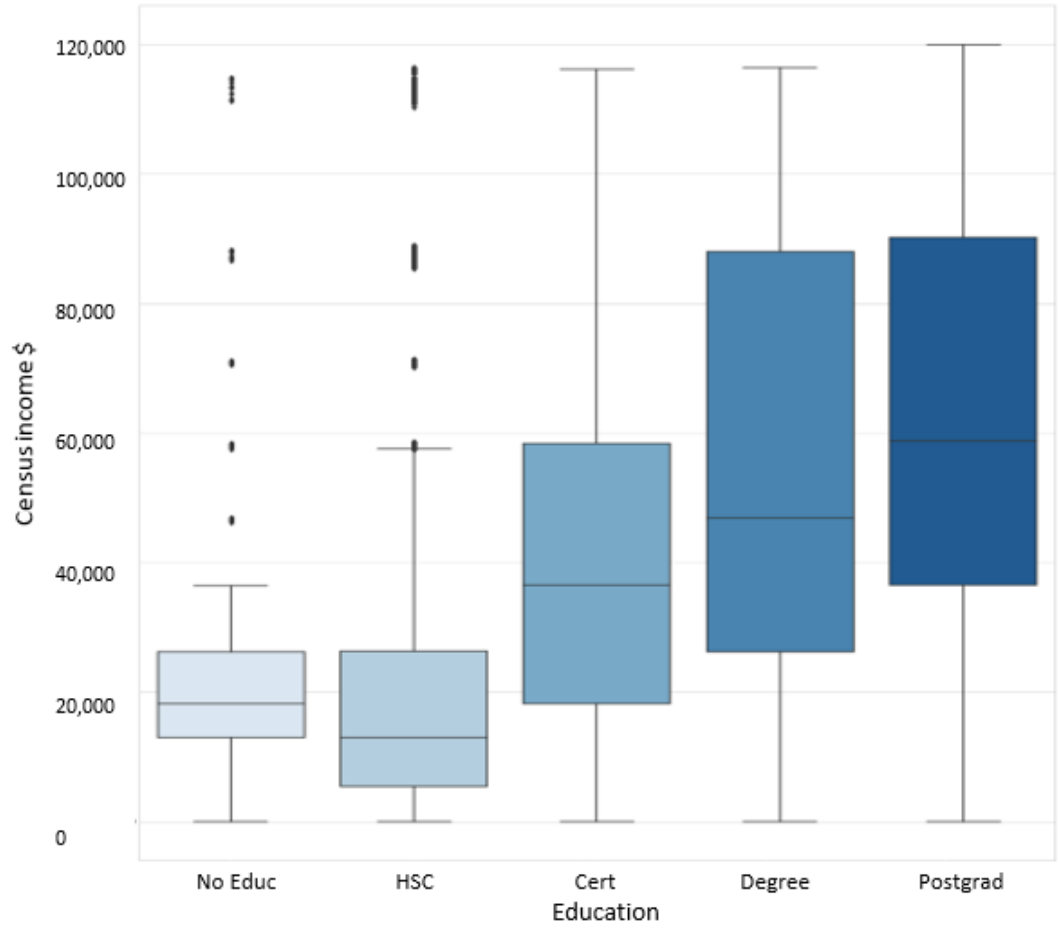


Figure 14: Census income by Occupation

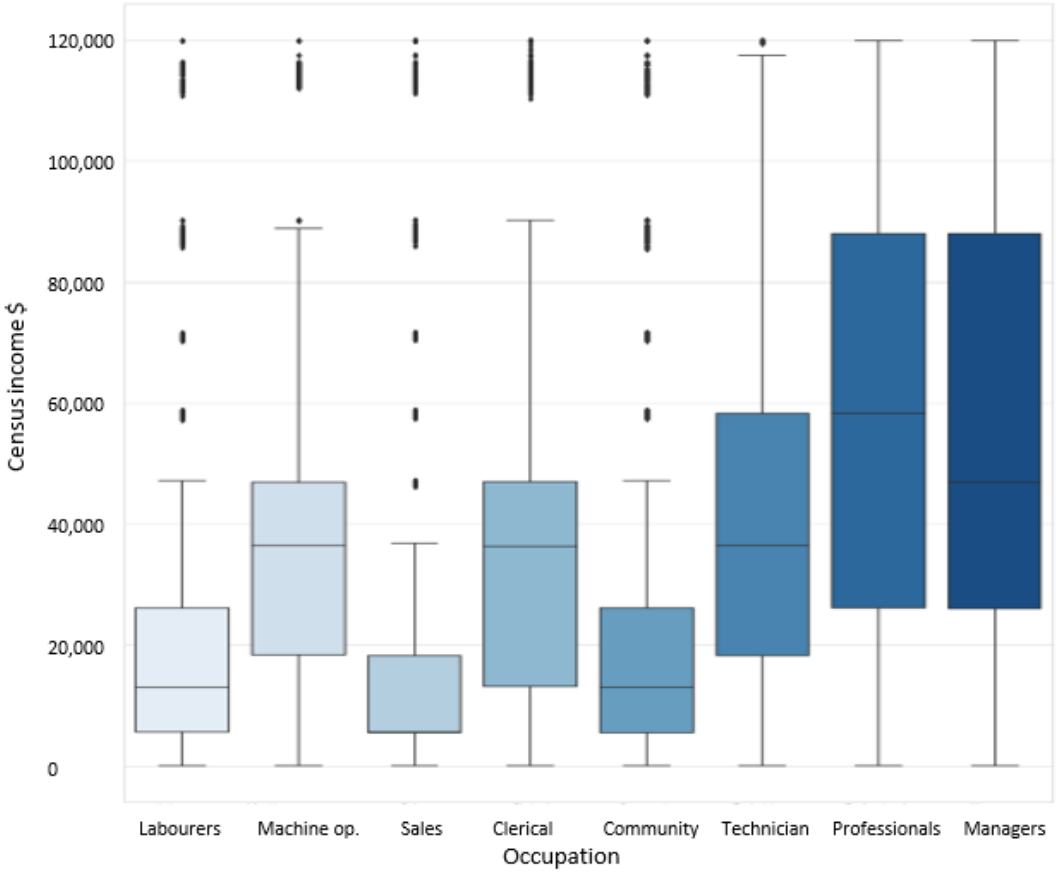
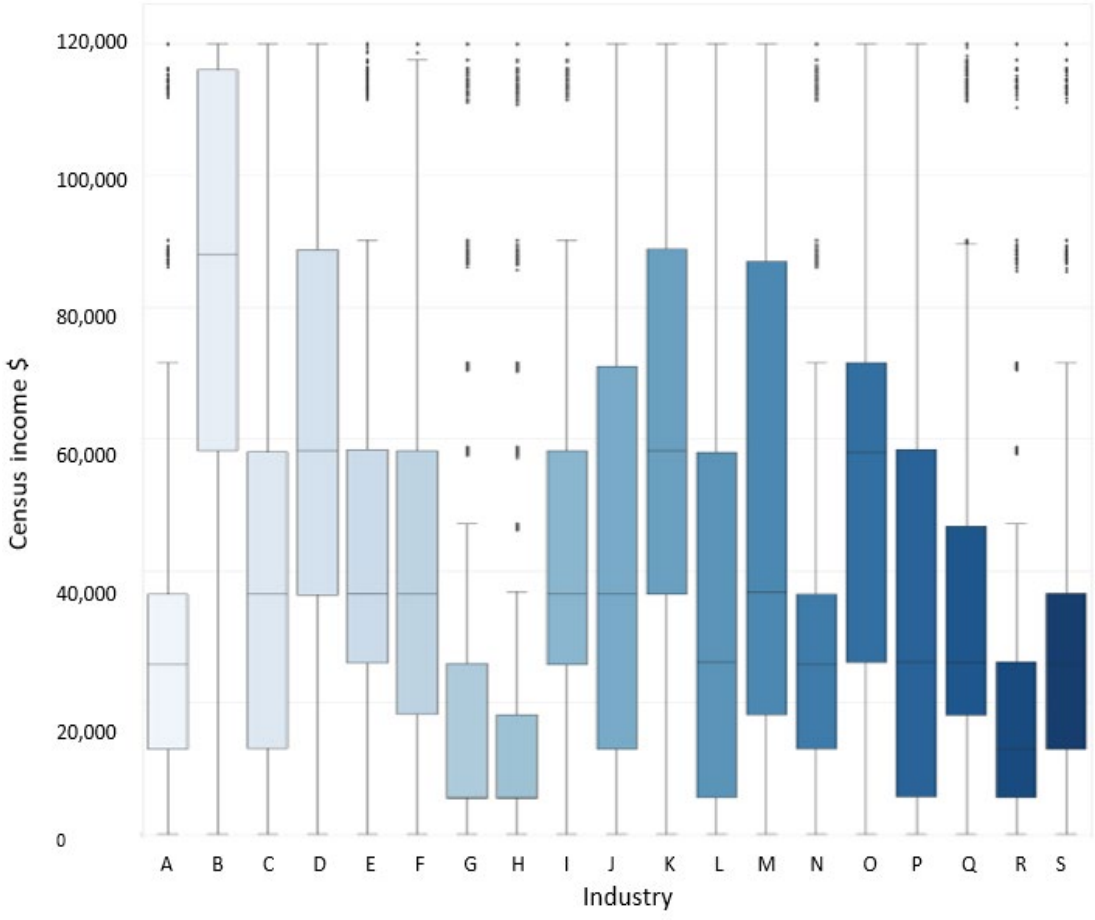
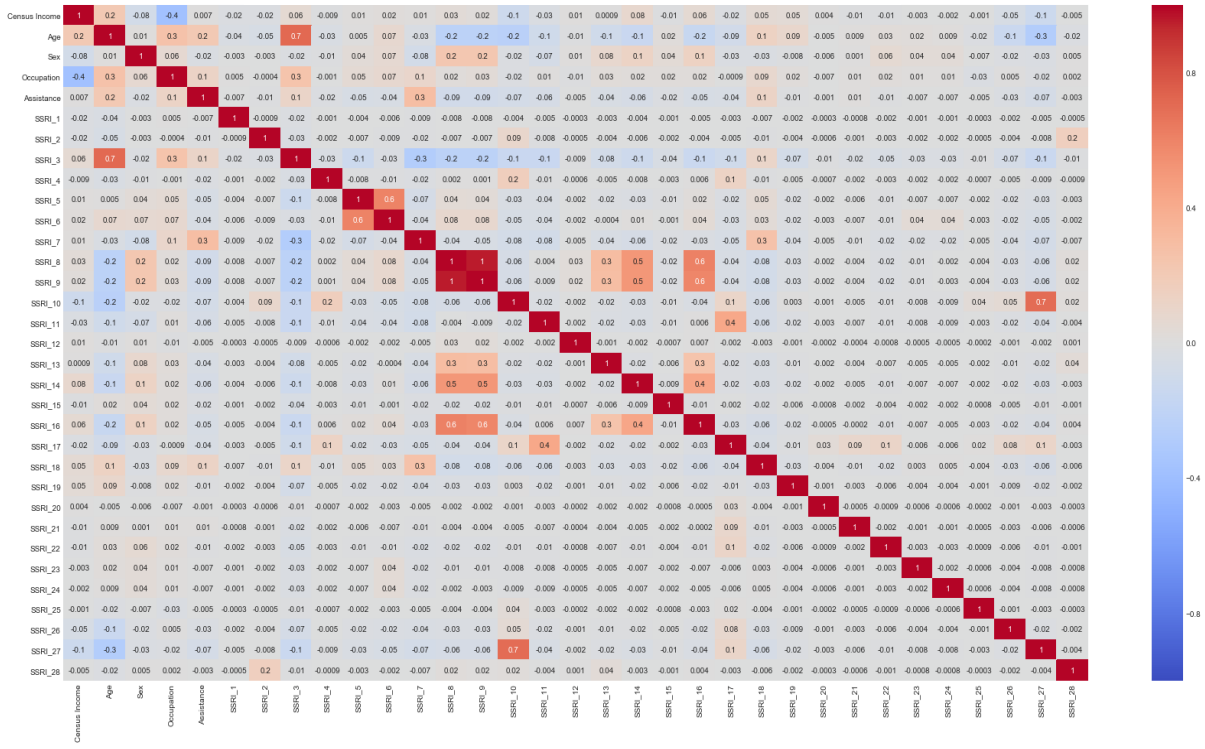


Figure 15: Census income by Industry



7.6 Correlation heatmap

Figure 16: Correlation heatmap



7.7 ML algorithm results obtained using 'first' technique.

Model	MAE	MSE	RMSE	R^2	Time (m)[27]
Linear Regression	163	65,728	256	0.34	0.15
Ridge Regression	169	69,243	263	0.31	0.10
Bayesian Regression	162	65,521	255	0.34	0.25
Decision Tree Regression	136	60,130	245	0.40	1.1
Random Forest Regression*	129	51,831	227	0.52	83
Extra Trees Regression	133	54,158	232	0.47	99
Gradient Boosted regression*	128	52,482	228	0.55	133
MLP Regression	126	49,273	221	0.58	310

[27] Time taken to run one iteration of the default model.

7.8 MLP model hyper-parameter tuning

Hidden layer sizes = [(50,), (100,), (200,), (50,50), (100,100), (200,200), (50,50,50), (100,100,100), (50,100,50)]

Activation = {'logistic', 'relu'}

Solver = {'adam', 'sgd'}

Alpha = [0.0001, 0.05, 0.1]

Learning rate = ['constant', 'adaptive']

7.9 Descriptive statistics of predicted income values

Table 7: Descriptive statistics of test and predicted values

	Test set	**Prediction**
Mean	15,860	15,120
Median	13,005	13,018
Percentiles		
1%	0	0
25%	5349	5499
50%	13,005	13,018
75%	18,206	18,174
90%	26,208	25,864
95%	46,603	42,479
99%	88,764	80,372

7.10 Lorenz curve of 2011 Synthetic income and HILDA distribution

Figure 17: Lorenz Curve for 2011 Synthetic Income

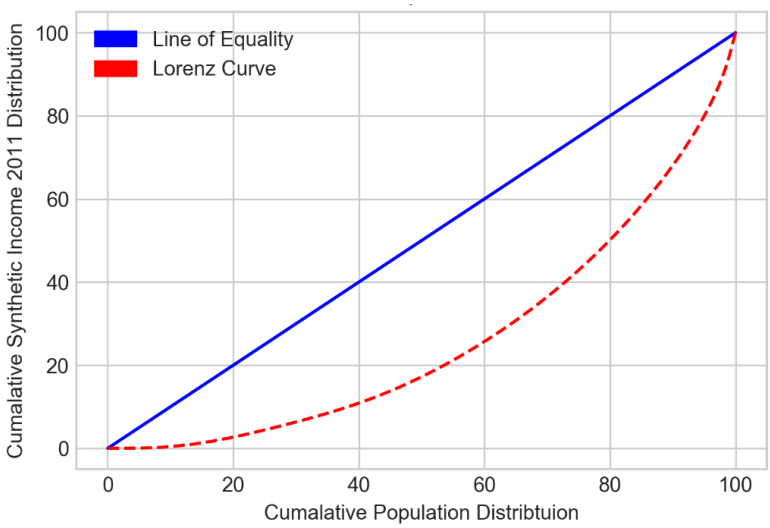


Figure 18: Lorenz Curve for 2011 HILDA Income