IARIW 2021

IARIW 2021

Monday 23 – Friday 27 August

Statistisk sentralbyrå
Statistics Norway

Measuring Inequality of Opportunity Using Sampling-based Shapley Decomposition: Application to Australian Data 2001-2017

Dongjie Wu

(University of Copenhagen)

Prasada Rao

(University of Queensland)

KK Tang

(University of Queensland)

Nicholas Rohde

(Griffith University)

Pravin Trivedi

(University of Queensland)

Paper prepared for the 36th IARIW Virtual General Conference
August 23-27, 2021
Session 17: Inequalities in Opportunity for Income, Education and Health II
Time: Thursday, August 26, 2021 [14:00-16:00 CEST]

Measuring inequality of opportunity using sampling-based Shapley decomposition: Application to Australian data 2001-2017 *

Dongjie Wu, Prasada Rao, Pravin Trivedi, Kam Ki Tang and Nicholas Rohde

In this paper, we propose an alternative approach to measure inequality of opportunity (IOP) using data from Australia between 2001 and 2017. We propose a sampling-based Shapley decomposition approach to measure the overall contribution of all observed circumstances to income inequality, and the contribution of each circumstance variable. This approach allows us to reduce the bias due to the sparse divisions associated with a large set of circumstances, to massively improve computational efficiency, and to provide both lower and upper bound estimates of IOP. Using this method, we find that inequality of opportunity in Australia ranges from around 5% to 20% on average during 2001 and 2017.

1 Introduction

Socio-economic inequality has been drawing public attention in recent years. The rising top one percent income share (Alvaredo et al., 2013) and the widening inequality during the COVID-19 pandemic (Patel et al., 2020) raise the question of the fairness of recent rising inequality. Inequalities due to differences in work ethics may be acceptable as they reflect the heterogeneity in work/leisure trade-off preferences among the population. Conversely, inequalities due to discrimination, or exogenously imposed barriers to success are much more likely to be unjust and socially harmful (Niehues and Peichl, 2011), as well as detrimental for economic growth (Marrero and Rodríguez, 2013). As such, policymakers are often advised to target these types of disparities (Ferreira, 2019).

Unravelling fair and unfair inequalities is conceptually challenging. However, the rapidly developing Inequality of Opportunity (IOP) literature offers a coherent framework to achieve that.

*Corresponding author: Dongjie Wu, Department of Sociology, University of Copenhagen, Copenhagen, Denmark. Email: dongjie.wu@soc.ku.dk

Developed principally by Roemer (1998), Arneson (1989), Cohen (1989) and Sen (1990), this literature seeks to decompose variations of outcomes like income into contributions from *circumstances* (factors that lie beyond individual control such as race, gender and family background) and *efforts* (variables that individuals are judged to be responsible for, such as their work ethic). The applied work in this area usually focus on the fraction of total inequality attributable to circumstances (see Ferreira and Gignoux, 2011 and Checchi and Peragine, 2010).

However, there are two main empirical challenges on measuring IOP. Firstly, with current empirical work methods, the IOP estimate is biased if the set of circumstances is large, and more so if circumstance variables are continuous variables. A general practice of measuring IOP is to divide samples into a limited number of types – each type representing a group of individuals sharing the same circumstances. With finite observations, an excessively fine division will result in few samples for each type and a biased estimate. Secondly, the current practice can only identify the effect of all variables on income variation but not the individual contributions of the variables and their relative importance. For instances, it has emerged that paternal education and socioeconomic status are especially important in transmitting inequalities over generations (see Corak, 2013 and Andersen, 2019). Knowing the precise contributions of each of the variables in concern to IOP may offer some clues about how best to combat inherited economic disadvantages (Tomes, 1981).

In this paper, we develop some computational techniques for measuring IOP, and apply them to Australian panel data from 2001-2017. Our method is based on the Shapley decomposition (see Shapley, 1953, Sastre and Trannoy, 2002 and Shorrocks, 2013). The method originated from the co-operative game theory, where a coalition of players are rewarded for their contributions towards a shared goal. By averaging over the marginal contributions of each factor, the Shapley decomposition exhaustively attributes the total inequality to the covariates. When applied to inequality analysis, the method can decompose an IOP estimate into positive, proportional contributions from individual covariates. In addition, the Shapley decomposition alleviates the bias due to a large set of circumstances and provides an upper bound for IOP that is smaller than the estimate from the non-parametric model adopted by many research (Checchi and Peragine, 2010).

The Shapley decomposition, however, is computationally burdensome to implement. Since contributions from a regression with k covariates need to be averaged over all possible permutations of the model, the total number of regressions to be estimate is a factorial function of

k. This means that timely computation is impossible even for modern computers if more than dozens of covariates are employed. However, by randomly sampling from the set of all model specifications, we can produce approximate Shapley decompositions that are fast yet accurate enough to be practically viable.

When applying Shapley decomposition to measure IOP, most literature uses a paramatric model (Björklund et al., 2012). In this paper, we implement the Shapley decomposition method using a parametric model and a non-parametric model. The Shapley implementation on a non-parametric model reduces the overestimation due to limited data and a large set of circumstances. Therefore, when applying the new methods to Australia data, in addition to provide a lower bound estimate of IOP like the study from Martinez et al. (2017), we provide an interval estimate showing that IOP in Australia is at least around 5% and at most 20% during 2001 to 2017.

The rest of the paper is structured as follows. Section 2 outlines some fundamental concepts in IOP measurement and explains our sampling-based decomposition method. Section 3 presents the simulation results using the methods introduced in section 2. Section 4 documents the application of the Shapley decomposition on the Australian data. Section 5 provides the conclusion.

2 Methodology

2.1 The conventional framework in measuring inequality of opportunity

To measure IOP, we adopt Roemer's (1998) framework, wherein income is determined by a set of circumstances and efforts. Assume that circumstances and effort are independent with each other (Checchi and Peragine, 2010) and circumstances are a set of discrete variables. With a sufficiently granular set of circumstances, the population can be divided into different "types", where all individuals within the same type share identical circumstances. Similarly, the population can also be stratified based on efforts into different "tranches", where all individuals within the same tranche exert the same level of effort. IOP can, therefore, be the inequality that occurs between types (as this captures inequalities identified by observed circumstances) or within tranches (as this captures variations attributable to unobserved circumstances). As circumstances are more readily measurable than efforts, most researchers prefer to use types to calculate IOP, which is known as the *ex ante* meausre of IOP[1].

Given Roemer's framework, IOP can be computed using a non-parametric approach (Checchi and Peragine, 2010). We denoted the whole set of "types" as $T = \{t_1, t_2, ..., t_K\}$. If there are N_k samples in t_k for $k \in \{1, ..., K\}$, we can replace each individual income $y_i^{t_k}$ into its type-average income $\bar{y}^{t_k}$. Given this counterfactual income $\bar{Y} = \{N_1\bar{y}^{t_1}, ..., N_K\bar{y}_{t_K}\}$, IOP can be computed as:

$$IOP_{nonpara} = \frac{I(\bar{Y})}{I(Y)} \tag{1}$$

where the numerator captures the inequality in opportunity, and the denominator the overall inequality. $IOP_{nonpara}$ ranges from zero to one; zero indicates that no income inequality is due to the contribution of circumstances, whereas one indicates that all income inequality is due to the contribution of circumstances. $I(\cdot)$ is an inequality measure that satisfies certain axioms such as scale independence and transfer sensitivity. Like most IOP research, this paper uses a path-independent inequality measure – mean log deviation (MLD).

An alternative approach is to use a parametric model. The most common one is the linear regression model:

$$y_i = C_i'\beta + \epsilon_i \tag{2}$$

where C_i is a vector of circumstances for individual i, y_i their income, β a parameter vector to be estimated, and the error term ϵ_i that captures unobservable factors including efforts.

The fitted value $\hat{y}_i = C_i'\hat{\beta}$ is then interpreted as the baseline income that an individual with the set of circumstances C_i' could be expected to achieve. Variations in $\hat{y}_i$ across $i = 1, ..., n$ therefore reflect differences in initial conditions. IOP can then be measured as:

$$IOP_{para} = \frac{I(\hat{y})}{I(y)} \tag{3}$$

Since we use a relatively large set of circumstances, results could be biased differently between the parametric and the non-parametric approach. Non-parametric estimates, while being flexible and intuitive, suffers from two problems. Firstly, since circumstance variables intersect, the set of cells required to exhaustively define types becomes extremely large very quickly, making non-parametric estimates vulnerable to the curse-of-dimensionality. As a result, they require either a very large dataset or some form of restrictions to be computationally feasible.

[1] The IOP calculated based on both circusmtances and effort are known as the *ex post* measure of IOP

Secondly, sparse cells offer little scope for capturing variation. For instance, if a cell contains a single income, the contributions of circumstances and efforts cannot be disentangled. On the other hand, a non-parametric model takes non-linearity into accounts. The model does not over-smooth the data and results in an underestimated IOP. In comparison, a parametric model such as a linear regression model, avoids the curse-of-dimensionality as long as the number of circumstances variable is significantly smaller than the number of observations. However, it could under-fit the data due to the assumption of the linearity.

2.2 Applying the Shapley decomposition to measure inequality of opportunity

To isolate the contributions from individual covariates and mitigate the bias given a large set of circumstances, we use the Shapley decomposition (or value) approach. The method was originally proposed to solve allocation problems in the cooperative game theory (Shapley, 1953). It can also be used to measure IOP. Suppose income y is exclusively determined by factors $K = \{E, C\}$ including effort E and circumstances C. To compute the Shapley value associated with factor $k \in K$, denoted SV_k, we first generate all unique permutations given the set of all circumstances and efforts $K = \{E, C\}$. We define this permutation set as Ω. In this case, Ω contains $|K|!$ number of elements. For each order r in Ω, we select the subset S_r preceding k and compute the inequality $I(S_r)$ given the subset S. We also compute the inequality $I(S_r+k)$ given the subset $\{S_r, k\}$. Therefore, for each order, the marginal contribution of k to I is $I(S_r + k)$ - $I(S_r)$.

The Shapley value is the average of such marginal contributions given all the orders of permutations of set Ω(Shorrocks, 2013):

$$SV_k = \frac{1}{|K|!} \sum_{r \in R} (I(S_r + k)) - I(S_r)) \tag{4}$$

Based on the Shapley value for each circumstance variable, one can compute the measure of IOP by summing up the Shapley values of all circumstance variables:

$$IOP_{shap} = \sum_{c_k \in C} SV_k \tag{5}$$

One issue for this computation of the Shapley value is how to compute the counterfactual income distribution given a subset S_r. To solve this problem, we use both a paramatric model and a non-paramatric model. The paramatric model with Shapley decomposition has been used in other studies (Björklund et al., 2012) for measuring IOP, while this paper proposes a non-paramatric model with the Shapley decomposition.

For the parametric model, we use a linear regression model described in equation (2) with circumstance variables and efforts as residuals. Since $\hat{y}_i = C_i'\hat{\beta}$ is simply a linear sum of circumstances weighted by a parameter vector, we can treat circumstances and efforts as income sources. Sastre and Trannoy (2002) propose two main methods to decompose an income distribution into income sources – a zero income distribution and equalized income distribution. The zero income distribution drops income sources excluded from subset S_r, while equalized income distribution only removes the variation of income sources excluded in subset S_r by replacing each income source with the average value. In this paper, we use the equalized income distribution with the linear regression model.

A non-parametric model can also be used for the Shapley decomposition. To compute the Shapley value for one factor k, instead of using the whole set of circumstances, we only use subsets S_r and $S_r + k$ to divide data for each order r into types.

In the conventional approach of the non-parametric model, the measure of IOP could be biased given limited observations and a large set of types. However, applying the Shapley decomposition to a non-parametric model provides a framework to handle this problem. This is because in the computation of the Shapley value, only a subset of circumstances is used to generate types in most of the combinations. For example, one combination could be that all circumstances affect income, so the sample is divided into the maximum number of types. In this case, the result could be the most biased and overestimated like the conventional non-parametric approach does, as more types could end up with fewer observations. Another combination could be that all circumstances have no effect on income, so the sample is divided into the minimum number of types, resulting in an underestimated value. However, most combinations locates in between. In the computation of SV_k, all these extreme combinations are averaged with other combinations with a subset of circumstances. Because of the averaging, the overestimation given limited observations and a large number of types is mitigated.

2.3 Sampling-based Shapley decomposition

One issue of the Shapley decomposition approach is that the calculation of one SV_k requires at least $2^{|K|}$ computations[2]. The computational burden increases exponentially as the number

[2] Another way to compute the Shapley value is to find all the subset in K without generating the set of all permutations. In this case, there are $2^{|K|}$ steps to compute one Shapley value. The approach we use with

of factors $|K|$ increases. To circumvent this computational complexity, Björklund et al. (2012), for instance, group selected circumstance variables and decompose income inequality according to the contributions of no more than 10 groups. Obviously, this solution scarifies the details on contributions of individual variables for higher computation efficiency.

We can avoid such a trade-off by applying a sampling-based approach to the Shapley decompositions (Castro et al., 2009).[3] In this approach, we randomly select m samples from the permutation set Ω and then compute the marginal contribution of $k \in K$ for each selected sample. The mean of the sampled marginal contributions is the sampled Shapley value for k (denoted as $\hat{SV}_k$), as the mean of all marginal contributions is the true Shapley value:

$$\hat{SV}_k = \frac{1}{m}\sum_{i=0}^{m} I(S_r^i + \{k\}) - I(S_r^i) \tag{6}$$

where $r^i \in \Omega$ is the order being selected.

Given the sample mean, we can also compute the sample variance:

$$Var(\hat{SV}_k) = \frac{\sum(I(S_r^i + \{k\}) - I(S_r^i) - (\hat{SV}_k))^2}{m-1} \tag{7}$$

The error gets smaller as m gets larger.

To select a proper sample size, we use eight factors to experiment with the computation of the sampled $\hat{SV}_k$. The result is shown in figure 1. The x-axis shows the sample size and the y-axis shows the average marginal contribution (i.e. the sampled $\hat{SV}_k$). We randomly draw samples given a sample size and compute the corresponding sampled $\hat{SV}_k$. We find that the sampled $\hat{SV}_k$ quickly converges to the real value, SV_k, as the sample size increases. In fact, $\hat{SV}_k$ is very already close to SV_k when the sample size reaches approximately 5,000. Therefore, in our implementation, we select 10,000 samples so that the Shapley approach can be computationally efficient with negligible bias.

the set of all permutations requires $|K|!$ steps. We use this approach because the Shapley value can be easily interpreted as the average of all marginal contributions. In either approach, the computational complexity grows exponentially.

[3] Another way to be more computationally efficient is to use a more efficient programming language. To implement the Shapley decomposition, we use a C programming package under Python.

3 The conventional methods versus the sampling Shapley decomposition: Comparison based on simulated data

Given a large set of circumstances and a limited number of observation, the measures of IOP could be substantially different using different methods. To compare with methods introduced in section 2, we generated a date set with 16 circumstances, each of which has two categories (e.g. male and female for the gender circumstance). In this case, there are 65,536 types in total. To make sure each type has enough observations. We generate 100 samples in each type, resulting in 6,553,600 observations.

We consider two data generation processes (DGPs). The first DGP is based on the linear model of equation 2. The second DGP addes some interaction terms of circumstances to the linear model, changing it into a non-linear model. Given both data, we are able to check how each method performs with or without non-linearity.

The methods we use are (i) the conventional parametric model (Conv.Para) stated in equation 2 and equation 3, (ii) the conventional non-parametric model (Conv.nonPara) stated in equation 1, (iii) the Shapley parametric model (Shap.Para), and (iv) the Shapley non-parametric model (Shap.nonPara). We apply these methods on either the whole simulated data or a proportion of the data to mimic the use of sampled data in reality. We consider the proportions of 0.1%, 1%, 5%, 10%, 50%, and 100% respectively. Due to the restriction of computational power, we only use the Shapley methods on the 0.1% and 1% samples.

Figure 2 presents the results for the IOP estimates using different proportion of simulated data. Panel (a) and (b) show the IOP estimates from the linear and non-linear models, respectively. In the simulation based on the linear DGP, the true IOP value is 0.55; in the simulation based on the non-linear DGP, the true IOP value is 0.58. The simulation results show that the parametric model and the non-parametric model converge as the proportion of the used data increases. For the simulated data with no interaction terms, the parametric model and the non-parametric model provides the same estimate of IOP; while for the simulated data with interaction terms, the parametric model underestimates IOP as it neglects interaction terms and treats the data generation process as a linear model.

For the Shapley method, we find that the parametric Shapley decomposition provides the same results as the conventional parametric model; while the non-parametric model provides lower estimate comparing to the conventional non-parametric model. In addition, the non-parametric Shapley estimates converge faster than the conventional non-parametric model. As the non-parametric model considers non-linearity, it can be viewed as an upper bound of IOP. As a result, the IOP estimates provided by the Shapley methods have a narrower interval comparing to the IOP provided by the conventional methods, because the Shapley methods mitigate the over-estimations due to limited observations and sparse types.

To conclude, the Shapley methods can provide a narrower interval with less bias from data with a large set of circumstances, where the lower bound estimate is the value of IOP if circumstances have linear effects on income, whereas the upper bound estimate is the value of IOP if circumstances have non-linear effects income.

4 The applications of the Shapley decomposition on HILDA 2001-2017

4.1 Data

The data we use in this paper comes from the Household, Income and Labour Dynamics in Australia (HILDA) Survey. HILDA is an approximately representative household-based longitudinal study started in 2001. It covers more than 6,500 households with about 20,000 individuals in Australia. It collects information related to a wide range of topics such as income dynamics, socio-economic status, education and health.

Our dataset covers the period from 2001 to 2017. Table 1 shows the summary statistics of the dataset for selected years. The variables we use include income and variables related to circumstances. We report circumstance variables in proportions except for the number of siblings. We focus on individuals aged from 15 to 65. In 2001, data from 13,252 individual observations were collected. However, due to the sample attrition, the sample size gradually decreases from 13,252 in 2001 to 12,184 in 2010. In 2011, HILDA added an additional sample containing more than 5,000 individual observations. Therefore, the sample size in 2011 increases to 15,912 and then gradually decreases to 15,343 in 2017.

To compute IOP, we use individual income as well as household equivalent income, the latter one is the household income per effective member and is calculated as the real household income divided by the square-root of household size. Most literature uses household income to measure income inequality. However, IOP is more related to individual characteristics like family backgrounds and parents' socioeconomic status. Therefore, we use individual income in this paper to measure IOP. We note that some research (e.g. Lefranc et al., 2008) uses household income to measure IOP. These studies treat household income as individual standard of livings. In this study, we also present the measure using household equivalent income. A comparison of the IOP estimated based on individual and household equivalent incomes allow us to examine how intra-household transfers may affect IOP.

For both incomes, we consider private income, gross income, and disposable income. Private income includes wages and salary, business income, investment income, regular private pensions and transfers and other irregular income. Gross income contains private income and government income and non-income support. These public supports include government pensions, parenting payments, family payments, government bonus payments, allowance and other government benefits. Disposable income is gross income net of taxes, which include the sum of income tax, the Medicare Levy, taxes on redundancy payments less offsets and dividend imputation credits. A comparison of the IOP estimated based on these three types of incomes allow us to examine how taxes and government transfers may exacerbate or mitigate IOP.

We use 13 circumstance variables including individuals' gender, migration status, family backgrounds and their parents' characteristics that, to divide the sample into 115,200 possible types. These variables are used in many empirical studies of IOP. For example, Checchi and Peragine (2010) use gender, geographic locations and parental educational level as circumstance, and Ferreira et al. (2008) use ethnicity, father's occupation, parents' education and birth region.

In Australia, individuals could be disadvantaged if they are female, migrants, refugees, or indigenous people. In our dataset, around 52% of individuals are female. Migrants consist of around 23% of the population. Among migrants, more than half of them come from a non-English speaking country. Around 2% of the population is Indigenous Australian, and another 1% are refugees.

We also include country of birth for respondents and their parents, respectively, as these variables might suggest respondents' migration status and their upbringing. For example, if both parents are from non-English speaking countries, children could live in a non-English speaking environment, which might have an impact on them. To capture the difference in parents' country of birth, we grouped this characteristic into three categories – at least one parent was born in Australian, both parents were born in foreign countries but at least one was born in an English-speaking country, both parents were born in non-English speaking countries. We found that around 17% of the respondents have both parents coming from non-English speaking countries, and around 12% of the respondents have at least one parent immigrating from an

English-speaking country.

In addition, we include variables related to respondents' family backgrounds. Around 34% of the respondents are the oldest child in the family. Around 80% of the respondents reported living with both own parents in childhood. Around 3% of individuals have no siblings, 20% one sibling, 27% two siblings, 28% three or four siblings, and 20% five or more siblings.

To capture parents' socioeconomic status (SES), we use variables that show whether each parent was employed and whether each parent worked as a labourer when the respondent was a child. On average, fathers' employment rates are around 84% and mothers' 50%.

In addition, we include age as a circumstance. Age is considered as a demographic variable in some literature. For example, Checchi and Peragine (2010) compute IOP for each age cohort, and Jusot et al. (2013) treat age cohort as a demographic variable in the regression. In this paper, we treat age as a circumstance because it is beyond individual's control and fits the definition of circumstances.

4.2 Income inequality in Australia based on Gini coefficients

We first measure income inequality in Australia from 2001 to 2017 using Gini coefficients. Results are shown in Figure 3. The inequality trend for individual income and household equivalent income are presented in the upper and lower panels, respectively. The solid line represents the computed Gini coefficients using private income, the dashed line with circle marks using gross income, and the dashed line with cross marks using the disposable income. In general, inequality in both individual income and household equivalent income decreased during 2001 and 2008 and increased during 2009 and 2017. However, these trends are negligible compared to the overall level of inequality. The Gini coefficient is around 0.3-0.33 in household disposable income and around 0.43-0.46 in individual disposable income. The differences between these two sets of figures indicate that intra-household transfers reduce inequaity by approximately 0.13 Gini points, or about one-third of individual income inequality.

For individual income, we find that redistribution by the public transfers reduce income inequality by roughly 0.06 Gini points, which is approximately an 11% decrease. The tax policy reduces income inequality by a further 0.04 in Gini points, which is a further 8% decrease. Household equivalent income inequality shows an even bigger reduction as a result of public transfers and taxation. The Gini coefficient reduces by about 0.1 points (or 18%) due to public transfers and another 0.05 points due to taxation (or 9%). These results suggest that government transfers and taxation are efficient in reducing income inequality, but government transfers reduce income inequality twice as much as taxation.

Our results are consistent with other findings in the literature. For example, Whiteford (2017) estimates that inequality in Australian household income is between 0.31 and 0.34 as measured by the Gini coefficient using Australian Bureau of Statistics (ABS) data between 2001 and 2014. His estimation using HILDA is around 0.3, which is slightly lower than the results from ABS data. Similar to Whiteford (2017), Kaplan et al. (2018) use ABS data and estimate that the Gini coefficients of disposable household income in Australia is around 0.30 in financial year 2004 and 0.32 in 2010.

4.3 The effect of differing number of circumstances on measuring IOP

To understand how differing opportunity sets contribute to these inequalities, we first use respectively the parametric and non-parametric models by replacing individual income with the average income of their types. In Figure 4, we test the performance of both models for different number of circumstances using both the Gini coefficient and MLD. The x-axis is the number of circumstances. For our dataset, the maximum number of circumstances is 16. The y-axis is the direct IOP measure.

If using the whole set of circumstances, IOP would be 30% based on the parametric model and 40% the non-parametric one. In both cases, since we divide the population into too many types, IOP could be over-estimated. As discussed previously, a common practice in the literature is to use a subset of observed circumstances to measure IOP. To investigate the implication of this practice, we set the number of circumstances from 1 to 15 and randomly draw a subset from the whole set of circumstances. In figure 4, the band represents the confidence interval corresponding to the draw given a number of circumstances. For example, if using six circumstances, we find that both the non-parametric model and the parametric model give similar results. However, the confidence intervals could range from around 14% to 21%. The differences between the parametric and non-parametric models widen as the number of circumstances increase further. The results imply that selections of different subsets of circumstances could lead to significant difference in IOP measures.

The conventional approach could also bias the results of the inequality contributions of individual circumstances. In Figure 5, we demonstrate the inequality contribution of gender in 2002 computed using the conventional approach and the Shapley value for gender. The figure shows how the contribution of gender changes related to the number of factors (including both circumstances and effort) excluding gender considered in the computation of the SV_k. If no other factor is considered in the computation, the contribution of gender is the difference between the inequality when gender is considered and the inequality when gender is not considered. In this case, the sample is only divided into two categories based on gender, which could lead to an over-estimation of the contribution of gender.

Figure 5 shows that the contribution of gender is 5% if no other factor is considered. On the other hand, if we consider many other observed factors, the contribution of gender is the difference when we have an additional slide of the data based on gender. This finer division could lead to gender contributing relatively much less to income inequality. As shown in Figure 5, the contribution of gender is close to 0 if all other factors are considered. However, after averaging the contributions associated with different number of factors are considered, that is, around 1.2% for the gender variable in 2002. The bias of a large set of types is mitigated by the averaging.

4.4 Inequality of opportunity in Australia during 2001 - 2017

Using the sampling-based Shapley approach, we are able to measure IOP using the whole set of circumstances with less over-estimation and bias due to the arbitrary selection of subsets. We compute IOP for disposable incomes during 2011 to 2017 using both the Shapley parametric and non-parametric approach and the conventional parametric and non-parametric approach, and present the results in Figure 6. For inequality measure, we use MLD for all empirical methods.

We find that IOP is around 5% if using parametric approaches. This result is close to Martinez et al. (2017)'s estimation using the conventional parametric model with the same dataset. Since the parametric model only assume a linear relationship between circumstances and income, the results could represent a lower bound of IOP in Australia. To compensate the oversimplified parametric model, our results from the non-parametric model shows the IOP estimates allowing non-linear relationships between income and circumstances. However, as we include a large set of circumstances, the conventional non-parametric model is likely to over-state IOP. Our results show that this figure is around 30%. After applying the Shapley decomposition on the non-parametric model, we reduce this over-stated IOP to around 20%. Although this figure could still overestimate IOP as our sampling size is around 10,000, it provides a much lower estimation comparing to the conventional non-parametric model. We can conclude from the non-parametric model that IOP is around 20% at most in Australia if we consider non-linear relationships between factors.

Zero income could affect inequality measured by MLD as MLD is more sensitive to low incomes. Therefore, we compare IOP measure with and without zero income as a robustness check. The results are shown in Figure 7. We find that the estimates of IOP for disposable incomes are similar with and without zero income. However, the estimates yield different results if we use private incomes especial for parametric models. This is probably because there are much more zero incomes in private incomes than in disposable incomes. When we include zero incomes in private incomes, the Shapley decomposition on the parametric model provides an overestimated IOP comparing to excluding zero incomes. Therefore, one should apply the Shapley approach to data including many zero outcome observations with caution.

4.5 The decomposed contributions of each factor to income inequality

Using the Shapley decomposition, we compute the contribution of each circumstance to income inequality. In terms of income, we use individual disposable income in 2017. The full results are shown in Table 2.

We find that IOP in 2017 ranges from 8.75% to 22.96%. The biggest contributor is gender, accounting for around 5 percentage points of the inequality in disposable individual income. The figure suggests a wide income gap between genders. The contribution of gender based on the parametric estimate is similar to that on the non-parametric estimate, which suggests that gender has few interactions with other factors.

Parents' occupation during respondents' childhood is another contributor to income inequality. Whether parents were employed and whether parents did labour work account for around 0.8 to 4.7 percentage points of income inequality in total. The gaps between the parametric and non-parametric estimate suggest that parents' occupation interacts with other factors. The estimate could be largely underestimated if only the linear relationship is assumed between income and circumstances.

Another big contributor to income inequality is age. Grouping individuals into five age cohorts, we find that age contributes 1.23 to 4 percentage points of income inequality. In this paper, we treat age as circumstances. Some literature also argue that age is a demographic characteristic and it should be excluded from circumstances. As we provide the individual contribution of each circumstance, one can exclude age, compute IOP and compare this result with other literature that excludes age.

The number of siblings also contributes to individual disposable income. Its contribution ranges from 0.4 to 2.46 percentage points. These contributions from number of siblings are consistent with Parr's (2006) finding that number of siblings affects education attainment, income, and wealth in Australia.

Note that the Shapley decomposition yields some negative results for the contribution of individual circumstances. For example, the contribution of individuals born in other English-speaking countries is -0.11 percentage points and parents born in non-English speaking countries is -0.03 percentage points. Since we adopt the equalized income distribution to compute the Shapley value with the parametric model, a negative figure suggests that these circumstances contribute to reducing income inequality.

5 Conclusion

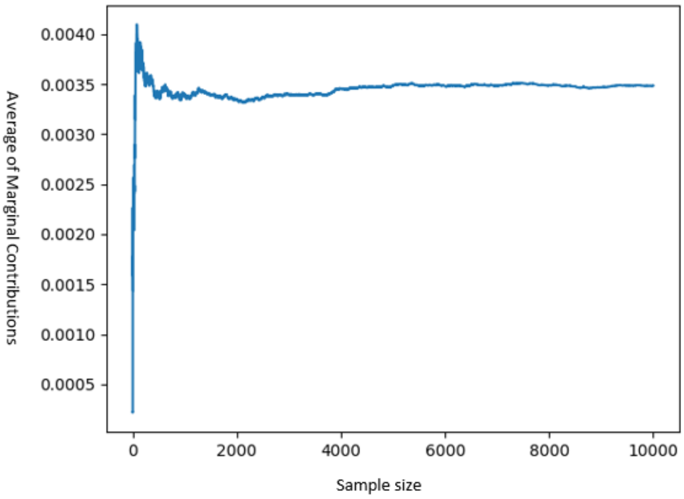
This paper has presented both a methodological and an applied contribution to the literature on inequality of opportunity. We have shown that with some modern computational innovations, the Shapley Value can be operationalised far more effectively in applied work, providing researchers a tool to decompose regressions and inequality metrics into additive contributions from explanatory variables. More importantly, the Shapley decomposition approach provides a less over-estimated IOP than the conventional non-parametric approach given a large set of circumstances, allowing researcher to consider more comprehensive sets of circumstances. The implication of this sort of innovation is potentially far-reaching; for example, Shorrocks (2013) suggests that Shapley decompositions could be presented alongside standard significance testing in regression models in all-purpose applied work.

Our empirical contribution relates to the income distribution in Australia since 2001. We provide an interval estimate of IOP using the Shapley approach. On the one hand, the lower bound represents the contribution of circumstances to income inequality based on linear relationships between circumstances and income; on the other hand, the upper bound indicates the contribution of circumstances based on non-linear relationships between circumstances and income. The interval estimates computed using the Shapley decomposition are narrower comparing to the estimates computed using the conventional parametric and non-parametric methods given a large set of circumstances, suggesting that the Shapley decomposition reduces the bias due to limited data.

Using our decomposition technique, we studied the evolution of economic inequality and the relationships between income and pre-determined factors such as migrant status, gender, and parental socio-economic level. We find that gender is the single largest source of income inequality in Australia.

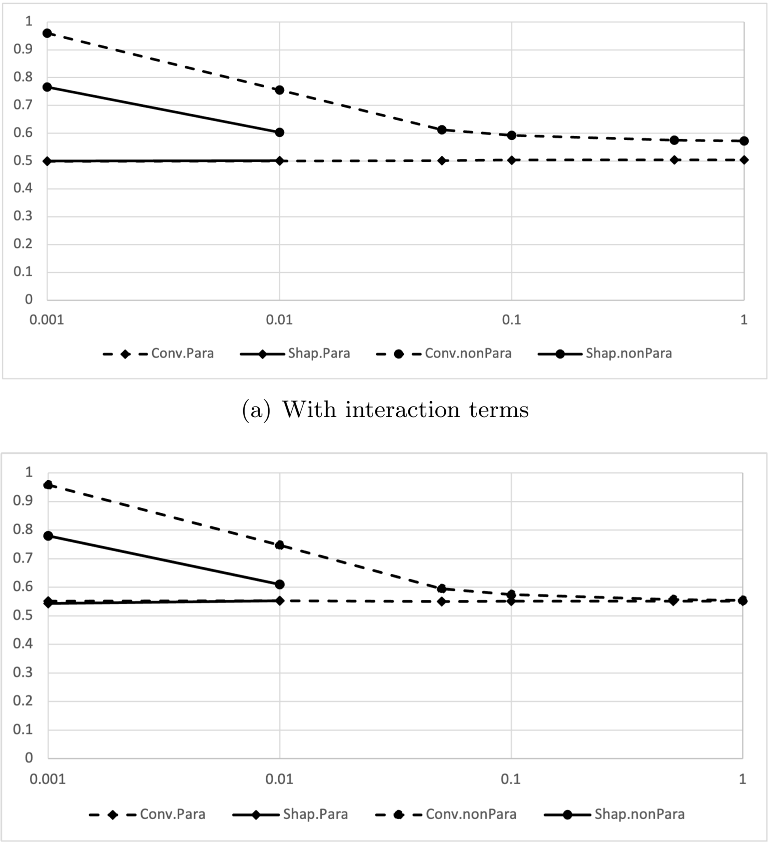
6 Figures and Tables

Figure 1: The Convergence of the sampling Shapley value (Based on MLD)



Source: Authors' calculation.

Figure 2: IOP estimates using simulated data (Conventional vs. Shapley methods)

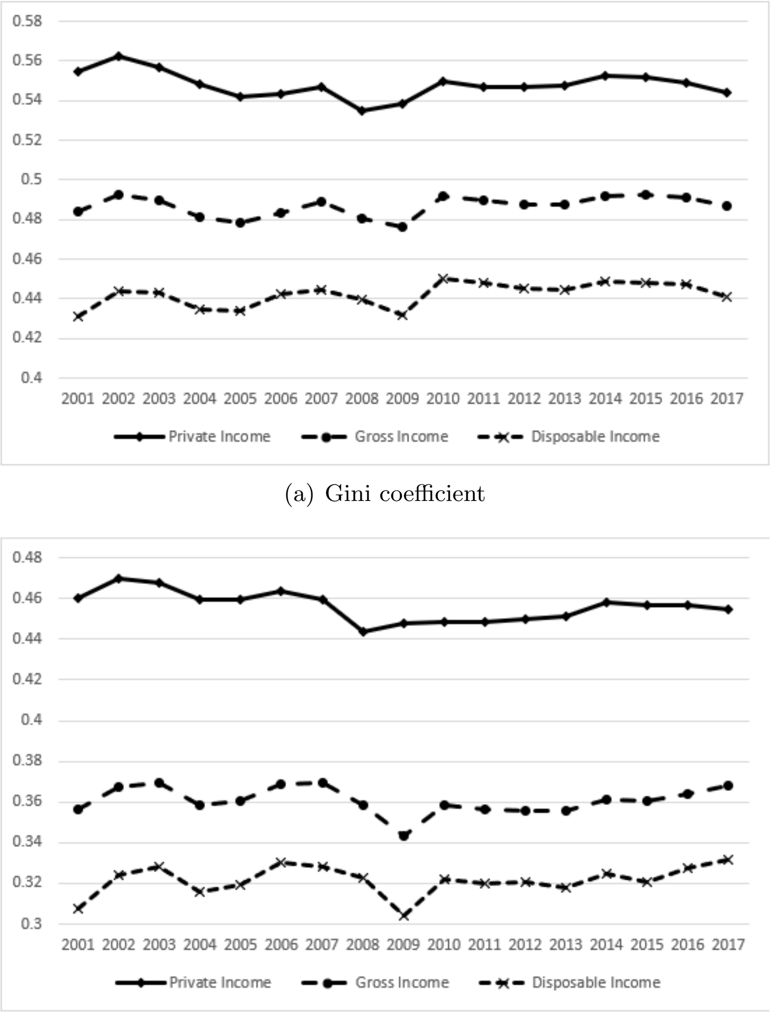


(a) With interaction terms

(b) No interaction terms

Source: Authors' calculation.

Figure 3: Income Inequality in Australia (2001 - 2017)

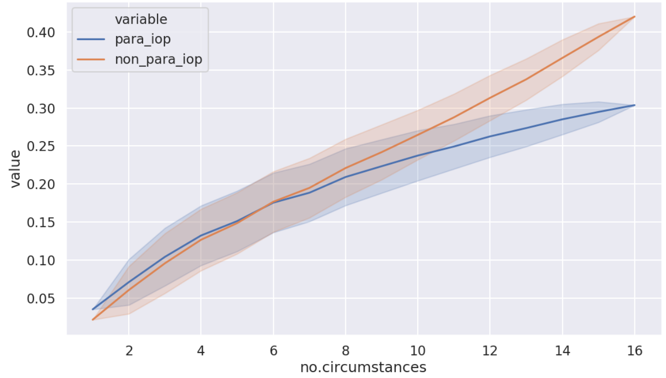


(a) Gini coefficient

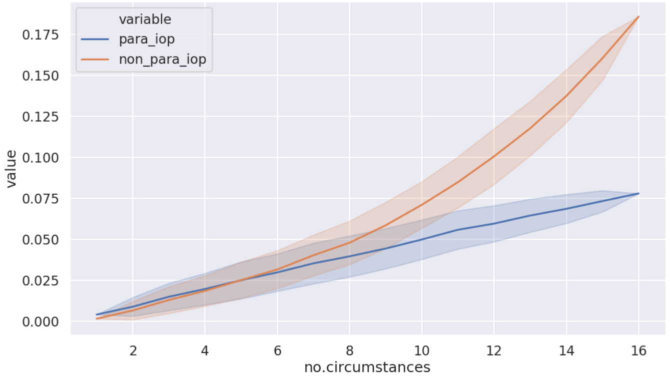
(b) Mean Log Deviation

Source: Authors' calculation.

Figure 4: Relationships between No. circumstances and IOR



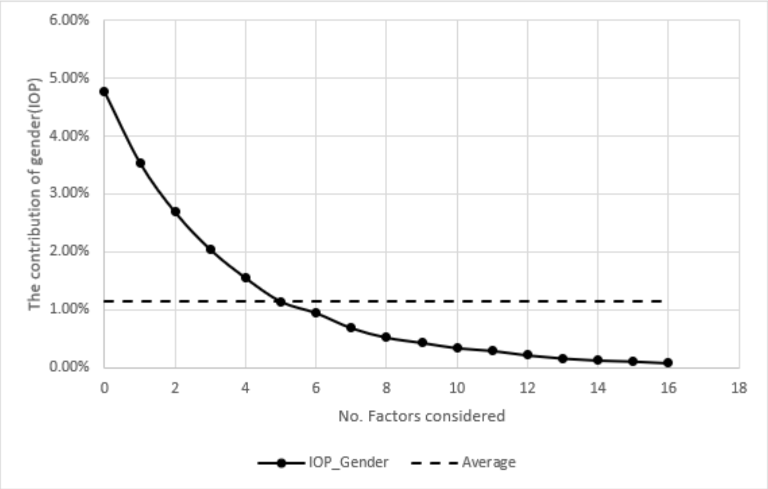
(a) Gini Coefficients



(b) Mean Log Deviation

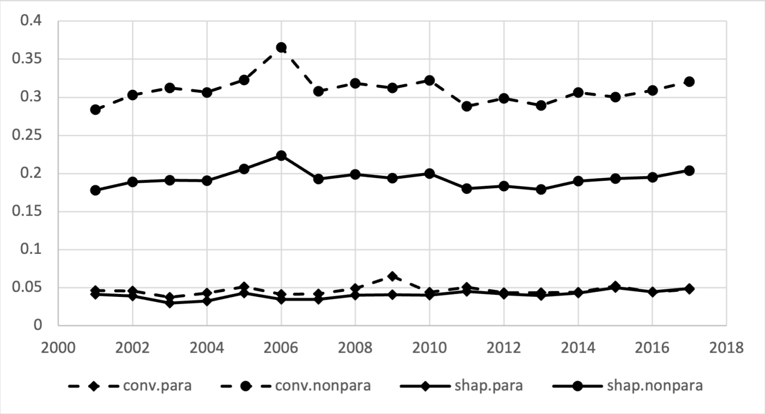
Source: Authors' calculation.

Figure 5: The computation of the Shapley value for gender with different No. factors in 2002)



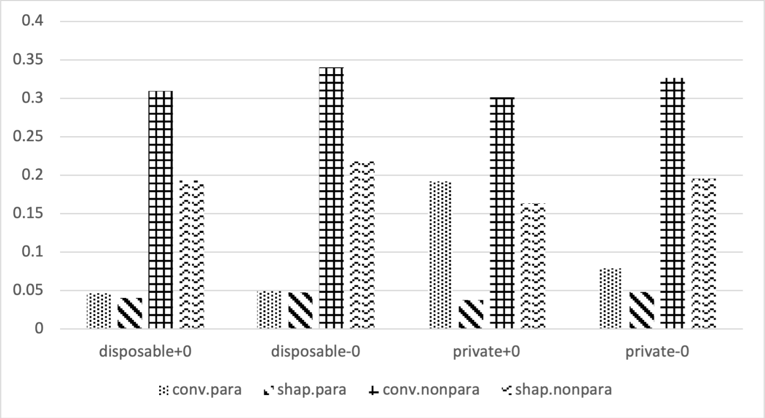
Source: Authors' calculation.

Figure 6: Trends of IOP in Australia: Conventional vs. Shapley methods



Source: Authors' calculation.

Figure 7: IOP estimates with and without zero incomes)



Source: Authors' calculation.

Table 1: Summary Statistics (2001-2017 Year Averaged)

Individual private income	61337.43
Individual gross income	65762.32
Individual disposable income	52840.48
Household private income	72876.49
Household gross income	77781.56
Household disposable income	62836.29
Female	0.52
Migration status(Respondents)	
English-speaking country	0.10
Non-English-speaking country	0.13
Aboriginal	0.02
Refugee	0.01
Migration status (Parents)	
English-speaking country (At least 1)	0.12
Non-English-speaking country (All parents)	0.17
Household conditions	
Is the oldest child in the family	0.34
Number of siblings	
One sibling	0.22
Two siblings	0.27
Three or four siblings	0.28
Five or more than five siblings	0.20
Living with both own parents	0.80
Parents' occupation status when respondents were children	
Employed (Father)	0.84
Employed (Mother)	0.50
Mother worked as a labourer	0.13
Father worked as a labourer	0.09
Age Group	
25 to 34	0.26
34 to 44	0.26
45 to 54	0.26
55 to 64	0.2
Observations	179465

Table 2: The contribution of circumstances to inequality in disposable income in 2017

	Shap.Para	Shap.nonPara
Female	5.08%	5.22%
Country of Birth (English)	−0.11%	0.60%
Country of Birth (Other)	0.18%	0.57%
Parents CoB (English)	0.14%	0.72%
Parents CoB (Other)	−0.03%	0.67%
Aboriginal	0.24%	0.75%
Refugee	0.06%	0.32%
Firstborn	0.15%	1.20%
One Sibling	0.07%	0.67%
Two Siblings	0.03%	0.50%
Three or Four Siblings	0.04%	0.59%
Five or More Siblings	0.26%	0.70%
Live with Parents	0.48%	1.62%
Employed (Father)	0.21%	1.09%
Employed (Mother)	0.18%	1.20%
Labourer (Father)	0.30%	1.18%
Labourer (Mother)	0.29%	1.21%
Age 25-34	0.16%	0.46%
Age 35-44	0.26%	1.69%
Age 45-54	0.27%	0.80%
Age 55-64	0.54%	1.14%
IOP	8.75%	22.96%

References

Alvaredo, F., Atkinson, A. B., Piketty, T., and Saez, E. (2013). The top 1 percent in international and historical perspective. *Journal of Economic perspectives*, 27(3):3–20.

Andersen, T. M. (2019). Social background, education, and inequality. *Economic Inquiry*, 57(3):1441–1459.

Arneson, R. J. (1989). Equality and equal opportunity for welfare. *Philosophical studies*, 56(1):77–93.

Björklund, A., Jäntti, M., and Roemer, J. E. (2012). Equality of opportunity and the distribution of long-run income in Sweden. *Social choice and welfare*, 39(2-3):675–696.

Castro, J., Gómez, D., and Tejada, J. (2009). Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730.

Checchi, D. and Peragine, V. (2010). Inequality of opportunity in Italy. *Journal of Economic Inequality*, 8(4):429–450.

Cohen, G. A. (1989). On the currency of egalitarian justice. *Ethics*, 99(4):906–944.

Corak, M. (2013). Income inequality, equality of opportunity, and intergenerational mobility. *Journal of Economic Perspectives*, 27(3):79–102.

Ferreira, F. and Gignoux, J. (2011). The measurement of inequality of opportunity: theory and an application to Latin America. *Review of Income and Wealth (forthcoming).*

Ferreira, F. H. (2019). Inequality as cholesterol: Attempting to quantify inequality of opportunity. Policy Research Talk.

Ferreira, F. H. G., Molinas Vega, J. R., de Barros, R., and Saavedra Chanduvi, J. (2008). *Measuring inequality of opportunities in Latin America and the Caribbean.* The World Bank.

Jusot, F., Tubeuf, S., and Trannoy, A. (2013). Circumstances and efforts: how important is their correlation for the measurement of inequality of opportunity in health? *Health economics*, 22(12):1470–1495.

Kaplan, G., La Cava, G., and Stone, T. (2018). Household economic inequality in Australia. *Economic Record*, 94(305):117–134.

Lefranc, A., Pistolesi, N., and Trannoy, A. (2008). Inequality of opportunities vs. inequality of outcomes: Are western societies all alike? *Review of Income and Wealth*, 54:513–546.

Marrero, G. A. and Rodríguez, J. G. (2013). Inequality of opportunity and growth. *Journal of Development Economics*, 104:107–122.

Martinez, A., Rampino, T., Western, M., Tomaszewski, W., and Roque, J. D. (2017). Estimating the Contribution of Circumstances that Reflect Inequality of Opportunities. *Economic Papers*, 36(4):380–400.

Niehues, J. and Peichl, A. (2011). Lower and upper bounds of unfair inequality: theory and evidence for Germany and the US. *Available at SSRN 1916583*.

Parr, N. (2006). Do children from small families do better? *Journal of Population Research*, 23(1):1.

Patel, J., Nielsen, F., Badiani, A., Assi, S., Unadkat, V., Patel, B., Ravindrane, R., and Wardle, H. (2020). Poverty, inequality and covid-19: the forgotten vulnerable. *Public health*, 183:110.

Roemer, J. E. (1998). *Equality of Opportunity*. Harvard University Press.

Sastre, M. and Trannoy, A. (2002). Shapley inequality decomposition by factor components: Some methodological issues. *Journal of Economics*, 77(1):51–89.

Sen, A. (1990). Development as capability expansion. *The community development reader*, pages 41–58.

Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317.

Shorrocks, A. F. (2013). Decomposition procedures for distributional analysis: a unified framework based on the Shapley value. *Journal of Economic Inequality*, 11(1):99.

Tomes, N. (1981). The family, inheritance, and the intergenerational transmission of inequality. *Journal of Political Economy*, 89(5):928–958.

Whiteford, P. (2017). Trends in income inequality in Australia. *Australian Quarterly*, 88(3):30–36.