Optimal income crossover for a two-class model using particle swarm optimization

Paulo H. dos Santos,[*] Igor D. S. Siciliani,[†] and M. H. R. Tragtenberg[‡]
Departamento de Física, Universidade Federal de Santa Catarina, Florianópolis, Santa Catarina, 88040-900, Brazil

(Received 9 December 2021; accepted 26 August 2022; published 13 September 2022)

Personal income distribution may exhibit a two-class structure, such that the lower income class of the population (85–98%) is described by exponential Boltzmann-Gibbs distribution, whereas the upper income class (2–15%) has a Pareto power-law distribution. We propose a method, based on a theoretical and numerical optimization scheme, which allows us to determine the crossover income between the distributions, the temperature of the Boltzmann-Gibbs distribution, and the Pareto index. Using this method, the Brazilian income distribution data provided by the National Household Sample Survey was studied. The data was stratified into two dichotomies (sex/gender and color/race), so the model was tested using different subsets along with accessing the economic differences between these groups. Last, we analyze the temporal evolution of the parameters of our model and the Gini coefficient discussing the implication on the Brazilian income inequality. In this paper, we propose an optimization method to find a continuous two-class income distribution, which is able to delimit the boundaries of the two distributions. It also gives a measure of inequality which is a function that depends only on the Pareto index and the percentage of people in the high-income region. We found a temporal dynamics relation, that may be general, between the Pareto and the percentage of people described by the Pareto tail.

DOI: 10.1103/PhysRevE.106.034313

I. INTRODUCTION

A long time ago, the economist Vilfredo Pareto identified a power-law behavior in the income distribution [1]. Pareto stated that the income probability density function describing this distribution is of the form

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

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 [2].

The Pareto power law was confirmed extensively on the upper income data from different countries [3,4] and was also found to describe wealth distribution [5]. In this paper 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 [6,7]. It is worth mentioning that the most used distribution for this region is the log-normal distribution; however, unlike the Boltzmann-Gibbs it is not a stationary distribution [8]. In a multiagent simulations context, its asymptotic states were capable of displaying Boltzmann-Gibbs as well as Pareto statistical behaviors [9,10]. Moreover, the two-class model arrives from a Fokker-Planck equation, considering an additive and multiplicative processes for the exponential and Pareto region, respectively [11]. Therefore, for income less than the crossover income, m_c, the distribution is given by

$$P(m) = \frac{a}{T}\exp\left(\frac{-m}{T}\right), \quad m < m_c, \tag{2}$$

where a is a normalization parameter and T is the "temperature" of the system.

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 (2–15%) follows a Pareto power-law distribution. 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 [12–14]. However, this choice is rather subjective, and therefore, the crossover income determined is not optimal.

We propose in this paper a method to determine the total income distribution defined by parts, thus the crossover income can be established optimally. First, we define a measure of goodness-of-fit statistics that will be minimized by a numerical algorithm called particle swarm optimization (PSO) with limited-memory Broyden-Fletcher-Goldfarb-Shanno bound (L-BFGS-B). We validate this method by studying the Brazilian income distribution using data from National Household Sample Survey (PNAD), an annual research available by the Brazilian Institute of Geography and Statistics (IBGE).

[*]psantos.fsc@gmail.com
[†]igorschoeller@gmail.com
[‡]marcelo.tragtenberg@ufsc.br

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 model. The second is the correlation between the Pareto index and the percentage of people that display the Pareto power-law behavior.

This paper is organized as follows. Section II explores the two-class complementary cumulative distribution function (CCDF) and its continuity. In Sec. III is derived the relation between the parameters of our model and the Gini coefficient. Section IV describes the L-BFGS-B particle swarm optimization and justifies our choice. We applied in Sec. V the PSO optimizer fitting our model into the Brazilian income distribution data of the total population and the stratified population (sex/gender and color/race), performing a cross validation with a re-sampling technique. Section VI explores the time evolution of the parameters of our model by fitting our model into the Brazilian income distribution in the years between 2001 and 2019. Conclusions are found in Sec. VII.

II. TWO-CLASS MODEL FOR INCOME DISTRIBUTION

We can define a two-class model for a country income distribution using Eqs. (1) and (2),

$$P(m) = \begin{cases} \frac{a}{T} e^{-m/T} & m < m_c, \\ b m^{-\alpha-1} & m \geqslant m, \end{cases} \tag{3}$$

equivalently CCDF

$$\int_m^\infty P(m')dm' \equiv \hat{C}(m) = \begin{cases} a[\exp(-m/T) - \exp(-m_c/T)] + \lambda, & m < m_c, \\ \lambda\left(\frac{m}{m_c}\right)^{-\alpha}, & m \geqslant m_c, \end{cases} \tag{4}$$

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 gives us

$$\lambda = 1 - a(1 - e^{-m_c/T}). \tag{5}$$

Setting $a = 1$ makes the Pareto distribution a correction to the exponential in the high-income tail. This makes the parameters of the model more interpretable and easier to optimize. So Eqs. (4) and (5) become

$$\hat{C}(m, \lambda, T, \alpha) = \begin{cases} \exp\left(\frac{-m}{T}\right), & m < m_c(\lambda, T), \\ \lambda\left(\frac{m}{m_c(\lambda,T)}\right)^{-\alpha}, & m \geqslant m_c(\lambda, T), \end{cases} \tag{6}$$

and

$$m_c(\lambda, T) = T \log(\lambda^{-1}). \tag{7}$$

The CCDF in Eq. (6) is the two-class model by Yakovenko [12]. 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.

In previous methods of the two-class model, the m_c is set by taking the intersection between the exponential fit (with $T = \langle m \rangle$) for the 85–98% 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 Fig. 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. The most commonly used method to determine the power-law income threshold is to plot the CCDF in a log-log scale and see where the behavior is linear. A more robust option is to use a goodness-of-fit function to optimize the threshold income [15].

In this paper we are going to determine the model parameters in Eq. (6) with a hybrid version of a numerical optimizer called particle swarm optimizer. To achieve this, starting with a given set of income values $\{m_i\}$, our method will predict a vector of parameters $\{\lambda_p, T_p, \alpha_p\}$, of the income distribution $\hat{C}(m, \lambda_p, T_p, \alpha_p)$ given by Eq. (6). With this function we are going to be able to minimize the root-mean-squared logarithmic error (RMSLE) applying a hybrid approach of PSO. This method will display continuity and the equality between the Pareto income threshold and m_c.

III. 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,

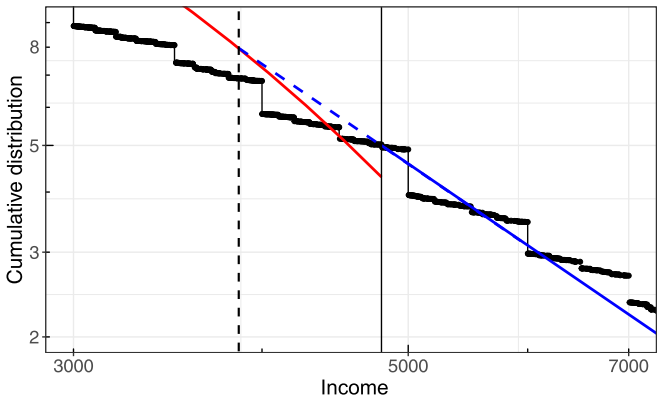


FIG. 1. Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of the data and the solid lines correspond to the fitted model described in Eq. (6). The red curve obeys Boltzmann-Gibbs distribution, the power-law distribution is characterized 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.

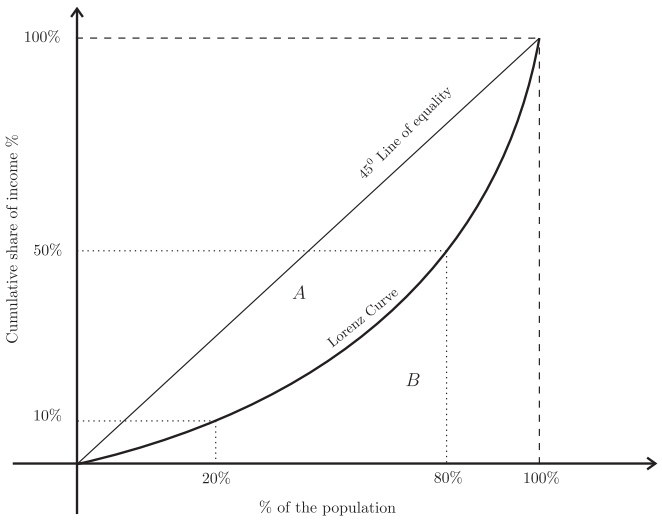


FIG. 2. The Lorenz curve framework (hypothetical data).

hence the Lorenz curve would follow a 45° line. As inequality increases, the Lorenz curve deviates from the line of equality as shown in Fig. 2.

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 $A + B = 0.5$.

We are interested in calculating the Gini coefficient given by the regression fit. In this study, we will call this indicator as the theoretical Gini coefficient, since it was calculated using the two-class model fit. Therefore, 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:

$$B = \int_0^1 L(m)dF(m) = \int_0^\infty L(m)P(m)dm, \tag{8}$$

where $P(m)$ is the probability density function, μ is the average income, and

$$L(m) = \frac{1}{\mu}\int_0^m xP(x)dx \tag{9}$$

is the percentage of total income by the population with up to income m. Simplifying Eq. (8) using integration by parts, the Gini coefficient becomes

$$G(C) = 1 - \frac{1}{\mu}\int_0^\infty C(m)^2 dm, \tag{10}$$

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

$$G(\lambda, \alpha) = 1 - \frac{(1-\lambda^2)/2 - (\lambda^2 \log\lambda)/(2\alpha - 1)}{(1-\lambda) - \lambda\log\lambda/(\alpha - 1)}, \tag{11}$$

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

$$G(\lambda, \alpha) = \frac{1}{2} + \left[\frac{\log(\lambda^{-1})}{(\alpha - 1)} - 1\right]\frac{\lambda}{2} + \mathcal{O}(\lambda^2 \log\lambda). \tag{12}$$

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.

IV. OPTIMIZATION OF THE TWO-CLASS MODEL WITH HYBRID-PSO

In this section we are going to detail the procedures to perform the particle swarm optimization to fit the empirical CCDF by the two-class model. To better differentiate the theoretical and empirical variables or statistic, we will use the following notation:

$$x \longrightarrow \text{for empirical variables or statistic,}$$

$$\hat{x} \longrightarrow \text{for theoretical variables or statistic.}$$

First, we are going to calculate the empirical CCDF. Taking a sample $m_1, m_2..., m_N$ of income drawn from a population, in this case Brazilian income, thus $m_{(1)} \geqslant m_{(2)} \geqslant ...m_{(N)}$ is the income order statistic. Accordingly, the empirical CCDF of the PNAD income data is defined as

$$C(m_n) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{m_{(i)} \geqslant m_n}, \tag{13}$$

$$\mathbb{1}_{m_{(i)} \geqslant m_n} = \begin{cases} 1, & m_{(i)} \geqslant m_n, \\ 0, & m_{(i)} < m_n, \end{cases} \tag{14}$$

where N is the total number of people in the data and $I[(m > m_i)]$ is the indicator function.

After finding the cumulative distribution of the data we need to define our loss function (quality measure). 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 [12,13]. 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

$$\langle \hat{m} \rangle_{\exp} = T - m_c\left[\exp\left(\frac{m_c}{T}\right) - 1\right]^{-1} \tag{15}$$

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

$$l_1 \overset{\text{def}}{=} \left| \frac{T - m_c\left[\exp\left(\frac{m_c}{T}\right) - 1\right]^{-1}}{\frac{1}{N_e}\sum_{i=1}^{N_e} m_{(i)}} - 1 \right|, \tag{16}$$

where 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

$$l_2 \stackrel{\text{def}}{=} \left| \frac{\lambda}{C(m_c)} - 1 \right|. \tag{17}$$

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

Let $\eta_n \geqslant \eta_{n-1} ... \geqslant \eta_1$ be an order statistic that follows $\eta_n = m_{(\lfloor n\frac{N}{k} \rfloor)}$, where $n = 1, 2, ..., k$ and $k < N$. This will be called class statistic and it divides the data into k income points. The notation $\lfloor . \rfloor$ denotes the floor function, which can be formally define as $\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \leqslant x\}$.

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

$$\begin{aligned} &\text{RMSLE}[C(\eta_n), \hat{C}(\eta_n, \boldsymbol{x})] \\ &= \sqrt{\frac{1}{k-1} \sum_{n=1}^{k-1} \{\log[C(\eta_n)] - \log[\hat{C}(\eta_n, \boldsymbol{x})]\}^2}, \end{aligned} \tag{18}$$

where $C(\eta_n)$ is the empirical complementary cumulative distribution, $\hat{C}(\eta_n, \boldsymbol{x})$ is the CCDF of the model and $\boldsymbol{x}$ 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 paper $k = 10\,000$.

Therefore using Eqs. (16)–(18) we define the loss function

$$\text{Loss}(\boldsymbol{x}) = \text{RMSLE}[C(\eta_n), \hat{C}(\eta_n, \boldsymbol{x})] + l_1 + l_2. \tag{19}$$

Now, a defined search space is needed, for the PSO to find a solution. To properly determine the crossover region, we are going to use the empirical crossover percentage $p = C(m_c)$, that is the empirical CCDF evaluated in the income m_c. Note that the income variable is not suited to separate the data since it has a lot of repeated values. With this percentage we calculate the crossover income m_c from a linear interpolation of the empirical CCDF [Eq. (13)]. The interpolation transforms the empirical cumulative distribution into a continuous distribution, thus a solution $m_c = C^{-1}(p)$ will not have discrete values. Therefore, given a T with Eq. (7) we can determine λ, then for a vector of parameters $\boldsymbol{x} = [C(m_c), \alpha, T]$ we can define a theoretical CCDF $[\hat{C}(m)]$ using Eq. (6), allowing us to derive a goodness-of-fit function.

Another information that needs to be provided, 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 $C(m_c) \in (0, 0.2]$. Therefore, the search space is a cuboid T in the parameter coordinates. The next step is to define the optimizer that will minimize the loss function (19).

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 $\{C(m_c), \alpha, T\}$) with regard to a given measure of quality [16]. Let $\{\boldsymbol{x}_t^i = (C(m_c)_t^i, \alpha_t^i, T_t^i)\}$ with $i = 1, 2, ...N_c$, where N_c is the number of candidates of parameters, be our set of candidate vectors in the t_{th} PSO 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. In the first iteration each candidate index is part of K randomly chosen sets $S_0 \in \{S_0^i\}_{i \in \{1,2...,N_c\}}$, this process defines the set 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. These neighbors sets 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 candidates indexes are used to inform a specific property of those to the i_{th} candidate. The exploration will use these sets to compose the next step of each candidate and it will be clarified in the next subsection. The 2007 standard PSO (SPSO2007) value of $K = 3$ and will be used for this paper.

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

$$\begin{aligned} \boldsymbol{v}_t^i = W\boldsymbol{v}_{t-1}^i + c_1\Lambda_1\left(\boldsymbol{P}_{t-1}^i - \boldsymbol{x}_{t-1}^i\right) \\ + c_2\Lambda_2(\boldsymbol{G}_{t-1}^i - \boldsymbol{x}_{t-1}^i), \end{aligned} \tag{20}$$

$$\boldsymbol{x}_t^i = \boldsymbol{x}_{t-1}^i + \boldsymbol{v}_t^i, \tag{21}$$

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

$$\boldsymbol{P}_t^i = \left\{ \boldsymbol{x} \in \left\{\boldsymbol{x}_\nu^i\right\} : \text{Loss}(\boldsymbol{x}) = \min_{z \in \{\boldsymbol{x}_\nu^i\}} \text{Loss}(z) \right\}, \tag{22}$$

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

$$\boldsymbol{G}_t^i = \left\{ \boldsymbol{x} \in C_t^i : \text{Loss}(\boldsymbol{x}) = \min_{z \in C_t^i} \text{Loss}(z) \right\}, \tag{23}$$

where $C_t^i \equiv \{\boldsymbol{P}_t^j\}_{j \in \{S_t^i\}}$ is the set of personal best of neighbors set of the i_{th} candidate in the t_{th} step and $\boldsymbol{G}_t^i$ is the best position in the C_t^i set. The parameters Λ_1 and Λ_2 are uniformly random with range [0,1]. The W, c_1, and c_2 are considered hyperparameters, so they will have a fixed value (or a behavior predetermined). In this paper $c_1 = c_2 = 1.7$ and W fall linearly in [0.7,0.4]. The initialization is done as follows:

$$\begin{aligned} \boldsymbol{x}_0^i &= \boldsymbol{U}_T, \\ \boldsymbol{v}_0^i &= \frac{\boldsymbol{U}_T - \boldsymbol{x}_0^i}{2}, \\ \boldsymbol{P}_0^i &= \boldsymbol{x}_0^i, \\ \boldsymbol{G}_0^i &= \left\{ \boldsymbol{x} \in \{\boldsymbol{P}_0^j\}_{j \in \{S_0^i\}} : \text{Loss}(\boldsymbol{x}) = \min_{z \in \{\boldsymbol{P}_0^j\}} \text{Loss}(z) \right\}, \end{aligned} \tag{24}$$

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

TABLE I. Mean, standard deviation, 95% confidence interval, and coefficient of variation of the model parameters, Gini coefficient, and RMSLE of the training and test sets.

2019 Data	Mean	σ	95% CI	CV
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}
Gini coefficient	0.578	0.001	0.576 // 0.581	2.31×10^{-3}
Train set RMLSE	0.1486	0.0008	0.1470 // 0.1501	5.31×10^{-3}
Test set RMLSE	0.1489	0.0017	0.1458 // 0.1525	1.12×10^{-2}

PSO was chosen for being able to work with nondifferentiable error function and discrete variables [17], which is needed since the two-class loss function is not differentiable. The standard PSO does not always converge to a good solution, so a hybrid approach was used [18]. 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 [19,20].

In short, to find the model parameters of Sec. II, first we need to calculate the cumulative probability distribution of income for the Brazilian population (13) and then use hybrid-PSO to fit the model by minimizing the value of the function (19). The results from these procedures will be displayed in the next section.

V. RESULTS AND CROSS-VALIDATION

In this section, we are going to perform cross validation with the bootstrap method to test the accuracy of the model [21]. 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 (BCV) [22]. 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, first, 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 hybrid-PSO to find the optimal parameters. Last, using the optimal parameters evaluated above, we calculate the RMSLE of each set. Last, 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.

In this section, we used from the data of the 2019 Continuous National Household Sample Survey (PNADc) available by the IBGE since 2012. The PNADc is a research that col-

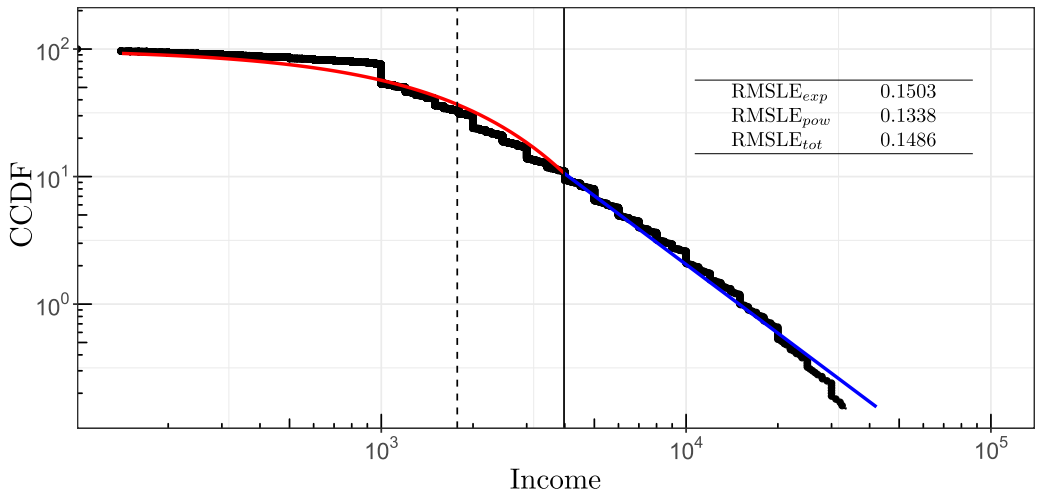


FIG. 3. Cumulative probability distribution of income on a log-log scale. The black points represent the cumulative distribution of the data from 2019 PNAD and the solid lines correspond to the fitted model described in Eq. (6). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution 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 $\lambda = (10.64 \pm 0.26)\%$ 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.

TABLE II. Inequality indicators and RMSLE for a stratified data.

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

lects data from a multitude of Brazilian social characteristics including labor, income and education. From PNADc data, we extracted the total monthly income, gender and color/race of each person in the year 2019. We neglected people without income and missing values. 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. For additional information one can visit the site [23]. After training the parameters of our model, to complete our set of indicators, we calculate the theoretical Gini coefficient following Sec. III.

As shown in Table I 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 coefficients 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.

Figure 3 displays the cumulative distribution fitted with our model with the parameters shown in Table I. 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 [Eq. (19)] is logarithmic, outliers whose $C(m)$ values are

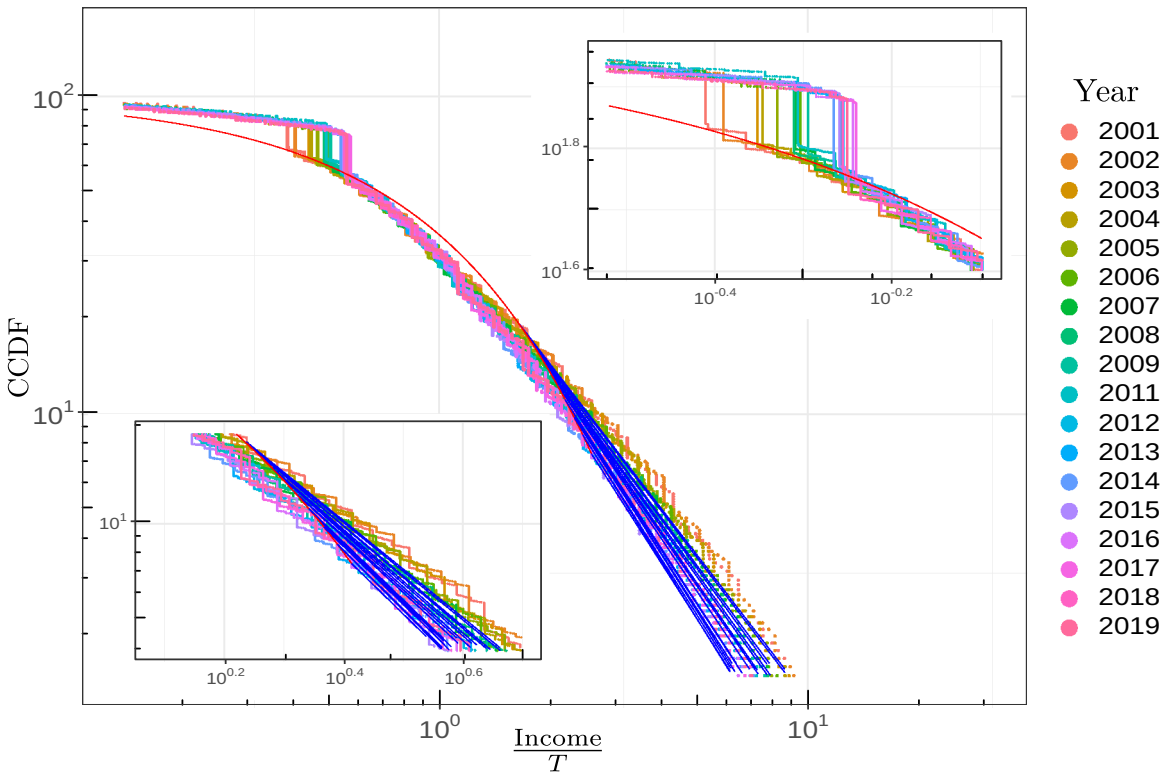


FIG. 4. Cumulative probability distribution of Income/T (income normalized by temperature) on a log-log scale. The colored points represent the cumulative distribution of the data from 2001 to 2019 PNAD and the solid lines correspond to the fitted model described in Eq. (6). The red curve obeys Boltzmann-Gibbs distribution and the power-law distribution is characterized by the blue curve.

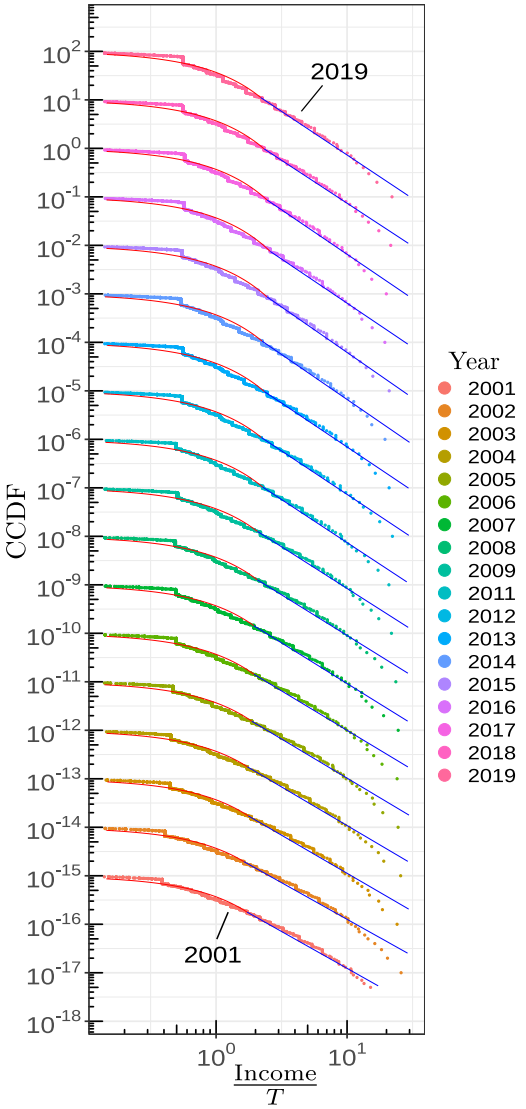


FIG. 5. Cumulative distribution functions constructed from the PNAD data from 2001 to 2019 and their fits with the theoretical distribution described in Eq. (6), shown in the log-log scale versus the normalized annual income Income/T (income normalized by temperature). Plots for different years are shifted vertically.

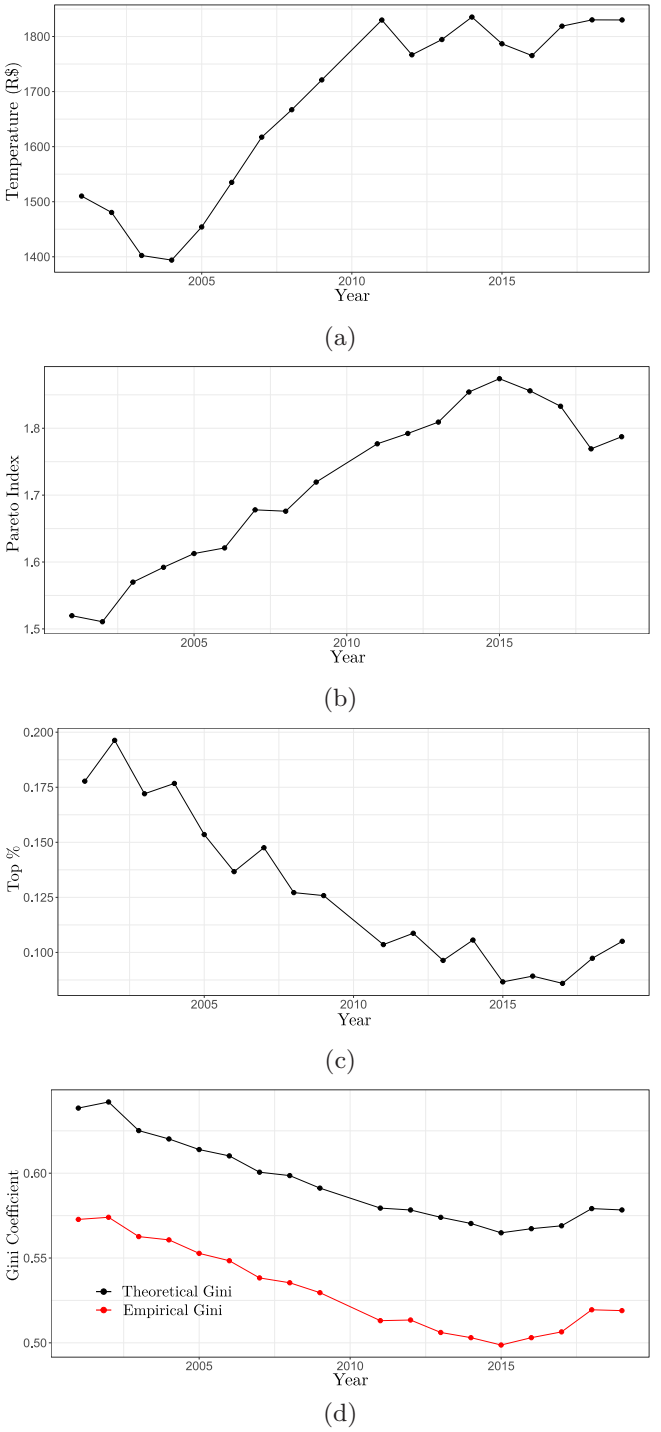


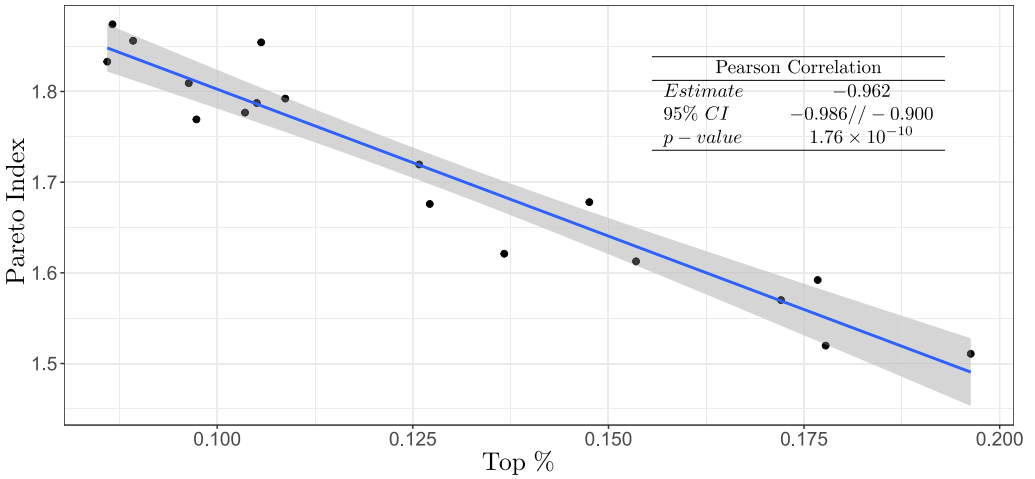
FIG. 6. Time series of the inequality indicators. (a) Deflated Temperature series; (b) Pareto index series; (c) top-percentage series; (d) Gini coefficient series. The black points are the theoretical Gini, Eq. (11), and the red points are the empirical Gini.

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 = $ R\$998) as can be seen in Fig. 3.

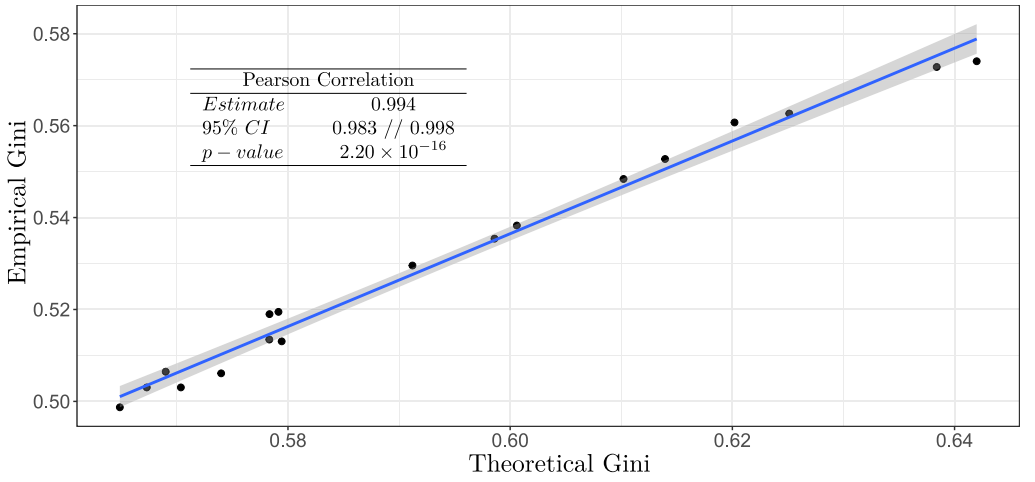
Comparing this method with a crossover income being determined with a fixed proportion for the Pareto tail, $\lambda = 5\%$, as seen in Ref. [13], we get a temperature equal to the mean, $T = 2067 \pm 7$, a Pareto index of 1.816 ± 0.006, and a $m_c = 6192 \pm 21$. This approach displays discontinuity between BG and Pareto regions, its training set error is 0.2257 ± 0.002 and the test set error is 0.2258 ± 0.002, which are significantly higher than with an optimal crossover (Table I).

We applied our model to the data of 2019 stratified into two dichotomies of the population allowing us to compare each indicator and test our model in different data distributions. The first dichotomy is the division by gender(man and woman) and the second is the division by race/color (black/brown/indigenous, BBI; white/yellow, WY) [24]. The results and validation can be seen in the Table II.

Looking at the gender dichotomy, there is a significant difference between the theoretical Gini coefficient. Man's income "temperature" is considerably higher than the woman's, but their Pareto and top-percentage indicators are lower, meaning that there is a less percentage of men in the Pareto



(a)



(b)

FIG. 7. Correlation and affine regression between indicators. (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.

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. BBI has the lowest top-percentage value, $\lambda = 6.42 \pm 0.12$, and the highest Pareto index, $\alpha = 2.01 \pm 0.02$, compared to all subsets. These results gives BBI the lowest Gini coefficient, indicating the highest equality within the subgroup.

VI. TEMPORAL EVOLUTION OF INEQUALITY INDICATORS

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 [13,25,26]. 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 analyze the Brazilian inequality over the years, we applied our model to describe the income distribution between 2001 and 2019. For our empirical data we used the National Household Sample Survey (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.

Last, we applied the optimization to fit the two-class model, Eq. (6), to the data of each year. We can access the temporal evolution of each indicator in the Figs. 6(a)–6(d). We

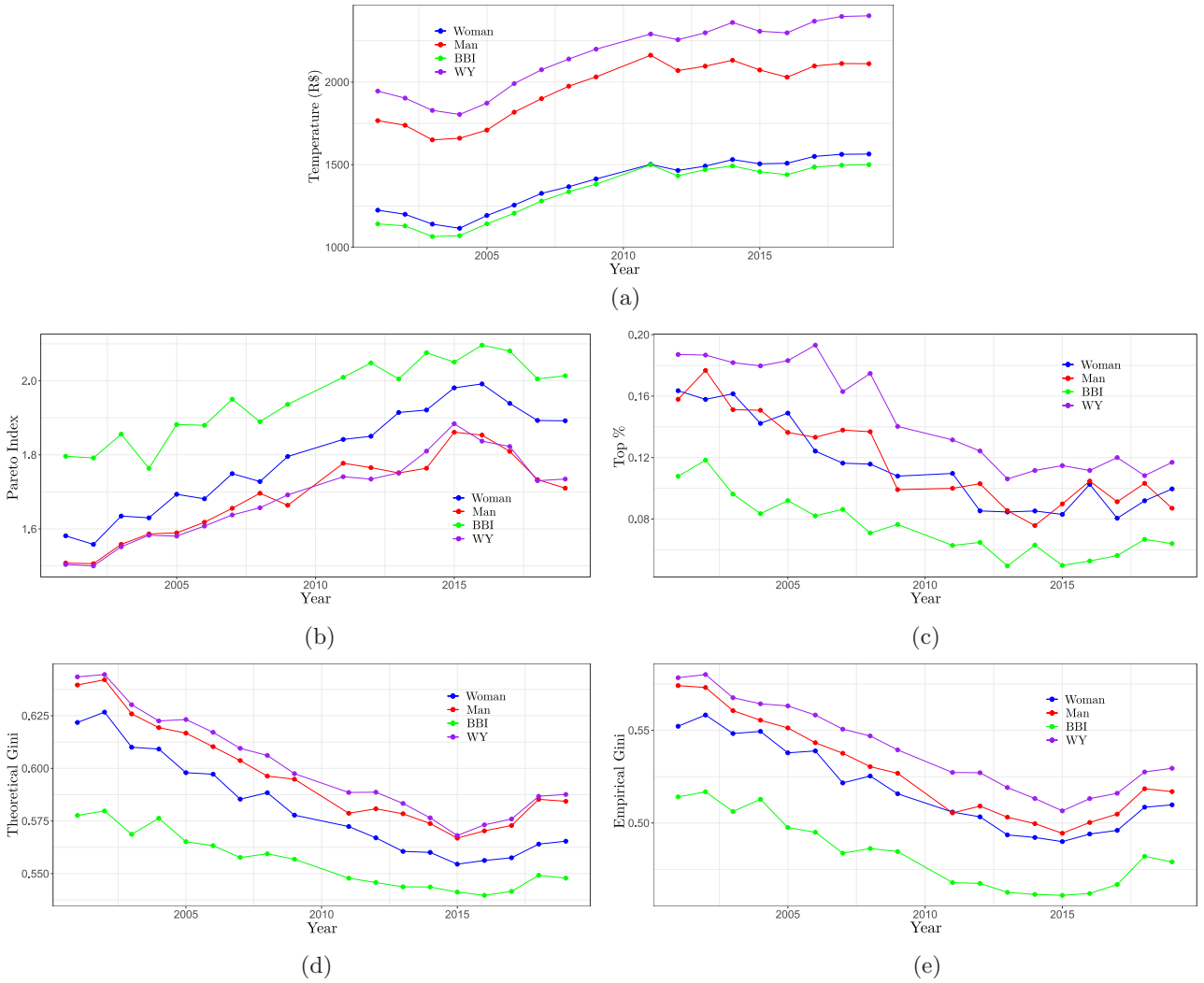


FIG. 8. Time series of the inequality indicators for the stratified data. (a) Deflated temperature series; (b) Pareto index series; (c) top-percentage series; (d) theoretical Gini coefficient series, Eq. (11); (e) empirical Gini coefficient series.

can see that every parameter of our model has an interpretable evolutionary trend. The CCDF and their fits can be seen in Figs. 4 and 5.

Compared to the temporal analysis from the U.S. [12] our results have less temporal stationary 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 Fig. 4. On the same figure in the bottom left subplot, corroborating with U.S. 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 Broad Consumer Price Indices (IPCA), available by IBGE. According to the Fig. 6(a) there is trend of an increase in the temperature.

The Pareto index and the top-percentage have a strong anti-correlation as can be seen in the Fig. 7(a). 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, 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 Fig. 6(d).

As discussed above, the model does not capture the minimum wage effect on the distribution as well as the extreme high income, top 0.01%, where the Pareto behavior breaks down. These observations coupled with the fact that the empirical estimator suffers from a downward bias when the distribution used is fat-tailed [27], explains why the theoretical Gini coefficient is higher than the empirical one. Although Fig. 6(d) 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 Fig. 7(b).

Figure 7(b) also shows the affine relation between the theoretical and the empirical 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

TABLE III. Anticorrelation of Pareto index with top-percentage and correlation of the theoretical and empirical Gini coefficients for the stratified 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

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 previous section, we also did the same time series analysis with a stratified data using the two dichotomies described in that section. The stratified 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 stratified data Figs. 8(a)–8(e) with the complete data Figs. 6(a)–6(d).

Now one can draw the same correlations using the stratified 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 III.

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

VII. CONCLUSIONS

The two-class model is a well-tested hypothesis for the income distribution, being built-on around two famous distributions: the exponential BG and the Pareto power law. It is important to remember that the exponential has stability in a multiagent system, which the log-normal distribution lacks.

Some previous studies have proposed the method to determine the crossover between the two distributions by using a log-log graph and manually trying to spot a discontinuity or a linear behavior. To our knowledge this paper provides for the first time a method to establish an optimal crossover income.

The optimal crossover method presented in this paper not only displays continuity, but also has a significantly lower RMSLE when comparing to a fixed proportion (5%) for the Pareto region. The optimization was cross validated by a bootstrap out-of-the-bag method, which had a good performance in the test sets.

Analyzing stratified data and comparing the dichotomies revealed a greater inequality in the privileged groups (male/white and yellow) compared with their respective counterparts. The black/brown group exhibited the most equality and the least proportion participating in the Pareto region, having only 6.42%.

Last, we analyze the temporal evolution of all indicators and draw two strong correlations. The first is the correlation between the theoretical Gini coefficient and the empirical Gini. The second is between the Top-percentage and Pareto index, which was found for the first time. These two correlations were also found in the stratified data, with the first having a strong correlation with low variance the second have an anticorrelation with more variance when we compare each subgroup. Further investigation in other countries is needed to generalize our findings using the Brazilian data.

The next step would be to implement this novel approach to other countries. Making simulations of a two-class model (define an empirical CCDF given a predetermined model), to determine a loss function that gives the best estimation of true value of the simulated model, is another step that would further validate an end-to-end method of fitting this model. The end goal would be to add a sample weighting and expansion, thus the distribution will have the correct sampling treatment. The sample weighting and expansion is a rather advanced topic of sampling statistics and usually dismissed in model regression.

ACKNOWLEDGMENT

P.H.S. acknowledges partial financial support from CAPES (BR).

[1] H. Moore, Cours d'économie politique, by Vilfredo Pareto, professeur à l'université de lausanne, vol. I, p. 430. 1896; vol. II, p. 426. 1897. Lausanne: F. Rouge, Ann. Amer. Acad. Politic. Soc. Sci. **9**, 128 (1897).

[2] B. M. Hill, A simple general approach to inference about the tail of a distribution, Ann. Stat. **3**, 1163 (1975).

[3] G. F. Shirras, The Pareto law and the distribution of income, Econ. J. **45**, 663 (1935).

[4] H. Aoyama, W. Souma, Y. Nagahara, M. P. Okazaki, H. Takayasu, and M. Takayasu, Pareto's law for income of individuals and debt of bankrupt companies, Fractals **08**, 293 (2000).

[5] M. Levy and S. Solomon, New evidence for the power-law distribution of wealth, Physica A **242**, 90 (1997).

[6] F. Clementi and M. Gallegati, Pareto's law of income distribution: Evidence for germany, the united kingdom, and the united states, in *Econophysics of Wealth Distributions* (Springer, Berlin, 2005), pp. 3–14

[7] A. Dragulescu and V. M. Yakovenko, Statistical mechanics of money, Eur. Phys. J. B **17**, 723 (2000).

[8] M. Kalecki, On the gibrat distribution, Econometrica **13**, 161 (1945).
[9] J. Gonzalez-Estevez, M. Cosenza, R. Lopez-Ruiz, and J. Sanchez, Pareto and Boltzmann–Gibbs behaviors in a deterministic multiagent system, Physica A **387**, 4637 (2008).
[10] A. Chatterjee, B. K. Chakrabarti, and S. Manna, Pareto law in a kinetic model of market with random saving propensity, Physica A **335**, 155 (2004).
[11] A. Banerjee and V. M. Yakovenko, Universal patterns of inequality, New J. Phys. **12**, 075032 (2010).
[12] A. Drăgulescu and V. M. Yakovenko, Exponential and power-law probability distributions of wealth and income in the united kingdom and the united states, Physica A **299**, 213 (2001).
[13] I. D. Siciliani and M. H. Tragtenberg, Kinetic theory and brazilian income distribution, Physica A **513**, 166 (2019).
[14] B. Oancea, T. Andrei, and D. Pirjol, Income inequality in romania: The exponential-Pareto distribution, Physica A **469**, 486 (2017).
[15] M. A. M. Safari, N. Masseran, and K. Ibrahim, Optimal threshold for Pareto tail modelling in the presence of outliers, Physica A **509**, 169 (2018).
[16] J. Kennedy and R. Eberhart, Particle swarm optimization, in *Proceedings of the International Conference on Neural Networks (ICNN)* (IEEE, Piscataway, NJ, 1995), Vol. 4, pp. 1942–1948.
[17] Y. Shi and R. Eberhart, A modified particle swarm optimizer, in *Proceedings of the IEEE International Conference on Evolutionary Computation* (IEEE, Piscataway, NJ, 1998), pp. 69–73.
[18] S. Li, M. Tan, I. W. Tsang, and J. T.-Y. Kwok, A hybrid PSO-BFGS strategy for global optimization of multimodal functions, IEEE Trans. Syst., Man, Cybernet. **41**, 1003 (2011).
[19] A. Skajaa, Limited memory BFGS for nonsmooth optimization, Master's thesis, 2010.
[20] J. Guo and A. Lewis, Nonsmooth variants of Powell's BFGS convergence theorem, SIAM J. Opt. **28**, 1301 (2018).
[21] B. Efron and R. Tibshirani, Improvements on cross-validation: The 632+ bootstrap method, J. Am. Stat. Assoc. **92**, 548 (1997).
[22] B. Efron, Estimating the error rate of a prediction rule: Improvement on cross-validation, J. Am. Stat. Assoc. **78**, 316 (1983).
[23] https://www.ibge.gov.br/en/statistics/social/labor/16833-monthly-dissemination-pnadc1.html?edicao=20780&t=o-que-e.
[24] Brown stands for mixed race and yellow stands for Asian.
[25] A. C. Silva and V. M. Yakovenko, Temporal evolution of the "thermal" and "superthermal" income classes in the U.S.A. during 1983–2001, Europhys. Lett. **69**, 304 (2005).
[26] W. Souma, Physics of personal income, in *Empirical Science of Financial Fluctuations* (Springer, Berlin, 2002), pp. 343–352.
[27] A. Fontanari, N. N. Taleb, and P. Cirillo, Gini estimation under infinite variance, Physica A **502**, 256 (2018).