



UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS FLORIANÓPOLIS
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA

Paulo Henrique dos Santos

**Studies of Income Distribution Functions in Econophysics Using Particle
Swarm Optimization**

Florianópolis
2025

Paulo Henrique dos Santos

**Studies of Income Distribution Functions in Econophysics Using Particle
Swarm Optimization**

Thesis submitted to the Programa de Pós-Graduação em Física of the
Universidade Federal de Santa Catarina for the qualifying exam.
Supervisor:: Prof. Marcelo Henrique Romano Tragtenberg, Dr.

Florianópolis
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Henrique dos Santos, Paulo

Estudos de funções de distribuição de renda em econofísica
usando otimização por enxame de partículas / Paulo Henrique
dos Santos ; orientador, Marcelo Henrique Romano
Tragtenberg, 2025.

107 p.

Tese (doutorado) - Universidade Federal de Santa
Catarina, Centro de Ciências Físicas e Matemáticas,
Programa de Pós-Graduação em Física, Florianópolis, 2025.

Inclui referências.

1. Física. 2. distribuição de renda. 3. Boltzmann-Gibbs.
4. distribuição de Pareto. 5. otimização por enxame de
partículas. I. Henrique Romano Tragtenberg, Marcelo. II.
Universidade Federal de Santa Catarina. Programa de Pós
Graduação em Física. III. Título.

Paulo Henrique dos Santos

**Studies of Income Distribution Functions in Econophysics Using Particle
Swarm Optimization**

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca
examinadora composta pelos seguintes membros:

Prof. Evaldo Mendonça Fleury Curado, Dr.
Instituição Centro Brasileiro de Pesquisas Físicas (CBPF)

Prof. Eraldo Sérgio Barbosa da Silva, Dr.
Instituição Universidade Federal de Santa Catarina (UFSC/CNM)

Prof. Fernando Fagundes Ferreira, Dr.
Instituição Universidade de São Paulo (USP/FSC)



Documento assinado digitalmente

Natalia Vale Asari

Data: 11/11/2025 17:32:40-0300

CPF: ***.497.091-**

Verifique as assinaturas em <https://v.ufsc.br>

Coordenação do Programa de
Pós-Graduação



Documento assinado digitalmente

MARCELO HENRIQUE ROMANO TRAGTENBERG

Data: 11/11/2025 12:44:00-0300

CPF: ***.769.958-**

Verifique as assinaturas em <https://v.ufsc.br>

Prof. Marcelo Henrique Romano
Tragtenberg, Dr.
Orientador

Florianópolis, 2025.

ACKNOWLEDGEMENTS

I would like to sincerely thank Professor Marcelo for his patient mentorship and insightful instruction, as well as for his occasional humorous remarks that made the learning process both enjoyable and memorable. I also appreciate the freedom he granted me to explore ideas, which was invaluable for my growth as a researcher.

I am grateful to Igor for his dedicated involvement throughout the development of this thesis. His active participation in meetings, careful attention to all aspects of the work, and insightful feedback were invaluable contributions that significantly enhanced the quality of this research. His commitment greatly enriched the entire process.

To my parents, whose enduring support, both emotional and financial, made this journey possible.

To Maria and Victor, for their advice in academic and organizational matters, and above all, for their constant presence as friends.

I also thank my peers, whose willingness to discuss ideas deepened my understanding and made this path more collaborative and enriching.

I am grateful to the members of the examination committee for their time, attention, and contributions to the evaluation of this work.

And to CAPES, for the essential financial support that sustained the development of this research.

*“When you can measure what you are speaking about, and express it in numbers, you know something
about it.”
(Lord Kelvin)*

RESUMO

O estudo apresenta um modelo contínuo de duas classes para a distribuição de renda no Brasil, que combina um regime exponencial de Boltzmann-Gibbs para a população de baixa renda e uma cauda de Pareto para a de alta renda, respectivamente. Utilizando microdados da PNAD anual de 2001 a 2011 e da PNAD Contínua de 2012 a 2019, aplicou-se um algoritmo de otimização por enxame de partículas para estimar três parâmetros principais: a renda de transição entre regimes, a temperatura efetiva da distribuição e o índice de Pareto. O modelo foi ajustado por meio de duas funções de perda, do erro médio quadrático logarítmico e da estimação de máxima verossimilhança ponderada, e validado com simulações de regressão. Os dados foram estratificados por sexo/gênero e cor/raça, permitindo avaliar desigualdades econômicas entre grupos. As análises temporais indicam que o coeficiente de Gini teórico, dependente apenas do índice de Pareto e da proporção de rendas altas, apresenta forte correlação com o Gini empírico, demonstrando que o modelo captura de forma consistente a dinâmica da desigualdade no período analisado. A análise de importância das variáveis, baseada em florestas aleatórias e árvores de inferência condicional, mostra que ocupação e idade são os principais preditores de renda, seguidos por educação, raça e sexo/gênero. Essa hierarquia revela que fatores ligados ao mercado de trabalho e à trajetória profissional explicam grande parte da variação da renda, enquanto características educacionais e demográficas modulam as diferenças entre indivíduos e grupos. Conclui-se que o modelo proposto representa adequadamente a estrutura da renda brasileira e constitui uma ferramenta sólida para estudos comparativos e análises longitudinais sobre desigualdade.

Palavras-chave: distribuição de renda; Boltzmann-Gibbs; distribuição de Pareto; Particle Swarm Optimization; crossover income; PNADc; importância de variáveis; simulação de regressão; coeficiente de Gini.

ABSTRACT

This study presents a continuous two-class model for income distribution in Brazil, combining an exponential Boltzmann-Gibbs regime for the low-income population and a Pareto tail for the high-income group, respectively. Using microdata from annual PNAD 2001–2011 and continuous PNAD from 2012–2019, a particle swarm optimization algorithm was applied to estimate three key parameters: the crossover income between regimes, the effective temperature of the distribution, and the Pareto index. The model was fitted using two loss functions, the root mean squared logarithmic error and weighted maximum likelihood estimation, and validated through regression simulations. The data were stratified by sex/gender and color/race, allowing the assessment of economic inequalities across groups. Temporal analyses indicate that the theoretical Gini coefficient, dependent only on the Pareto index and the proportion of high-income households, shows a strong correlation with the empirical Gini, demonstrating that the model consistently captures the dynamics of inequality over the analyzed period. The variable importance analysis, based on random forests and conditional inference trees, shows that occupation and age are the main income predictors, followed by education, race, and sex/gender. This hierarchy reveals that labor market and professional trajectory factors explain much of the income variation, while educational and demographic characteristics modulate differences among individuals and groups. It is concluded that the proposed model adequately represents Brazil's income structure and constitutes a robust tool for comparative and longitudinal studies on inequality.

Keywords: income-distribution; Boltzmann-Gibbs; Pareto distribution; Particle Swarm Optimization; crossover income; PNADc; variable importance; regression simulation; Gini coefficient.

RESUMO EXPANDIDO

INTRODUÇÃO

Apresentamos um modelo contínuo de duas classes para a distribuição de renda pessoal no Brasil, combinando um regime exponencial de Boltzmann-Gibbs (BG) para a classe de baixa renda (85–98% dos domicílios) com uma cauda de potência de Pareto para a classe de alta renda (2–15%). O modelo de duas classes é amplamente reconhecido por sua capacidade de caracterizar distribuições de renda ao unir um regime exponencial com uma cauda de potência. A componente exponencial possui propriedades de estabilidade em sistemas multiagentes que a alternativa log-normal não apresenta, tornando-a uma escolha natural para modelar a renda da maior parte da população.

Pesquisas anteriores geralmente determinaram o ponto de transição entre os regimes examinando gráficos log-log e identificando visualmente mudanças de inclinação ou descontinuidades. Em contraste, este estudo introduz um procedimento automatizado e orientado por dados para determinar a renda de transição ótima, eliminando a subjetividade presente na literatura tradicional e garantindo continuidade entre os regimes.

Além disso, ao utilizar microdados da PNAD anual de 2001 a 2011 e da PNAD Contínua de 2012 a 2019, propomos uma abordagem metodológica que integra amostragem ponderada, permitindo análises mais fiéis à população brasileira e a comparação com estudos prévios da econofísica. A estrutura desenvolvida também possibilita uma análise desagregada por sexo/gênero e cor/raça, ampliando a compreensão sobre as desigualdades econômicas entre grupos sociais.

METODOLOGIA

A estimativa dos parâmetros do modelo foi conduzida por meio de um novo esquema de otimização baseado em Particle Swarm Optimization (PSO). Três parâmetros principais foram estimados: (i) a renda de transição entre os regimes BG e Pareto; (ii) a temperatura efetiva da distribuição BG; e (iii) o índice de Pareto.

Para a otimização, utilizamos duas funções de perda complementares: Root Mean Squared Logarithmic Error (RMSLE), aplicada em amostragem simples, possibilitando comparações diretas com estudos anteriores; e Maximum Likelihood Estimation (MLE), incorporando pesos amostrais estratificados da Continuous National Household Sample Survey (PNADc), garantindo representatividade populacional.

Simulações de regressão foram realizadas com dados sintéticos gerados a partir do modelo de duas classes, definindo uma Complementary Cumulative Distribution Function (CCDF) empírica com parâmetros conhecidos. Essa etapa permitiu avaliar a capacidade de recuperação dos parâmetros verdadeiros sob cada função de perda. Os resultados mostraram que a função MLE apresentou estimativas mais próximas dos valores reais, ainda que a RMSLE tenha demonstrado desempenho consistente.

O ponto de transição entre os regimes foi otimizado para assegurar continuidade na distribuição e minimizar o erro em RMSLE, substituindo limiares fixos e subjetivos, como o tradicional corte de 5%. A validação do procedimento foi realizada via *bootstrap out-of-bag*, com desempenho robusto em amostras fora da base de treinamento.

Por fim, os dados foram estratificados segundo sexo/gênero e cor/raça, permitindo análises comparativas entre subgrupos populacionais, bem como a avaliação temporal dos parâmetros estimados e do coeficiente de Gini obtido a partir do modelo.

RESULTADOS E DISCUSSÃO

A análise mostrou que o coeficiente de Gini teórico, derivado apenas da fração da população na cauda Pareto e do índice de Pareto, apresentou correlação fortemente positiva com o Gini empírico calculado diretamente a partir dos microdados da PNADc. Essa concordância indica que o modelo de duas classes captura com elevada fidelidade os principais aspectos da desigualdade de renda no Brasil, tornando o Gini teórico um substituto confiável para análises comparativas, quando a estrutura de duas classes se ajusta adequadamente aos dados.

Adicionalmente, observou-se uma relação empírica relevante entre a fração de indivíduos na cauda Pareto (“fração de alta renda”) e o próprio índice de Pareto ao longo do tempo, sugerindo um padrão dinâmico de concentração de renda com potencial de generalização internacional.

Na análise preditiva da renda individual, utilizamos algoritmos de *Floresta Aleatória* com correção de viés via Conditional Inference Trees (CIT) e Conditional Permutation of Predictors (CPP). Três especificações foram testadas:

- **Modelo abrangente:** inclui ocupação (Brazilian Classification of Occupations (CBO)), idade, escolaridade, estado (UF), raça e sexo/gênero;
- **Modelo centrado na educação:** exclui variáveis ocupacionais e geográficas;
- **Modelo centrado no indivíduo:** inclui apenas atributos demográficos.

No modelo abrangente, ocupação e idade emergem como os preditores mais relevantes. Quando variáveis ocupacionais e de UF são removidas, a escolaridade passa a dominar a explicação da renda. Já no modelo centrado no indivíduo, raça e sexo/gênero tornam-se os principais determinantes. A aplicação do coeficiente de Theil's U demonstrou que variáveis demográficas ganham importância relativa à medida que fatores estruturais são removidos.

A análise segmentada entre os regimes revelou diferenças marcantes:

- No **regime exponencial (BG)**, a escolaridade é o principal fator de predição, com UF ganhando destaque no modelo completo;

- No **regime Pareto**, a influência da escolaridade diminui e sexo/gênero torna-se o preditor mais relevante, seguido de raça.

Entre os subgrupos, verificou-se que homens e pessoas brancas ou amarelas apresentaram maior concentração de renda na cauda Pareto, refletindo desigualdade interna mais acentuada. O grupo preto/pardo, por outro lado, mostrou a menor fração de indivíduos na cauda (6,42%) e a maior equidade, resultado consistente nas análises ponderadas e não ponderadas.

CONSIDERAÇÕES FINAIS

O modelo de duas classes proposto oferece um arcabouço automatizado e de ponta a ponta para caracterizar distribuições de renda, eliminando a necessidade de inspeções visuais e pontos de corte arbitrários. O procedimento de otimização proposto não apenas garante continuidade na distribuição ajustada, mas também reduz substancialmente o erro RMSLE em comparação com abordagens tradicionais de cauda fixa.

Ao incorporar ponderação amostral e expansão, asseguramos o tratamento adequado da estrutura de amostragem estratificada da PNADc, garantindo que as estimativas representem corretamente a população. Essa abordagem é particularmente relevante em análises de desigualdade, frequentemente prejudicadas por omitir o uso de pesos amostrais.

Entre as principais contribuições do estudo, destacam-se:

- Um procedimento automatizado de determinação da renda de transição que garante continuidade da distribuição e resulta em menor RMSLE em comparação com abordagens com cauda fixa;
- Identificação de uma relação temporal entre o índice de Pareto e a fração de domicílios de alta renda, com potencial de generalização para além do Brasil;
- Um modelo de duas classes capaz de capturar a desigualdade de renda com elevada fidelidade, com o coeficiente de Gini teórico — dependente apenas da fração de alta renda e do índice de Pareto — fortemente correlacionado com o Gini empírico;
- Uma análise de importância de variáveis que revela maior poder preditivo da escolaridade no regime exponencial, e maior influência de características demográficas (sexo/gênero e raça) no regime Pareto.

Apesar de seu caráter inovador, reconhecemos limitações. A superfície de perda não é garantidamente convexa, o que torna a busca pelo mínimo global desafiadora, e as regressões via PSO são computacionalmente intensivas, aumentando o custo dos ciclos de estimação. Algumas direções futuras são propostas:

1. Aplicar o método de otimização de ponto de transição a outras pesquisas nacionais, permitindo comparações internacionais de desigualdade;
2. Ampliar a análise temporal das ondas ponderadas da PNADc;
3. Explorar estratégias computacionais que acelerem a convergência do PSO sem comprometer a precisão na identificação do mínimo global.

LIST OF FIGURES

Figure 1 – Cumulative probability distribution of income on a log-log scale for PNADc 2019 data. The black points represent the cumulative distribution of the data and the solid lines correspond to the fitted model described in Equation (2.9). The red curve obeys - distribution, the power-law distribution is characterized by the blue curve and the dashed blue line is its extrapolation. The income crossover is represented by the vertical dashed line, whereas the Pareto threshold by the vertical solid line.	28
Figure 2 – The Lorenz curve framework (hypothetical data).	32
Figure 3 – Discrete Lorenz curve. The points represent the data, the dashed black line represents the line of equality and the dashed red line the boundaries of each trapezoid.	35
Figure 4 – Simulated data with impurities on a cumulative probability on a log-log scale. The black line shows the final two-class distribution with all impurities, using parameters $\lambda = 0.075$, $\alpha = 1.0$, and $T = 1750$. The red line represents the two-class distribution without impurities, while the blue line depicts the normal distribution used for the 1% impurity.	47
Figure 5 – Two-class regression with impurities for the simulated data cumulative probability on a log-log scale with parameters $\lambda = 0.075$, $\alpha = 1.0$ and $T = 1750$. The black points represent the cumulative distribution of the simulated data CCDF and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve. The temperature is highlighted by the dashed vertical line and the crossover income is represented by the solid vertical line, which separates the Boltzmann-Gibbs and Pareto regions of income. The table in the top right corner contains information on both the parameters and the value of the regression's loss function.	48

Figure 6 – Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of the data from 2019 National Household Sample Survey (PNAD) and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve. The temperature is highlighted by the dashed vertical line, at value R\$ 1775 $\pm$ 3, and the crossover income is represented by the solid vertical line at m_c = R\$ 3977 $\pm$ 42, which separates the Boltzmann-Gibbs and Pareto regions of income. Therefore λ = (10.64 $\pm$ 0.26)% is the top income percentage of people which obeys a Pareto power-law with index α = 1.789 $\pm$ 0.011. Also, the RMSLE of the original data set calculated in each part of the distribution and total can be found in the top right table.	53
Figure 7 – Cumulative probability distribution of Income/ T (income normalized by temperature) on a log-log scale. The colored points represent the cumulative distribution of the data from 2001 to 2019 PNAD and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve.	57
Figure 8 – Cumulative distribution functions constructed from the PNAD data from 2001 to 2019 and their fits with the theoretical distribution described in Equation (2.12), shown in the log-log scale versus the normalized annual income Income/ T (income normalized by temperature). Plots for different years are shifted vertically.	58
Figure 9 – Time series of the inequality indicators.	59
Figure 10 – Correlation and affine regression between indicators.	60
Figure 11 – Time series of the inequality indicators for the subgroup data.	61

Figure 12 – Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of data from the 2019 PNADc, while the solid lines the fitted model described in Equation (2.12). The red curve corresponds to the Boltzmann-Gibbs distribution, and the blue curve the power-law distribution. The temperature is denoted by the dashed vertical line, with a value of R\$ 2010 $\pm$ 18, while the crossover income by the solid vertical line at m_c = R\$ 5035 $\pm$ 47, marking the boundary between the Boltzmann-Gibbs and Pareto regions of income. In the upper right corner, you'll find all the model's parameters in a table, where λ represents the top income percentage of individuals, yielding (8.16 $\pm$ 0.25)%, and adhering to a Pareto power-law with an index of α = 1.5 $\pm$ 0.046.	64
Figure 13 – An example of a regression tree where X_1 is a numerical covariate, while X_2 and X_3 are categorical.	74
Figure 14 – The image in the left is the how the a decision tree partition the space while in the right you have the conditional permutation grid partition. The solid lines indicates a separation, the grey lines are ignored and the dashed lines are the parts that was extended due from the tree to the conditional permutation due to the grid being made of a complete split of the predictor space in regards to X_1 which is the covariate that we are calculating the importance or is not part of $\mathcal{Z}_j$	77
Figure 15 – This figure presents two side-by-side tables: the normal permutation on the left, and the conditional permutation on the right. The conditional permutation is based on the $\mathcal{Z}_{j^*}$ set of variables. The a and b represent grid cells of $\mathcal{Z}_{j^*}$, where in the conditional permutation, values are permuted only within each cell.	77
Figure 16 – Comparison of variable importance within the overall population of PNADc.	82
Figure 17 – The Theil's U matrix displays the entropy coefficient $U(X Y)$ values for the set of predictor variables.	84
Figure 18 – Comparison of variable importance within the exponential region of PNADc.	85
Figure 19 – The Theil's U matrix displays the entropy coefficient $U(X Y)$ values for the set of predictor variables for the exponential region.	86
Figure 20 – Comparison of variable importance within the pareto region of PNADc.	87
Figure 21 – The Theil's U matrix displays the entropy coefficient $U(X Y)$ values for the set of predictor variables for the pareto region.	88

LIST OF TABLES

Table 1 – Comparative statistics resulting from the regression grid run using both the RMSLE and MLE methods. The table includes mean percentage error for temperature (in R\$), top-percentage (%), and Pareto index. Additionally, regression statistics such as Kolmogorov-Smirnov and Mean Least Square values are presented.	49
Table 2 – Mean, standard deviation, 95% confidence interval and coefficient of variation of the model parameters, Theoretical Gini coefficient, and RMSLE of the training and test sets.	52
Table 3 – Inequality indicators and RMSLE for a subgroup data.	54
Table 4 – Anticorrelation of Pareto index with top-percentage and correlation of the theoretical and empirical Gini coefficients for the subgroup data. .	61
Table 5 – Inequality indicators for a subgroup data.	65
Table 6 – Description of variables. The ranges and levels correspond to the active group in the PNADc 2019 dataset.	107

LIST OF ABBREVIATIONS AND ACRONYMS

BBI	Black, Brown and Indigenous Group
BGD	Boltzmann-Gibbs Distribution
CBO	Brazilian Classification of Occupations
CCDF	Complementary Cumulative Distribution Function
CIT	Conditional Inference Trees
CPI	Conditional Permutation Importance
CPP	Conditional Permutation of Predictors
CV	Coefficient of Variation
EI	Educational Issue
FI	Full Issue
IBGE	Brazilian Institute of Geography and Statistics
IPCA	Extended National Consumer Price Index
KS	Kolmogorov-Smirnov Statistic
L-BFGS-B	Limited-Memory Broyden-Fletcher-Goldfarb-Shanno with Box Constraints
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error
OOB	Out-of-Bag Set
PI	Personal Issue
PNAD	National Household Sample Survey
PNADc	Continuous National Household Sample Survey
PPS	Probability Proportional to Size
PSO	Particle Swarm Optimization
PSO-BFGS	Particle Swarm Optimization with Limited-Memory BFGS and Box Constraints
PSUs	Primary Sampling Units
RMSLE	Root Mean Squared Logarithmic Error
SPSO2007	2007 Standard for Particle Swarm Optimization
WY	White and Yellow Group

LIST OF SYMBOLS

T	Temperature
λ	Top Percentage
α	Pareto Coefficient
vec	Vectorization Operator
$\lfloor \cdot \rfloor$	Floor Function
$\mathbb{1}$	Indicator Function

CONTENTS

1	INTRODUCTION	21
1.1	GENERAL OBJECTIVES	21
1.2	SPECIFIC OBJECTIVES	21
1.3	PREFACE	22
2	TWO-CLASS MODEL FOR INCOME DISTRIBUTION	26
2.1	PREVIOUS METHODS	27
2.2	EMPIRICAL CUMULATIVE DISTRIBUTION	28
2.3	PROPOSED METHOD	29
3	GINI COEFFICIENT	32
3.1	THE THEORETICAL GINI COEFFICIENT	33
3.2	THE EMPIRICAL GINI COEFFICIENT UNWEIGHTED	34
3.3	THE EMPIRICAL GINI COEFFICIENT WITH WEIGHTS	35
4	HYBRID-PSO	37
4.1	PARTICLE SWARM OPTIMIZATION	37
4.2	SEARCH SPACE	37
4.2.1	Adaptive Topology	38
4.2.2	PSO Steps	39
4.2.3	Initialization	40
4.3	Limited-Memory Broyden-Fletcher-Goldfarb-Shanno with Box Constraints (L-BFGS-B)	40
5	LOSS FUNCTION	43
5.1	TYPE OF LOSS FUNCTION	43
5.1.1	Mean Squared Error	43
5.1.2	Maximum Likelihood	44
5.2	VALIDATION USING GENERATED DATA FROM INVERSE TRANS- FORM SAMPLING	46
6	REGRESSION ON THE PNAD SIMPLE SAMPLING DATA AND CROSS- VALIDATION	51
6.1	CONTINUOUS NATIONAL HOUSEHOLD SAMPLE SURVEY DATA OF 2019	51
6.2	SUBGROUP DATA OF 2019	53
7	TEMPORAL EVOLUTION OF INEQUALITY INDICATORS	56
8	REGRESSION ON THE PNAD STRATIFIED SAMPLING DATA . . .	63
9	IMPORTANCE OF VARIABLES	67
9.1	CONDITIONAL INFERENCE TREES	69
9.2	VARIABLE SELECTION	70
9.2.1	Standardization for $c_{\max}$:	72

9.2.2	Standardization for c_{quad}:	72
9.3	SPLITTING CRITERIA	73
9.4	CONDITIONAL PERMUTATION OF PREDICTORS	74
9.5	CONDITIONAL RANDOM FOREST	78
10	VARIABLE IMPORTANCE OF PNADC	80
10.1	EXPONENTIAL REGION	84
10.2	PARETO REGION	86
11	CONCLUSIONS AND PERSPECTIVES	90
	References	93
	APPENDIX A – HYBRID PSO ALGORITHM	100
	APPENDIX B – BFGS SUMMARY	101
	APPENDIX C – THE NATIONAL HOUSEHOLD SAMPLE SURVEY	102
C.1	SURVEY DESIGN EVOLUTION	102
C.1.1	Traditional PNAD (2001-2011)	102
C.1.2	PNADc (2012-2019)	102
C.2	SAMPLING METHODOLOGY	103
C.2.1	Traditional PNAD Sampling Design (2001-2011)	103
C.2.2	PNADc Sampling Design (2012-2019)	103
C.2.3	Rotation Scheme	104
C.3	CALIBRATION TO POPULATION BENCHMARKS	104
C.3.1	Calibration Methodology	104
C.3.2	Traditional PNAD Calibration	104
C.3.3	PNADc Calibration	104
C.4	VARIANCE ESTIMATION AND REPLICATE WEIGHTS	105
C.4.1	Bootstrap Methodology	105
C.4.2	Microdata Structure	105
C.4.3	Variance Estimation Procedure	105
C.4.3.0.1	<i>Use of Replicate Weights for Complex Statistics</i>	106
C.5	PNADC MICRODATA INCOME AND DEMOGRAPHIC INDICATORS	106

1 INTRODUCTION

1.1 GENERAL OBJECTIVES

The primary objective of this work is to develop a robust and consistent end-to-end method for fitting a two-class income distribution model to empirical data, while optimally determining the crossover point between the exponential and power-law regimes during the fitting process. This model should yield interpretable parameters that capture income inequality, offer the flexibility to handle weighted data, and contribute to a more rigorous econophysics approach (SICILIANI; TRAGTENBERG, 2019).

1.2 SPECIFIC OBJECTIVES

The specific objectives can be stated as follows:

- Analyze the income inequality across racial and gender subgroups by applying the two-class model, with a particular focus on the income distribution of both the total population and subgroups (sex/gender and color/race), extending methodologies used in previous studies (SICILIANI; TRAGTENBERG, 2019; TAO et al., 2019).
- Fit the model to the 2019 Brazilian income distribution data using the PSO optimizer (KENNEDY; EBERHART, 1995) and the Mean Squared Error (MSE) approach with simple sampling, while performing cross-validation with re-sampling techniques (EFRON, 1983; EFRON; TIBSHIRANI, 1997) to ensure robustness.
- Extend the analysis by fitting the model to a time series of Brazilian income data from 2001 to 2019, using the same MSE approach. Investigate the time evolution of the model's parameters and their implications for understanding trends in income inequality, as works for US data (SILVA, A. Christian; YAKOVENKO, 2004b, 2004a).
- Refine model fitting for subgroups using the MLE approach (MYUNG, 2003) with stratified sampling (KOLENIKOV, 2010), enabling a more precise analysis of population subgroups based on race and gender.
- Implement a variable importance analysis using conditional inference trees (HOTHORN et al., 2006), identifying the socio-demographic characteristics that most significantly influence class membership. Use random forest methodology (BREIMAN, 2001) and conditional variable importance measures (STROBL et al., 2008; WEI et al., 2015) with Theil's U (uncertainty coefficient) to examine the interaction between variables.

- Apply advanced numerical optimizers, such as the Particle Swarm Optimization with Limited-Memory BFGS and Box Constraints (PSO-BFGS) (BYRD et al., 1995), to improve model fitting and stability, considering both original and modified PSO algorithms (SHI; EBERHART, 1998a; LI et al., 2011).
- Perform simulations to evaluate the efficacy of different loss functions (MSE and MLE) using generated data, ensuring that the model accurately recovers pre-known parameters through statistical simulation frameworks (ALFONS et al., 2010), further validating the model's reliability.

1.3 PREFACE

A long time ago, the economist Vilfredo Pareto identified a power-law behavior in the income distribution (MOORE, 1897). Pareto stated that the income probability density function describing this distribution is of the form:

$$P(m) = bm^{-\alpha-1}, \quad (1.1)$$

where m denotes the income, α is known as the Pareto index ranging between 1 and 3 and b is a normalization constant. Later it was found that the Pareto law is suited for representing just the upper tail of the income distribution (HILL, 1975). The income where the Pareto tail starts is known as the **Pareto income threshold**.

The Pareto power-law was confirmed extensively on the upper income data from different countries, including evidence from India (SHIRRAS, 1935), Japan (AOYAMA et al., 2000), the United States, and was also found to describe wealth distribution effectively (Levy; LEVY; SOLOMON, 1997). In this work this high income region will be defined by the **top-percentage** indicator, which is the percentage of the population that follows Pareto behavior.

The Boltzmann-Gibbs Distribution (BGD), in the classical kinetic theory, is the most probable energy distribution of a gas with elastic collisions in thermal equilibrium. It was later found to be very useful for modeling income distribution for the low and middle-income class, by setting energy to be the money of the agents (CLEMENTI; GALLEGATI, 2005; DRAGULESCU; YAKOVENKO, 2000). The statistical mechanics approach to money distribution has been extensively studied (DRAGULESCU; YAKOVENKO, 2000), with applications showing exponential distributions in the lower income ranges (DRAGULESCU; YAKOVENKO, 2001; SILVA, A Christian; YAKOVENKO, 2004a). Therefore, for income less than the crossover income, m_c , the distribution is given by

$$P(m) = \frac{a}{T} e^{-m/T}, \quad (1.2)$$

where a is a normalization parameter and T is the 'temperature' of the system, acting as the characteristic income scale of the exponential regime. In practice, T corresponds

to the mean income in that regime — so it directly determines how quickly the tail of the distribution decays. A larger T results in a shallower drop-off (greater income dispersion), while a smaller T leads to a steeper decline.

Consequently, the personal income distribution can be considered as a two-class structure, as the lower class of the population (85-98%) is described by exponential BG distribution, whereas the upper class (15-2%) follows a Pareto power-law distribution (DRAGULESCU; YAKOVENKO, 2001; OANCEA et al., 2017; BANERJEE; YAKOVENKO, 2010). This two-class behavior has been observed across multiple countries and time periods (SILVA, A. Christian; YAKOVENKO, 2004b; SICILIANI; TRAGTENBERG, 2019; TAO et al., 2019). The most used method to determine the crossover between these two regions is to use a fixed proportion for the Pareto tail based on a log-log graph, where the Pareto region will present a linear behavior (DRAGULESCU; YAKOVENKO, 2001; SICILIANI; TRAGTENBERG, 2019; OANCEA et al., 2017). However, this choice is rather subjective, and therefore, the crossover income determined is not optimal.

We propose in this work a novel method to determine the total income distribution defined by parts, introducing a crossover income optimally established. Firstly, it is defined a measure of goodness-of-fit statistics that will be minimized by a numerical algorithm called PSO (KENNEDY; EBERHART, 1995) combined with L-BFGS-B (BYRD et al., 1995). This hybrid version of PSO strategies have shown effectiveness in global optimization problems (LI et al., 2011), with L-BFGS-B method being particularly suited for bound constrained optimization. We validate this method by studying the Brazilian income distribution using data from PNAD and PNADc, annually available by the Brazilian Institute of Geography and Statistics (IBGE), more on the data in Appendix C.

By optimally determining the crossover income, the percentage of people in the Pareto region will vary appropriately with the data, enabling us to analyze its temporal dynamics with greater precision than previous methods that relied on fixed or arbitrarily chosen crossover points. Among our findings, we highlight two of them, obtained from the study of the temporal evolution of the Brazilian income distribution. The first is the correlation between the Gini coefficient calculated with the data and the one calculated with the two-class model (FONTANARI et al., 2018). The second is the correlation between the Pareto index and the percentage of people that display the Pareto power-law behavior, building on previous work on universal patterns of inequality (BANERJEE; YAKOVENKO, 2010).

One aspect that is often overlooked in the econophysics literature is how to handle stratified data. Stratified sampling is widely used in many countries to manage sampling costs while ensuring the sample reflects desired characteristics (KOLENIKOV, 2010). Stratified sampling divides the population into subgroups to ensure accurate, cost-effective, and representative samples (ALFONS et al., 2010). For example, a na-

tional survey might use disproportionate stratified sampling, stratifying by state and urban/rural areas, then sampling households within each stratum with different sampling probabilities. If rural areas represent 20% of the population but researchers oversample them to comprise 40% of the survey sample (to improve precision for rural estimates), the rural observations would then receive weights of 0.5 to correct for this oversampling, ensuring the final results reflect the true 20% rural population share. Therefore, developing methods that effectively account for this type of sample is crucial for achieving accurate analysis.

This work is organized as follows. Chapter 2 explores the two-class CCDF and its continuity. In Chapter 3, the relation between the parameters of our model and the Gini coefficient is derived, building on theoretical foundations. Chapter 4 describes the L-BFGS-B Particle Swarm Optimization and justifies our choice, considering both the original PSO algorithm and its modifications. Chapter 5 defines two loss functions — one based on MSE and the other on MLE — and presents simulations demonstrating their efficacy using data generated from inverse transform sampling with pre-known parameters.

Chapters 6 and 7 both use the MSE approach with simple sampling. In Chapter 6, we apply the PSO optimizer to fit our model to the 2019 Brazilian income distribution data for both the total population and subgroups (sex/gender and color/race), performing cross-validation with a re-sampling technique. Chapter 7 focuses on the time evolution of the model's parameters from 2001 to 2019, extending previous work on temporal evolution of income classes.

In Chapters 8 and 9, we switch to the MLE approach with stratified sampling. Chapter 8 presents the results from the fitting process, where errors on the estimated parameters are assessed using replicated bootstrap weights (RUST; RAO, 1996). In Chapter 9, we apply the PSO optimizer to the same data and introduce a variable importance algorithm based on conditional inference trees, using random forest methodology and conditional variable importance measures. We use Theil's U (uncertainty coefficient) to analyze interactions between variables, following comprehensive approaches to variable importance analysis. Conclusions are presented in Chapter 10.

2 TWO-CLASS MODEL FOR INCOME DISTRIBUTION

The BGD, in the classical kinetic theory, is the most probable energy distribution of a gas with elastic collisions in thermal equilibrium. This distribution is of the form,

$$P(e_i) = \frac{a}{T} \exp\left(-\frac{e_i}{T}\right), \quad (2.1)$$

where e_i is the energy of the i_{th} state, T is the temperature and a is a normalization constant. The constraints used to derive BGD by maximum entropy probability distributions is the conservation of the total energy and the number of particles:

$$N = \sum_i n_i \quad (2.2)$$

$$E = \sum_i n_i e_i. \quad (2.3)$$

The derivation of the BGD consists in utilizing these constraints to maximize the entropy using lagrangian multipliers. Because the derivation of the BGD is very general and only requires two conservation constraints, one may expect that it can be applied to study the distribution in economic systems. In this application Dragulescu and Yakovenko proposed that one of the conservative quantities along with the number of the population is money (DRAGULESCU; YAKOVENKO, 2000). Therefore to describe a financial system, one can set the money as the analogue of the energy and the number of agents (population) is the analogue to the number of particles. The exponential distribution was found to be very useful for modeling income distribution for the low and middle-income class, by setting energy to be the money of the agents (CLEMENTI; GALLEGATI, 2005; DRAGULESCU; YAKOVENKO, 2000).

Unfortunately, the BGD does not describe the high income part of the income distribution. Thus to properly describe the whole distribution, one can correct the model using the Pareto distribution for the high income region. The resulting distribution is called the two-class model and consists of the BGD for the low to middle income class and Pareto distribution for the high income class.

Using two disjoint populations in which the probability distribution of income has a different functional form was extensively used to describe economical systems (BEE, 2015; PREDA; CIUMARA, et al., 2006). There is little doubt about the validity of the Pareto distribution to describe empirical data in the asymptotically high income, this concept is also known as Pareto weak law (MANDELBROT, 1960).

It is worth mentioning that the log-normal distribution is the most commonly used model for this region; however, unlike the Boltzmann-Gibbs distribution, it is not stationary, meaning its statistical properties such as mean and variance change over

time rather than remaining constant, as noted by Kalecki (KALECKI, 1945). This non-stationarity arises in contexts such as geometric Brownian motion (BLACK; SCHOLLES, 1973), multiplicative stochastic processes (REDNER, 1990), and Markov chain models (SUTTON, 1997).

In a multi-agent simulations context, its asymptotic states were capable of displaying Boltzmann-Gibbs as well as Pareto statistical behaviors (GONZALEZ-ESTEVEZ et al., 2008; CHATTERJEE et al., 2004). Moreover, the two-class model arrives from a Fokker-Planck equation, considering an additive and multiplicative processes for the exponential and Pareto region, respectively (BANERJEE; YAKOVENKO, 2010).

However in order to fit the two-class model into empirical data, a method to determine how to define the two regions and how to perform the regression needs to be provided. In this context, the Section 2.1 will clarify the problems of previous methods. In the Section 2.3 is proposed a version of the two class models that display continuity and define precisely the regions.

To better differentiate the theoretical and empirical variables or statistic, will be used the following notation

$x \longrightarrow$ for theoretical variables or statistic

$\hat{x} \longrightarrow$ for empirical variables or statistic

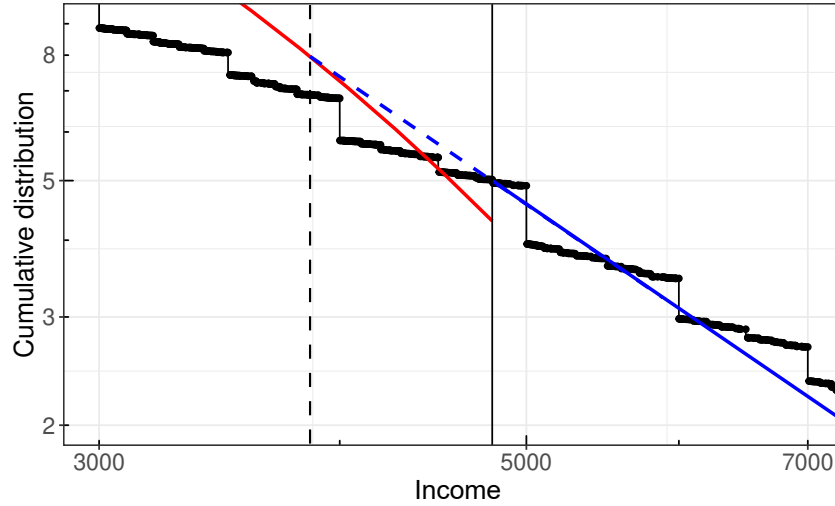
2.1 PREVIOUS METHODS

In previous methods the two distributions were fitted separately, having a lot of unique problems depending on the empirical data fitted. Between these problems the lack of continuity is one the most problematic, since it introduces a lot of choices before being able to use an optimization algorithm.

Usually for the BGD region, Equation (2.1), the normalization constant a was set to 1 and the T was the average of income (SILVA, A Christian; YAKOVENKO, 2004a; DRAGULESCU; YAKOVENKO, 2001). Some papers treat the parameters T more carefully by setting using an exponential regression, after defining the exponential region (SKVORTSOV et al., 2007). Since the BGD does not describe the high income behaviour, one can use the Pareto distribution to fit this region. However before is needed to determine the Pareto income threshold. The most commonly used method to determine this threshold is to plot the CCDF in a log-log scale and see where the behavior is linear (SICILIANI; TRAGTENBERG, 2019; OANCEA et al., 2017). A more robust option is to use a goodness-of-fit function to optimize the threshold income (SAFARI et al., 2018).

Since previous methods of the two-class model fit the two distributions separately, the m_c is determined by taking the intersection between the exponential fit (with

$T = \langle m \rangle$ for the 98–85% poorer people and a power-law fit in the 2–15% richer region determined by the Pareto income threshold, sometimes needing an extrapolation as shown in Figure 1. Notice that the m_c will not always be equal to this threshold, and, as stated before, CCDF will not be continuous in all cases.



Source: Own Author.

Figure 1 – Cumulative probability distribution of income on a log-log scale for PNADc 2019 data. The black points represent the cumulative distribution of the data and the solid lines correspond to the fitted model described in Equation (2.9). The red curve obeys - distribution, the power-law distribution is characterized by the blue curve and the dashed blue line is its extrapolation. The income crossover is represented by the vertical dashed line, whereas the Pareto threshold by the vertical solid line.

2.2 EMPIRICAL CUMULATIVE DISTRIBUTION

Before we introduce the proposed method, we are going to derive the empirical CCDF, since the determination of the crossover region depends on it. Considering a sample

$m_1, m_2, \dots, m_N$ of income drawn from a population with weights $w_1, w_2, \dots, w_N$, in this case Brazilian income, thus $m_{(1)} \geq m_{(2)} \geq \dots m_{(N)}$ is the income order statistic. Accordingly, the empirical CCDF of the PNAD income data is defined as

$$C(m_n) = \frac{1}{N} \sum_{i=1}^N w_{(i)} \mathbb{1}(m_{(i)} \geq m_n) \quad (2.4)$$

$$\mathbb{1}(m_{(i)} \geq m_n) = \begin{cases} 1 & m_{(i)} \geq m_n \\ 0 & m_{(i)} < m_n \end{cases}, \quad (2.5)$$

where N is the total number of people in the data, $\mathbb{1}$ is the indicator function, and $w_{(i)}$ represents the ordered weights, arranged according to the income order statistic.

2.3 PROPOSED METHOD

In this section we are going to define a two-class model that displays continuity in the CCDF. We can define a two-class model for a country income distribution using Equations (1.1,1.2)

$$P(m) = \begin{cases} \frac{a}{T} e^{-m/T} & m < m_c \\ bm^{-\alpha-1} & m \geq m_c \end{cases}, \quad (2.6)$$

equivalently CCDF

$$\int_m^\infty P(m') dm' \equiv C(m) = \begin{cases} a \left[\exp\left(\frac{-m}{T}\right) - \exp\left(\frac{-m_c}{T}\right) \right] + \lambda & m < m_c \\ \lambda \left(\frac{m}{m_c}\right)^{-\alpha} & m \geq m_c \end{cases}, \quad (2.7)$$

where, by continuity, $\lambda = \frac{b}{\alpha} m_c^{-\alpha}$ is the top percentage of income that follows the Pareto behavior, hence λ is the top-percentage indicator mentioned before. The normalization give us

$$\lambda = 1 - a \left[1 - \exp\left(\frac{-m_c}{T}\right) \right] \quad (2.8)$$

Setting $a = 1$ transforms the Pareto distribution into a correction for the exponential distribution in the high-income tail. In this context, the exponential distribution represents a normalized probability if it extended across all income levels. This adjustment enhances the interpretability of the model's parameters and simplifies the optimization process. So Equations (??) become

$$C(m, \lambda, T, \alpha) = \begin{cases} \exp\left(\frac{-m}{T}\right) & m < m_c(\lambda, T) \\ \lambda \left(\frac{m}{m_c(\lambda, T)}\right)^{-\alpha} & m \geq m_c(\lambda, T) \end{cases}, \quad (2.9)$$

and

$$m_c(\lambda, T) = T \log(\lambda^{-1}). \quad (2.10)$$

The CCDF in (2.9) is the two-class model by Yakovenko (SILVA, A. Christian; YAKOVENKO, 2004b). Two things to notice, first is the change in the parameters of the model that now are given by (λ, T, α) , the second is the guarantee of the theoretical CCDF continuity, which is not always the case when the two distributions are fitted separately.

To accurately determine the crossover region, we use the empirical crossover percentage $p = \hat{C}(m_c)$, where $\hat{C}$ is the empirical CCDF evaluated at income m_c . Since

the income variable contains repeated values, it is not ideal for directly separating the data. To address this, we compute m_c through linear interpolation of the empirical CCDF, yielding a continuous function C' . This interpolated CCDF is constructed by linearly interpolating between each pair $[m_{(i)}, m_{(i+1)}]$, resulting in an $(N-1)_{th}$ piecewise function defined over the continuous domain $m \in \mathbb{R}$. This approach transforms the empirical distribution into a continuous one, allowing the solution $m_c = (C')^{-1}(p)$ to take non-discrete values. Hence, $p = C'(m_c)$ is no longer constrained by the discrete data points. Given a value of T as defined in (2.10), we can then determine $\lambda \equiv C(m_c)$. For a parameter vector $\mathbf{x} = (p, \alpha, T)$, we define a theoretical distribution under the two-class model, which enables the construction of a goodness-of-fit function. Consequently, the piecewise separation Equations (2.6,2.9), representing the theoretical probability and complementary cumulative functions, becomes:

$$P(m) = \begin{cases} \frac{a}{T} e^{-m/T} & C'(m) > p \\ bm^{-\alpha-1} & C'(m) \leq p \end{cases}, \quad (2.11)$$

$$C(m, \lambda, T, \alpha) = \begin{cases} \exp\left(\frac{-m}{T}\right) & C'(m) > p \\ \lambda \left(\frac{m}{m_c(\lambda, T)}\right)^{-\alpha} & C'(m) \leq p \end{cases}. \quad (2.12)$$

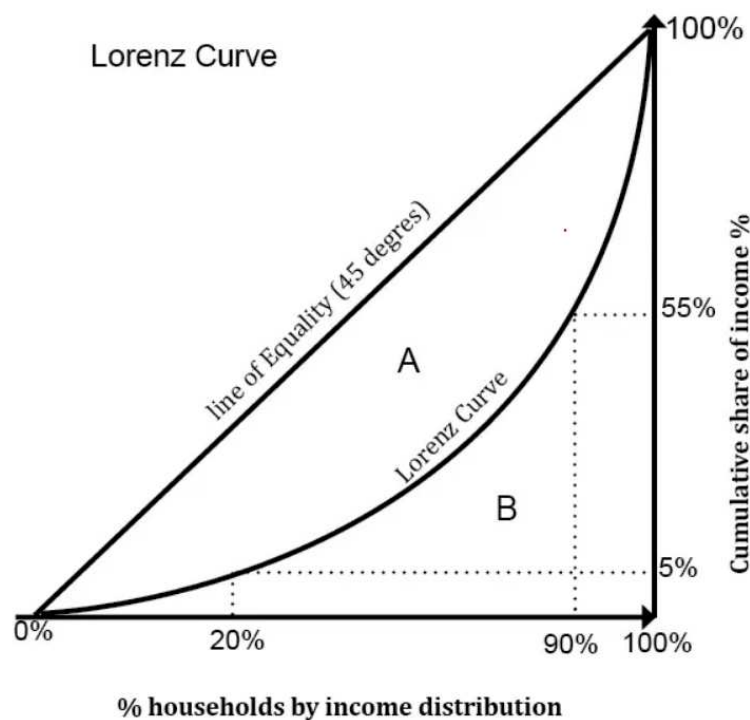
In this study, we will determine the two-class model using a hybrid version of a numerical optimizer called PSO. Starting with a given set of income values $\{m_i\}$, our method will predict a vector of parameters $\{\lambda_p, T_p, \alpha_p\}$ for the income distribution model defined by Equations (2.11,2.12). This definition of the model enables us to minimize a yet-to-be-defined loss function using the hybrid PSO technique, and ensures continuity and equality between the Pareto income threshold and m_c .

3 GINI COEFFICIENT

The Gini Coefficient, the most popular measure of income inequality, is derived from the Lorenz curve.

The Lorenz curve shows the percentage of total income earned by the cumulative percentage of the population. In a perfect income equality, the 25% lowest income of the population would earn 25% of the total income, the 50% lowest income of the population would earn 50% of the total income, hence the Lorenz curve would follow a 45° line. As inequality increases, the Lorenz curve deviates from the line of equality as shown in Figure 2, where we can see that the 20% lowest income of the population is only earning 5% of the total income and the 90% lowest income is only earning 55%.

The Gini coefficient is defined by the area between the equality line and the Lorenz curve divided by the total area below the equality line, that is, $G = A/(A + B)$. It is also equal to $1 - 2B$ due to the fact that $A + B = 0.5$.



Source: Own Author.

Figure 2 – The Lorenz curve framework (hypothetical data).

In the next section we are going to see the continuous definition of the Gini coefficient, useful for when one wants the Gini coefficient determined by the model named theoretical Gini coefficient.

3.1 THE THEORETICAL GINI COEFFICIENT

For a continuous income probability the Lorenz curve $L(F)$ can be represented as a parametric function in $L(m)$ and $F(m)$, where F is the cumulative distribution and m is the income. The value of the area B can be found by integration:

$$B = \int_0^1 L(m) dF(m) = \int_0^\infty L(m) P(m) dm, \quad (3.1)$$

where $P(m)$ is the probability density function and letting μ denote the average income then

$$L(m) = \frac{1}{\mu} \int_0^m x P(x) dx, \quad (3.2)$$

is the percentage of total income by the population with up to income m . Simplifying Equation (3.1) using integration by parts, the Gini coefficient becomes

$$G(C) = 1 - \frac{1}{\mu} \int_0^\infty C(m)^2 dm, \quad (3.3)$$

where μ is the average income of the $P(m)$ distribution and $C(m) = 1 - F(m)$ is the complementary cumulative distribution.

Using the formula above, an exponential distribution has a Gini coefficient of 0.5. Therefore, for the two-class model, the Gini coefficient is a good indicator of how much the Pareto correction affects inequality. Other important property of Gini coefficient in this context is that it can be written as

$$G(\lambda, \alpha) = 1 - \frac{(1 - \lambda^2)/2 - (\lambda^2 \log \lambda)/(2\alpha - 1)}{(1 - \lambda) - \lambda \log \lambda/(\alpha - 1)}, \quad (3.4)$$

and expanding previous equation around $\lambda = 0$, we arrive at

$$G(\lambda, \alpha) = \frac{1}{2} + \left[\frac{\log(\lambda^{-1})}{(\alpha - 1)} - 1 \right] \frac{\lambda}{2} + \mathcal{O}(\lambda^2 \log \lambda). \quad (3.5)$$

Hence, the Gini coefficient only depends on the Pareto index and the percentage of people that belong to the Pareto distribution.

Our model parameters and this theoretical Gini coefficient will be the set of indicators for analysis of inequality.

3.2 THE EMPIRICAL GINI COEFFICIENT UNWEIGHTED

For a discrete income probability without weights the Lorenz curve $L(F)$ a similar parametric representation can be done, but instead of the m parameter we will have the income rank index parameter, $k = 1, 2, \dots, N$, which is the index of the income order statistic $m_{(1)} \geq m_{(2)} \geq \dots m_{(N)}$, thus the k index counts each person in increasing order of income.

In Figure 3 is shown the discrete Lorenz curve, where are observed the trapezoids that compose the area B . So, the value of area B can be determine by the summation of trapezoids

$$B = \sum_{k=1}^N (L_k + L_{k-1}) \frac{\Delta F_k}{2} = \frac{1}{2N} \sum_{i=1}^N (L_k + L_{k-1}), \quad (3.6)$$

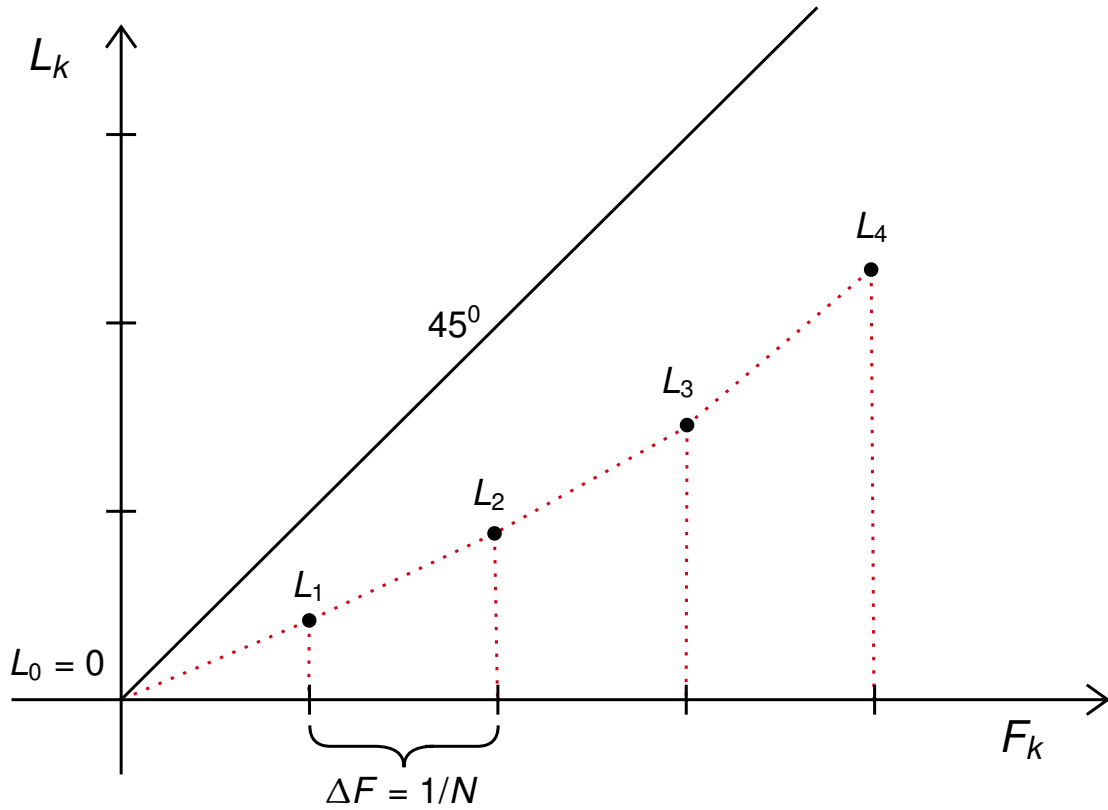
where

$$L_i = \frac{1}{M} \sum_{k=1}^i m_{(k)}, \quad L_0 = 0, \quad (3.7)$$

being $m_{(k)}$ the income order statistic of the k_{th} person and M is the summed income of all populations.

Therefore the empirical Gini coefficient is defined as

$$G = 1 - \frac{1}{N} \sum_{k=1}^N (L_k + L_{k-1}). \quad (3.8)$$



Source: Own Author.

Figure 3 – Discrete Lorenz curve. The points represent the data, the dashed black line represents the line of equality and the dashed red line the boundaries of each trapezoid.

3.3 THE EMPIRICAL GINI COEFFICIENT WITH WEIGHTS

For weighted data, we need to note that ΔF_k in Equation (3.6) represents the normalized weights that were previously $1/N$ for every k . Additionally, we have to consider the weighted Lorenz curve $L_i = \frac{\sum_{k=1}^i w_k m_{(k)}}{\sum_{k=1}^N w_k m_{(k)}}$. Thus, Equation (3.8) becomes:

$$G = 1 - \sum_{k=1}^N w_k (L_k + L_{k-1}). \quad (3.9)$$

Now that we have defined both empirical and theoretical Gini coefficients, we can analyze how well they correlate with each other. This will be investigated in Chapter 7. In the next chapter, we will define how to implement the hybrid PSO for our use case.

4 HYBRID-PSO

In this chapter we are going to detail the procedures to perform the Particle Swarm Optimization with L-BFGS-B to fit the two-class model. This chapter will focus on the inner workings of the search algorithm, while the next chapter will define the loss functions that this algorithm aims to minimize.

4.1 PARTICLE SWARM OPTIMIZATION

The PSO is a computational method that optimizes a problem by iteratively trying to improve a set of candidates in the parameter space (in our case $\{\lambda, \alpha, T\}$) with regard to a given measure of quality (KENNEDY; EBERHART, 1995). Let $\{\mathbf{x}_t^i = (\lambda_t^i, \alpha_t^i, T_t^i)\}$ with $i = 1, 2, \dots, N_c$, where N_c is the number of candidates of parameters, be our set of candidate vectors in the t_{th} step. Each candidate is treated as a solution of the problem. Thus the optimization will be derived by the search of the parameter space with N_c candidates. The best solution will minimize the quality measure.

PSO was chosen for being able to work with a non-differentiable and non-convex error function (SHI; EBERHART, 1998a), which is needed since the two-class loss function is not differentiable.

4.2 SEARCH SPACE

Now, a defined search space is needed for the PSO to find a solution. As discussed at the end of Chapter 2, to properly determine the crossover region, the use of the empirical crossover percentage $p = C'(m_c)$ is needed; that is, the empirical CCDF evaluated at the income m_c . Moreover, the search space parameter vector is defined as $\{C'(m_c), \alpha, T\}$. Considering these search parameters, to define a two-class model, we need to retrieve λ , which is done by using Equation (2.10), where $m_c = (C')^{-1}(p)$ and C' is the interpolated empirical cumulative distribution.

Another information that needs to be provided, in order to define the search space is its range. For our case $T \in [\frac{\langle m \rangle}{2}, 2\langle m \rangle]$, $\alpha \in [1, 3]$ and $C'(m_c) \in (0, 0.2]$. The T range is chosen to ensure the threshold parameter remains proportional to the mean income $\langle m \rangle$, providing reasonable bounds for the distribution's support. The α and $C'(m_c)$ ranges are set to be less restrictive than typical empirical observations ($\alpha \in [1, 2]$ and $C'(m_c) \in [0.02, 0.15]$ for two-class models), allowing the optimizer greater flexibility while maintaining physically meaningful parameter bounds. Therefore the search space is a cuboid in the parameter coordinates. The next step is to define the optimizer that will minimize a loss function.

4.2.1 Adaptive Topology

In the first iteration each candidate index randomly choose K times to be part of sets $S_0 \in \{S_0^i\}_{i \in \{1,2,\dots,N_c\}}$, this process defines the initial sets of neighbors of all candidates. S_{t-1}^i is the neighbors set of the i_{th} candidate in the t_{th} step and it contains the neighbors index and its own index (i). One candidate index can randomly choose to participate in a set repeatedly times and duplicated indexes are removed, thus S^i size may vary averaging a little less than $K + 1$. These sets of neighbors are redefined by the same random process in each step that the algorithm didn't improve the best solution between the history of all the candidates.

This sets S_t^i containing randomly chosen candidate indexes are used to inform the personal best solution of those to the i_{th} candidate. The personal best solution will be defined in the next section.

The exploration will use sets of neighbors to compose the next step of each candidate and it will be clarified in the next subsection. 2007 Standard for Particle Swarm Optimization (SPSO2007) value of $K = 3$ and will be used for this study.

This process is also known as adaptive topology and it is exemplified as follow:

Ex:

For $K = 2$, $t = 0$ and candidates 1, 2 and 3:

- candidate 1 randomly chose to participate in the neighbors sets of candidate 2 and 3.
- candidate 2 randomly chose to participate in the neighbors sets of candidate 1 and 3.
- candidate 3 randomly chose to participate in the neighbors sets of candidate 1 and 1 again.

So the sets were defined as follows:

$$\rightarrow S_0^1 = \{1, 2, 3\}$$

$$\rightarrow S_0^2 = \{1, 2\}$$

$$\rightarrow S_0^3 = \{1, 2, 3\}$$

4.2.2 PSO Steps

The exploration is done by each candidate making steps that are influenced by the direction to the best neighbors position (the best position between its set of neighbors and itself), the candidate best position in its step history and its last step direction. For each candidate i , the step t is determined by:

$$\mathbf{v}_t^i = W\mathbf{v}_{t-1}^i + c_1\Lambda_1(\mathbf{P}_{t-1}^i - \mathbf{x}_{t-1}^i) + c_2\Lambda_2(\mathbf{G}_{t-1}^i - \mathbf{x}_{t-1}^i) \quad (4.1)$$

$$\mathbf{x}_t^i = \mathbf{x}_{t-1}^i + \mathbf{v}_t^i, \quad (4.2)$$

where $\mathbf{x}$ is the vector position, $\mathbf{v}$ is analog to the velocity,

$$\mathbf{P}_t^i = \{\mathbf{x} \in \{\mathbf{x}_{\nu}^i\} : \text{Loss}(\mathbf{x}) = \min_{\mathbf{z} \in \{\mathbf{x}_{\nu}^i\}} \text{Loss}(\mathbf{z})\} \quad (4.3)$$

with $\nu = 0, 1, \dots, t$ is the personal best in regard to the quality measure,

$$\mathbf{G}_t^i = \{\mathbf{x} \in C_t^i : \text{Loss}(\mathbf{x}) = \min_{\mathbf{z} \in C_t^i} \text{Loss}(\mathbf{z})\} \quad (4.4)$$

$C_t^i \equiv \{\mathbf{P}_t^j\}_{j \in \{S_t^i\}}$ is the set of personal best of neighbors set of the i_{th} candidate in the t_{th} step and $\mathbf{G}_t^i$ is the best position in the C_t^i set. The parameters Λ_1 and Λ_2 are uniformly random variables with range $[0, 1]$, introducing stochastic behavior that prevents premature convergence. The W , c_1 and c_2 are hyperparameters with predetermined values. We set $c_1 = c_2 = 1.7$ to balance the cognitive component (particle's attraction to its personal best position) and social component (particle's attraction to the neighborhood's

best position). The inertia weight W decreases linearly from 0.7 to 0.4 over iterations (SHI; EBERHART, 1998b), promoting exploration in early stages and exploitation during convergence.

4.2.3 Initialization

After define the neighbors sets the initialization is done as follows:

$$\begin{cases} \mathbf{x}_0^i = \mathbf{U}_T \\ \mathbf{v}_0^i = \frac{\mathbf{U}_T - \mathbf{x}_0^i}{2} \\ \mathbf{p}_0^i = \mathbf{x}_0^i \\ \mathbf{G}_0^i = \{\mathbf{x} \in \{\mathbf{p}_0^j\}_{j \in \{S_0^i\}} : \text{Loss}(\mathbf{x}) = \min_{\mathbf{z} \in \{\mathbf{p}_0^j\}} \text{Loss}(\mathbf{z})\} \end{cases}, \quad (4.5)$$

where $\mathbf{U}_T$ is a random vector inside the cuboid search space drawn according to the uniform distribution.

A simplified pseudo algorithm can be found in the Appendix A. For each candidate the 9th line apply a condition statement that detects if the step taken in the 8th had a improvement from the personal best and then in the 10th line is done a local search by the L-BFGS-B algorithm. Thus the local search is only done when the PSO algorithm finds a new personal best for a given candidate.

4.3 L-BFGS-B

The standard PSO does not always converge to a good solution, so a hybrid approach was used (LI et al., 2011). This hybrid approach utilizes L-BFGS-B steps to improve the candidates local search ability (exploitation). Even though BFGS is a quasi-Newton method it can be used for nonsmooth optimization (SKAJAA, 2010; GUO; LEWIS, 2018). First, will be given a description of the BFGS algorithm, since the L-BFGS-B version can easily be expanded upon.

The BFGS algorithm is one the most successful of the quasi-Newton methods and is often applied to optimize on top of another method enhancing the local search capabilities.

For the BFGS, given a function $\text{Loss}(\mathbf{x})$ such that $\mathbf{x} \in \mathbb{R}^n$ the algorithm minimizes $\text{Loss}(\mathbf{x})$ iteratively by a sequence of t steps $\mathbf{x}_t$ such that

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t \mathbf{d}_t \quad (4.6)$$

where $\mathbf{d}_t$ is a vector that gives the search direction and α_t is the learning rate. For each step t , $\mathbf{d}_t$ is obtained by

$$\mathbf{d}_t = -\mathbf{H}_t^{-1} \nabla \text{Loss}(\mathbf{x}_t) \quad (4.7)$$

where H_t^{-1} is inverse Hessian at the t_{th} step. In our case the gradients of the Loss were determined numerically instead, a procedure that was explored in (BRÄUNINGER, 1981) for the feasibility as well as the convergence.

Since BFGS is quasi-Newton method the determination of the Hessian is done iteratively using the secant equation

$$\nabla \text{Loss}(\mathbf{x}_{t+1}) \approx \nabla \text{Loss}(\mathbf{x}_t) + H_{t+1}(\mathbf{x}_{t+1} - \mathbf{x}_t), \quad (4.8)$$

then the rank-2 approximation (BROYDEN, 1970), which is used specifically for the BFGS, is of the form

$$H_{t+1} = H_t + \frac{q_t q_t^T}{q_t^T p_t} - \frac{H_t p_t p_t^T H_t^T}{p_t^T H_t^T p_t}. \quad (4.9)$$

where $p_t = \mathbf{x}_{t+1} - \mathbf{x}_t$ and $q_t = \nabla \text{Loss}(\mathbf{x}_{t+1}) - \nabla \text{Loss}(\mathbf{x}_t)$.

However, in order to calculate Equation (4.7) we are interested in the inverse of Hessian, then using Sherman-Morrison-formula and doing some algebra we get

$$H_{t+1}^{-1} = \left(I - \frac{p_t q_t^T}{q_t^T p_t} \right) H_t^{-1} \left(I - \frac{q_t p_t^T}{q_t^T p_t} \right) + \frac{p_t p_t^T}{q_t^T p_t}. \quad (4.10)$$

Now we need to define α_t . If we were to calculate the Hessian exactly α_t would be 1, so for this case this parameter is dismissed. However, since we are approximating the Hessian we are going to only rely on its direction d_t , and then use the Golden ratio linear search to find the optima in that given direction. We can see a summarization of this algorithm in the Appendix B.

Limited-memory-BFGS or L-BFGS is a version the of BFGS that sacrifice a time and accuracy for a less taxing memory usage (NOCEDAL, 1980). This is done by storing the vectors $\{p_t, q_t\}_{t \in 1, 2, \dots, r}$ and calculating the inverse Hessian iterating the Equation (4.10) r times.

The L-BFGS-B was used since it handles the cuboid constraint on the vector $\mathbf{x}$ (BYRD et al., 1995). The method works by identifying fixed and free components at every step (using a gradient method), and then using the L-BFGS method on the free components only to get higher accuracy, and then repeating the process.

Using this algorithm in conjunction with PSO makes it so that we can gain exploitation and the consistency of a quasi-Newton method with the ability of the exploration done by the PSO steps, allowing us to minimize our chances of being stuck at a local minimum.

In summary, to determine the model parameters discussed in Section 2.3, we first need to calculate the cumulative probability distribution of income for the Brazilian population (2.4). Then, we use PSO-BFGS to fit the model by minimizing a loss function. The next chapter will define two types of loss functions used for the regression.

5 LOSS FUNCTION

In this chapter, we will derive the loss functions used in PSO to perform regression. The chapter is divided into two parts: RMSLE and the MLE approach. Both methods we will develop considering weighted data, since we can apply to non-weighted data setting weights to 1.

5.1 TYPE OF LOSS FUNCTION

First, we will examine the RMSLE, which measures the average squared difference between observed and predicted values. Then, we will explore the MLE approach, used for parameter estimation to maximize the likelihood of observed data.

5.1.1 Mean Squared Error

We will start with the mean squared error approach. In the literature this is done separately for exponential and Pareto regions, fitted by minimizing a goodness-of-fit function after passing through a logarithmic transform (TAO et al., 2019; SICILIANI; TRAGTENBERG, 2019).

Before specifying our loss function we will define two regularization terms to ensure meaningful values of T and λ . So taking the theoretical income average for the exponential regime

$$\langle \hat{m} \rangle_{exp} = T - e^{-\frac{m_c}{T}} (m_c + T) \quad (5.1)$$

then the regularization being added to the loss function to ensure the equality between the theoretical and empirical average for the exponential region is of the form

$$l_1 \stackrel{\text{def}}{=} \left| \frac{T - e^{-\frac{m_c}{T}} (m_c + T)}{\frac{1}{N_e} \sum_{i=1}^{n_e} w_{(i)} m_{(i)}} - 1 \right|, \quad (5.2)$$

where $N_e = \sum_{i=1}^{n_e} w_{(i)}$ and n_e is the greatest rank statistic (index of the ordered income) belonging to the exponential region. Therefore, the parameter T that comes out of the process of optimization with this regularization can be interpreted as the estimation of the average income of the data, in the hypothetical case of an exponential distribution without any Pareto tail.

In parallel, note that can be a difference between the $\lambda = \hat{C}'(m_c)$ and $C(m_c)$, which are the model top percentage and the data percentage of people gaining more than m_c , respectively. This is addressed by the second regularization term

$$l_2 \stackrel{\text{def}}{=} \left| \frac{\lambda}{C'(m_c)} - 1 \right|. \quad (5.3)$$

Thus, the parameter λ will always be equivalent to the percentage of people that have the power-law behavior given by the data.

Let $\eta_n \geq \eta_{n-1} \dots \geq \eta_1$ be an order statistic that follows $\eta_n = m_{(\lfloor \frac{nN}{k} \rfloor)}^1$ where $n = 1, 2, \dots, k$ and $k < N$. This will be called class statistic and it divides the data into k income points.

Now we can define the measure of quality as a RMSLE between the data and the model using the η_n statistics

$$\text{RMSLE}(C'(\eta_n), C(\eta_n, \mathbf{x})) = \sqrt{\frac{1}{k-1} \sum_{n=1}^{k-1} \left\{ \log[C'(\eta_n)] - \log[C(\eta_n, \mathbf{x})] \right\}^2}, \quad (5.4)$$

where $C'(\eta_n)$ is the interpolated empirical complementary cumulative distribution, $C(\eta_n, \mathbf{x})$ is the CCDF of the model and $\mathbf{x}$ is the parameter vector. Using the class statistic not only helps with the computational load, but also gives consistency to the CCDF precision used in the loss function. With $m_{(n)}$ statistics the precision of the CCDF was tied to the number of the population, having N points. The set η_n instead has $k - 1$ points, which is independent on the population number. Note that the η_k was excluded from the loss function since $\log[C(\eta_k)]$ diverges. In this work $k = 10000$.

Therefore using Equations (5.2), (5.3) and (5.4) we define the loss function

$$\text{Loss}(\mathbf{x}) = \text{RMSLE}(C'(\eta_n), \hat{C}(\eta_n, \mathbf{x})) + l_1 + l_2. \quad (5.5)$$

5.1.2 Maximum Likelihood

Other type of loss function that will be used is the MLE (MYUNG, 2003). This statistical approach determines the parameters of a predefined probability distribution based on empirical data. Maximizing a likelihood function implies that, for an statistical model, the selected parameters gives the highest probability of encountering the empirical data. This method's ability to be executed numerically makes for a good choice to fit jointed probability distribution.

Given an sample data $m_1, m_2, \dots, m_n$ from a population, a parametric density function $P(m; \vec{\theta})$, where θ is the parameters, the probability of the data being collected by this probability function is

$$\mathcal{L}(m; \vec{\theta}) = \prod_{i=1}^n P(m_i; \vec{\theta}) \quad (5.6)$$

This joint probability of the dataset is know as likelihood function. The formula presented above assumes that the sampling process was conducted independently and using simple random sampling without any weighting applied.

¹ The notation $\lfloor \cdot \rfloor$ denotes the floor function, which can be formally define as $\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \leq x\}$

We proceed to determine the model parameters values maximizing the likelihood function across the parameter space. Thus, considering a set of probability density $\{P(m; \vec{\theta}) : \vec{\theta} \in \vec{\Theta}\}$, where $\vec{\Theta}$ represents the parameter space, we aim to maximize Equation (5.6) searching the best-fitting probability density. The optimal parameters can be express as

$$\hat{\vec{\theta}} = \underset{\vec{\theta} \in \vec{\Theta}}{\operatorname{argmax}} \mathcal{L}(m; \vec{\theta}) \quad (5.7)$$

In the context of weighted data, the equation denoted by (5.6) can be expressed as follows:

$$\mathcal{L}(m; \vec{\theta}) = \prod_{i=1}^n P(m_i; \vec{\theta})^{w_i}, \quad (5.8)$$

upon taking the logarithm of both sides, we obtain

$$\log[\mathcal{L}(m; \vec{\theta})] = \sum_{i=1}^n w_i \log[P(m_i; \vec{\theta})]. \quad (5.9)$$

The equation above is referred to as the weighted logarithm likelihood function (MARKATOU et al., 1997), which is a crucial tool for optimizing parameters. When this equation is applied to the statistical model outlined in Equation (2.6), it yields the subsequent set of equations:

$$\hat{\alpha} = \frac{\hat{n}_2}{\sum_{i=\hat{k}}^N w_{(i)} \log \frac{m_{(i)}}{\hat{m}_c}}, \quad (5.10)$$

$$\hat{T} = \sum_{i=1}^{\hat{k}-1} w_{(i)} \frac{m_{(i)}}{\hat{n}_1} + \frac{\hat{n}_2}{\hat{n}_1} \hat{m}_c, \quad (5.11)$$

where

$$\hat{n}_1 = \sum_{i=1}^{\hat{k}-1} w_{(i)}, \quad \hat{n}_2 = \sum_{i=\hat{k}}^N w_{(i)}, \quad (5.12)$$

and considering income-ordered data, $\hat{k}$ is the rank number that defines the crossover between the two regions. The equations (5.10) and (5.11) will be used as quadratic regularization together with the equation $\lambda \cong C(m_c)$, which indicates that the empirical percentage of people in the Pareto distribution approximates its theoretical counterpart. This approach defines the loss function as:

$$\text{loss}(m; \vec{\theta}) = \sum_{i=1}^n w_i \log[P(m_i; T, \alpha, \lambda)^{-1}] + l_1 + l_2 + l_3, \quad (5.13)$$

$$l_1 \stackrel{\text{def}}{=} \left(1 - \frac{T}{\sum_{i=1}^{\hat{k}-1} \frac{w_{(i)} m_{(i)}}{\hat{n}_1} + \frac{\hat{n}_2}{\hat{n}_1} m_c} \right)^2 \quad (5.14)$$

$$l_2 \stackrel{\text{def}}{=} \left[1 - \frac{\alpha \sum_{i=\hat{k}}^N w_{(i)} \log\left(\frac{m_{(i)}}{m_c}\right)}{\hat{n}_2} \right]^2 \quad (5.15)$$

$$l_3 \stackrel{\text{def}}{=} \left(\frac{\lambda}{C(m_c)} - 1 \right)^2. \quad (5.16)$$

With the loss functions and optimization framework defined, we can now apply them to the data. However, before doing so, we'll validate the process using the inverse cumulative function. This allows us to create data with known parameters, which we can then use to calculate the errors in our parameter estimations and assess the efficacy of our regression models.

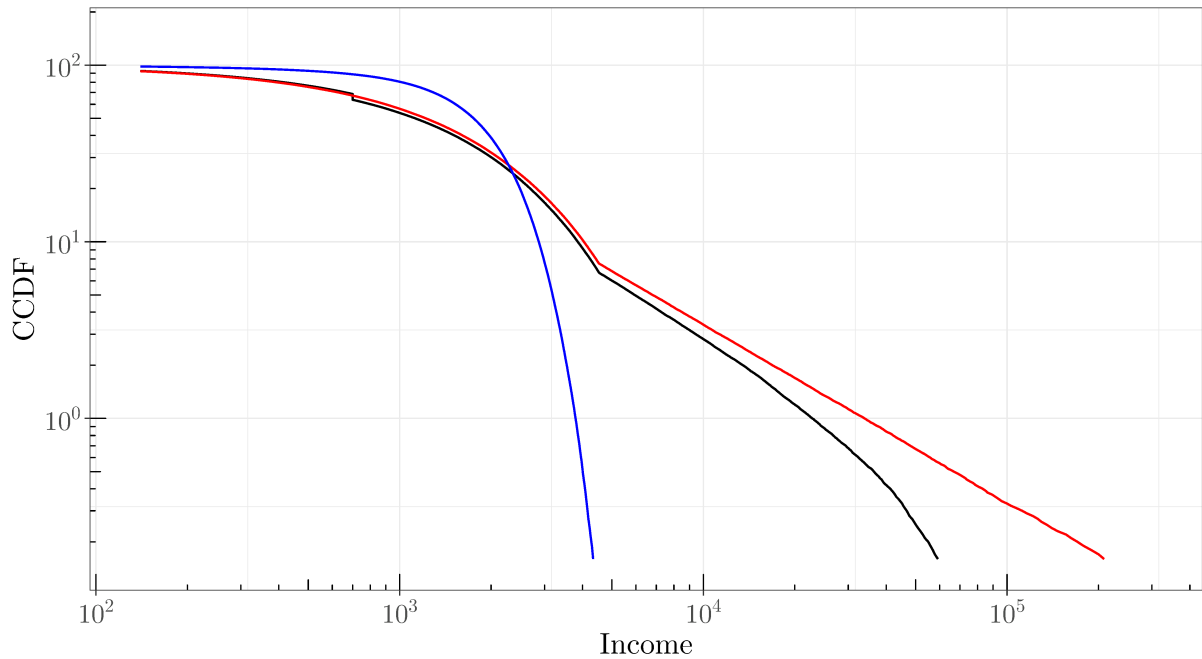
5.2 VALIDATION USING GENERATED DATA FROM INVERSE TRANSFORM SAMPLING

In order to assess the efficacy of the regression models developed in the previous section, we conducted a simulation analysis. This analysis involved generating data through inverse transform sampling, while introducing impurities using the simframe framework (ALFONS et al., 2010). Subsequently, we trained both models on a set of generated datas, and compared them based on metrics like mean squared error and the Kolmogorov-Smirnov statistic. Moreover, given that we have access to the true values of the parameters due to data generation, we also compared the errors in the parameters. For simplicity, this section will use unweighted data.

Firstly, the inverse transform sampling of the two-class distribution was employed to generate the datasets. After data generation, variability was introduced through a 1% normal distribution impurity, 0.1% non-assignment impurity, and a 5% minimum wage effect. This implies that 1% of the data was randomly replaced with values from a normal distribution (mean T and standard deviation $T/2$), 0.1% of the data was replaced with missing values to simulate incomplete data, and 5% of the data was

randomly chosen to be replaced by the minimum wage. An upper income limit was imposed due to the restricted availability of data at the highest income levels for the Brazilian survey data (MORGAN, 2017). These impurity levels were selected to balance realism with statistical detectability, introducing plausible data artifacts without substantially altering the underlying distribution or producing excessively high Kolmogorov-Smirnov test statistics to the point of undermining the desired distribution.

The comparison between the normal distribution used to introduce impurity, the original two-class model distribution, and the distribution after applying all the mentioned impurities is shown in Figure 4. The figure highlights the effects of the impurities applied to the simulated data, demonstrating how these modifications alter the original two-class distribution. The imposed upper income limit is clearly visible in the final distribution, which closely mirrors the characteristics of the Brazilian survey data at higher income levels.



Source: Own Author.

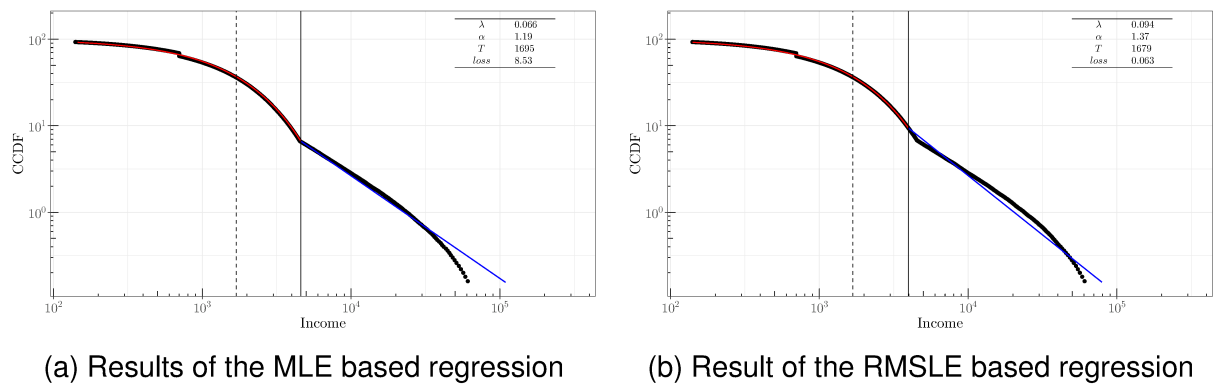
Figure 4 – Simulated data with impurities on a cumulative probability on a log-log scale. The black line shows the final two-class distribution with all impurities, using parameters $\lambda = 0.075$, $\alpha = 1.0$, and $T = 1750$. The red line represents the two-class distribution without impurities, while the blue line depicts the normal distribution used for the 1% impurity.

To define the grid of parameter values used in the simulation, we consider the following sets for each parameter: Let $\Lambda = \{0.15, 0.10, 0.05, 0.01\}$ be the set of λ values, $A = \{1.2, 1.4, 1.6, 1.8\}$ be the set of α values, and $\mathcal{T} = \{1500, 1750, 2000\}$ be the set of Temperature values.

The simulation iterates over all combinations of $\lambda \in \Lambda$, $\alpha \in A$, and $T \in \mathcal{T}$, where:

$$\text{Parameter grid} = \{(\lambda, \alpha, T) \mid \lambda \in \Lambda, \alpha \in A, T \in \mathcal{T}\} \quad (5.17)$$

We now will apply the simulation to the losses defined in the previous section, one based on RMSLE (5.5), while the new method employed a loss function based on MLE (5.13). In the Figure 5 we can see the results of one of this regression in the simulated data, that follow $\lambda = 0.075$, $\alpha = 1.0$ and $T = 1750$ and the upper income limit is set as the 95% quantile of the values within the Pareto region.



Source: Own Author.

Figure 5 – Two-class regression with impurities for the simulated data cumulative probability on a log-log scale with parameters $\lambda = 0.075$, $\alpha = 1.0$ and $T = 1750$. The black points represent the cumulative distribution of the simulated data CCDF and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve. The temperature is highlighted by the dashed vertical line and the crossover income is represented by the solid vertical line, which separates the Boltzmann-Gibbs and Pareto regions of income. The table in the top right corner contains information on both the parameters and the value of the regression's loss function.

Table 1 – Comparative statistics resulting from the regression grid run using both the RMSLE and MLE methods. The table includes mean percentage error for temperature (in R\$), top-percentage (%), and Pareto index. Additionally, regression statistics such as Kolmogorov-Smirnov and Mean Least Square values are presented.

Parameters	Coefficient of variation	
	MLE	RMSLE
Temperature (R\$)	0.039	0.039
Top-percentage (%)	0.14	0.24
Pareto index	0.23	0.30
Regression Statistics		
	MLE	RMSLE
Kolmogorov-Smirnov statistic	0.0215	0.0215
Mean Least Square	0.0002	0.0002

The Table 1 encapsulates the outcomes of the simulation study that involved the comparison of two regression methods. The table provides the coefficients of variation for the parameters, as well as regression statistics. The results favor the new method, showing lower percentage error rates and comparable performance in terms of the Kolmogorov-Smirnov Statistic (KS) and MSE.

Furthermore, the novel method not only yields better regression outcomes but is also independent of the binning process. This is particularly noteworthy, as the binning process, as illustrated in Equation (5.5), can lead to varied results.

6 REGRESSION ON THE PNAD SIMPLE SAMPLING DATA AND CROSS-VALIDATION

In this chapter, we will perform cross-validation using the bootstrap method to test the model's accuracy (EFRON; TIBSHIRANI, 1997). Both this chapter and the following one will use non-weighted data and the RMSLE loss function approach, since is the approach used in our first article (SANTOS et al., 2022). As mentioned above, the bootstrap sampling is performed before calculating the cumulative distribution, that is, picking a random sample of the income data, with replacement, then calculating the cumulative distribution. This approach not only will give us the ability to estimate model parameters errors as well as will allow us to do an out of the bootstrap cross-validation (EFRON, 1983). On average, random sampling with replacement includes $(1 - e^{-1})\% \approx 63.2\%$ of the original data in each bootstrap set of samples and hence the rest allow us to define out-of-sample test sets.

In this context, firstly, we define a pair of sets by bootstrap sampling: training set and test set. Then, with the CCDF of the training set, we use the PSO-BFGS to find the optimal parameters. Lastly, using the optimal parameters evaluated above, we calculate the RMSLE of each set. Lastly, we calculate the RMSLE of each set using the optimal parameters evaluated above. Repeat this process for R pairs of training and test sets, in this study, we have chosen $R = 1000$. This procedure gives us the ability to see how well our model can fit the CCDF of an income data set, which came from the same distribution of the training set but was not in the training process.

6.1 CONTINUOUS NATIONAL HOUSEHOLD SAMPLE SURVEY DATA OF 2019

We utilized data from the 2019 PNADc, made available by the IBGE since 2012. The PNADc is a research initiative that gathers information on various social characteristics in Brazil, encompassing aspects such as labor, income, and highest education attained. From the PNADc data, we extracted details on the total monthly income, gender, and color/race of each individual for the year 2019. Individuals without income and those with missing values were excluded, focusing on the economically active group. This preprocessing yields a final sample of $N = 266,231$ records. For a comprehensive overview of the dataset, please consult Appendix C.

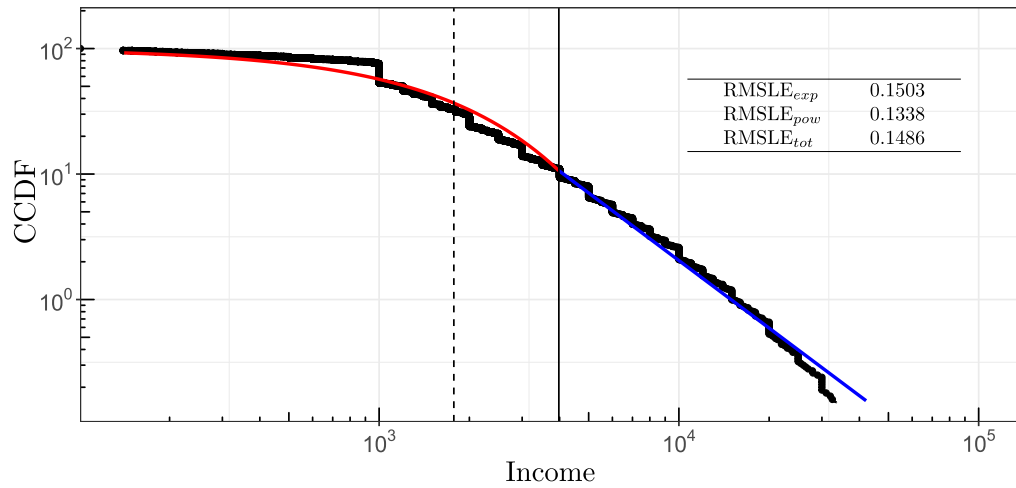
After training the parameters of our model, in order to complete our set of indicators, we calculate the theoretical Gini coefficient following Section 3.1.

As shown in Table 2 the mean RMSLE of the test set is close to the training set RMSLE, so the model has little bias. The parameters of the model have small Coefficient of Variation (CV), and the highest is that of the top-percentage. This is due to the discontinuity of the empirical income data at the crossover.

Table 2 – Mean, standard deviation, 95% confidence interval and coefficient of variation of the model parameters, Theoretical Gini coefficient, and RMSLE of the training and test sets.

2019 Data				
	<i>mean</i>	σ	95% <i>CI</i>	<i>CV</i>
Crossover Income (R\$)	3977	42	3900 // 4000	1.06×10^{-2}
Top-percentage (%)	10.64	0.26	10.42 // 11.17	2.45×10^{-2}
Temperature (R\$)	1775	3.40	1769 // 1782	1.92×10^{-3}
Pareto index	1.789	0.011	1.767 // 1.782	6.04×10^{-3}
Theoretical Gini coefficient	0.578	0.001	0.576 // 0.581	2.31×10^{-3}
Train set RMSLE	0.1486	0.0008	0.1470 // 0.1501	5.31×10^{-3}
Test set RMSLE	0.1489	0.0017	0.1458 // 0.1525	1.12×10^{-2}

Figure 6 displays the cumulative distribution fitted with our model with the parameters shown in Table 2. These results were estimated by bootstrapping the data, calculating the cumulative distribution for each bootstrap set and then fitting the model numerically with PSO. Analyzing the RMSLE between the model with estimates of the parameters and the original data set in parts, we can identify that the Pareto region have a greater error, which is expected since Pareto part does not capture the top 0.01% very well, and since the error Equation (5.5) is logarithmic, outliers whose values are close to 0 are amplified. In the exponential part, a fraction of the error is due to the minimum wage effect (low income discontinuity at $m = \text{R\$ } 998$) as can be seen in Figure 6.



Source: Own Author.

Figure 6 – Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of the data from 2019 PNAD and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve. The temperature is highlighted by the dashed vertical line, at value R\$ 1775 ± 3 , and the crossover income is represented by the solid vertical line at $m_c = \text{R\$ } 3977 \pm 42$, which separates the Boltzmann-Gibbs and Pareto regions of income. Therefore $\lambda = (10.64 \pm 0.26)\%$ is the top income percentage of people which obeys a Pareto power-law with index $\alpha = 1.789 \pm 0.011$. Also, the RMSLE of the original data set calculated in each part of the distribution and total can be found in the top right table.

Comparing this method with a crossover income being determined with a fixed proportion for the Pareto tail, $\lambda = 5\%$, as seen in reference (SICILIANI; TRAGTENBERG, 2019), we get a temperature equals to the mean, $T = 2067 \pm 7$, a Pareto index of 1.816 ± 0.006 and a $m_c = 6192 \pm 21$. This approach results in a discontinuity between BG and Pareto regions, with training and test errors of 0.2257 ± 0.002 and 0.2258 ± 0.002 , respectively — significantly higher than those with the optimal crossover (Table 2).

6.2 SUBGROUP DATA OF 2019

We applied our model to the data separated into two dichotomies of the population allowing us to compare each indicator and test our model in different subgroup data distributions. The first dichotomy is the division by gender (man and woman) and the second is the division by race/color (Black, Brown and Indigenous Group (BBI) - White and Yellow Group (WY))¹. The results and validation can be seen in the Table 3.

¹ Brown stands for mixed race and yellow stands for asian.

Table 3 – Inequality indicators and RMSLE for a subgroup data.

Subgroup table for 2019 data		
Category: Sex/Gender		
Subgroup	Man	Woman
Crossover Income (R\$)	4930 ± 13	3503 ± 25
Top-percentage (%)	8.97 ± 0.59	9.93 ± 0.16
Temperature (R\$)	2047 ± 6	1517 ± 4
Pareto index	1.72 ± 0.02	1.89 ± 0.02
Theoretical Gini coeff.	0.584 ± 0.002	0.565 ± 0.002
Train set RMSLE	0.155 ± 0.001	0.155 ± 0.001
Test set RMSLE	0.156 ± 0.002	0.155 ± 0.002
Category: Race/Color		
Subgroup	WY	BBI
Crossover Income (R\$)	4987 ± 55	3995 ± 26
Top-percentage (%)	11.76 ± 0.30	6.42 ± 0.12
Temperature (R\$)	2329 ± 7	1455 ± 3
Pareto index	1.74 ± 0.01	2.01 ± 0.02
Theoretical Gini coeff.	0.588 ± 0.002	0.547 ± 0.001
Train set RMSLE	0.137 ± 0.001	0.177 ± 0.001
Test set RMSLE	0.138 ± 0.002	0.178 ± 0.002

Looking at the gender dichotomy, there is a significant difference between the theoretical Gini coefficient. Men income "temperature" is considerably higher than the women one, but their Pareto and top-percentage indicators are lower, meaning that there is a less percentage of man in the Pareto Region, but their inequality in this region is higher. Remember that, for the two-class model, the temperature does not affect the theoretical Gini coefficient, so the properties of the Pareto region completely define the inequality.

Analyzing the color/race dichotomy, we get a higher contrast in their inequality indicators compared to the gender dichotomy. WY has Top-percentage and Temperature considerably higher than BBI. The other indicators have similar values for both groups. BBI has the lowest Top-percentage value, $\lambda = 6.42 \pm 0.12$, and the highest Pareto index, $\alpha = 2.01 \pm 0.02$, from the two dichotomies. BBI has also the lowest Gini coefficient, indicating a lower inequality within this subgroup. The next chapter will use the same approach to derive the time series of the two-class model.

7 TEMPORAL EVOLUTION OF INEQUALITY INDICATORS

As previously mentioned, this chapter will develop the temporal analysis of the two-class model using non-weighted data and the RMSLE loss function approach. The time evolution of inequality indicators, in the context of the two-class model or the lognormal-Pareto model, is a subject of interest in the literature (SICILIANI; TRAGTENBERG, 2019; SILVA, A Christian; YAKOVENKO, 2004a; SOUMA, 2002). However, the crossover income is usually fixed or determined by a log-log graph. In this paper, we introduce a formal approach to determine the temporal evolution of the optimal crossover income.

To be able to analyse the Brazilian inequality over the years, we applied our model to describe the income distribution between 2001 – 2019. For our empirical data we used the PNAD for the years 2001 to 2011, and for 2012 to 2019 we used the PNADc. The PNAD is the predecessor of the PNADc and was discontinued in 2015. We gave priority to PNADc in the years that the two survey programs were running since PNADc gives a broader territorial coverage and larger population sample. We choose the first annual interview of PNADc since it contains the income of all sources. Following the same data cleaning procedure, we neglected people without income and missing values.

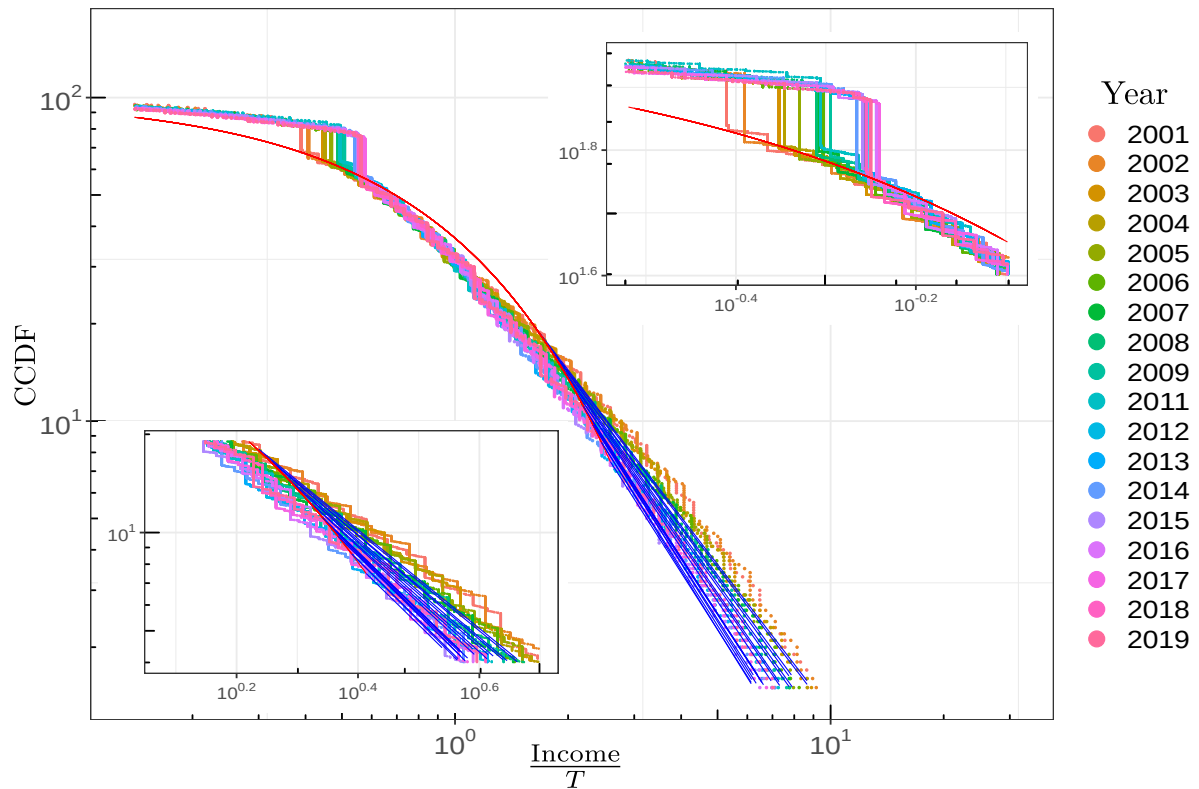
Lastly, we applied the optimization to fit the two-class model, Equation (2.12), to the data of each year. We can access the temporal evolution of each indicator in the Figs. 9a–9d. We can see that every parameter of our model has an interpretable evolutionary trend. The CCDF and their fits can be seen in Figs. 7 and 8.

Compared to the temporal analysis from the US (DRAGULESCU; YAKOVENKO, 2001) our results have less temporal stationarity in the exponential part. This is mainly due to the effect of the minimum wage and can be seen more clearly on the top right subplot of Figure 7. On the same figure in the bottom left subplot, corroborating with US results, the Pareto region is shown to have more temporal variability than the exponential region.

The temperature reflects the income power of the lower to middle class. This parameter was deflated to the 2019 currency using the Extended National Consumer Price Index (IPCA)¹, available by IBGE. According to the Figure 9a there is a trend of an increase in the temperature.

The Pareto index and the top-percentage have a strong anticorrelation as can be seen in the Figure 10a. From 2002 to 2015, there is an increase in the Pareto index and a decrease in the top-percentage. This means that, according to our survey data,

¹ The IPCA is the official inflation index in Brazil, calculated monthly by the IBGE. It reflects the change in prices of a representative basket of goods and services — such as food, housing, transport, and health — consumed by households with monthly incomes between 1 and 40 minimum wages, across 13 metropolitan areas. The IPCA serves as the primary benchmark for inflation targeting by the Central Bank and is also used for indexing wages, rents, and government bonds (IBGE, 2020).



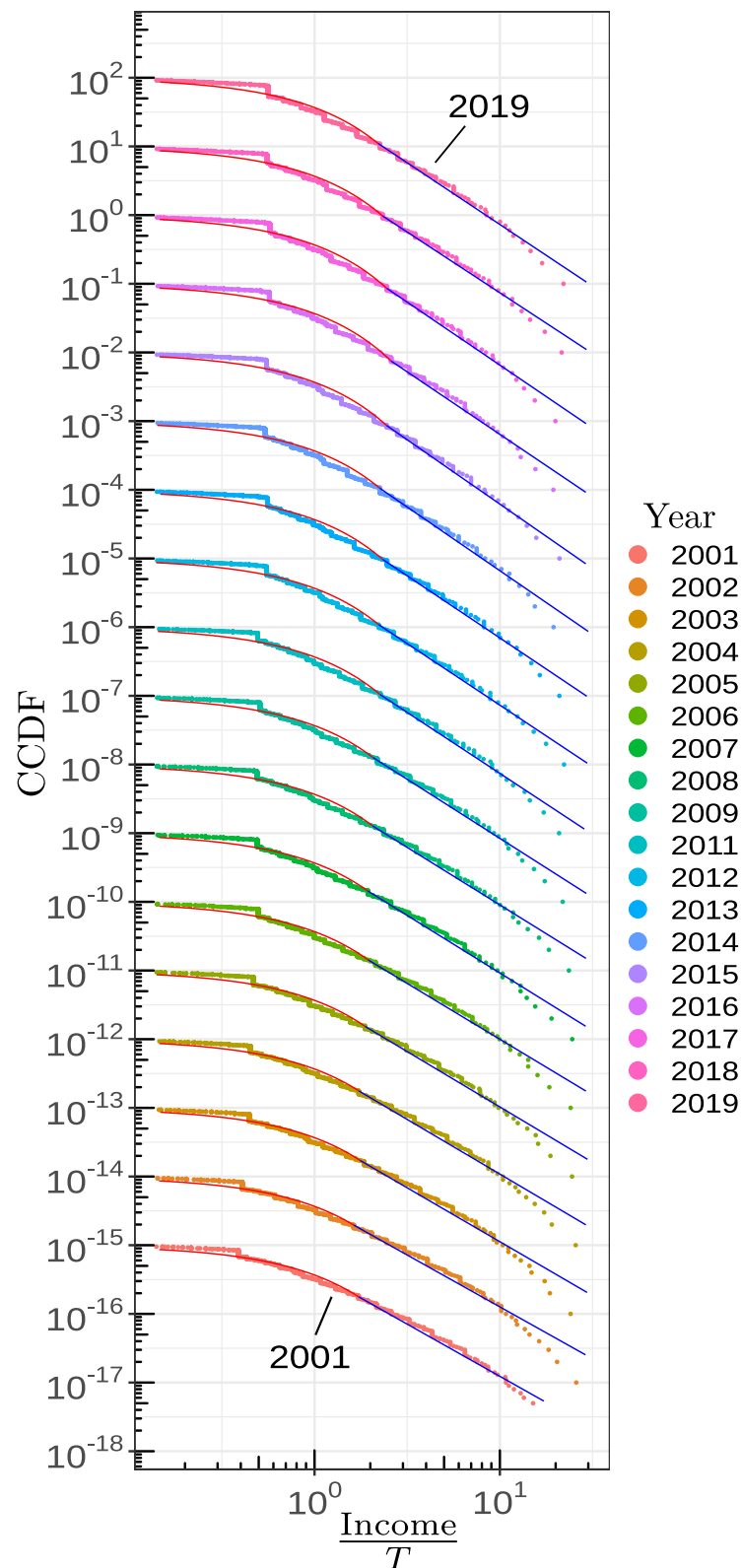
Source: Own Author.

Figure 7 – Cumulative probability distribution of Income/T (income normalized by temperature) on a log-log scale. The colored points represent the cumulative distribution of the data from 2001 to 2019 PNAD and the solid lines correspond to the fitted model described in Equation (2.12). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve.

we are in the presence of an income redistribution process. And it is corroborated by the decrease of the Gini coefficient in this time range, see Figure 9d.

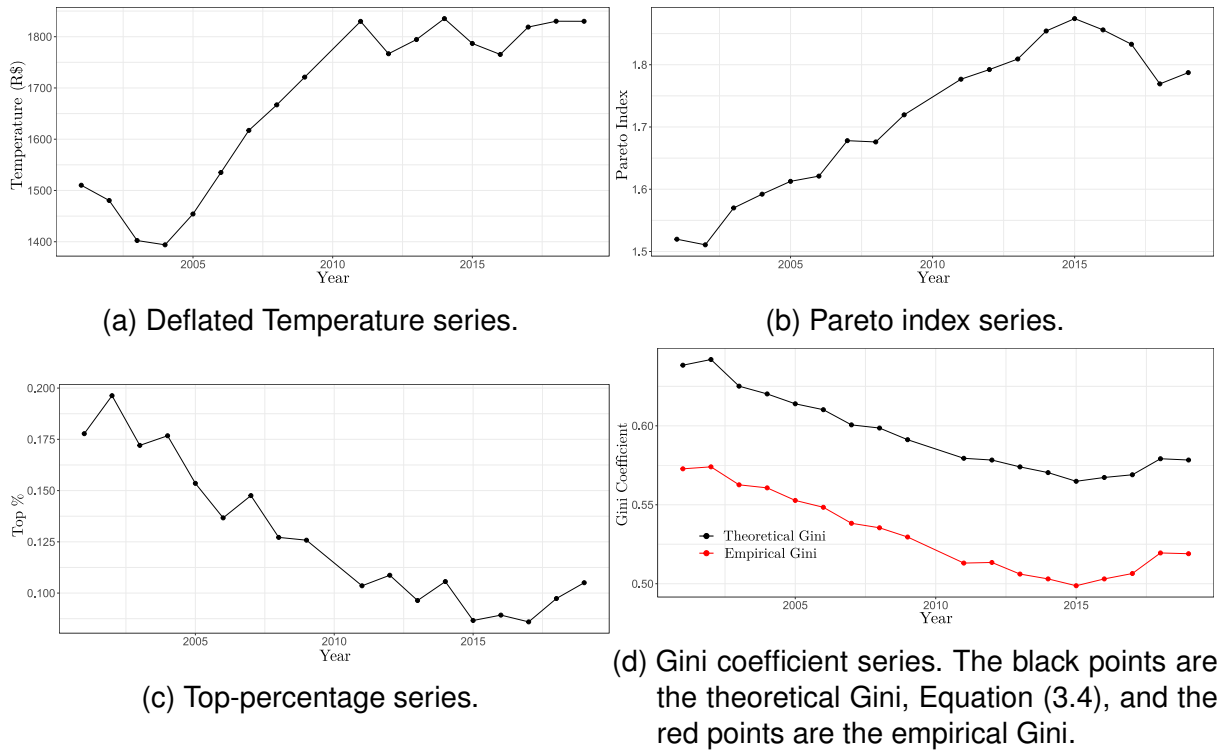
The two-class model reproduces the central part of the distribution but misses two features: the mass at the minimum wage and the instability of the very top tail ($\approx$ top 0.01%), where Pareto behavior breaks down. In fat-tailed settings, the empirical Gini tends to be downward biased and converges very slowly (FONTANARI et al., 2018), whereas parametric approaches (such as RMSLE or MLE-based estimates) can reduce this bias. This helps explain — though not fully account for — why the model estimated Gini is higher than the empirical one: the discrepancy arises less from the model "inflating" inequality than from the empirical estimator understating it under fat tails. Moreover, although Figure 9d shows an apparent big difference between theoretical and empirical Gini coefficients, it can be shown that they are very strongly correlated as can be seen in the Figure 10b.

Figure 10b also shows the affine relation between the theoretical and the empir-



Source: Own Author.

Figure 8 – Cumulative distribution functions constructed from the PNAD data from 2001 to 2019 and their fits with the theoretical distribution described in Equation (2.12), shown in the log-log scale versus the normalized annual income Income/T (income normalized by temperature). Plots for different years are shifted vertically.



Source: Own Author.

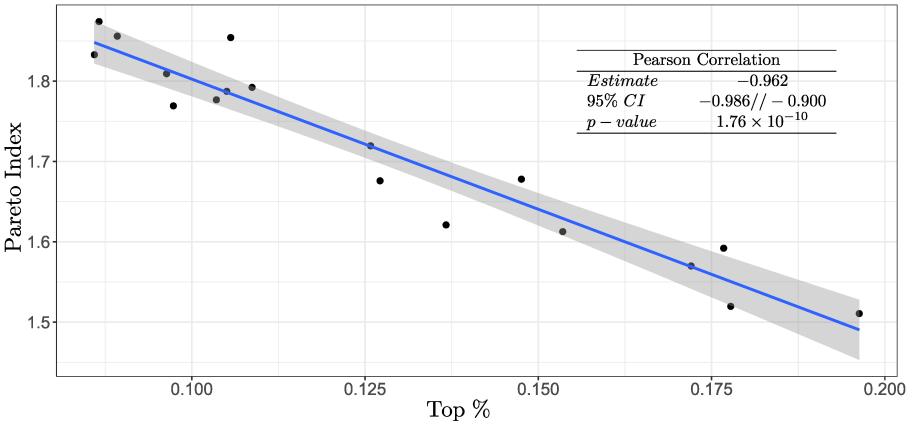
Figure 9 – Time series of the inequality indicators.

ical Gini coefficients, as well as the Pearson correlation. The correlation is $\rho = 0.994$ with a p-value 2.20×10^{-16} , which means that there is high evidence of the strong correlation between the two coefficients. The affine regression $G_e = (-0.07 \pm 0.02) + (1.01 \pm 0.03)G_t$ between theoretical and empirical Gini has a residual standard error of 0.0029. Since the theoretical and empirical Gini coefficients have a strong correlation, we can conclude that the theoretical Gini can also be used as a measure of inequality, if the data can be well explained by the two-class model.

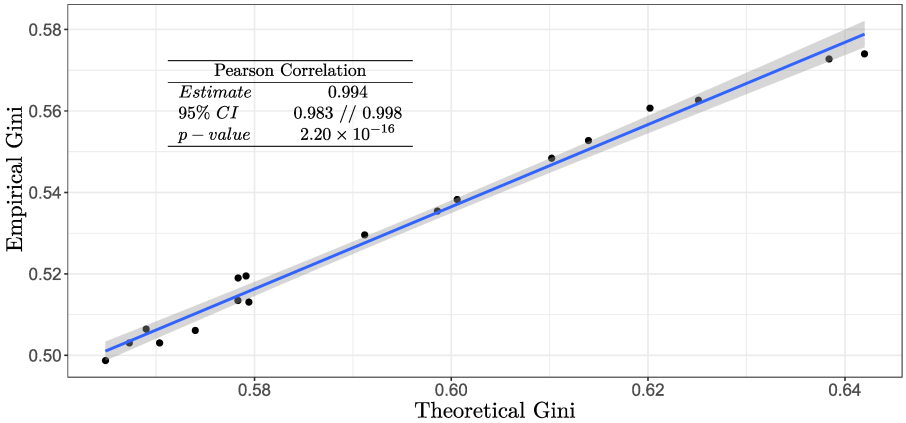
Like the Section 6.2, we also did the same time series analysis with a subgroup data using the two dichotomies described in that section. The subgroup data time series has similar behavior between each subgroup. These time series of the subgroup data also present a similar behavior to the complete data set, as can be seen comparing the subgroup data Figures 11a–11e with the complete data Figures 9a–9d.

Now one can draw the same correlations using the subgroup data. This will give us if these correlations are likely to be general or a specific correlation of the Brazilian time series data. The results of the Pearson correlation are shown in the Table 4.

According to our results, the Empirical and Theoretical Gini coefficients correlation is a really stable correlation, not varying much when switching the subgroup. The anti-correlation between the Pareto index and the top-percentage has more variance and, with exception of the woman subgroup, is weaker for the subgroup data when



(a) Correlation and affine regression with the Pareto index as a function of the top-percentage.

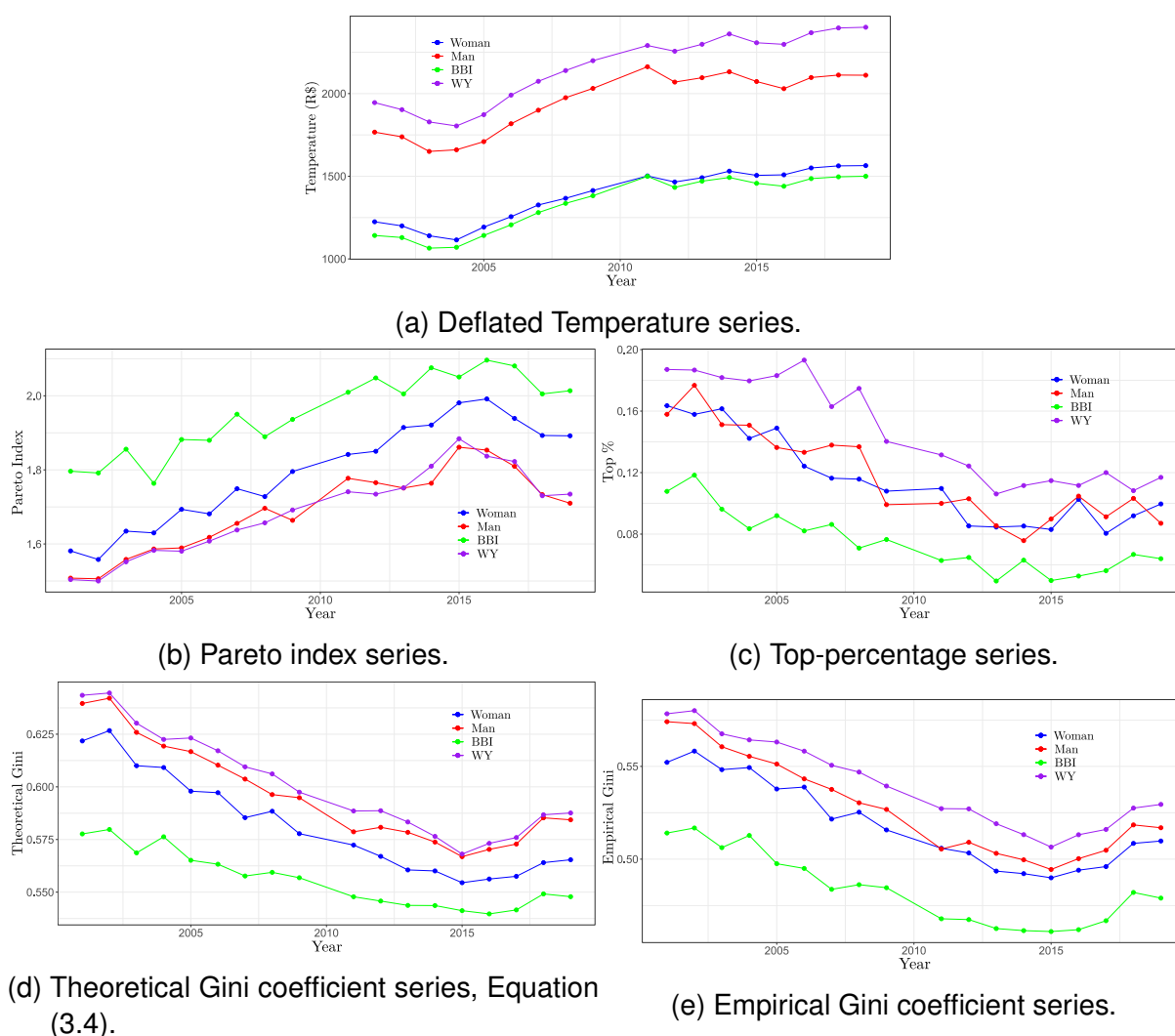


(b) Correlation and affine regression with theoretical Gini as a function of the empirical Gini.

Source: Own Author.

Figure 10 – Correlation and affine regression between indicators.

compared to the original data. The next two chapters will present an analysis using stratified data, resulting in a more robust regression for the two-class model.



Source: Own Author.

Figure 11 – Time series of the inequality indicators for the subgroup data.

Table 4 – Anticorrelation of Pareto index with top-percentage and correlation of the theoretical and empirical Gini coefficients for the subgroup data.

	Pearson Correlation	
	Pareto X Top %	Empirical X Theoretical Gini
Man	-0.858 ± 0.128	0.995 ± 0.025
Woman	-0.922 ± 0.097	0.985 ± 0.043
WY	-0.902 ± 0.108	0.996 ± 0.024
BBI	-0.861 ± 0.127	0.983 ± 0.046

8 REGRESSION ON THE PNAD STRATIFIED SAMPLING DATA

In this chapter, we will conduct regression analysis utilizing weighted data of the PNADc 2019. Additionally, we will compute the variability of each parameter using replicated weights provided by the dataset (KOLENIKOV, 2010).

Stratified sampling is a statistical method that improves the accuracy and representativeness of a sample by dividing a population into homogeneous subpopulations, or strata, based on specific characteristics like race, gender identity, or location. Each member of the population is assigned to exactly one stratum. Researchers then sample each stratum using another probability sampling method, such as simple random sampling or cluster sampling, to estimate statistical measures accurately for each subpopulation.

This method is particularly useful when a population's characteristics are diverse, ensuring that every characteristic is properly represented in the sample. It enhances the generalizability and validity of the study while avoiding biases like undercoverage bias. Stratified sampling reduces sampling error and can yield a weighted mean with lower variability than a simple random sample, without requiring an increase in sample size, which can be costly.

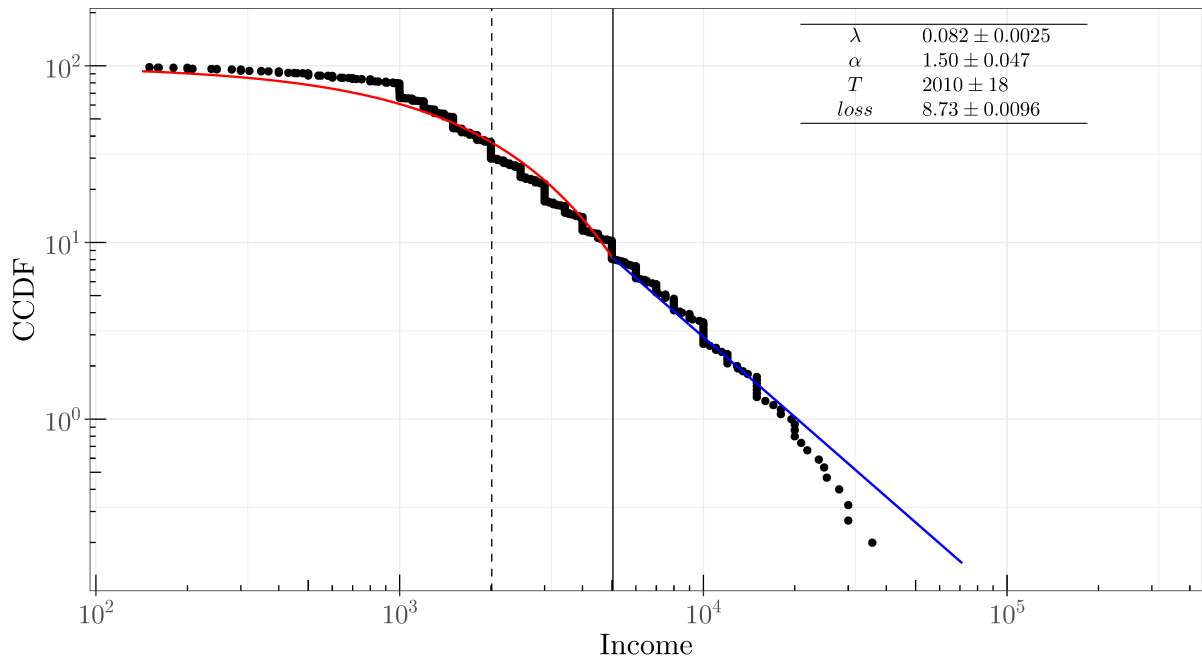
The PNADc in Brazil uses stratified sampling approach to ensure robust survey representation and accuracy. Stratification includes administrative divisions such as states and metropolitan regions, as well as geographical and spatial considerations to cover diverse population characteristics. This method optimizes sample spread across territories, enhances data collection efficiency, and controls costs, supporting precise estimation across various demographic and economic indicators.

PNADc data sampling plan also includes replicated weights, which enable a single sample to simulate multiple samples. This feature allows for the estimation of standard errors while maintaining the intricacies of the complex sample design. Specifically, we utilize 200 sets of weights, each applied to a different sample. This approach involves conducting regressions on datasets generated from each set of weights to assess the model's variability. The formula for the standard error of each parameter is of the form:

$$\Delta\theta_i = \frac{1}{\sqrt{R}} |\hat{\theta}_i^{(r)} - \hat{\theta}| \quad (8.1)$$

where R represents the number of replicated weight sets, $\hat{\theta}_i^{(r)}$ denotes the i th-parameter estimation using the r replicated weights, and $\hat{\theta}$ represents the same estimation using the original weights. This formula is regarded as the mean squared error version of the estimator, as it quantifies the variance comparing the estimation of the replicated weights with the original set of weights.

Figure 12 showcases the CCDF of the PNADc 2019, modeled using the two-



Source: Own Author.

Figure 12 – Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of data from the 2019 PNADc, while the solid lines the fitted model described in Equation (2.12). The red curve corresponds to the Boltzmann-Gibbs distribution, and the blue curve the power-law distribution. The temperature is denoted by the dashed vertical line, with a value of R\$ 2010 ± 18 , while the crossover income by the solid vertical line at $m_c = \text{R\$ } 5035 \pm 47$, marking the boundary between the Boltzmann-Gibbs and Pareto regions of income. In the upper right corner, you'll find all the model's parameters in a table, where λ represents the top income percentage of individuals, yielding $(8.16 \pm 0.25)\%$, and adhering to a Pareto power-law with an index of $\alpha = 1.5 \pm 0.046$.

class distribution. Evaluation of the model's performance yields a KS statistic of 0.072 and an MLS of 0.0060, both calculated in comparison to the CCDF. Apologies for the ambiguity. Based on the result of the parameter λ , we can partition the data into two classes, creating a class variable. This variable will be the target used for importance analysis in the Chapter 9. Additionally, the theoretical Gini coefficient for this model is 0.619 ± 0.010 .

We apply our model on the 2019 dataset, categorized into two population dichotomies, facilitating the comparison of indicators and the validation of our model across diverse data distributions. The initial dichotomy separates the population by gender (male and female), while the second dichotomy categorizes them by race/color (black/brown/indigenous - BBI - and white/yellow - WY). The detailed results and validation can be found in Table 5.

Table 5 – Inequality indicators for a subgroup data.

Subgroup table for 2019 stratified data		
Category: Sex/Gender		
Subgroup	Man	Woman
Crossover Income (R\$)	5000 ± 20	5000 ± 25
Top-percentage (%)	9.98 ± 0.69	6.49 ± 0.23
Temperature (R\$)	2170 ± 27	1828 ± 24
Pareto index	1.51 ± 0.04	1.64 ± 0.05
Theoretical Gini coeff.	0.625 ± 0.01	0.586 ± 0.008
Category: Race/Color		
Subgroup	WY	BBI
Crossover Income (R\$)	6120 ± 54	5340 ± 49
Top-percentage (%)	10.14 ± 0.36	4.18 ± 0.11
Temperature (R\$)	2673 ± 27	1682 ± 10
Pareto index	1.51 ± 0.04	1.93 ± 0.04
Theoretical Gini coeff.	0.626 ± 0.01	0.545 ± 0.003

Comparing the results of Table 3 from Chapter 6 despite variations, the fundamental conclusions regarding the comparison of gender and race/color dichotomies on inequality are align across both. The change in value stems from the use of the weights aswell as update in the loss function now utilizing a MLE approach instead of MSE.

In our analysis of the new approach, concerning the gender dichotomy, men display greater income inequality. This is evident in their higher top-percentage indicators and lower Pareto index compared to women. Additionally, men exhibit a higher income temperature than women.

Looking at the color/race dichotomy, we have a more pronounced contrast in inequality indicators compared to the gender dichotomy. The WY subgroup shows considerably higher top-percentage and Temperature values than the BBI subgroup. BBI displays the lowest top-percentage value, $\lambda = 4.18 \pm 0.12$, and the highest Pareto index, $\alpha = 1.68 \pm 0.02$, compared to all subsets, leading to the lowest Gini coefficient and indicating the highest level of equality within the subgroup.

For the non-weighted data, we examined the temporal evolution of the model parameters from 2001 to 2019, as detailed in Chapter 7. In contrast, the analysis of the weighted data focuses on a different aspect of the income distribution, namely the relative importance of the model parameters, which will be addressed in the following chapter. This change in perspective is motivated by the fact that incorporating sample weights provides a more accurate representation of the population structure, making it particularly suitable for evaluating the influence of each parameter in the overall income composition, since the stratification process was carried out accounting for some of the variables analyzed (see Appendix C).

9 IMPORTANCE OF VARIABLES

The random forest is a machine learning algorithm highly used for classification and regression problems, and it can also provide a measure of variable importance (BREIMAN, 2001). The great advantage of this measure of importance variable compared to univariate methods such as Information gain, Chi-squared test, and Fisher's Score lies in its ability to account for the interaction between predictor variables.

The most advanced variable importance method for random forest is permutation accuracy importance. However, it still exhibits bias towards correlated predictors and those with more levels or of numeric type. To address these issues, two methods have been proposed: unbiased CIT, which involves growing each individual tree in a specific manner to avoid bias of covariates variables (HOTHORN et al., 2006), and Conditional Permutation Importance (CPI), which imposes restrictions on the permutation process (STROBL et al., 2008). In the latter, permutation of a predictor variable only occurs within predefined groups defined by a grid of the predictor variables determined in the random forest trained model.

A decision tree is a popular supervised learning algorithm used for both classification and regression tasks. The essence of a decision tree lies in its ability to recursively split data into smaller, more homogeneous groups based on the values of input features. At each node of the tree, the algorithm selects the feature that best separates the data, creating branches that lead to distinct classes or more accurate predictions of a target variable.

In a decision tree, the splitting process is guided by certain criteria designed to maximize the distinction between groups, such as Gini impurity optimization, whether in terms of classification or predicting a numerical outcome. Additionally, decision trees must employ strategies to control their growth and stop at an appropriate point, often using stopping criteria based on tree depth or the number of leaves.

In the context of regression, we define the covariate space $\mathcal{X}$ with the covariate vector $\mathbf{X} = (X_1, \dots, X_m)$, where each vector is taken from a sample space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_m$. For a specific observation, such as Y_i , it is associated with a corresponding set of covariates $X_{i1}, X_{i2}, \dots, X_{im}$. The relationship between the response variable Y and the covariates X is modeled using recursive partitioning, where the response Y is hypothetically conditioned on a function of the covariates. This approach involves partitioning the covariate space $\mathcal{X}$ into r disjoint cells $G_1, \dots, G_r$, allowing for the creation of a model fitted on a training sample $\mathcal{D}$ that consists of N independent and identically distributed observations of both Y and X :

$$\mathcal{D} = \{(\mathbf{X}_i, Y_i) \mid i = 1, \dots, N\}, \quad (9.1)$$

where $\mathbf{X}_i = (X_{i1}, \dots, X_{im})$.

The recursive binary partitioning algorithm can be expressed with an indicator function that assigns a value of 1 if an observation is within a given node A and 0 otherwise

$$1_A(i) = \begin{cases} 1 & \text{if } i \in A, \\ 0 & \text{if } i \notin A. \end{cases}, \quad \text{for } i \in \{1, 2, \dots, N\}. \quad (9.2)$$

The definition of the subset A is constrained by the available covariates, and its composition evolves through the sequence of splits leading to the current node. Specifically, the subset A represents the accumulated conditions imposed by all preceding splits that define the path from the root to this node. For a binary split:

- **For numerical covariates:** The split is defined by a threshold value t such that:

$$\mathcal{D}^L = \{(\mathbf{X}_i, Y_i) \in \mathcal{D} \mid X_{ij} \leq t\}, \quad \mathcal{D}^R = \{(\mathbf{X}_i, Y_i) \in \mathcal{D} \mid X_{ij} > t\}$$

Here, $\mathcal{D}^L$ and $\mathcal{D}^R$ represent the left and right subsets of data, respectively. The split is entirely defined by the value of the chosen numerical covariate X_j and is constrained by the available values of this covariate.

- **For categorical covariates:** The split divides the data based on the membership of a category:

$$\mathcal{D}^L = \{(\mathbf{X}_i, Y_i) \in \mathcal{D} \mid X_{ij} \in A\}, \quad \mathcal{D}^R = \{(\mathbf{X}_i, Y_i) \in \mathcal{D} \mid X_{ij} \notin A\}$$

Here, A is a subset of categories or levels for the categorical covariate X_j . The restriction ensures that data is split into exactly two groups based on category membership. Any subsequent split within this branch must operate within the reduced set of categories defined by B . Thus, A evolves as:

$$A = B_1 \cap B_2 \cap \dots \cap B_m \quad (9.3)$$

where $B_1, B_2, \dots, B_m$ are the subsets defined by the sequence of splits that led to the current node.

Thus, each node in the tree a indicator function defined above, where elements are set to 1 for observations that belong to the node and 0 for those that do not. Typically, case weights replace the indicator function to account for weighted data, assigning a weight of 0 to observations outside the node. However, in this algorithm, weighting will be handled by the random forest component. A key aspect of the algorithm is the process of selecting the covariate that exhibits the strongest association with the response variable Y . Once a covariate is selected, it can be split into two disjoint sets, which allows for a recursive refinement of the model. This splitting process is fundamental to the construction of partition-based regression models, enabling the creation

of interpretable decision trees that effectively capture the relationships between the response variable and the covariates. Thus, the algorithm effectively partitions the covariate space $\mathcal{X}$, where each partition cell G is associated with an indicator function. The following algorithm implements recursive binary partitioning:

1. For the current node, check the stopping criteria. If the criteria are not met, select the j^* th covariate X_{j^*} with the strongest association to Y , according to the variable selection criteria.
2. Given the split criteria choose a set $A^* \subset X_{j^*}$ in order to divide X_{j^*} into two disjoint sets A^* and $X_{j^*} \setminus A^*$. Define the indicator function for each observation in the node as

$$1_{\text{left}}(i) = 1_{X_{j^*} \in A^*}(i) \quad \text{and} \quad 1_{\text{right}}(i) = 1_{X_{j^*} \notin A^*}(i) \quad (9.4)$$

for all $i = 1, \dots, n$, taking the value 1 if the condition is true (the observation belongs to the node) and 0 otherwise.

3. Recursively repeat steps 1 and 2 on each new created node defined by the left and right indicator functions that defines the two subgroups of observations.

With these steps the algorithm induces a partition $\{G_1, \dots, G_r\}$ of the covariate space $\mathcal{X}$, where each cell $G \in \{G_1, \dots, G_r\}$ is associated with an indicator function.

For traditional decision trees, the variable selection is typically based on criteria such as maximizing the reduction in impurity (e.g., using Gini impurity or information gain), while the stopping criteria often involve conditions like reaching a minimum node size, achieving a maximum tree depth, or when further splitting no longer significantly improves the model's performance.

While decision trees are easy to interpret and highly effective, they can sometimes favor features in a biased manner. This is where conditional inference trees offer a more refined approach to selecting features and stopping the tree's growth.

9.1 CONDITIONAL INFERENCE TREES

CIT build upon the basic idea of decision trees but address some of the inherent biases that can arise in traditional approaches. One of the key challenges with standard decision trees is that features with many unique values or categories tend to be overrepresented in splits, which can lead to biased models. Conditional inference trees mitigate this issue by introducing statistical hypothesis testing into the feature selection process, ensuring that splits are both unbiased and statistically significant (HOTHORN et al., 2006).

The core of this approach lies in testing the null hypothesis for each potential split. Formally, we can consider the hypothesis H_0 , which states that the distribution

of the target variable $D(Y)$ is independent of the input feature X , meaning that the conditional distribution $D(Y|X)$ does not significantly differ from $D(Y)$. In other words, if the null hypothesis holds, the feature does not provide meaningful information about the target, and the tree will not split on it.

CITs systematically test this null hypothesis at each node, using permutation tests to evaluate whether splitting on a given feature would lead to a statistically significant improvement in the model. If the null hypothesis is rejected, meaning there is a significant relationship between the feature and the target variable, the tree selects that feature for splitting. This ensures that only features that contribute meaningfully to improving the model are used. If no feature passes this significance test, the tree stops growing at that node.

By relying on this statistical framework, conditional inference trees avoid the biases associated with traditional decision trees and reduce the likelihood of overfitting, making them particularly effective for high-dimensional data where feature selection is critical.

9.2 VARIABLE SELECTION

As we pointed in the introduction of this chapter, the separation of variable selection and splitting procedure into steps 1 and 2 of the algorithm is key for the construction of interpretable tree structures that do not suffer from a systematic tendency towards covariates with many possible splits or many missing values. In addition, a statistically motivated and intuitive stopping criterion can be implemented: We stop when the global null hypothesis of independence between the response and any of the m covariates cannot be rejected at a pre-specified nominal level α .

In order to do so one must define a we need to decide whether there is any information about the response variable covered by any of the m covariates. In each node identified by an indicator $1(X)$, the global hypothesis of independence is formulated in terms of the m partial hypotheses

$$H_{0j} : D(Y|X_j) = D(Y) \quad (9.5)$$

with the global null hypothesis

$$H_0 = \bigcap_{j=1}^m H_{0j}. \quad (9.6)$$

When we are not able to reject H_0 at a pre-specified level α , we stop the recursion. If the global hypothesis can be rejected, we measure the association between Y and each of the covariates X_j , $j = 1, \dots, m$, by test statistics or P-values indicating the deviation from the partial hypotheses H_{0j} .

Let $\mathcal{D} = \{(Y_i, X_{i1}, \dots, X_{im}) : i = 1, \dots, n\}$ denote the learning sample consisting of n independent observations of the response variable Y and m covariates $X_1, \dots, X_m$. The symmetric group of all permutations of the elements $(1, \dots, n)$ is denoted by $S(\mathcal{D}, I)$, where I represents the null hypothesis of independence between Y and the covariates. We measure the association between Y and X_j , $j = 1, \dots, m$, by linear statistics of the form

$$T_j(\mathcal{D}, 1) = \text{vec} \left(\sum_{i=1}^n \mathbf{1}(i) g_j(X_{ij}) h(Y_i, (Y_1, \dots, Y_n)) \right) \in \mathbb{R}^{p_j q} \quad (9.7)$$

where $g_j : X_j \rightarrow \mathbb{R}^{p_j}$ is a non-random transformation of the covariate X_j . The influence function

$$h : Y \times Y^n \rightarrow \mathbb{R}^q \quad (9.8)$$

depends on the responses $(Y_1, \dots, Y_n)$ in a permutation symmetric way. For numerical variables, g_j and h can be chosen as the identity function, $f(x) = x$. For categorical variables with levels $k \in K$, one option is to use the unit vector $f(k) = \mathbf{e}_K(k)$, where $\mathbf{e}_K(k) \in \mathbb{R}^K$. The p_j is the dimension of the predictor X_j and the q is the dimension of the target variable. A $p_j \times q$ matrix is converted into a $p_j q$ column vector by column-wise combination using the vec^1 operator.

To calculate the P-value, we need the distribution of the test statistic under the null hypothesis H_0 , but in this case the joint distribution of Y and X_j is unknown. Therefore, we use permutation tests by fixing covariates X_j and generating all possible permutations of the responses to create the null distribution of the test statistic. The conditional expectation $\mu_j \in \mathbb{R}^{p_j q}$ and covariance $\Sigma_j \in \mathbb{R}^{p_j q \times p_j q}$ of $T_j(\mathcal{D}, 1)$ under the null hypothesis H_0 , conditioned on all possible permutations $\sigma \in S(\mathcal{D}, 1)$ of the responses, are derived in the referenced work (STRASSER; WEBER, 1999) and are of the form:

$$\mu_j = E(T_j(\mathcal{D}, 1) | S(\mathcal{D}, 1)) = \text{vec} \left(\sum_{i=1}^n \mathbf{1}(i) g_j(X_{ij}) \right) E(h | S(\mathcal{D}, 1))^\top \quad (9.9)$$

$$\begin{aligned} \Sigma_j = V(T_j(\mathcal{D}, 1) | S(\mathcal{D}, 1)) &= \frac{\mathcal{N}^1}{\mathcal{N}^1 - 1} V(h | S(\mathcal{D}, 1)) \otimes \left(\sum_i \mathbf{1}(i) g_j(X_{ij}) \otimes \mathbf{1}(i) g_j(X_{ij})^\top \right) \\ &\quad - \frac{1}{\mathcal{N}^1 - 1} V(h | S(\mathcal{D}, 1)) \otimes \left(\sum_i \mathbf{1}(i) g_j(X_{ij}) \right) \otimes \left(\sum_i \mathbf{1}(i) g_j(X_{ij}) \right)^\top \end{aligned} \quad (9.10)$$

¹ The vec operator stacks the columns of a matrix on top of one another to produce a column vector. Formally, if A is a $m \times n$ matrix with columns $a_1, a_2, \dots, a_n$, then $\text{vec}(A) = [a_1 \ a_2 \ \dots \ a_n]^\top$, resulting in a $mn \times 1$ column vector.

where $\mathcal{N}^1 = \sum_{i=1}^n \mathbf{1}(i)$ denotes the sum of the indicator function that will depend on the node, $\otimes$ is the Kronecker product, and the conditional expectation of the influence function is

$$E[h|S(\mathcal{D}, \mathbf{1})] = \frac{1}{\mathcal{N}^1} \sum_i \mathbf{1}(i) h(Y_i, (Y_1, \dots, Y_n)) \in \mathbb{R}^q \quad (9.11)$$

with corresponding $q \times q$ covariance matrix

$$\begin{aligned} V(h|S(\mathcal{D}, \mathbf{1})) &= \frac{1}{\mathcal{N}^1} \sum_i \mathbf{1}(i) \{h(Y_i, (Y_1, \dots, Y_n)) - E[h|S(\mathcal{D}, \mathbf{1})]\} \\ &\quad \times \{h(Y_i, (Y_1, \dots, Y_n)) - E[h|S(\mathcal{D}, \mathbf{1})]\}^\top \end{aligned} \quad (9.12)$$

Now to perform hypothesis testing, we first standardize the observed multivariate statistic $T \in \mathbb{R}^{pq}$ using its conditional expectation μ and covariance matrix Σ . Standardization ensures that the statistic is scaled in a way that allows for meaningful comparisons, regardless of the original units or scale of each covariate. The two common approaches to standardizing the statistic are through the $c_{\max}$ and c_{quad} test statistics.

9.2.1 Standardization for $c_{\max}$:

For the $c_{\max}$ statistic, standardization is done by first centering the statistic T around its conditional expectation μ . This is achieved by subtracting μ from t , resulting in the centered statistic:

$$T_{\text{centered}} = T - \mu \quad (9.13)$$

Next, each element of the centered statistic is scaled by the standard deviation, which is derived from the covariance matrix Σ . Specifically, the elements of T_{centered} are divided by the square root of the corresponding diagonal elements of Σ . This ensures that the components of the statistic have unit variance. The standardized $c_{\max}$ statistic is then calculated as the maximum of the absolute values of the standardized components:

$$c_{\max}(T, \mu, \Sigma) = \max_{k=1, \dots, pq} \left(\frac{T_k - \mu_k}{\sqrt{\Sigma_{kk}}} \right) \quad (9.14)$$

This procedure transforms the statistic into a form where the maximum deviation from the expected value is measured, scaled by the variability of each component.

9.2.2 Standardization for c_{quad} :

For the c_{quad} statistic, the standardization process also starts with centering the statistic T around its mean μ by subtracting μ using (9.13).

However, instead of scaling each element individually, the scaling is done using the entire covariance structure through the generalized inverse Σ^+ . The quadratic form c_{quad} is then computed as the Mahalanobis distance between T and μ , which accounts for both the variance and the covariance between the components of T . This is given by:

$$c_{\text{quad}}(T, \mu, \Sigma) = (T - \mu)^\top \Sigma^+ (T - \mu) \quad (9.15)$$

The Mahalanobis distance is a measure of how far T is from μ in terms of the multivariate distribution of the covariate. A large value indicates that T deviates significantly from μ , while a small value suggests that T is close to the expected value. We will use this statistic because, although it requires more computation due to the need for calculating the Moore-Penrose inverse, Σ^+ , of Σ , it offers the advantage of producing a Chi-square distribution under a permutation test. This makes it possible to efficiently calculate the P-value (RASCH, 1995).

The P-value not only tests the null hypothesis giving us a stopping criteria, but also enables unbiased variable selection. This is because P-values for the conditional distribution of the test statistics $c(T_j(\mathcal{D}, 1), \mu_j, \Sigma_j)$ can be compared across covariates measured on different scales. In step 1 of the generic algorithm, the covariate selected for the split is the one with the smallest P-value, i.e., the covariate X_{j^*} where $j^* = \text{argmin}_{j=1, \dots, m} P_j$. The P-value P_j is calculated as

$$P_j = \mathbb{P}_{H_{0j}} \left(c(T_j(\mathcal{D}, 1), \mu_j, \Sigma_j) \geq c(T_j(\sigma), \mu_j, \Sigma_j) \mid \sigma \in S(\mathcal{D}, 1) \right) \quad (9.16)$$

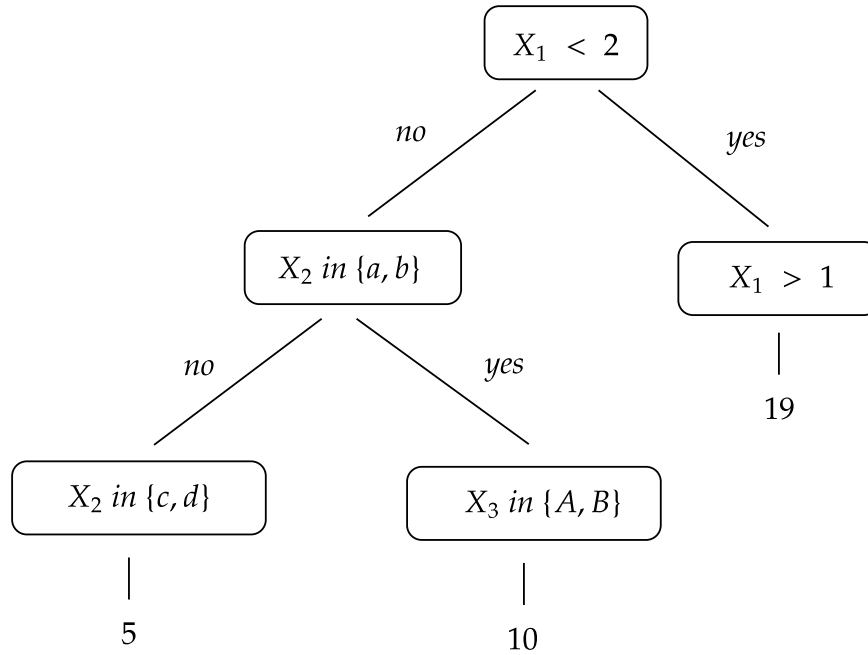
where H_{0j} is the null hypothesis, and the P-value represents the probability of observing a test statistic as extreme as or more extreme than the observed value, based on the distribution under the null hypothesis obtained through permutations.

9.3 SPLITTING CRITERIA

In step 1 of the algorithm, we select a covariate, and in step 2, we perform a binary split based on this covariate. The quality of the split is evaluated using a two-sample test statistic, which compares the response variable Y between two groups: one where the covariate X_{j^*} is in a subset A , and one where it is not. The optimal split is determined by maximizing the test statistic over all possible subsets of A . To prevent overfitting or pathological splits, we may impose restrictions on the number of subsets or the size of the groups. The split is performed by maximizing the following expression:

$$A^* = \arg \max_A \left(c_{\text{quad}}(T_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A) \right) \quad (9.17)$$

where $\mu_{j^*}^A$ and $\Sigma_{j^*}^A$ are the mean and covariance of the permutation within subset A , while $T_{j^*}^A$ is defined in Equation (9.7), and g_j is the indicator function $1_{X_{j^*} \in A}$. These two steps complete the description of the conditional inference tree-growing procedure.



Source: Own Author.

Figure 13 – An example of a regression tree where X_1 is a numerical covariate, while X_2 and X_3 are categorical.

Figure 13 illustrates a regression tree, where the numerical covariate X_1 and the categorical covariate X_2 are each selected twice. Selecting a covariate multiple times can happen with categorical covariates that have more than two levels or with numerical covariates.

After activating the stop criteria, which means we could not reject the null hypothesis for all covariates based on a predefined significance level α , we obtain the partition of the data $\{G_1, \dots, G_r\}$, which are also known as the sets of the leaf nodes. The values of the leaf nodes are defined by the target values within each G_j . For a regression tree, the value is typically a statistic such as the mean, while for a classification tree, it is usually the mode.

9.4 CONDITIONAL PERMUTATION OF PREDICTORS

With the model trained, we can now evaluate the significance of the variable for predicting the target variable.

To assess the permutation importance of a specific feature X_{j^*} in a predictive model, we begin by creating a modified dataset through a process of permuting the values of X_{j^*} across all samples. Let the test dataset be $\mathcal{T} = \{\mathbf{X}, Y\}$, where $\mathbf{X}$ is the covariates and Y is the target variable of the test dataset. Permuting the values of X_{j^*} results in a new dataset $\mathcal{T}^{\text{perm}} = \{\mathbf{X}, Y\}$, where $\tilde{\mathbf{X}}$ denotes the feature matrix with X_{j^*} permuted while all other features remain unchanged.

To clarify the permutation process for individual samples, consider the feature vector of observation i , denoted as $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})$, where p is the total number of features. After permuting the values of X_{j^*} , we obtain a new feature vector:

$$\mathbf{X}_i^{(j^*)} = (X_{i1}, \dots, X_{ij^*-1}, X_{p(i)j^*}, X_{ij^*+1}, \dots, X_{im}) \quad (9.18)$$

Here, $\mathbf{X}_i^{(j^*)}$ is the permuted value of feature X_{j^*} for observation i , while all other components of the feature vector remain unchanged. This modification isolates the effect of permuting X_{j^*} and allows us to evaluate its impact on the model's predictions.

The model's predictions on the permuted data are represented as $\hat{Y}_i^{(j^*)} = M(\mathbf{X}_i^{(j^*)})$, where M denotes the prediction function for tree. By comparing the performance of the model on the permuted dataset with the baseline performance, we quantify the importance of the feature X_{j^*} .

In the context of classification trees, a prediction error R can be defined as:

$$R = \sum_{i \in \mathcal{T}} \frac{I(\hat{Y}_i \neq Y_i)}{|\mathcal{T}|}. \quad (9.19)$$

After permuting the values of feature X_{j^*} , the corresponding prediction error becomes:

$$R^{(j^*)} = \sum_{i \in \mathcal{T}} \frac{I(\hat{Y}_i^{(j^*)} \neq Y_i)}{|\mathcal{T}|}. \quad (9.20)$$

For regression tree, the prediction error before permutation is given by:

$$R = \sum_{i \in \mathcal{T}} \frac{(\hat{Y}_i - Y_i)^2}{|\mathcal{T}|}, \quad (9.21)$$

and after permutation of feature X_{j^*} , it becomes:

$$R^{(j^*)} = \sum_{i \in \mathcal{T}} \frac{(\hat{Y}_i^{(j^*)} - Y_i)^2}{|\mathcal{T}|}. \quad (9.22)$$

The permutation importance for feature X_{j^*} is then quantified as the difference:

$$PI(j^*) = R^{(j^*)} - R. \quad (9.23)$$

This quantifies the change in prediction error due to permuting X_{j^*} , reflecting its importance in the model's predictive performance. Considering (9.18), if $p(i)$ represents a simple permutation, the method corresponds to simple permutation importance. On the other hand, if $p(i)$ represents a conditional permutation, the result corresponds to conditional permutation importance (DEBEER; STROBL, 2020). The conditional permutation can be executed by the following procedure:

1. For the j^* th covariate X_{j^*} find a limited set of covariates $\mathcal{Z}_{j^*} \subseteq \{X_j \mid j \in \{1, \dots, m\} \setminus \{j^*\}\}$ with the strongest association to X_{j^*} .

This association is determined by testing the null hypothesis of independence between X_{j^*} (treated as the response variable Y) and each X_j (for $j \neq j^*$) in the set $\{X_j \mid j \in \{1, \dots, m\} \setminus \{j^*\}\}$. Specifically, for each X_j , the hypothesis is:

$$H_{0j} : D(X_{j^*} \mid X_j) = D(X_{j^*}), \quad (9.24)$$

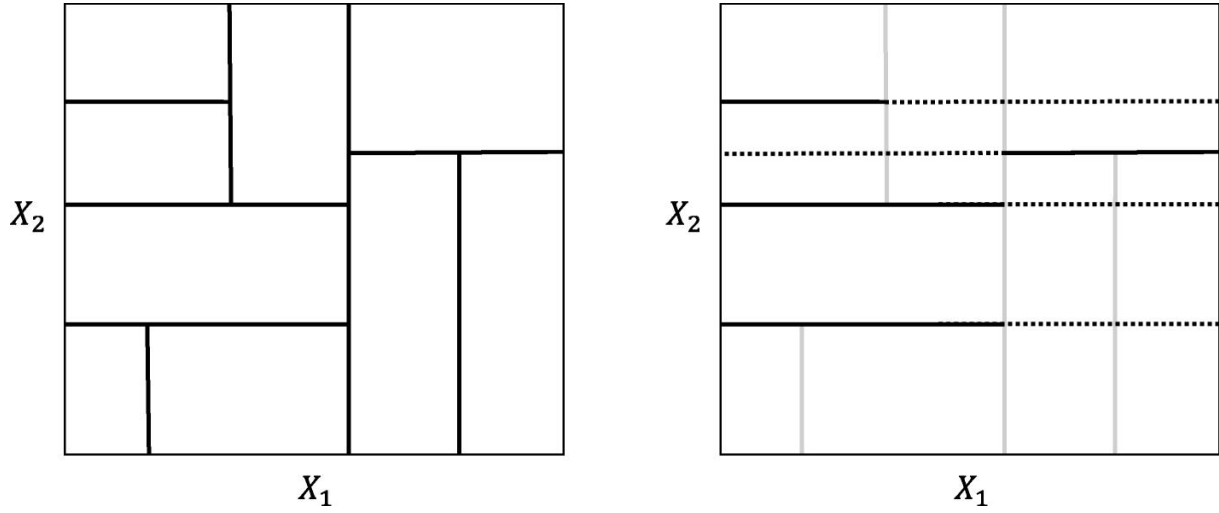
where $D(\cdot)$ denotes the distribution of the random variable. The null hypothesis H_{0j} states that the conditional distribution of X_{j^*} given X_j is the same as the marginal distribution of X_{j^*} , indicating no dependence between the two variables.

The subset $\mathcal{Z}_{j^*}$ is then defined as the set of covariates for which the null hypothesis H_{0j} is rejected at a pre-specified significance level α , indicating a statistically significant association with X_{j^*} .

This approach ensures that only the covariates that provide meaningful information about X_{j^*} are included in $\mathcal{Z}_{j^*}$. This will be repeated for every covariate to define $\mathcal{Z}_{j^*}$ for each, using the full dataset and without depending on the specifics of the tree process. The Table 15 presents a comparison between normal permutation and conditional permutation.

2. Define a grid that partitions the predictor space based on the split points of the covariates in $\mathcal{Z}_{j^*}$. Permute the values of X_{j^*} **conditionally within each partition** of this grid, defined split points in of the tree for the predictors in $\mathcal{Z}_{j^*}$, and combine these splits to construct a multi-dimensional grid that partitions the predictor space. In this grid, each split completely bisects the predictor space, as opposed to the original tree-growing partitioning, where splits only divide limited subspaces, as can be seen in Figure 14. Permute the values of X_{j^*} **conditionally within each partition** of this grid. By performing the permutation within the grid defined by $\mathcal{Z}_{j^*}$, we mitigate biases towards X_{j^*} , even if it exhibits a high correlation with the covariates in $\mathcal{Z}_{j^*}$.

3. Compute (9.23) using the ****conditionally permuted**** values of X_{j^*} for each variable.



Source: Article (DEBEER; STROBL, 2020).

Figure 14 – The image in the left is the how the a decision tree partition the space while in the right you have the conditional permutation grid partition. The solid lines indicates a separation, the grey lines are ignored and the dashed lines are the parts that was extended due from the tree to the conditional permutation due to the grid being made of a complete split of the predictor space in regards to X_1 which is the covariate that we are calculating the importance or is not part of $\mathcal{Z}_j$.

Figure 15 – This figure presents two side-by-side tables: the normal permutation on the left, and the conditional permutation on the right. The conditional permutation is based on the $\mathcal{Z}_{j^*}$ set of variables. The a and b represent grid cells of $\mathcal{Z}_{j^*}$, where in the conditional permutation, values are permuted only within each cell.

Y	X_{j^*}	$\mathcal{Z}_{j^*}$
Y_1	X_{2j^*}	$\mathcal{Z}_{1j^*}$
$\vdots$	$\vdots$	$\vdots$
Y_5	X_{14j^*}	$\mathcal{Z}_{5j^*}$
$\vdots$	$\vdots$	$\vdots$
Y_n	X_{5j^*}	$\mathcal{Z}_{Nj^*}$

Y	X_{j^*}	$\mathcal{Z}_{j^*}$
Y_1	X_{3j^*}	$\mathcal{Z}_{1j^*} = a$
Y_3	X_{1j^*}	$\mathcal{Z}_{3j^*} = a$
Y_7	X_{7j^*}	$\mathcal{Z}_{7j^*} = a$
$\vdots$	$\vdots$	$\vdots$
Y_2	X_{15j^*}	$\mathcal{Z}_{2j^*} = b$
Y_{15}	X_{19j^*}	$\mathcal{Z}_{15j^*} = b$
Y_{19}	X_{2j^*}	$\mathcal{Z}_{19j^*} = b$
$\vdots$	$\vdots$	$\vdots$

9.5 CONDITIONAL RANDOM FOREST

Now that we have defined CIT and the CPI, we can define the random forest algorithm that uses these two components to calculate variable importance.

In a random forest, N_{tree} trees are constructed by randomly selecting a subset of rows (observations) with replacement for each tree, a process known as bootstrap sampling. This sampling can also be performed based on the case weights of the data, allowing observations with higher weights to have a greater chance of being included in the sample. This procedure can be used to incorporate the effect of stratified sampling into the algorithm, enabling it to emphasize certain observations more strongly during training. Moreover, for each split in a tree, a random subset of covariates (features) is selected to determine the best split, adding further randomness and reducing overfitting.

Each tree is trained on its respective bootstrap sample, leaving some observations out, referred to as the Out-of-Bag Set (OOB) sample. The OOB sample can then be used to estimate prediction errors and evaluate variable importance for each tree, making random forests a useful ensemble learning method for testing variable importance. Therefore one can use the OOB to calculate the CPI of the random forest algorithm transforming (9.23):

$$PI^{(j^*)} = \frac{\sum_{t=1}^{N_{\text{tree}}} PI_t^{(j^*)}}{N_{\text{tree}}}, \quad (9.25)$$

where the $PI_t^{(j^*)}$ is the importance of the covariate X_{j^*} for the t -th tree.

10 VARIABLE IMPORTANCE OF PNADC

Now that we have defined variable importance, we can apply the methods discussed in the previous chapter to the PNADC data. In this chapter, we will train a random forest using CIT and then calculate CPI to evaluate each predictor's relevance. Our target variable is VD4022, which represents total income from all sources, including aid programs. We will assess variable importance both for the overall population and separately within the Pareto and exponential regions identified in Chapter 8.

The PNADC dataset encompasses five primary themes of variables, including original variables and derivative variables which are synthesized from multiple original variables. These themes are: identity and control, general characteristics, education characteristics, job characteristics, housing characteristics, and income from other sources. We excluded the last two themes due to a high frequency of non-response in their variables. From the first theme we got the UF variable which are the Brazil federative units (states). From the second category, general characteristics, we selected three variables: age, sex, and Race. From the third category, we extracted the variable representing the highest level of education attained. Lastly from the fourth one, job characteristics, we got the great groups of the CBO, available from variable *VD4011*¹ of the PNADC data. The details of all variables can be encountered in the Appendix C.

The selection of variables for this study — UF, age, sex, race, education level, and great occupational groups from CBO — is strongly supported by the literature on socioeconomic inequality and the labor market in Brazil. Race significantly influences educational attainment and social mobility, as shown by longitudinal PNAD analyses covering the 1980–2007 period, which document persistent gaps between white, mixed-race, and Black populations despite overall improvements (MARTELETO, 2012). Studies using PNAD data also find that the wage returns to a university degree are lower for Black workers compared to white counterparts, underscoring structural inequality (VAZ, 2020). The classification of occupations using CBO provides a structured and standardized way to capture individuals' positions in the labor market, making it a natural choice for analyzing patterns of socioeconomic stratification and employment dynamics. Controlling for UF is essential to account for significant intra- and interstate variability.

It is important to emphasize that both race and sex, as used in this study, have major social dimensions that extend beyond their biological outcomes. The concept

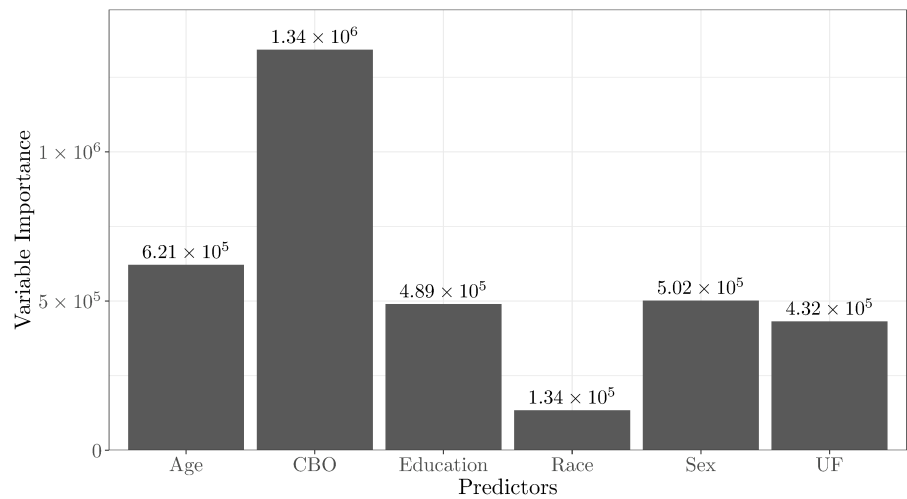
¹ The variable *VD4011* from the PNADC classifies the worker's main occupation according to the great occupational groups of the CBO (MINISTRY OF LABOR AND EMPLOYMENT (MTE), 2002). The CBO, established by the Ministry of Labor in 2002, is the official system for categorizing occupations in the Brazilian labor market based on the tasks performed. The *VD4011* levels cover a wide range of occupations, from directors and managers to elementary occupations, which correspond to some of the 10 CBO groups, and include two additional categories: "not well defined occupation" and "not applicable." This classification is essential for socio-labor and economic analyses, as it allows comparison across occupational strata with different qualification levels and social status.

of race adopted here follows the notion of "social race" (*raça social*), introduced to Brazilian scholarship by Charles Wagley and Marvin Harris and further discussed in the literature on race relations in Brazil (WAGLEY, 1952; HARRIS, 1964; ALFREDO, 1999). In this perspective, race should not be viewed solely as a biological fact but as a social construct — "forms of identity based on an erroneous biological idea, yet socially effective in constructing, maintaining, and reproducing differences and privileges" (ALFREDO, 1999).

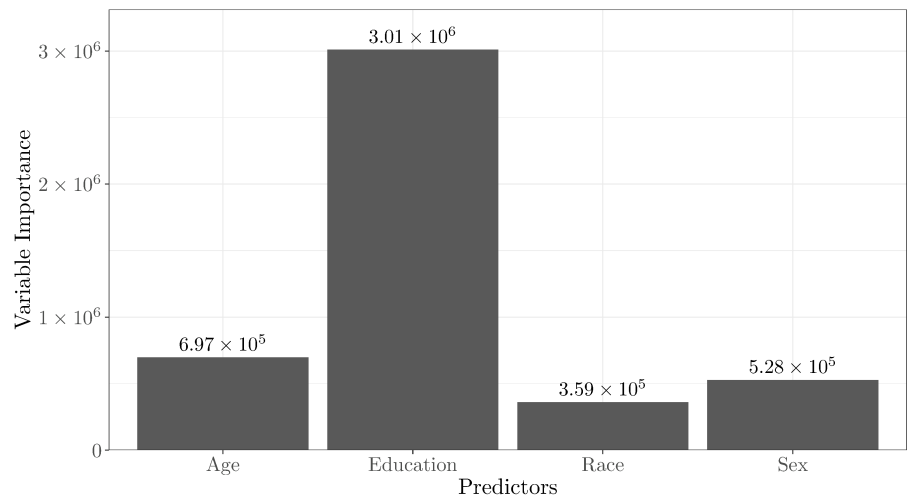
Similarly, the category of sex, although grounded in biological markers, is also shaped by social norms, expectations, and institutions. Feminist and gender studies have long argued that the meanings attached to sex and gender are historically and culturally contingent (BUTLER, 1990). In this sense, both *race* and *sex* categories in surveys such as the PNADc reflects social perceptions, hierarchies, and historically embedded power relations rather than genetic distinctions.

The variable education has a 10.97% non-response rate. To address this, we will recategorize education into two levels: "graduated" and "non-graduated". Missing values will be imputed with the mode, which corresponds to the "non-graduated" level. The "graduated" category will include the original levels of graduation, higher-level specialization, master's degree, and doctorate degree. The "non-graduated" category will include all other levels. Additionally, we will transform the numerical variable age into a categorical variable to calculate the uncertainty coefficient $U(X|Y)$, allowing us to derive a Theil's U matrix for further analysis. This was achieved by creating age bins of 10-year intervals: (0, 9], [10, 19], ..., [99, ∞).

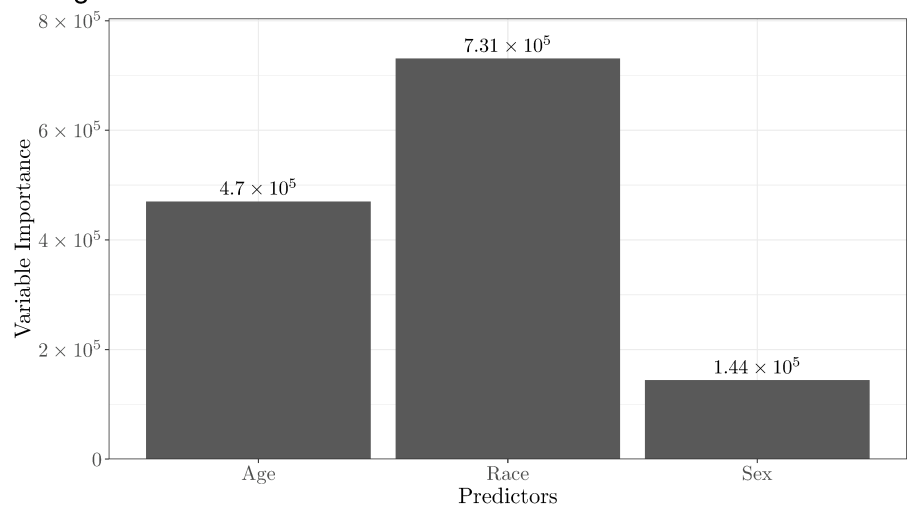
Initially, we applied a variable importance algorithm to all predictor variables previously mentioned; this will be referred to as the Full Issue (FI). Subsequently, we extended the application of this algorithm to include only the general characteristics category (age, sex, and Race) along with the highest education attained variable; this will be referred to as the Educational Issue (EI). Finally, we applied the algorithm using only the variables within the general characteristics category, this will be referred to as the Personal Issue (PI). The results are shown in Figure 16.



(a) FI: Variable importance ranking including all predictors.



(b) EI: Variable importance ranking including age, sex, Race, and the highest level of education attained.



(c) PI: Variable importance ranking for age, sex, and Race.

Source: Own Author.

Figure 16 – Comparison of variable importance within the overall population of PNADc.

In Figure 16, we present the results of various analyses of the overall population. For the FI variable importance analysis, CBO emerges as the most relevant predictor, followed by age, sex, education, UF and, lastly, Race.

This finding of the FI is corroborated by a study in the US, which used signal-to-noise ratios to rank the influence of income against various demographic and socio-economic factors (FIX, 2019). Similarly, research employing principal component analysis in the US supports these conclusions (LAZAR, 2004). Additionally, still on US, an ensemble meta-algorithm study using boosting techniques to reduce bias and variance further confirms these findings (CHEN, 2021).

When CBO and UF are removed from the analysis, the importance shifts markedly, which are shown in the EI analysis. education emerges as the most important variable, significantly surpassing age and sex, which were previously higher. Race remains the least important variable, and age remains more important than sex. These results highlight that, when not considering the CBO and UF predictors, educational attainment has a stronger influence on predicting income, suggesting it may capture broader economic effects than other predictors.

The last analysis on the overall population is the PI, which places Race as the most influential variable followed by age and sex as the least influential. This pattern suggests that, considering only the characteristics variables, Race and age differences play a more critical role than gender in influencing the dataset outcomes.

For further analysis of this rankings, we will use an an statistics called, Theil's U or uncertainty coefficient, to get the information relation of all predictors variables.

Given a sample of variable set $\mathcal{X}$ and $\mathcal{Y}$, we can define the uncertainty coefficient as

$$U(X|Y) = \frac{I(X; Y)}{H(X)}, \quad (10.1)$$

where,

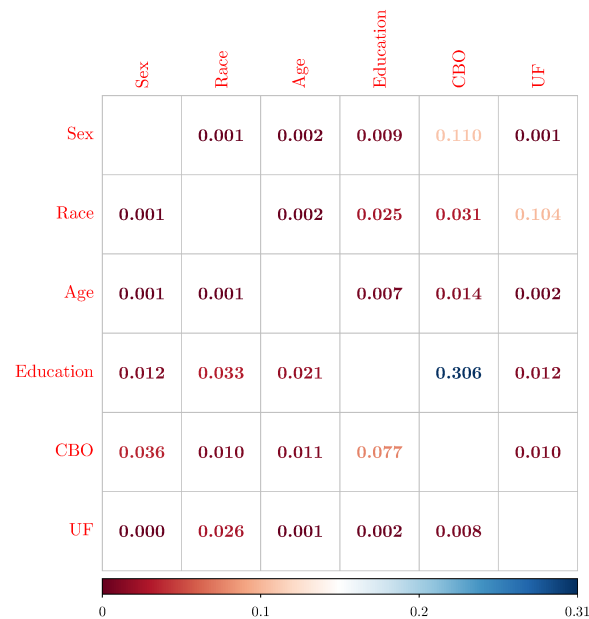
$$H(X) = - \sum_{x \in \mathcal{X}} P_X(x) \log P_X(x), \quad (10.2)$$

is the entropy,

$$I(X; Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{(X, Y)}(x, y) \log \left(\frac{P_{(X, Y)}(x, y)}{P_X(x) P_Y(y)} \right) \quad (10.3)$$

$P_X(x)$ is the probability of getting x from the set $\mathcal{X}$ and $P_{(X, Y)}(x, y)$ is the jointed probability. All this probabilities are calculated considering the weights of the data.

Following Figure 17 illustrates the extent to which each variable Y gives information about the X variable. Higher values indicate that Y has a greater fraction of information attained from the X variable.



Source: Own Author.

Figure 17 – The Theil's U matrix displays the entropy coefficient $U(X|Y)$ values for the set of predictor variables.

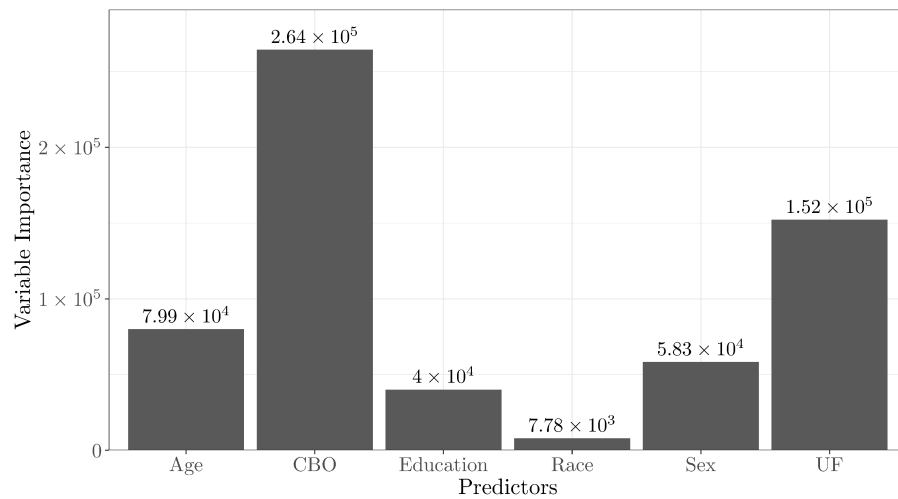
Considering the results of the Figure 17 we can now partially elucidate the disparities between the analysis: FI, EI and PI.

Going from the FI to the EI analysis, education advanced from being the fourth-ranked variable to the top rank in terms of importance. This can be partly explained by the fact that the uncertainty coefficient $U(X = \text{education} | Y = \text{CBO}) = 0.306$ is the highest. Thus, with the removal of CBO from the analysis, education's importance in predicting the target variable increased significantly.

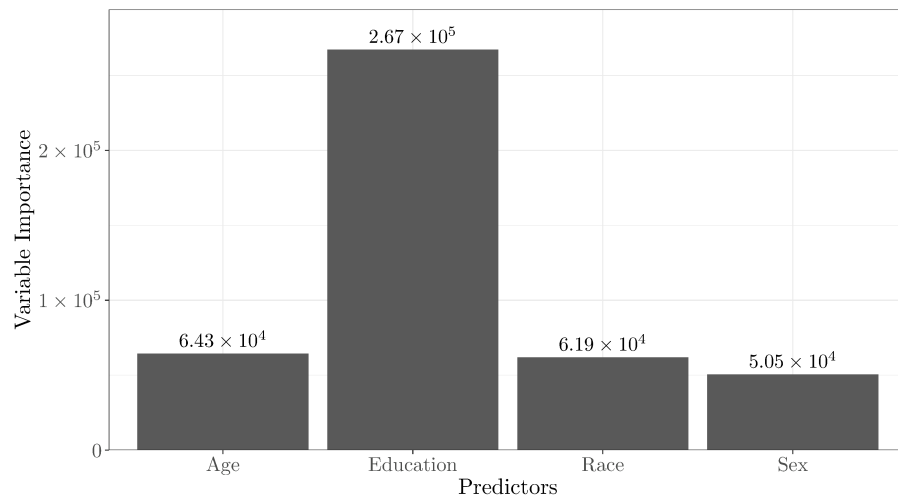
Now, moving from the EI to the PI analysis, Race shows a trend similar to that of education. With an uncertainty coefficient of $U(X = \text{Race} | Y = \text{education}) = 0.025$, it has the highest value among the PI predictors. Consequently, in line with the previous comparison, the removal of education from the analysis significantly increased the importance of Race in predicting the target variable.

10.1 EXPONENTIAL REGION

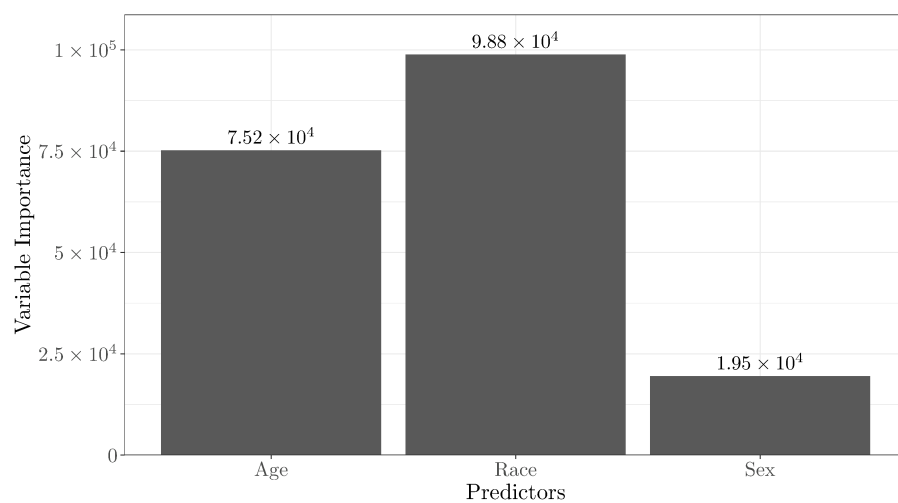
We also conducted the same analysis for each region, starting with the exponential region. The results of all types of variable importance analyses are presented in Figure 18.



(a) FI: Variable importance ranking including all predictors.



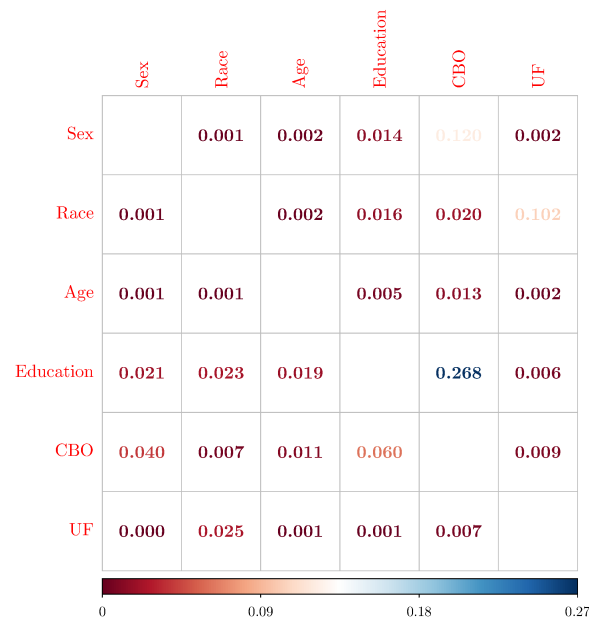
(b) EI: Variable importance ranking including age, sex, Race, and the highest level of education attained.



(c) PI: Variable importance ranking for age, sex, and Race.

Source: Own Author.

Figure 18 – Comparison of variable importance within the exponential region of PNADc.



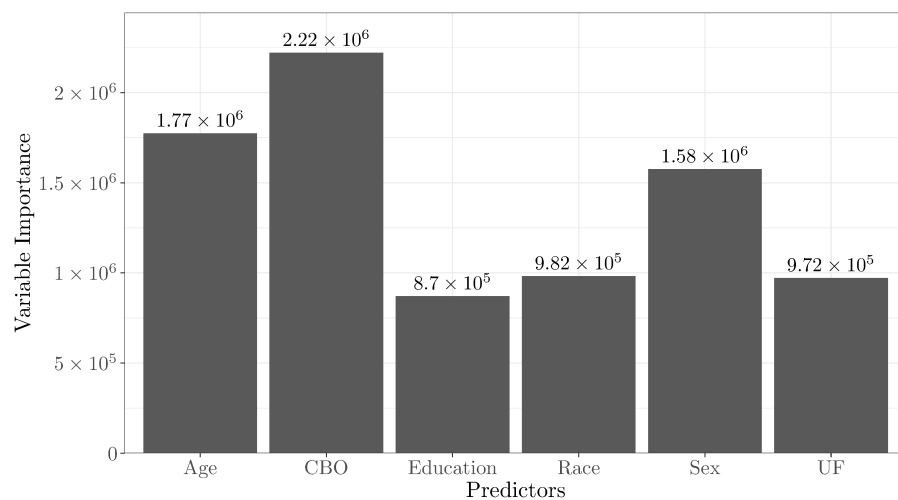
Source: Own Author.

Figure 19 – The Theil's U matrix displays the entropy coefficient $U(X|Y)$ values for the set of predictor variables for the exponential region.

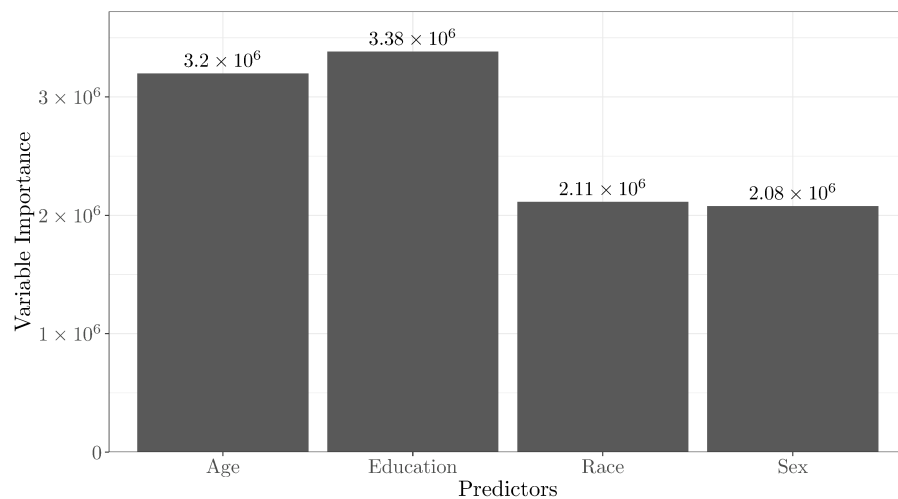
When considering only the exponential region, we observe similar behavior for education when transitioning from the FI to the EI analysis. One notable difference is the increased importance of the UF variable in the exponential region. Race does not exhibit the same behavior when transitioning from the EI to the PI analysis. As can be seen in Figure 19 despite the uncertainty coefficient $U(X = \text{Race} | Y = \text{education}) = 0.016$ being the highest, it is not significantly higher than the other coefficients in the education column. Thus, when education is introduced, Race remains more important than sex in the exponential region.

10.2 PARETO REGION

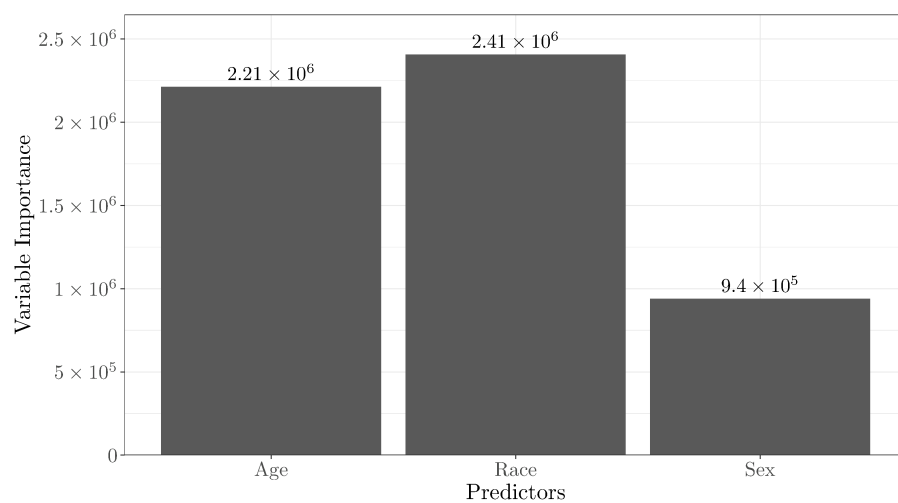
Now for the Pareto region the Figure 20 below shows the results of all types of variable importance analyses.



(a) FI: Variable importance ranking including all predictors.



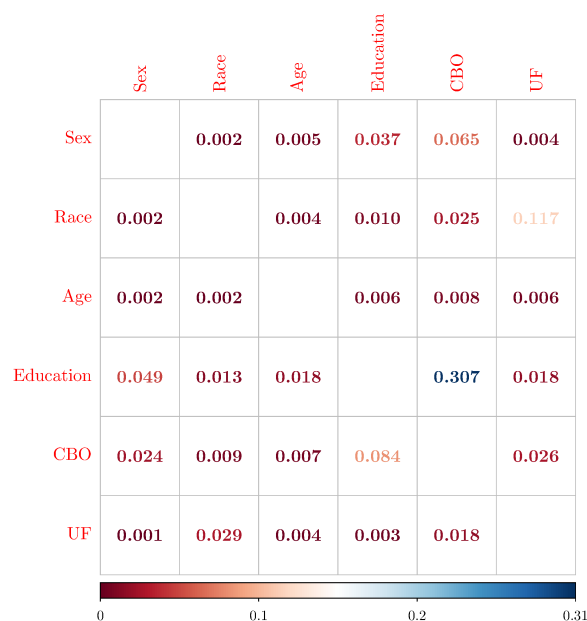
(b) EI: Variable importance ranking including age, sex, Race, and the highest level of education attained.



(c) PI: Variable importance ranking for age, sex, and Race.

Source: Own Author.

Figure 20 – Comparison of variable importance within the pareto region of PNADc.



Source: Own Author.

Figure 21 – The Theil's U matrix displays the entropy coefficient $U(X|Y)$ values for the set of predictor variables for the pareto region.

In the Pareto region, the results show a lower degree of importance for education. In the FI analysis, education ranked last, while in the EI analysis, it moved to the top rank again, but not by a considerable margin. Additionally, sex is more important than Race in the FI analysis but is less important in the other two analyses, even despite having the highest uncertainty coefficient, $U(X = \text{sex} | Y = \text{CBO}) = 0.065$, as it can be seen in Figure 21. This difference in behaviour indicates that sex has a higher importance in predicting wage when considering CBO. However, when CBO is removed in the EI and PI analyses, the absence of these distinct subgroup patterns reduces the significance of sex as a predictive variable. This suggests that sex in the FI analysis has a partial importance tied to its interaction with CBO, highlighting the role of occupational segregation in amplifying gender differences in the Pareto region. Also, the highest entropy coefficient is $U(X = \text{education} | Y = \text{CBO}) = 0.307$. This indicates that, similar to previous analyses in other regions, education shows the same behavior while transitioning from FI to EI analyses.

11 CONCLUSIONS AND PERSPECTIVES

The two-class model is a well-established framework for characterizing income distributions by combining an exponential (BG) regime with a Pareto power-law tail. Notably, the exponential component enjoys stability properties in multi-agent systems that the log-normal alternative does not offer, making it a natural choice for modeling the bulk of the population's earnings.

Prior research has located the transition between these regimes by examining log-log plots and visually identifying a change in slope or a discontinuity. In contrast, our study is the first to introduce an automated, data-driven procedure for pinpointing the optimal crossover income, thereby removing subjective judgment from the process.

We utilized two types of loss functions: RMSLE for simple sampling and MLE for stratified sampling. The first approach facilitates easy comparison with previous studies on PNADc, as, to our knowledge, all prior econophysics research on PNADc used simple sampling. The second approach incorporates the weights from PNADc, which uses stratified sampling, to provide an analysis that more accurately represents the entire population.

Using both RMSLE and MLE loss functions, we ran simulations of the two-class income model to see which loss criterion more accurately recovers known parameters. For each trial, we generated a synthetic CCDF from a preset model (with true parameter values) and then fitted both loss functions to that data. The results show that MLE consistently produces the closest estimates to the true parameters, although RMSLE also delivers strong, reliable performance.

The first approach led to several key findings. Our optimized crossover not only produces a continuous distribution but also achieves a substantially lower RMSLE compared with a fixed 5 % Pareto threshold. We validated the optimization via out-of-bag bootstrap cross-validation, which performed well on held-out samples. Disaggregating by sex and race/skin color revealed that male and White/yellow subgroups experience higher inequality and a larger Pareto share. In contrast, the Black/brown group shows the greatest equality and the smallest Pareto participation only 6.42 %.

We also tracked the temporal evolution of key metrics across successive PNADc waves, uncovering two novel correlations in Brazilian income data: one linking the theoretical and empirical Gini coefficients, and another associating the Pareto tail fraction ("Top %") with the Pareto exponent itself.

The end goal was to incorporate sample weighting and expansion to ensure proper treatment of the sampling distribution. In stratified sampling, weights adjust the data to account for varying representation across strata, which often correspond to geographic regions. This adjustment allows the sample to accurately reflect the entire population, even when some strata are under- or overrepresented. Sample weighting

and expansion are advanced topics in survey statistics and are often overlooked in regression models. In our second approach, applying these techniques yielded racial and sex subgroup results that closely match those from the unweighted analysis.

Lastly, we assessed variable importance in income prediction using conditional-inference random forests, supported by both CIT and CPP. This analysis evaluated three model specifications — FI, EI, and PI — to isolate the effects of occupation, education, and personal characteristics respectively.

In the FI model, occupation (CBO) and age emerged as the two most important predictors. When we excluded CBO and state (UF) in the EI model, education became the dominant factor. In the PI model, focused solely on personal characteristics, race proved to be the most influential variable. Additionally, Theil's U coefficient offered insight into the relationships among predictors, showing how their relative importance shifts depending on the modeling context.

We also analyzed each region of the two-class model separately: the exponential and Pareto regions. In the exponential region, education remained the dominant predictor in the EI model, while in the FI model, state (UF) gained importance compared to the overall population analysis. In the Pareto region, however, education's influence in the EI model decreased, and sex emerged as a stronger predictor. These contrasting patterns highlight how the drivers of income differ between the exponential and Pareto regions.

Despite providing an end-to-end, automated fitting framework, the loss surface is not guaranteed to be convex; finding the global minimum can be challenging. Moreover, the PSO regressions are computationally intensive, which slows down error-estimation loops considerably.

FUTURE DIRECTIONS

1. Apply our crossover optimization method to other national surveys (for example, European or North American datasets) to compare cross-country inequality patterns.
2. Extend the temporal analysis of weighted PNADc data.
3. Explore algorithmic strategies that deliver faster convergence yet still reliably locate the global minimum.

REFERENCES

- ALFONS, Andreas; TEMPL, Matthias; FILZMOSER, Peter. An object-oriented framework for statistical simulation: The R package simFrame. **Journal of Statistical Software**, v. 37, p. 1–36, 2010.
- ALFREDO, Guimarães A. S. Raça e os estudos de relações raciais no Brasil. **Novos Estudos CEBRAP**, São Paulo, n. 54, p. 147–156, 1999.
- AOYAMA, Hideaki; SOUMA, Wataru; NAGAHARA, Yuichi; OKAZAKI, Mitsuhiro P; TAKAYASU, Hideki; TAKAYASU, Misako. Pareto's law for income of individuals and debt of bankrupt companies. **Fractals**, World Scientific, v. 8, n. 03, p. 293–300, 2000.
- BANERJEE, Anand; YAKOVENKO, Victor M. Universal patterns of inequality. **New Journal of Physics**, IOP Publishing, v. 12, n. 7, p. 075032, 2010.
- BEE, Marco. Estimation of the lognormal-pareto distribution using probability weighted moments and maximum likelihood. **Communications in Statistics-Simulation and Computation**, Taylor & Francis, v. 44, n. 8, p. 2040–2060, 2015.
- BLACK, Fischer; SCHOLES, Myron. The pricing of options and corporate liabilities. **Journal of Political Economy**, v. 81, n. 3, p. 637–654, 1973.
- BRÄUNINGER, J. A variable metric algorithm for unconstrained minimization without evaluation of derivatives. **Numerische Mathematik**, Springer, v. 36, n. 4, p. 359–373, 1981.
- BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.
- BROYDEN, Charles George. The convergence of a class of double-rank minimization algorithms 1. general considerations. **IMA Journal of Applied Mathematics**, Oxford University Press, v. 6, n. 1, p. 76–90, 1970.
- BUTLER, Judith. **Gender Trouble: Feminism and the Subversion of Identity**. New York: Routledge, 1990.
- BYRD, Richard H; LU, Peihuang; NOCEDAL, Jorge; ZHU, Ciyou. A limited memory algorithm for bound constrained optimization. **SIAM Journal on scientific computing**, SIAM, v. 16, n. 5, p. 1190–1208, 1995.

CHATTERJEE, Arnab; CHAKRABARTI, Bikas K; MANNA, S. S. Pareto law in a kinetic model of market with random saving propensity. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 335, n. 1-2, p. 155–163, 2004.

CHEN, Li-Pang. Supervised learning for binary classification on us adult income. **Journal of Modeling and Optimization**, v. 13, n. 2, p. 80–91, 2021.

CLEMENTI, Fabio; GALLEGATI, Mauro. Pareto's law of income distribution: Evidence for Germany, the United Kingdom, and the United States. In: **ECONOPHYSICS of wealth distributions**. [S.l.]: Springer, 2005. P. 3–14.

DEBEER, Dries; STROBL, Carolin. Conditional permutation importance revisited. **BMC bioinformatics**, Springer, v. 21, p. 1–30, 2020.

DRAGULESCU, Adrian; YAKOVENKO, Victor M. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 299, n. 1-2, p. 213–221, 2001.

DRAGULESCU, Adrian; YAKOVENKO, Victor M. Statistical mechanics of money. **The European Physical Journal B-Condensed Matter and Complex Systems**, Springer, v. 17, n. 4, p. 723–729, 2000.

EFRON, Bradley. Estimating the error rate of a prediction rule: improvement on cross-validation. **Journal of the American statistical association**, Taylor & Francis, v. 78, n. 382, p. 316–331, 1983.

EFRON, Bradley; TIBSHIRANI, Robert. Improvements on cross-validation: the 632+ bootstrap method. **Journal of the American Statistical Association**, Taylor & Francis, v. 92, n. 438, p. 548–560, 1997.

FIX, Blair. Personal income and hierarchical power. **Journal of Economic Issues**, Taylor & Francis, v. 53, n. 4, p. 928–945, 2019.

FONTANARI, Andrea; TALEB, Nassim Nicholas; CIRILLO, Pasquale. Gini estimation under infinite variance. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 502, p. 256–269, 2018.

GONZALEZ-ESTEVEZ, J.; COSENZA, M. G.; LOPEZ-RUIZ, R.; SANCHEZ, J. R. Pareto and Boltzmann–Gibbs behaviors in a deterministic multi-agent system. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 387, n. 18, p. 4637–4642, 2008.

GUO, Jiayi; LEWIS, A. S. Nonsmooth variants of Powell's BFGS convergence theorem. **SIAM Journal on Optimization**, SIAM, v. 28, n. 2, p. 1301–1311, 2018.

HARRIS, Marvin. **Patterns of Race in the Americas**. New York: Walker and Company, 1964.

HILL, Bruce M. A simple general approach to inference about the tail of a distribution. **The annals of statistics**, JSTOR, p. 1163–1174, 1975.

HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased recursive partitioning: A conditional inference framework. **Journal of Computational and Graphical statistics**, Taylor & Francis, v. 15, n. 3, p. 651–674, 2006.

IBGE. **PNAD - Nota Técnica**. [S.l.], 2015. Available from: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv98887.pdf>.

IBGE. **PNAD Contínua - Pesquisa Nacional por Amostra de Domicílios Contínua: Metodologia**. [S.l.], 2025. Available from: https://biblioteca.ibge.gov.br/visualizacao/livros/liv102172_notas_tecnicas.pdf.

IBGE. **Sistema Nacional De Índices De Preços Ao Consumidor - SNIPC**. [S.l.: s.n.], 2020. Available from: https://ftp.ibge.gov.br/Precos_Indices_de_Precos_ao_Consumidor/Sistema_de_Indices_de_Precos_ao_Consumidor/Notas_Tecnicas/snipc_nota_tecnica_2020_01.pdf.

KALECKI, Micha. On the Gibrat distribution. **Econometrica: Journal of the Econometric Society**, JSTOR, p. 161–170, 1945.

KENNEDY, James; EBERHART, Russell. Particle swarm optimization. In: IEEE. PROCEEDINGS of ICNN'95-International Conference on Neural Networks. [S.l.: s.n.], 1995. P. 1942–1948.

KOLENIKOV, Stanislav. Resampling variance estimation for complex survey data. **The Stata Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 10, n. 2, p. 165–199, 2010.

LAZAR, Alina. Income prediction via support vector machine. In: ICMLA. [S.l.: s.n.], 2004. P. 143–149.

LEVY, Moshe; SOLOMON, Sorin. New evidence for the power-law distribution of wealth. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 242, n. 1-2, p. 90–94, 1997.

LI, Shutao; TAN, Mingkui; TSANG, Ivor W; KWOK, James Tin-Yau. A hybrid PSO-BFGS strategy for global optimization of multimodal functions. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, IEEE, v. 41, n. 4, p. 1003–1014, 2011.

MANDELBROT, Benoit. The Pareto-Levy law and the distribution of income. **International economic review**, JSTOR, v. 1, n. 2, p. 79–106, 1960.

MARKATOU, Marianthi; BASU, Ayanedranath; LINDSAY, Bruce. Weighted likelihood estimating equations: The discrete case with applications to logistic regression. **Journal of Statistical Planning and Inference**, Elsevier, v. 57, n. 2, p. 215–232, 1997.

MARTELETO, Leticia J. Educational inequality by race in Brazil, 1982–2007: structural changes and shifts in racial classification. **Demography**, Springer, v. 49, p. 337–358, 2012.

MINISTRY OF LABOR AND EMPLOYMENT (MTE). **Brazilian Occupational Classification – CBO 2002: Introduction**. [S.l.], 2002. Available from: <https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/cbo/servicos/downloads/livro-1-portal-cbo.pdf>.

MOORE, H. L. Cours d'Économie Politique. By VILFREDO PARETO, Professeur à l'Université de Lausanne. Vol. I. Pp. 430. 1896. Vol. II. Pp. 426. 1897. Lausanne: F. Rouge. **The ANNALS of the American Academy of Political and Social Science**, Sage Publications Sage CA: Thousand Oaks, CA, v. 9, n. 3, p. 128–131, 1897.

MORGAN, Marc. Extreme and persistent inequality: new evidence for Brazil combining national accounts, surveys and fiscal data, 2001-2015, 2017.

MYUNG, In Jae. Tutorial on maximum likelihood estimation. **Journal of mathematical Psychology**, Elsevier, v. 47, n. 1, p. 90–100, 2003.

NOCEDAL, Jorge. Updating quasi-Newton matrices with limited storage. **Mathematics of computation**, v. 35, n. 151, p. 773–782, 1980.

OANCEA, Bogdan; ANDREI, Tudorel; PIRJOL, Dan. Income inequality in Romania: The exponential-Pareto distribution. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 469, p. 486–498, 2017.

PREDA, Vasile; CIUMARA, Roxana, et al. On composite models: Weibull-Pareto and Lognormal-Pareto. A comparative study. **Romanian Journal of Economic Forecasting**, Citeseer, v. 3, n. 2, p. 32–46, 2006.

RAO, J. N. K.; WU, C. F. J. Resampling inference with complex survey data. **Journal of the American Statistical Association**, v. 83, n. 401, p. 231–241, 1988. DOI: 10.1080/01621459.1988.10478591. Available from: <https://doi.org/10.1080/01621459.1988.10478591>.

RASCH, DAMK. Mathematische Statistik, Joh. **Ambrosius Barth. Heidelberg, Leipzig**, 1995.

REDNER, Sidney. Random multiplicative processes: An elementary tutorial. **Am. J. Phys**, v. 58, n. 3, p. 267–273, 1990.

RUST, Keith F.; RAO, J. N. K. Variance estimation for complex surveys using replication techniques. **Statistical methods in medical research**, Sage Publications Sage CA: Thousand Oaks, CA, v. 5, n. 3, p. 283–310, 1996.

SAFARI, Muhammad Aslam Mohd; MASSERAN, Nurulkamal; IBRAHIM, Kamarulzaman. Optimal threshold for Pareto tail modelling in the presence of outliers. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 509, p. 169–180, 2018.

SANTOS, Paulo H. dos; SICILIANI, Igor D. S.; TRAGTENBERG, M. H. R. Optimal income crossover for a two-class model using particle swarm optimization. **Physical Review E**, APS, v. 106, n. 3, p. 034313, 2022.

SHI, Yuhui; EBERHART, Russell. A modified particle swarm optimizer. In: IEEE. 1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360). [S.l.: s.n.], 1998a. P. 69–73.

SHI, Yuhui; EBERHART, Russell. A modified particle swarm optimizer. In: IEEE. 1998 IEEE international conference on evolutionary computation proceedings. IEEE world congress on computational intelligence (Cat. No. 98TH8360). [S.l.: s.n.], 1998b. P. 69–73.

SHIRRAS, G Findlay. The Pareto Law and the distribution of income. **The Economic Journal**, JSTOR, v. 45, n. 180, p. 663–681, 1935.

SICILIANI, Igor D. S.; TRAGTENBERG, M. H. R. Kinetic theory and Brazilian income distribution. **Physica A: Statistical Mechanics and its Applications**, Elsevier, v. 513, p. 166–174, 2019.

SILVA, A Christian; YAKOVENKO, Victor M. Temporal evolution of the "thermal" and "superthermal" income classes in the USA during 1983–2001. **EPL (Europhysics Letters)**, IOP Publishing, v. 69, n. 2, p. 304, 2004a.

SILVA, A. Christian; YAKOVENKO, Victor M. Temporal evolution of the "thermal" and "superthermal" income classes in the USA during 1983–2001. **EPL (Europhysics Letters)**, IOP Publishing, v. 69, n. 2, p. 304, 2004b.

SKAJAA, Anders. Limited memory BFGS for nonsmooth optimization. **Master's thesis**, Citeseer, 2010.

SKVORTSOV, Grigorii G; YAKOVENKO, Marina V; CASTRO, Pascal M; FUKIN, Georgy K; CHERKASOV, Anton V; CARPENTIER, Jean-François; TRIFONOV, Alexander A. **Lanthanide Borohydride Complexes of Bulky Guanidinate Ligands [(Me₃Si)₂NC(N-Cy)₂]₂Ln (μ-BH₄)₂Li (THF)₂ (Ln= Nd, Sm, Yb): Synthesis, Structure and Catalytic Activity in Lactide Polymerization**. [S.l.]: Wiley Online Library, 2007.

SOUMA, Wataru. Physics of personal income. In: EMPIRICAL science of financial fluctuations. [S.l.]: Springer, 2002. P. 343–352.

STRASSER, Helmut; WEBER, Christian. On the asymptotic theory of permutation statistics, 1999.

STROBL, Carolin; BOULESTEIX, Anne-Laure; KNEIB, Thomas; AUGUSTIN, Thomas; ZEILEIS, Achim. Conditional variable importance for random forests. **BMC bioinformatics**, Springer, v. 9, p. 1–11, 2008.

SUTTON, John. Gibrat's legacy. **Journal of economic literature**, JSTOR, v. 35, n. 1, p. 40–59, 1997.

TAO, Yong; WU, Xiangjun; ZHOU, Tao; YAN, Weibo; HUANG, Yanyuxiang; YU, Han; MONDAL, Benedict; YAKOVENKO, Victor M. Exponential structure of income inequality: evidence from 67 countries. **Journal of Economic Interaction and Coordination**, Springer, v. 14, n. 2, p. 345–376, 2019.

VAZ, Daniela Verzola. Family background, returns to education and racial inequality in brazil. **Cadernos de Pesquisa**, SciELO Brasil, v. 50, p. 845–864, 2020.

WAGLEY, Charles. **Race and Class in Rural Brazil**. Paris: UNESCO, 1952. Available from: <https://unesdoc.unesco.org/ark:/48223/pf0000135515>.

WEI, Pengfei; LU, Zhenzhou; SONG, Jingwen. Variable importance analysis: A comprehensive review. **Reliability Engineering & System Safety**, Elsevier, v. 142, p. 399–432, 2015.

APPENDIX A – HYBRID PSO ALGORITHM

Algorithm 1 Hybrid-Pso

```

1: Initialize variables:  $\mathbf{x}_i$ ,  $\mathbf{v}_i$ ,  $\mathbf{P}_i$  and  $\mathbf{G}$ ; according to Eq.
2: for every  $i$  candidate: do
3:    $y_i \leftarrow \text{Loss}(\mathbf{P}_i)$ ;
4: end for
5: repeat
6:   for every  $i$  candidate: do
7:     Determine  $\mathbf{G}$  given the current topology  $S_i$ 
8:      $\mathbf{v}_i \leftarrow W\mathbf{v}_i + c_1r_1(\mathbf{P}_i - \mathbf{x}_i) + c_2r_2(\mathbf{G} - \mathbf{x}_i)$ ;
9:      $\mathbf{x}_i \leftarrow \mathbf{x}_i + \mathbf{v}_i$ ;
10:    if  $f(\mathbf{x}_i) < y_i$  then
11:       $\mathbf{x}'_i \leftarrow \text{L-BFGS-B-search}(\mathbf{x}_i)$ ;
12:       $\mathbf{p}_i \leftarrow \mathbf{x}'_i$ 
13:       $y_i \leftarrow \text{Loss}(\mathbf{x}'_i)$ 
14:    end if
15:  end for
16: until The stopping condition is true;

```

APPENDIX B – BFGS SUMMARY

Algorithm 2 BFGS

- 1: Initialize Inverse Hessian $B_0 = I$; and the step $t = 0$;
 - 2: **repeat**
 - 3: Get the search direction $d_t = -B_t \nabla L(\mathbf{x}_t)$;
 - 4: Make a line search $\mathbf{x}_{t+1} = \mathbf{x}_t + \alpha_t d_t$; (α_t given by Golden section search)
 - 5: $p_t \leftarrow \mathbf{x}_{t+1}; q_t \leftarrow \nabla L(\mathbf{x}_{t+1}) - \nabla L(\mathbf{x}_t)$;
 - 6: $B_{t+1} = \left(I - \frac{p_t q_t^T}{q_t^T p_t} \right) B_t \left(I - \frac{q_t p_t^T}{p_t^T q_t} \right) + \frac{p_t p_t^T}{q_t^T q_t}$
 - 7: $t \leftarrow t + 1$
 - 8: **until** The stopping condition is true;
-

APPENDIX C – THE NATIONAL HOUSEHOLD SAMPLE SURVEY

The PNAD represents Brazil's primary source of socio-economic and demographic statistics at the household level, conducted by the IBGE. The PNAD program for annual data has employed two distinct methodological approaches over its operational history: the traditional PNAD (1992-2015) and the continuous PNAD (2012-2024). Each approach represents a fundamentally different sampling strategy and operational design, reflecting evolving needs for labor market monitoring and socio-demographic analysis in Brazil. In this study we used 2001 to 2011 data from the traditional PNAD and 2012 to 2019 data from the PNADc.

The traditional PNAD operated as an annual cross-sectional survey with a multi-stage stratified cluster design, while PNADc introduced a continuous quarterly design with a sophisticated household rotation scheme. This methodological evolution reflects IBGE's response to demands for more timely labor market indicators and enhanced geographical coverage, particularly in rural areas that were previously underrepresented in the sampling frame (IBGE, 2025).

C.1 SURVEY DESIGN EVOLUTION

C.1.1 Traditional PNAD (2001-2011)

The traditional PNAD was conducted annually as a sample household survey of private households across Brazil. The survey was designed as a national survey providing representative results for Brazil as a whole, its regions, individual states, and major metropolitan areas (IBGE, 2015).

The primary objective of the traditional PNAD was to produce fundamental socio-economic and demographic indicators encompassing population characteristics, education levels, labor market conditions, income distribution, and housing conditions. These indicators served as essential tools for analyzing Brazil's socio-economic development.

C.1.2 PNADc (2012-2019)

In contrast to its predecessor, the PNADc represents an ongoing trimester survey that replaced the traditional PNAD beginning in 2012 (IBGE, 2025). The continuous design focuses specifically on monitoring labor force dynamics and related socio-demographic variables on an ongoing basis, providing quarterly and annual indicators of employment, unemployment, education, migration, and related topics.

A significant improvement in the PNADc is its more extensive geographical coverage. The survey was implemented nationwide, including all rural areas, from its incep-

tion in 2012. This represents a substantial improvement over the pre-2004 traditional PNAD, which had excluded certain rural areas from its sampling frame.

C.2 SAMPLING METHODOLOGY

C.2.1 Traditional PNAD Sampling Design (2001-2011)

The traditional PNAD was implemented via a **three-stage probabilistic sample** of households, based on the 2010 Demographic Census territorial divisions and census-tract mesh (IBGE, 2015):

- **First Stage (Primary Sampling Units - Primary Sampling Units (PSUs)):** Municipalities were classified into two categories:
 - *Self-representing units (SRs):* Selected with certainty.
 - *Non-self-representing units (NSRs):* Stratified, then selected *with replacement* via Probability Proportional to Size (PPS) using the 2010 Census resident population as the size measure.
- **Second Stage (Secondary Sampling Units):** Within each selected municipality, census tracts (setores censitários) were selected *with replacement* via PPS, using the number of households recorded in the 2010 Census as the size measure.
- **Third Stage (Ultimate Sampling Units):** Within each selected tract, private households and housing units in collective households were listed and a simple random sample of households was drawn with equal probability.

C.2.2 PNADc Sampling Design (2012-2019)

The PNADc employs a two-stage stratified cluster design drawing on the 2010 Population Census master sample (IBGE, 2025):

- **Stage 1 (Primary Sampling Units - PSUs):** The IBGE aggregates neighboring small census tracts into PSUs, each containing at least 60 occupied households; larger tracts remain as standalone PSUs. PSUs are stratified by state and metropolitan status. Within each stratum, PSUs are selected *with replacement* using PPS, where the measure of size is the number of occupied households from the 2010 Census.
- **Stage 2 (Ultimate Sampling Units - Households):** Within each selected UPA, a simple random sample without replacement of 14 occupied private dwellings is drawn from updated household listings.

Each quarter, the PNADc interviews approximately 211.000 households across some 16.000 PSUs.

C.2.3 Rotation Scheme

A distinctive feature of PNADc is its sophisticated rotation scheme. Each selected household participates in five consecutive quarters following a "1-2(5)" pattern: households are interviewed, skip two quarters, then interviewed again, continuing this pattern for a total of five interviews.

This design ensures that in any given quarter, the sample comprises five rotation groups representing households in their 1st, 2nd, 3rd, 4th, or 5th interview. Only permanent private households are eligible for inclusion; seasonal and collective households are excluded from the sampling frame.

C.3 CALIBRATION TO POPULATION BENCHMARKS

C.3.1 Calibration Methodology

IBGE applies calibration techniques to the base weights of both PNAD surveys to ensure alignment with official population projections, maintaining consistency with demographic totals derived from CENSO projections (IBGE, 2015, 2025).

C.3.2 Traditional PNAD Calibration

For the traditional PNAD, calibration was performed after adjusting for selection probabilities and nonresponse. Expansion factors (*fator de expansão*) underwent post-stratification or calibration using known population margins by state, sex, age group, and urban/rural status, based on IBGE's population projections.

These population projections were periodically revised, with IBGE adopting the "AiBi" method from 2008 onward, later transitioning to component-based demographic models.

C.3.3 PNADc Calibration

Each quarter and within each rotation group, the base sampling weight (the inverse of inclusion probability adjusted for nonresponse, variable V1028) is calibrated by *raking* to known population margins from IBGE's quarterly population estimates.

The margins include domains defined by state, metropolitan status, sex and age group. Raking (iterative proportional fitting) replaces the earlier post-stratification approach, adjusting the weights so that sample margins exactly match the auxiliary totals, subject to bounds and preserving equal weights among residents of the same household.

C.4 VARIANCE ESTIMATION AND REPLICATE WEIGHTS

C.4.1 Bootstrap Methodology

To enable variance estimation that properly accounts for the complex survey design and calibration procedures, IBGE provides replicate weights based on a bootstrap method, formally introduced in PNADc from third quarter 2021.

The replicate weights are constructed using the Rao-Wu bootstrap method specifically designed for stratified multistage designs. For each replicate, PSUs are resampled with replacement within strata, and household weights are re-calibrated to the same population benchmarks used for the full sample. This approach captures both design effects and the impact of calibration in variance estimation (RAO; WU, 1988).

C.4.2 Microdata Structure

In the microdata files, the variance estimation components are structured as follows:

- The **main calibrated weight** is stored in variable V1032
- The **replicate weights** are stored in variables following the pattern: V1032001, V1032002, ..., V1032200
- There are exactly **200 replicate weights** available for variance estimation

C.4.3 Variance Estimation Procedure

The standard procedure for variance estimation using replicate weights follows these steps:

1. Compute the point estimate using V1032 (the full-sample weight).
2. Compute the same estimate across all 200 replicate weights.
3. Calculate the variance as the sample variance across the 200 replicate estimates, scaled by the appropriate factor.

This replicate-based approach supersedes earlier linearization methods and represents the current official standard for PNADc design-based statistical inference. This method ensures that the variance estimation is consistent for every user of the PNADc microdata, and more importantly, it is consistent with the official IBGE variance estimation methodology. Moreover, the bootstrap method allows for the estimation of the standard error of complex statistics or model-based estimators, such as regression coefficients or variance estimates, which is often not feasible with the linearization method.

When calculating variance using replicate weights, the Mean Squared Error (MSE) approach is employed for PNADc. The formula for the this variance estimation can be represented as follows:

$$\hat{Var}(\hat{\theta}) = \frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}_r - \hat{\theta})^2 \quad (C.1)$$

where $\hat{\theta}_r$ is the estimate using the r th replicate weight, $\hat{\theta}$ is the full-sample estimate computed with the original weight (variable V1032), and $R = 200$ is the total number of replicates.

Each replicate weight represents a slight perturbation of the original sample under your survey design; the deviations $\hat{\theta}_r - \hat{\theta}$ show how much the estimate would shift across these pseudo-samples. By squaring and averaging those deviations (with the $1/(R-1)$ factor for unbiasedness), you obtain a design-based estimate of the variance.

C.4.3.0.1 Use of Replicate Weights for Complex Statistics

This replicate-based variance estimation approach is highly flexible and allows for the computation of variances for complex statistics or model-based estimators, such as regression coefficients or other summary statistics. The MSE framework is particularly useful when dealing with such complex estimators, as it provides an estimate of both the variance and the bias of the estimator. This is often not achievable using traditional linearization techniques.

In summary, variance estimation using replicate weights with MSE = TRUE allows for a more robust and accurate measure of uncertainty in survey-based estimates, consistent with the methodology recommended by IBGE for the PNADc.

C.5 PNADC MICRODATA INCOME AND DEMOGRAPHIC INDICATORS

The PNADc microdata variables are organized with codes, where the prefix "V" designates a pure variable directly extracted from the survey, and "VD" represents a composed variable, typically a linear equation involving pure variables. For the total monthly income, we utilized the variable VD4022¹, which signifies the total income from all sources, including aid programs, among others. The variables employed in this study are V2007 (sex), V2009 (age), V2010 (race/skin color), V3009A (highest level of education attained), VD4011 (major occupational group based on the Brazilian Classification of Occupations — CBO), and UF (federative unit or state of residence). For detailed information about each variable used in this study, please refer to the Table 6. Although these are standard variables across survey years, the numeric ranges and number of observations presented in this section refer specifically to the 2019 dataset

(Annual, Visit 1), which comprises 443,790 observations, of which 266,231 constitute the active group, reporting positive income; the remainder have zero or missing income.

Table 6 – Description of variables. The ranges and levels correspond to the active group in the PNADc 2019 dataset.

Variable	Name	Type	Levels or Range
V2007	Sex	Categorical	Men, Women
V2009	Age	Numerical	[0, 113]
V2010	Race	Categorical	Black, Brown, Indigenous, White, Yellow, Disregarded
V3009A	Education	Categorical	Pre-school, Literacy class, Youth and adult literacy, Elementary, Middle school (1st cycle), Regular elementary school or 1st grade, Youth and adult education or 1st grade supplementary, High school (2nd cycle), Regular high school or 2nd grade, Youth and adult education or 2nd grade supplementary, Undergraduate, Specialization (postgraduate), Master's degree, Doctorate, Not applicable
VD4011	Occupation	Categorical	Technicians and associate professionals; Professionals in science and intellectual fields; Elementary occupations; Skilled workers in construction, mechanics, and related trades; Service workers and shop/market sales workers; Clerical support workers; Machine operators and assemblers; Armed forces, police officers and firefighters; Directors and managers; Skilled agricultural, forestry, hunting and fishing workers; Ill-defined occupations
UF	State	Categorical	Rondônia, Acre, Amazonas, Roraima, Pará, Amapá, Tocantins, Maranhão, Piauí, Ceará, Rio Grande do Norte, Paraíba, Pernambuco, Alagoas, Sergipe, Bahia, Minas Gerais, Espírito Santo, Rio de Janeiro, São Paulo, Paraná, Santa Catarina, Rio Grande do Sul, Mato Grosso do Sul, Mato Grosso, Goiás, Distrito Federal
VD4022	Income	Numerical	[1, 350000]

¹ PNADc microdata variables are organized with codes, the prefix V is for a pure variable that is extracted directly from the survey and the VD is for the composed variable, usually a linear equation of pure variables. For the total monthly income we use the variable VD4022, which is the total income from all sources, including aid programs among others. This variable is only present in the first annual interview of PNADc.