mathematics

MDPI

Article

Three-Part Composite Pareto Modelling for Income Distribution in Malaysia

Muhammad Hilmi Abdul Majid *, Kamarulzaman Ibrahim and Nurulkamal Masseran

Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; kamarulz@ukm.edu.my (K.I.); kamalmsn@ukm.edu.my (N.M.)
* Correspondence: hilmi.majid@ukm.edu.my

Abstract: Income distribution models can be useful for describing the economic properties of a population. In this study, three-part composite Pareto models are fitted to the income distribution in Malaysia for the years 2007, 2009, 2012, 2014, and 2016. The three-part composite Pareto models divide the population into three parts, each following a different distribution model. The lower part follows the inverse Pareto distribution, the upper part follows the Pareto distribution, and the middle part follows another unspecified distribution model. For application in income data, the use of Gaussian mixture distribution is proposed for the middle part, making the inverse Pareto–Gaussian mixture-Pareto distribution model semi-parametric. From the model, it is found that the levels of income inequality in the lower and upper income groups decrease over the period of study. Additionally, the proportion of data following the inverse Pareto distribution in the model is highly correlated with the official absolute poverty incidence.

Keywords: composite model; income distribution; income inequality; Pareto distribution

MSC: 62P20

Citation: Majid, M.H.A.; Ibrahim, K.; Masseran, N. Three-Part Composite Pareto Modelling for Income Distribution in Malaysia. *Mathematics* **2023**, *11*, 2899. https://doi.org/10.3390/math11132899

Academic Editors: Paolo Pagnottoni, Domenico Scopelliti and Alessandro Bitetto

Received: 15 May 2023
Revised: 26 June 2023
Accepted: 26 June 2023
Published: 28 June 2023

1. Introduction

The research on income distribution has been around for a long time, starting from Vilfredo Pareto's observation on income in 1896. Even though the topic has been discussed at length, it is still relevant and important, as income is heavily related to the well-being of a country. For example, it is found that income inequality is highly correlated with the number of criminal activities [1,2] and countries with low income inequality have healthier citizens mentally and physically [3–6]. High income inequality between groups in a population may also cause political instability [7] and is considered one of the main causes of the racial riot in Malaysia in 1969 [8].

There have been many models proposed for describing income distribution. For example, the lognormal, Weibull, gamma, Dagum, beta distribution of the second kind, and Singh–Maddala distributions have been used to model the income distribution of the whole population. However, these distributions may not fit the upper and lower tails of the income distribution well. Dagum [9] and Singh and Maddala [10] for example have noted that the lognormal or gamma distributions alone are not enough to describe the upper and lower tails of the income distribution well. On the other hand, Pareto distribution has been used extensively to model the upper tail of the income distribution in various countries [11–15] and the inverse Pareto distribution can be used for the lower tail as both the upper and lower tails of income distribution exhibit power-law behaviour [16,17]. In general, the power-law behaviour is observed for the top 5–10% of the population [18]. In the context of the household income in Malaysia, previous studies have identified the power-law behaviour in the upper tail of the income distribution [19–22]. Additionally, Safari et al. [23] have noted that the inverse Pareto distribution is a good model to be used for lower income groups in Malaysia.

The three-part composite Pareto model can be seen as an extension of the two-part composite Pareto model introduced by Cooray and Ananda [24]. This composite model is also known as the spliced distribution in some literature [25,26]. In the introduction of the two-part composite Pareto model, Cooray and Ananda have used the Pareto distribution for the upper tail and the lognormal distribution for the rest of the data. After the model was introduced, there have been many advances and varieties of the two-part composite Pareto model [27–34]. As for the three-part composite Pareto model, Mendes and Lopes [35] have used a composite model that uses *t*-distribution for the middle part and two generalized Pareto distributions both for the lower and upper parts of the data. Luckstead and Devadoss [36] and Luckstead et al. [37] on the other hand have used the inverse Pareto distribution for the lower part, lognormal distribution for the middle part, and Pareto distribution for the upper part for the cities size distribution in the US and India. Wiegand and Nadarajah [38] have also used the three-part composite Pareto model with Pareto type IV distribution for the upper part and lognormal, gamma, beta Weibull, or Pareto type IV for the lower and middle part of the data for categorizing companies based on their market value, sales, assets, and profits. The two-part composite Pareto model has been applied to the income data [39], but to the authors knowledge, the three-part composite Pareto model has not been proposed and used to describe the income distribution.

Since the Pareto distribution fits well for the upper tail and the inverse Pareto distributions is suitable for the lower tail of the household income data, a model that combines these two distributions together with another distribution for the middle part of the data may be useful. In this paper, we propose the use of the three-part composite Pareto models for income distribution that describe lower, middle, and upper parts of the data using separate distributions: inverse Pareto distribution for the lower part, an unspecified distribution for the middle part, and Pareto distribution for the upper part of the income data. For the middle part of the data, this paper proposed the usage of Gaussian mixture distribution. By combining these three distributions, the three-part composite Pareto model can divide the population into three categories: the lower, middle, and upper income groups. Further analysis on each of these categories can be performed by studying the properties of the respective distribution for the group. This approach of combining three separate distributions is different as compared to the practice used in other literature, which which analyses each part separately [14,40–42].

The choice of using Pareto and inverse Pareto distributions in the composite model is due to their properties that fit with the upper and lower parts of the income distribution together their simplicity. While there are other distributions that can be used, for example Pareto Type II–IV, Generalized Pareto, or Generalized Extreme Value distributions for the upper tail, using these distributions increases the complexity of the composite model. Additionally, the shape parameters in the Pareto and inverse Pareto distributions are useful for measuring income inequalities, as discussed in Section 2.2. The Lorenz curve and Gini index used for measuring the income inequality model will also be shown. And finally, the model is then applied to the household income data in Malaysia for the years 2007, 2009, 2012, 2014, and 2016.

This paper is organized as follows. Section 2 discusses the methodologies used in the study. This includes the three-part composite Pareto model, the Lorenz curve and Gini index, the semi-parametric three part composite Pareto model, as well as the pseudo-likelihood approach used to estimate the parameters in the model. Section 3 focuses on the application of the three-part composite Pareto model to the income distribution in Malaysia. Then finally, Section 4 concludes the paper.

2. Methodologies

2.1. Three-Part Composite Pareto Model

In the three-part composite Pareto (3PCP) model, the data are divided into three parts, the lower, upper, and middle parts, each following a different distribution model.

The lower part of the data follows the inverse Pareto distribution with probability density function (PDF)

$$f_{IP}(x|\tau_1, \alpha_1) = \frac{\alpha_1 x^{\alpha_1 - 1}}{\tau_1^{\alpha_1}}, \text{ for } 0 < x < \tau_1, \tag{1}$$

where τ_1 and α_1 are the threshold and shape parameters of the inverse Pareto distribution, respectively. The inverse Pareto distribution is also called the power function distribution in some literature [43]. The upper part of the data follows the Pareto distribution with PDF:

$$f_P(x|\tau_2, \alpha_2) = \frac{\alpha_2 \tau_2^{\alpha_2}}{x^{\alpha_2 + 1}}, \text{ for } x > \tau_2, \tag{2}$$

where τ_2 and α_2 are the threshold and shape parameters of the Pareto distribution, respectively. And finally, the middle part of the data follows another distribution that is not specified or fixed. Figure 1 shows a graphical representation of the PDF of a 3PCP model with Gaussian mixture for the middle part of the data. In the figure, observations with values between 0 and τ_1 are modelled by the inverse Pareto distribution, observations with values between τ_1 and τ_2 are modelled by the Gaussian mixture distribution, and observations with values greater than τ_2 are modelled by the Pareto distribution.

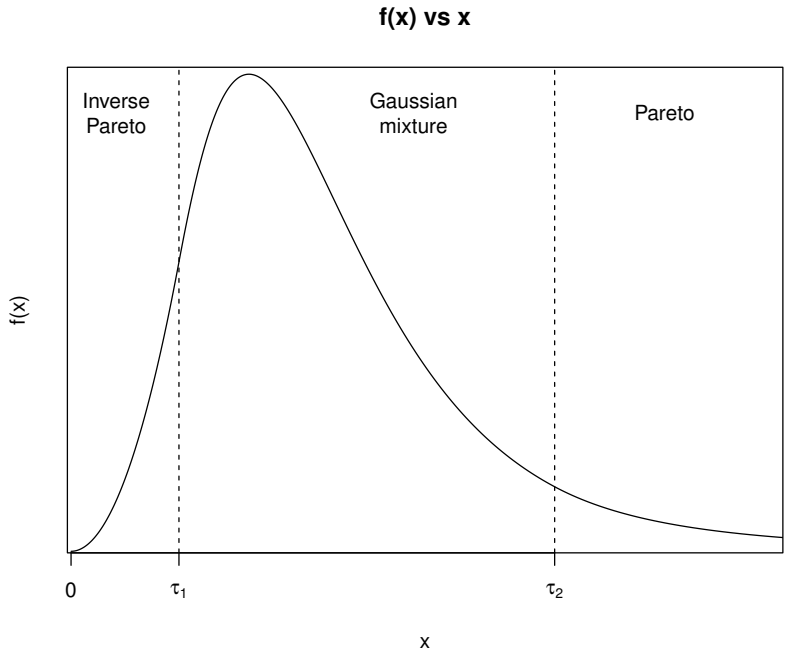

Figure 1. A graphical example of the PDF of the 3PCP model with Gaussian mixture distribution for the middle data.

Let $h(x|\eta)$ and $H(x|\eta)$ be the PDF and cumulative distribution function (CDF) of the middle part of the data with parameter η, respectively. Combining the three distributions for each part gives the following PDF for the 3PCP model:

$$f(x|\theta) = \begin{cases} \rho_1 f_{IP}(x|\tau_1, \alpha_1), & \text{for } 0 < x < \tau_1, \\ (1 - \rho_1 - \rho_2) \dfrac{h(x|\eta)}{H(\tau_2|\eta) - H(\tau_1|\eta)}, & \text{for } \tau_1 \leq x \leq \tau_2, \\ \rho_2 f_P(x|\tau_2, \alpha_2), & \text{for } x > \tau_2, \end{cases} \tag{3}$$

where θ is the collection of all parameters in the model, ρ_1 is the proportion of data in the lower part following the inverse Pareto distribution, and ρ_2 is the proportion of data in the upper part following the Pareto distribution.

The PDF in Equation (3) indicates that there are two threshold parameters, τ_1 and τ_2. Any observation with a value less than τ_1 follows the inverse Pareto distribution with PDF $f_{IP}(x|\tau_1, \alpha_1)$. Any observation with a value between τ_1 and τ_2 follows the middle part distribution with PDF $h(x|\eta)$. And lastly, for any observation with a value greater than τ_2, it follows the Pareto distribution with PDF $f_P(x|\tau_2, \alpha_2)$. Because $f_{IP}(x|\tau_1, \alpha_1)$, $h(x|\eta)$, and $f_P(x|\tau_2, \alpha_2)$ are all PDFs, then $\int_0^\infty f(x|\theta)\,dx$ must be equal to 1. The PDF in Equation (3) can also be considered as a mixture distribution, except the distributions do not overlap each other.

The CDF for 3PCP model can be calculated simply by integrating the PDF in Equation (3) to obtain

$$F(x|\theta) = \begin{cases} \rho_1 \left(\dfrac{x}{\tau_1}\right)^{\alpha_1}, & \text{for } 0 < x < \tau_1, \\ \rho_1 + (1-\rho_1-\rho_2)\dfrac{H(x|\eta) - H(\tau_1|\eta)}{H(\tau_2|\eta) - H(\tau_1|\eta)}, & \text{for } \tau_1 \le x \le \tau_2, \\ 1 - \rho_2 \left(\dfrac{\tau_2}{x}\right)^{\alpha_2}, & \text{for } x > \tau_2. \end{cases} \tag{4}$$

Moreover, the quantile function for the model is

$$F^{-1}(u|\theta) = \begin{cases} \tau_1 \left(\dfrac{u}{\rho_1}\right)^{1/\alpha_1}, & \text{for } 0 < u < \rho_1, \\ H^{-1}\left[H(\tau_1) + \dfrac{(u-\rho_1)[H(\tau_2) - H(\tau_1)]}{1-\rho_1-\rho_2}\middle|\eta\right], & \text{for } \rho_1 \le u \le 1-\rho_2, \\ \tau_2 \left(\dfrac{\rho_2}{1-u}\right)^{1/\alpha_2}, & \text{for } 1-\rho_2 < u < 1. \end{cases} \tag{5}$$

The overall mean when $\alpha_2 > 1$ is given in the equation below:

$$\mu_X = \frac{\rho_1 \alpha_1 \tau_1}{\alpha_1 + 1} + \frac{1-\rho_1-\rho_2}{H(\tau_2|\eta) - H(\tau_1|\eta)} \int_{\tau_1}^{\tau_2} x h(x|\eta)\,dx + \frac{\rho_2 \alpha_2 \tau_2}{\alpha_2 - 1}. \tag{6}$$

If $\alpha_2 \le 1$, then the integral $\int_{\tau_2}^\infty x f_P(x|\tau_2, \alpha_2)\,dx$ diverges and $\mu_X = \infty$.

However, note that the PDF in Equation (3) may not be continuous or differentiable. Additional constraints are required if continuity and differentiability are desired. The continuity of the PDF can be achieved by setting

$$\rho_1 = \frac{\alpha_2 \tau_1 h(\tau_1|\eta)}{\alpha_1 \alpha_2 [H(\tau_2|\eta) - H(\tau_1|\eta)] + \alpha_1 \tau_2 h(\tau_2|\eta) + \alpha_2 \tau_1 h(\tau_1|\eta)}, \tag{7}$$

and

$$\rho_2 = \frac{\alpha_1 \tau_2 h(\tau_2|\eta)}{\alpha_1 \alpha_2 [H(\tau_2|\eta) - H(\tau_1|\eta)] + \alpha_1 \tau_2 h(\tau_2|\eta) + \alpha_2 \tau_1 h(\tau_1|\eta)}. \tag{8}$$

And as for the differentiability of the PDF, Equations (7) and (8) must be satisfied together with

$$\alpha_1 = \frac{h(\tau_1|\eta) + \tau_1 h'(\tau_1|\eta)}{h(\tau_1|\eta)}, \tag{9}$$

and

$$\alpha_2 = \frac{-h(\tau_2|\eta) - \tau_2 h'(\tau_2|\eta)}{h(\tau_2|\eta)}, \tag{10}$$

where $h'(x|\eta)$ is the first derivative of $h(x|\eta)$ with respect to x.

2.2. Lorenz Curve and Gini Index

Lorenz curve and Gini index are commonly used tools to measure the level of income inequality in a population [18,44]. The general formula that can be used to calculate the Lorenz curve for a population with a distribution function is [45]

$$\mathrm{LC}(u) = \frac{1}{\mu_X} \int_0^u F^{-1}(y|\theta)\, dy, \tag{11}$$

where μ_X is the overall mean and $F^{-1}(y|\theta)$ is the quantile function. The value of $\mathrm{LC}(u)$ for a specific u refers to the proportion of cumulative wealth or income earned by the lowest u proportion of the population. Let

$$A(u) = \int_{\rho_1}^u H^{-1}\left[H(\tau_1) + \frac{(y-\rho_1)[H(\tau_2) - H(\tau_1)]}{1-\rho_1-\rho_2}\middle|\eta\right] dy. \tag{12}$$

Then, it can be shown that for the 3PCP model,

$$\mathrm{LC}(u) = \begin{cases} \dfrac{1}{\mu_X}\dfrac{\rho_1\alpha_1\tau_1}{\alpha_1+1}\left(\dfrac{u}{\rho_1}\right)^{1+1/\alpha_1}, & \text{for } 0 < u < \rho_1, \\ \dfrac{1}{\mu_X}\left[\dfrac{\rho_1\alpha_1\tau_1}{\alpha_1+1} + A(u)\right], & \text{for } \rho_1 \le u \le 1-\rho_2, \\ \dfrac{1}{\mu_X}\left[\dfrac{\rho_1\alpha_1\tau_1}{\alpha_1+1} + A(1-\rho_2) + \dfrac{\rho_2\alpha_2\tau_2}{\alpha_2-1}\left(1 - \left(\dfrac{\rho_2}{1-u}\right)^{1/\alpha_2-1}\right)\right], & \text{for } 1-\rho_2 < u < 1. \end{cases} \tag{13}$$

Using the obtained Lorenz curve function in Equation (13), the Lorenz curve can be plotted on a unit square where the x-axis is the population proportion, u, and the y-axis is the proportion of cumulative wealth or income, $\mathrm{LC}(u)$, and will be compared to the 45° equality line. The closer the Lorenz curve is to the 45° equality line, the lower the level of income inequality.

The Gini index on the other hand is a numerical measure calculated using the Lorenz curve that can be used to assess the level of income inequality. The value of Gini index is between 0 and 1 where the higher the Gini index is, the higher the level of income inequality. Using the Lorenz curve in Equation (13), it can be shown that

$$\begin{aligned} \mathrm{Gini} &= 1 - 2\int_0^1 \mathrm{LC}(u)\, du, \\ &= 1 - \frac{2}{\mu_X}\left[\frac{\rho_1^2\alpha_1^2\tau_1}{(\alpha_1+1)(2\alpha_1+1)} + \frac{\rho_1\alpha_1\tau_1(1-\rho_1)}{\alpha_1+1} + \int_{\rho_1}^{1-\rho_2} A(u)\, du \right. \\ &\qquad \left. + \rho_2 A(1-\rho_2) + \frac{\rho_2^2\alpha_2\tau_2}{\alpha_2-1} - \frac{\rho_2^2\alpha_2^2\tau_2}{(\alpha_2-1)(2\alpha_2-1)}\right]. \end{aligned} \tag{14}$$

The integral $\int_{\rho_1}^{1-\rho_2} A(u)\, du$ in the expression above may require a numerical method, for example the trapezoidal rule, to approximate its value. Using Equation (14) above, the income inequality for the whole population can be measured. A high Gini index value shows a high level of income inequality, whereas a low Gini index value shows a low level of income inequality.

Note that the Lorenz curve and Gini index can both be calculated empirically or using a non-parametric approach, without having to specify a distribution model for the income. In general, if the number of observations in the data is large, the Lorenz curve and Gini index calculated empirically or using a non-parametric approach provide good estimates of inequality measures. However, the Lorenz curve and Gini index calculated using the underlying distribution model have the advantage when the sample size is small, and provide more reliable estimates as compared to the non-parametric approach [46]. This

is true for any distribution model, including the 3PCP model, provided that the distribution model fits the data adequately.

Additionally, it can be shown that the Gini index for the inverse Pareto distribution with shape parameter α_1 is [23]

$$\text{Gini} = \frac{1}{2\alpha_1 + 1}, \tag{15}$$

whereas for Pareto distribution with shape parameter α_2, its Gini index is [43]

$$\text{Gini} = \frac{1}{2\alpha_2 + 1}. \tag{16}$$

From these two equations, we can then use the values of α_1 and α_2 in the 3PCP model to evaluate the income inequalities in the lower and upper data, respectively. For example, if the value of α_1 is high, this indicates a high level of income inequality in the lower part of the data. On the other hand, if α_1 is low, then the level of income inequality in the lower part of the data is low, and similarly for α_2 for the upper part of the data. Note that comparisons on the income inequalities using the shape parameters can be made for different datasets with different proportions of the upper and lower data, as long as the proportions are not too small. As shown in Equations (15) and (16), the Gini index depends on the shape of the distribution, and not on the threshold parameters or the proportions of data.

2.3. Semi-Parametric Three-Part Composite Pareto Model

A problem might occur when using commonly used models for income distribution, for example lognormal, gamma, or Weibull distributions, for the middle part of the data. Suppose for example, the overall income distribution comes from a lognormal distribution. Then, when the 3PCP model with PDF in Equation (3) is applied to the data with $h(x|\eta)$ be the lognormal distribution, we would expect ρ_1 and ρ_2 to be zero as the inverse Pareto and Pareto distributions are not required for describing the income distribution, and lognormal distribution alone is enough for the whole data. With that, information regarding the lower and upper income earners will be lost and the 3PCP model is not useful.

To overcome this problem, we can set the middle part to follow a semi-parametric model, for example the Gaussian mixture with k components. We can set

$$h(x|\mathbf{r}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{j=1}^{k} \frac{r_j f_N(x|\mu_j, \sigma_j^2)}{F_N(\tau_2|\mu_j, \sigma_j^2) - F_N(\tau_1|\mu_j, \sigma_j^2)}, \qquad \text{for } \tau_1 \le x \le \tau_2, \tag{17}$$

and

$$H(x|\mathbf{r}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{j=1}^{k} \frac{r_j \left[F_N(x|\mu_j, \sigma_j^2) - F_N(\tau_1|\mu_j, \sigma_j^2)\right]}{F_N(\tau_2|\mu_j, \sigma_j^2) - F_N(\tau_1|\mu_j, \sigma_j^2)}, \qquad \text{for } \tau_1 \le x \le \tau_2, \tag{18}$$

where $f_N(x|\mu_j, \sigma_j^2)$ and $F_N(x|\mu_j, \sigma_j^2)$ are the PDF and CDF of a normal distribution with mean μ_j and variance σ_j^2, respectively, and r_j is the weight for the jth component in the mixture model with $\sum_{i=1}^{k} r_j = 1$ and $r_j > 0$ for all $j = 1, \cdots, k$. The number of components k is not specified and depends on the data themselves. The AIC and BIC values can be used to find the value of k that gives the lowest AIC and BIC values. Additionally, notice that from Equation (18), $H(\tau_1|\eta) = 0$ and $H(\tau_2|\eta) = 1$. In this paper, the model that uses this specification is called the inverse Pareto–Gaussian mixture-Pareto (IP-GM-P) model.

The usage of Gaussian mixture for the middle part of the data can give space for the lower and upper part of the data to be modelled by the inverse Pareto and Pareto distributions, respectively. In general, a normal distribution is not suitable to be used for income distribution due to its properties. For example, a normal distribution is symmetric and covers the whole real number, whereas the income distribution is commonly skewed

to the right with heavy upper tail and with positive values only. When the Gaussian mixture distribution is used for the middle part of the data, the Pareto and inverse Pareto distributions are both required to fit the upper and lower parts of the data, respectively. Thus, information about the upper and lower data in the form of the Pareto and inverse Pareto distributions are not lost.

The finite mixture models have been used extensively to model the whole income distribution and to separate income groups within the population [47]. Some finite mixture models that have been used for the whole income distribution include the Gaussian mixture model [48,49], the gamma mixture model [50], and the lognormal mixture model [51]. But as mentioned, heavy tail distribution such as the lognormal, gamma, or Weibull distributions including their mixture models, should be avoided for the middle part of the data when the 3PCP model is to be fitted to income distribution. To our knowledge, the mixture model has not been used to model middle-class income specifically. The choice of using the Gaussian mixture model for the middle part of the 3PCP model is due to the properties of the normal distribution that is not suitable for income distribution and that any continuous distribution can be fitted by the Gaussian mixture model with a large enough number of components [52].

By substituting Equation (17) with Equation (3), the PDF for the IP-GM-P model is

$$f(x|\theta) = \begin{cases} \rho_1 f_{IP}(x|\tau_1, \alpha_1), & \text{for } 0 < x < \tau_1, \\ (1-\rho_1-\rho_2) \displaystyle\sum_{j=1}^{k} \frac{r_j f_N(x|\mu_j, \sigma_j^2)}{F_N(\tau_2|\mu_j, \sigma_j^2) - F_N(\tau_1|\mu_j, \sigma_j^2)}, & \text{for } \tau_1 \le x \le \tau_2, \\ \rho_2 f_P(x|\tau_2, \alpha_2), & \text{for } x > \tau_2. \end{cases} \tag{19}$$

Additionally, the overall mean for the IP-GM-P model can be written as

$$\mu_X = \frac{\rho_1 \alpha_1 \tau_1}{\alpha_1 + 1} + \frac{\rho_2 \alpha_2 \tau_2}{\alpha_2 - 1} + \sum_{j=1}^{k} \left\{ \frac{r_j(1-\rho_1-\rho_2)}{F_{2,j} - F_{1,j}} \left\{ \mu_j \left[\Phi\left(\frac{\tau_2 - \mu_j}{\sigma_j} \right) - \Phi\left(\frac{\tau_1 - \mu_j}{\sigma_j} \right) \right] + \frac{\sigma_j}{\sqrt{2\pi}} \left[\exp\left\{ -\frac{(\tau_1 - \mu_j)^2}{2\sigma_j^2} \right\} - \exp\left\{ -\frac{(\tau_2 - \mu_j)^2}{2\sigma_j^2} \right\} \right] \right\} \right\}, \tag{20}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution.

For the IP-GM-P model, it can be shown that the $A(u)$ function in Equation (12) can be written as

$$A(u) = \sum_{j=1}^{k} \left\{ \frac{r_j(1-\rho_1-\rho_2)}{F_N(\tau_2|\mu_j, \sigma_j^2) - F_N(\tau_1|\mu_j, \sigma_j^2)} \left\{ \mu_j \left[\Phi\left(\frac{u^* - \mu_j}{\sigma_j} \right) - \Phi\left(\frac{\tau_1 - \mu_j}{\sigma_j} \right) \right] + \frac{\sigma_j}{\sqrt{2\pi}} \left[\exp\left\{ -\frac{(\tau_1 - \mu_j)^2}{2\sigma_j^2} \right\} - \exp\left\{ -\frac{(u^* - \mu_j)^2}{2\sigma_j^2} \right\} \right] \right\} \right\}, \tag{21}$$

where

$$u^* = H^{-1}\left(\frac{u - \rho_1}{1 - \rho_1 - \rho_2} \middle| \eta \right), \tag{22}$$

and $H^{-1}(u|\eta)$ is the quantile function of the Gaussian mixture. Since $H(x|\eta)$ is an increasing function, calculating u^* is easy, for example by using bisection method.

Using Equations (20) and (21), the Lorenz curve and Gini index using IP-GM-P model can be calculated using Equations (13) and (14), respectively. The integral $\int_{\rho_1}^{1-\rho_2} A(u)\, du$ in Equation (14) requires a numerical method, for example the trapezoidal rule, to approximate it. The approximation is fast and easy as the integral is bounded.

2.4. Statistical Methods for Complex Survey Data

In complex survey data, samples in the survey are given different weights depending on the size of target population and the size of the samples collected. These weights, when available, should be included in analysis to improve accuracy and to avoid bias in the results [53]. To include sample weights in the parameter estimation of the model, the pseudo-likelihood approach can be used.

Let x_i be the income of ith household in the sample with weight w_i and n be the sample size. The weight is scaled such that the total weight is n using the following expression:

$$w_i = \frac{nw_i^*}{\sum_{i=1}^n w_i^*} \tag{23}$$

where w_i^* is the unscaled sample weight. Then, the pseudo-likelihood function of the data can be written as

$$\tilde{L} = \prod_{i=1}^{n} [f(x_i|\theta)]^{w_i}. \tag{24}$$

Notice that if $w_i = 1$ for all i, as seen in simple random sampling, then the pseudo-likelihood function is the regular likelihood function. The maximum pseudo-likelihood estimate can be defined as the parameters that results with the highest value for the pseudo-likelihood function [54]:

$$\hat{\theta} = \arg\max_{\theta} \left\{ \prod_{i=1}^{n} [f(x_i|\theta)]^{w_i} \right\}. \tag{25}$$

Unfortunately, due of the complexity of the 3PCP model, the analytical form of the solution is not possible. In this paper, the `mle2` function in `bbmle` R package is used to estimate the parameters. This function uses the `optim` optimizer in R and gives the numerical estimate for the values of parameters that maximize the log pseudo-likelihood.

To perform a goodness-of-fit test, the modified Kolmogorov–Smirnov (KS) test will be used. The KS goodness-of-fit test is used to determine whether data fits the model by comparing the empirical CDF with the theoretical CDF. If the p-value of the test is lower than the significance level, then the null hypothesis that the model fits the data will be rejected. Since the sample weights are included in the analysis, the test statistic for this test is modified such that

$$D_n = \frac{\sum_{i=1}^n w_i}{\sqrt{\sum_{i=1}^n w_i^2}} \max_x |F_n(x) - F(x)|, \tag{26}$$

where $F_n(x)$ is the weighted empirical CDF and $F(x)$ is the theoretical CDF. Observe that if $w_i = 1$ for all i, then D_n in the expression above reduces to the regular KS test statistic. It has been shown that D_n in Equation (26) converges weakly to the KS distribution as $n \to \infty$ [55,56].

As for finding the best number of components, k, the pseudo-likelihood based BIC values are used. The formula for this information criterion is as follows:

$$\text{BIC} = d \log(n) - 2 \log(\hat{L}), \tag{27}$$

where d is the number of parameters in the model and $\hat{L}$ is the value of the pseudo-likelihood function using the maximum pseudo-likelihood estimate. The model with the lowest BIC values is more preferable. The consistency of the pseudo-likelihood-based BIC has been established by Xu et al. [57].

3. Application to Income Distribution in Malaysia

3.1. Household Income Survey

The data used in this paper are from the Household Income and Basic Amenities Survey (HIS & BA) conducted by the Department of Statistics Malaysia. Twice every

five years, the Department of Statistics Malaysia would conduct this survey to collect information related to the economic well-being of the citizens in Malaysia. In this paper, the household income and its size will be used to study the changes in household income in Malaysia. Five datasets are used: household income for the years 2007, 2009, 2012, 2014, and 2016. These datasets are obtained from the Bank Data UKM through its agreement with the Department of Statistics Malaysia. The data consist of at least 12,000 households in each dataset. The monthly gross income, household size, and weight of each sample in the data are used to model the income distribution by the 3PCP model and using the pseudo-likelihood approach. The monthly gross income is first equivalized by dividing the income by the square root of the household size. This square root equivalization is used in many studies to take into account the household size when considering the economic status of a household [58,59].

3.2. Application of the Model

The 3PCP models are applied to the HIS data. Originally, the 3PCP models with lognormal, gamma, or Weibull distributions for the middle data are applied to the HIS data. It is found that the inverse Pareto–lognormal–Pareto model with continuous but not differentiable PDF fits all five datasets based on the KS test statistics and the lowest BIC values as compared to other models. However, for some datasets, the estimated values $\hat{\rho}_1$ and $\hat{\rho}_2$ are found to be too small, with the smallest value being 0.0054 followed by 0.0074, which cannot be interpreted as proportion of poor and rich subpopulations, respectively. The estimated proportions are also inconsistent throughout the five datasets. This may occur because the lognormal distribution is already a good fit for some of the data, without needing the inverse Pareto and Pareto distributions in the model.

This is where the semi-parametric IP-GM-P model can be useful to make sure the lower and upper parts of the data are modelled by the inverse Pareto and Pareto distributions, respectively. The IP-GM-P model used is specified such that its PDF is continuous but not differentiable by specifying the values of ρ_1 and ρ_2 as in Equations (7) and (8). Adding the differentiability condition to the IP-GM-P model causes the number of components in the Gaussian mixture to increase just to make the PDF differentiable. While differentiability condition is more realistic, it is not useful for the IP-GM-P model.

When applying the IP-GM-P model, the number of components k for each dataset is first determined by using $k = 1, 2, 3$ and 4 and finding the value of k that gives the lowest BIC value. It is found that $k = 2$ gives the lowest BIC value for HIS data for the years 2007, 2009, and 2012 whereas $k = 3$ gives the lowest BIC value for HIS data for the years 2014 and 2016. Table 1 shows the estimated values for α_1, α_2, ρ_1, ρ_2, τ_1, and τ_2 for IP-GM-P model. The table also shows the p-values for the KS goodness-of-fit test. Based on the very large p-values for all five datasets, the IP-GM-P model has successfully fit with all of them. This is not unexpected because of the large number of parameters in the IP-GM-P model that helps with fitting the model.

Table 1. Estimated parameter values for IP-GM-P model, together with the p-values for the KS goodness-of-fit test.

Year	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\rho}_1$	$\hat{\rho}_2$	$\hat{\tau}_1$	$\hat{\tau}_2$	p-Value
2007	3.2515	2.3674	0.0781	0.0757	464	4074	0.9782
2009	3.7424	2.3995	0.0615	0.0819	475	4431	0.9115
2012	4.4722	2.4513	0.0361	0.0996	518	4956	0.7948
2014	4.4227	2.6125	0.0296	0.0287	681	10179	0.9919
2016	4.5953	2.6405	0.0245	0.0597	759	8261	0.9306

3.3. Income Inequality Using IP-GM-P Model

The income inequality of the data can be measured using the values of α_1 and α_2, as well as the Lorenz Curve and Gini index. Looking at the estimated values for ρ_1 in Table 1, the IP-GM-P model estimated that around 2.45% to 7.81% of the population belongs to the

lower income group. Note that the estimated values also drop from 2007 to 2016. Here, comparisons are made on the proportions and not the threshold parameters, as proportions are unit-free. If comparisons were made using threshold parameters, the inflation effect must be taken into consideration. Additionally, from the table, the estimated value for α_1 generally increases from 3.25 in 2007 to 4.59 in 2016. Since the Gini index for the lower data is inversely related to α_1, these values indicate that in general, the level of income inequality for the lower income group decreases from 2007 to 2016. Figure 2 shows the changes in α_1 and α_2 from 2007 to 2016. From the figure, it can be inferred that the level of income inequality for the lower income group decreases from 2007 to 2012 (as the value of α_1 increases), and does not change much from 2012 to 2016.

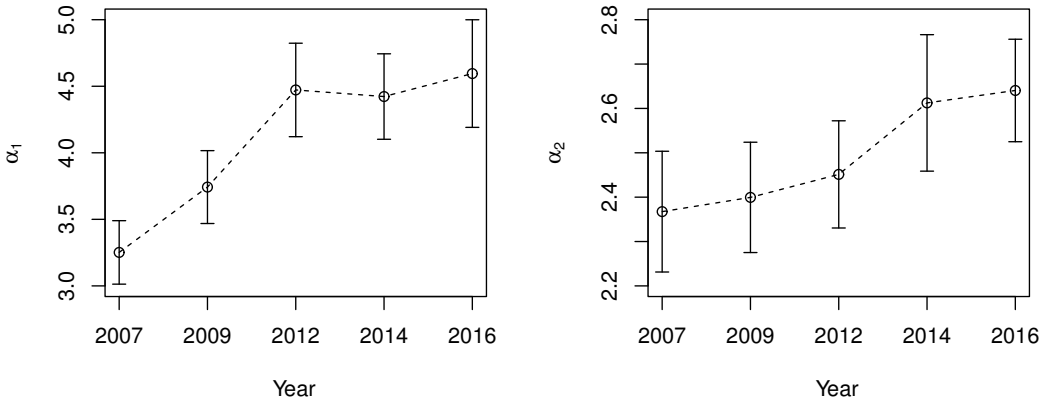

Figure 2. The estimated values for α_1 and α_2 using the IP-GM-P model together with their 95% confidence intervals.

For the upper income group, Table 1 shows that the estimated proportion of the upper income group in the population is around 2.87% to 9.96%. There does not seem to be any trend for the changes in ρ_2. As for α_2, the table shows that the its estimated value generally increases from 2.37 in 2007 to 2.64 in 2016. The changes in α_2 are also shown in Figure 2. Similar to the lower income group, the increase of α_2 indicates that the level of income inequality for the upper income group decreases over the period of time.

Figure 3 shows the Lorenz curve for household income for all five datasets obtained by using the IP-GM-P model. This Lorenz curve is obtained by substituting Equations (20) and (21) into Equation (13). From the figure, it can be observed that the Lorenz curve moves closer to the equality line from 2007 to 2016. Additionally, the Gini index is obtained by substituting Equations (20) and (21) into Equation (14) and using the trapezoidal rule to estimate the integral in Equation (14). It is found that the Gini index obtained by using IP-GM-P model for HIS data for the years 2007, 2009, 2012, 2014, and 2016 are 0.4434, 0.4406, 0.4267, 0.4051, and 0.3929, respectively. The decrease in the Gini index together with the increase in proximity of the Lorenz curve to the equality line suggest that, overall, the level of income inequality in Malaysia decreased from 2007 to 2016.

3.4. Comparison with Official Poverty Rate

The proportions of household in the lower income group represented by the inverse Pareto distribution in the IP-GM-P are compared to the official poverty incidences published by the Department of Statistics Malaysia [60]. There are two types of poverty used by the Department of Statistics Malaysia. The first type is the absolute poverty that includes households with income lower than a minimum threshold called poverty line income. According to the Department of Statistics Malaysia [61], the poverty line income is the minimum income required for a household to satisfy the basic needs of its members that has been identified through research conducted by the Economic Planning Unit, Prime Minister's Department and the Department of Statistics Malaysia in collaboration with the United Nations Development Programme. The second type of poverty is the relative

poverty defined as households with income less than half of the median household income of the population.

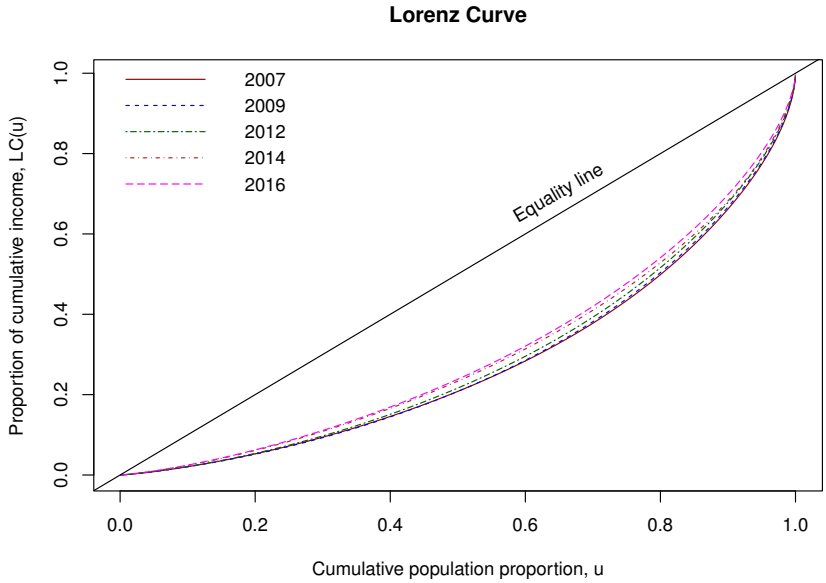

Figure 3. Lorenz curve for the household income data in Malaysia using IP-GM-P model.

Table 2 shows the percentage of household represented by the inverse Pareto distribution in the IP-GM-P model together with the official poverty incidences published by the Department of Statistics Malaysia for the years 2007, 2009, 2012, 2014, and 2016. Overall, the absolute poverty incidence decreases over the period of time and no trend can be observed for the relative poverty incidence. It is also noted that the relative poverty incidence is much higher as compared to the absolute poverty incidence.

Table 2. Percentages of lower income data explained by the inverse Pareto distribution and the official poverty incidences in Malaysia.

Year	IP-GM-P	Absolute Poverty Incidence	Relative Poverty Incidence
2007	7.81	3.6	17.4
2009	6.15	3.8	19.3
2012	3.61	1.7	19.2
2014	2.96	0.6	15.6
2016	2.45	0.4 [1]	15.9

[1] Using 2005 method. This value is 7.6% if 2019 method is used.

Overall, Table 2 shows that the percentage of household income modelled by the inverse Pareto distribution is between the absolute poverty incidence and relative poverty incidence. Additionally, the percentage of household modelled by the inverse Pareto distribution also decreases from 2007 to 2016, and the same can be observed for the absolute poverty incidence. Figure 4 shows the relationship between the percentage of the lower income group and the absolute and relative poverty incidences. Based on the figure, it can be observed that the percentage of lower income group seems to be linearly related to the absolute poverty incidence, with the high correlation coefficient. However, no relationship can be observed between the percentage of the group and the relative poverty incidence.

Although the percentage of the lower income group is not exactly equal to any of the two poverty incidences reported by the Department of Statistics Malaysia, there is a strong relationship between this percentage and the absolute poverty incidence based on the high correlation coefficient. This indicates that ρ_1 in the IP-GM-P model may be related to the absolute poverty incidence, and can be used to determine the absolute poverty incidence

without the need to determine the poverty line income, which may require additional time and cost.

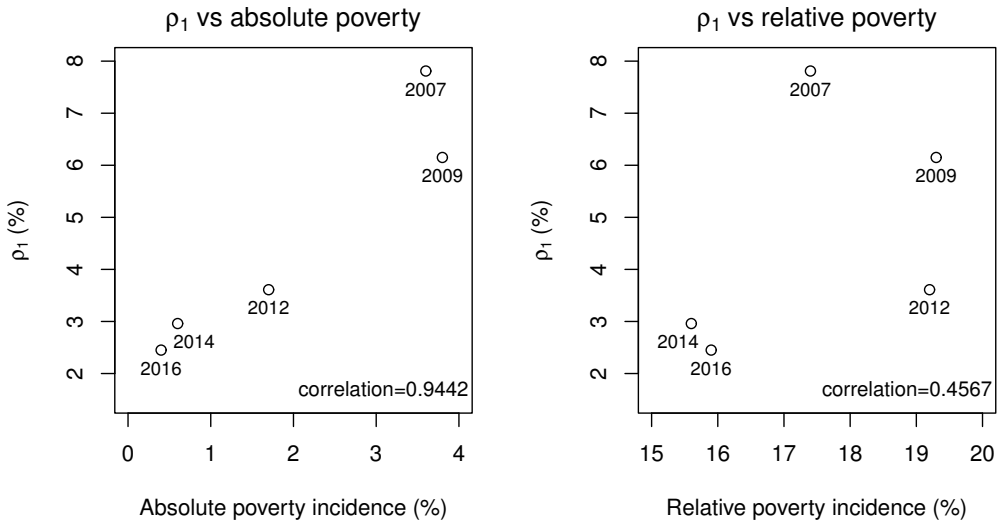

Figure 4. Comparison between ρ_1 obtained from IP-GM-P model and the official poverty incidences.

4. Conclusions

This paper proposes the use of a three-part composite Pareto (3PCP) model to be applied to the income distribution. The 3PCP model is a combination of inverse Pareto distribution for the lower part of the data, Pareto distribution for the upper part of the data, and another unspecified distribution for the middle part of the data. The general form of the probability density function (PDF) as well as the constraints required for the PDF to be continuous and differentiable are also given. Additionally, the Lorenz curve and Gini index for the 3PCP model are given. For the middle part of the data, this paper proposes to use a semi-parametric approach by using the Gaussian mixture distribution. This inverse Pareto–Gaussian mixture-Pareto (IP-GM-P) distribution model has the benefit that it allows lower and upper parts of the data to be described by the inverse Pareto and Pareto distributions, respectively.

The main advantage of the 3PCP model is that the model divides the population into three categories—the lower, middle, and upper income groups—and analyses them simultaneously, unlike previous literature that analyses each group separately. Additionally, the shape parameters in the Pareto and inverse Pareto distributions give insight on the levels of income inequality in the upper and lower income groups, respectively. Knowing how the income inequality changes in the lower and upper income groups may help policy makers in making decisions. Additionally, it is found, at least for the Malaysian household income, that the proportion of data following the inverse Pareto distribution is highly correlated with the official absolute poverty incidence. Therefore, the 3PCP model can be used to estimate the absolute poverty incidence in a country without having to find the poverty line income, which can be difficult.

However there are some challenges to the 3PCP model. First, due to the model complexity, the parameter estimation process can be difficult. In this paper, the parameters are estimated numerically which may not give reliable results. In some cases, several initial values were used to find the maximum likelihood estimates and there is no guarantee that the numerically estimated values are the ones that maximize the likelihood. Additionally, the lower, middle, and upper income groups derived from the 3PCP model may not align with the definition used by the governments and policy makers. In many countries, the income groups are defined by the quantiles, for example lower income earners are those in the bottom 40% of the population. The classification based on the quantiles are easier to be understood by the general public, compared to estimates found using the 3PCP model.

For future work, the performance of the 3PCP model must be assessed for other countries and not just Malaysia. It would be interesting to see if the 3PCP model can explain properties of income distribution in other countries. This paper focuses on the household income in Malaysia due to data availability and to make comparison with poverty incidence based on poverty line income. We expect the 3PCP model to fit income distribution of other countries. Comparison on the proportions of data following the Pareto and inverse Pareto distributions based on the 3PCP model can also be made for different countries. Furthermore, the robustness of the 3PCP model may be explored further, but we expect that the robustness of the 3PCP model to be similar to the Pareto distribution for data with extreme outliers. Robust estimators for the 3PCP model may also be developed.

Author Contributions: Conceptualization, M.H.A.M.; methodology, M.H.A.M.; software, M.H.A.M.; validation, K.I.; formal analysis, M.H.A.M.; writing—original draft preparation, M.H.A.M.; writing—review and editing, M.H.A.M. and N.M.; supervision, K.I. and N.M.; funding acquisition, N.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Universiti Kebangsaan Malaysia grant number DIP-2022-002.

Data Availability Statement: Restrictions apply to the availability of these data. Data is owned by a 3rd party and was obtained from the Bank Data UKM and the Department of Statistics Malaysia. Data can be requested via https://www.dosm.gov.my/portal-main/article/data-request, accessed on 17 April 2023.

Acknowledgments: The authors would like to thank the Bank Data UKM and the Department of Statistics Malaysia for providing the data used in this study. The authors would also like to thank the anonymous reviewers for their inputs and comments.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3PCP	Three-part composite Pareto model
BIC	Bayesian information criterion
CDF	Cumulative distribution function
HIS	Household income survey
IP-GM-P	Inverse Pareto-Gaussian mixture-Pareto
KS	Kolmogorov–Smirnov
PDF	Probability density function

References

1. Latimaha, R.; Bahari, Z.; Ismail, N.A. Examining the linkages between street crime and selected state economic variables in Malaysia: A panel data analysis. *J. Ekon. Malays.* **2019**, *53*, 59–72.
2. Wang, S. Hate Crime Analysis based on Artificial Intelligence Methods. *E3S Web Conf.* **2021**, *251*, 01062. [CrossRef]
3. Wilkinson, R.G.; Pickett, K.E. Income inequality and population health: A review and explanation of the evidence. *Soc. Sci. Med.* **2006**, *62*, 1768–1784. [CrossRef]
4. Babones, S.J. Income inequality and population health: Correlation and causality. *Soc. Sci. Med.* **2008**, *66*, 1614–1626. [CrossRef]
5. Patel, V.; Burns, J.K.; Dhingra, M.; Tarver, L.; Kohrt, B.A.; Lund, C. Income inequality and depression: A systematic review and meta-analysis of the association and a scoping review of mechanisms. *World Psychiatry* **2018**, *17*, 76–89. [CrossRef]
6. Dewan, P.; Rørth, R.; Jhund, P.S.; Ferreira, J.P.; Zannad, F.; Shen, L.; Køber, L.; Abraham, W.T.; Desai, A.S.; Dickstein, K.; et al. Income Inequality and Outcomes in Heart Failure. *JACC Heart Fail.* **2019**, *7*, 336–346. [CrossRef]
7. Posner, R.A. Equality, Wealth, and Political Stability. *J. Law Econ. Organ.* **1997**, *13*, 344–365. [CrossRef]
8. Ravallion, M. Ethnic inequality and poverty in Malaysia since May 1969. Part 1: Inequality. *World Dev.* **2020**, *134*, 105040. [CrossRef]
9. Dagum, C. *A New Model of Personal Income Distribution: Specification and Estimation*; Economie Appliquee; 1977; pp. 413–437.
10. Singh, S.K.; Maddala, G.S. A Function for Size Distribution of Incomes. *Econometrica* **1976**, *44*, 963–970. [CrossRef]
11. Drăgulescu, A.; Yakovenko, V.M. Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Phys. A Stat. Mech. Appl.* **2001**, *299*, 213–221. [CrossRef]

12. Fujiwara, Y.; Souma, W.; Aoyama, H.; Kaizoji, T.; Aoki, M. Growth and fluctuations of personal income. *Phys. A Stat. Mech. Appl.* **2003**, *321*, 598–604. [CrossRef]
13. Jenkins, S.P. Pareto Models, Top Incomes and Recent Trends in UK Income Inequality. *Economica* **2017**, *84*, 261–289. [CrossRef]
14. Oancea, B.; Andrei, T.; Pirjol, D. Income inequality in Romania: The exponential-Pareto distribution. *Phys. A Stat. Mech. Appl.* **2017**, *469*, 486–498. [CrossRef]
15. Oancea, B.; Pirjol, D.; Andrei, T. A Pareto upper tail for capital income distribution. *Phys. A Stat. Mech. Appl.* **2018**, *492*, 403–417. [CrossRef]
16. Reed, W.J. The Pareto, Zipf and other power laws. *Econ. Lett.* **2001**, *74*, 15–19. [CrossRef]
17. Reed, W.J. The Pareto law of incomes—An explanation and an extension. *Phys. A Stat. Mech. Appl.* **2003**, *319*, 469–486. [CrossRef]
18. Chakrabarti, B.K.; Chakraborti, A.; Chakravarty, S.R.; Chatterjee, A. Income and wealth distribution data for different countries. In *Econophysics of Income and Wealth Distributions*; Cambridge University Press: Cambridge, UK, 2013; pp. 7–34. [CrossRef]
19. Razak, F.A.; Shahabuddin, F.A. Malaysian Household Income Distribution: A Fractal Point of View. *Sains Malays.* **2018**, *47*, 2187–2194. [CrossRef]
20. Safari, M.A.M.; Masseran, N.; Ibrahim, K. Optimal threshold for Pareto tail modelling in the presence of outliers. *Phys. A Stat. Mech. Appl.* **2018**, *509*, 169–180. [CrossRef]
21. Safari, M.A.M.; Masseran, N.; Ibrahim, K. A robust semi-parametric approach for measuring income inequality in Malaysia. *Phys. A Stat. Mech. Appl.* **2018**, *512*, 1–13. [CrossRef]
22. Safari, M.A.M.; Masseran, N.; Ibrahim, K. On the identification of extreme outliers and dragon-kings mechanisms in the upper tail of income distribution. *J. Appl. Stat.* **2019**, *46*, 1886–1902. [CrossRef]
23. Safari, M.A.M.; Masseran, N.; Ibrahim, K.; AL-Dhurafi, N.A. The power-law distribution for the income of poor households. *Phys. A Stat. Mech. Appl.* **2020**, *557*, 124893. [CrossRef]
24. Cooray, K.; Ananda, M. Modeling actuarial data with a composite lognormal-Pareto model. *Scand. Actuar. J.* **2005**, *2005*, 321–334. [CrossRef]
25. Albrecher, H.; Beirlant, J.; Teugels, J.L. *Reinsurance: Actuarial and Statistical Aspects*; Statistics in Practice; John Wiley & Sons: Hoboken, NJ, USA, 2017.
26. Klugman, S.A.; Panjer, H.H.; Willmot, G.E. *Loss Models: From Data to Decisions*, 5th ed.; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2019.
27. Preda, V.; Ciumara, R. On composite models: Weibull-Pareto and Lognormal-Pareto. A comparative study. *Rom. J. Econ. Forecast.* **2006**, *3*, 32–46.
28. Scollnik, D.P.M. On composite lognormal-Pareto models. *Scand. Actuar. J.* **2007**, *2007*, 20–33. [CrossRef]
29. Cooray, K. The Weibull–Pareto Composite Family with Applications to the Analysis of Unimodal Failure Rate Data. *Commun. Stat. Theory Methods* **2009**, *38*, 1901–1915. [CrossRef]
30. Teodorescu, S.; Vernic, R. On composite Pareto models. *Math. Rep.* **2013**, *15*, 11–29.
31. Bakar, S.; Hamzah, N.; Maghsoudi, M.; Nadarajah, S. Modeling loss data using composite models. *Insur. Math. Econ.* **2015**, *61*, 146–154. [CrossRef]
32. Calderín-Ojeda, E. A Note on Parameter Estimation in the Composite Weibull–Pareto Distribution. *Risks* **2018**, *6*, 11. [CrossRef]
33. Aminzadeh, M.S.; Deng, M. Bayesian predictive modeling for Inverse Gamma-Pareto composite distribution. *Commun. Stat. Theory Methods* **2019**, *48*, 1938–1954. [CrossRef]
34. Benatmane, C.; Zeghdoudi, H.; Shanker, R.; Lazri, N. Composite Rayleigh-Pareto distribution: Application to real fire insurance losses data set. *J. Stat. Manag. Syst.* **2021**, *24*, 545–557. [CrossRef]
35. Mendes, B.V.d.M.; Lopes, H.F. Data driven estimates for mixtures. *Comput. Stat. Data Anal.* **2004**, *47*, 583–598. [CrossRef]
36. Luckstead, J.; Devadoss, S. Pareto tails and lognormal body of US cities size distribution. *Phys. A Stat. Mech. Appl.* **2017**, *465*, 573–578. [CrossRef]
37. Luckstead, J.; Devadoss, S.; Danforth, D. The size distributions of all Indian cities. *Phys. A Stat. Mech. Appl.* **2017**, *474*, 237–249. [CrossRef]
38. Wiegand, M.; Nadarajah, S. New composite distributions for modeling industrial income and wealth per employee. *Phys. A Stat. Mech. Appl.* **2018**, *492*, 1901–1908. [CrossRef]
39. Majid, M.H.A.; Ibrahim, K. Composite Pareto Distributions for Modelling Household Income Distribution in Malaysia. *Sains Malays.* **2021**, *50*, 2047–2058. [CrossRef]
40. Clementi, F.; Gallegati, M. Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States. In *Econophysics of Wealth Distributions: Econophys-Kolkata I*; Chatterjee, A., Yarlagadda, S., Chakrabarti, B.K., Eds.; Springer: Milano, Italy, 2005; pp. 3–14. [CrossRef]
41. Banerjee, A.; Yakovenko, V.M.; Di Matteo, T. A study of the personal income distribution in Australia. *Phys. A Stat. Mech. Appl.* **2006**, *370*, 54–59. [CrossRef]
42. Cowell, F.A.; Victoria-Feser, M.P. Robust stochastic dominance: A semi-parametric approach. *J. Econ. Inequal.* **2007**, *5*, 21–37. [CrossRef]
43. Kleiber, C.; Kotz, S. *Statistical Size Distributions in Economics and Actuarial Sciences*; Wiley Series in Probability and Statistics; John Wiley & Sons Inc.: Hoboken, NJ, USA, 2003. [CrossRef]

44. Kakwani, N.C.; Podder, N. Efficient Estimation of the Lorenz Curve and Associated Inequality Measures from Grouped Observations. In *Modeling Income Distributions and Lorenz Curves*; Chotikapanich, D., Ed.; Springer: New York, NY, USA, 2008; pp. 57–70. [CrossRef]
45. Gastwirth, J.L. A General Definition of the Lorenz Curve. *Econometrica* **1971**, *39*, 1037. [CrossRef]
46. Jorda, V.; Sarabia, J.M.; Jäntti, M. Inequality Measurement with Grouped Data: Parametric and Non-Parametric Methods. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2021**, *184*, 964–984. [CrossRef]
47. Schneider, M.P.A.; Scharfenaker, E. Mixing it up: The case for finite mixture models to study the distribution of income. *Eur. Phys. J. Spec. Top.* **2020**, *229*, 1685–1704. [CrossRef]
48. Anderson, G.; Farcomeni, A.; Pittau, M.G.; Zelli, R. A new approach to measuring and studying the characteristics of class membership: Examining poverty, inequality and polarization in urban China. *J. Econom.* **2016**, *191*, 348–359. [CrossRef]
49. Anderson, G.; Pittau, M.; Zelli, R.; Thomas, J. Income Inequality, Cohesiveness and Commonality in the Euro Area: A Semi-Parametric Boundary-Free Analysis. *Econometrics* **2018**, *6*, 15. [CrossRef]
50. Chotikapanich, D.; Griffiths, W.E. Estimating Income Distributions Using a Mixture of Gamma Densities. In *Modeling Income Distributions and Lorenz Curves*; Chotikapanich, D., Ed.; Springer: New York, NY, USA, 2008; pp. 285–302. [CrossRef]
51. Lubrano, M.; Ndoye, A.A.J. Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach. *Comput. Stat. Data Anal.* **2016**, *100*, 830–846. [CrossRef]
52. Marron, J.S.; Wand, M.P. Exact Mean Integrated Squared Error. *Ann. Stat.* **1992**, *20*, 712–736. [CrossRef]
53. Pfeffermann, D. The Role of Sampling Weights When Modeling Survey Data. *Int. Stat. Rev. Rev. Int. De Stat.* **1993**, *61*, 317. [CrossRef]
54. Chambers, R.L.; Skinner, C.J. *Analysis of Survey Data*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2003.
55. Janczura, J.; Weron, R. *Goodness-of-Fit Testing for Regime-Switching Models*; Working Paper; 2010.
56. Janczura, J.; Weron, R. Goodness-of-fit testing for the marginal distribution of regime-switching models with an application to electricity spot prices. *AStA Adv. Stat. Anal.* **2013**, *97*, 239–270. [CrossRef]
57. Xu, C.; Chen, J.; Mantel, H. Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data. *Surv. Methodol.* **2013**, *39*, 303–321.
58. OECD. *In It Together: Why Less Inequality Benefits All*; OECD Publishing: Paris, France, 2015. [CrossRef]
59. Congressional Budget Office. *Projected Changes in the Distribution of Household Income, 2016 to 2021*; Congressional Budget Office: Washington, DC, USA, 2019.
60. Department of Statistics Malaysia. *Estimates of Household Income and Poverty Incidence in Malaysia 2020*; Technical Report; Department of Statistics Malaysia: Putrajaya, Malaysia, 2021.
61. Department of Statistics Malaysia. *Household Income and Basic Amenities Survey Report 2016*; Technical Report; Department of Statistics Malaysia: Putrajaya, Malaysia, 2017; ISSN 2232-1012.