Sains Malaysiana 55(3)(2026): 503-515
http://doi.org/10.17576/jsm-2026-5503-12

Robust Estimation of Inverse Pareto Distribution: Insights into Malaysian Lower-Income Groups

(Anggaran Teguh Taburan Pareto Songsang: Pandangan tentang Kumpulan Berpendapatan Rendah di Malaysia)

MUHAMMAD FAHEM MUSA[1], MOHD AZMI HARON[1,*], MUHAMMAD ASLAM MOHD SAFARI[2,3] & ZAILAN SIRI[1]

[1]Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[3]Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Received: 22 January 2025/Accepted: 19 February 2026

ABSTRACT

The lower tail data of the income distribution can be well described by the inverse Pareto distribution, which is a useful statistical model. Researchers often use the maximum likelihood approach in evaluating the model's shape parameter. However, this approach might be unreliable as extreme values data points can easily affect it negatively, which may cause inaccurate estimates. This paper introduces a robust estimator known as repeated median estimator for the inverse Pareto distribution. We also demonstrate its efficiency by measuring its asymptotic relative efficiency. The robustness of the method is assessed by the breakdown point and the influence function. The Monte Carlo simulation studies are then conducted to measure its performance as compared to its counterparts. It is found that our proposed estimator outperforms the maximum likelihood, method of moments, and method of product spacings based on the simulation studies. We then apply the inverse Pareto distribution, leveraging the repeated median estimator, to model the lower-income data from the Household Income Surveys in Malaysia for the years 2012, 2014, 2016, 2019, and 2022. To compute the income disparity of low-income households in Malaysia, the parametric Lorenz curve is fitted using the inverse Pareto model, and the Gini coefficient is estimated. This confirms that the proposed approach achieves meaningful practical performance improvements over conventional estimation techniques in the presence of outliers.

Keywords: Inverse Pareto distribution; Malaysian lower income; Monte Carlo simulation; repeated median estimator; robust estimation

ABSTRAK

Taburan Pareto songsang ialah model statistik yang berkesan dalam menerangkan data ekor bawah pengagihan pendapatan. Penyelidik kebiasaannya bergantung kepada penganggar kebolehjadian maksimum dalam menganggarkan parameter bentuk model ini. Walau bagaimanapun, kaedah ini mungkin tidak berkesan kerana ia boleh dipengaruhi dengan mudah oleh nilai melampau yang boleh menyebabkan anggaran yang tidak tepat. Dalam kajian ini, kami memperkenalkan penganggar teguh yang dikenali sebagai penganggar median berulang. Kami juga menunjukkan kecekapannya dengan mengukur kecekapan relatif asimtotik. Keteguhan penganggar ini dinilai oleh nilai pecahan dan fungsi pengaruh. Kajian simulasi Monte Carlo kemudiannya dijalankan untuk mengukur prestasinya berbanding dengan penganggar lain. Didapati bahawa penganggar yang kami cadangkan mengatasi kebolehjadian maksimum, kaedah momen dan kaedah jarak produk berdasarkan kajian simulasi. Kajian ini kemudiannya menggunakan taburan Pareto songsang dengan memanfaatkan penganggar median berulang untuk memodelkan data pendapatan rendah daripada Tinjauan Pendapatan Isi Rumah di Malaysia bagi tahun 2012, 2014, 2016, 2019 dan 2022. Untuk mengukur jurang pendapatan isi rumah berpendapatan rendah di Malaysia, keluk Lorenz dipadankan dan pekali Gini dianggarkan berdasarkan model Pareto songsang. Ini mengesahkan bahawa pendekatan yang dicadangkan mencapai peningkatan prestasi praktikal yang signifikan berbanding dengan teknik anggaran konvensional walaupun terdapat pencilan.

Kata kunci: Anggaran teguh; pendapatan rendah Malaysia; penganggar median berulang; simulasi Monte Carlo; taburan Pareto songsang

INTRODUCTION

Studies on the Pareto principle have sparked much interest in the recent few decades from interdisciplinary sectors such as economic science and financial analysis (Abd Raof et al. 2022; Alfons, Templ & Filzmoser 2013; Filimonov & Sornette 2015; Safari, Masseran & Ibrahim 2018; Safari

et al. 2024). The study by Razak and Shahabuddin (2018) evidenced that the upper tails of the Malaysian gross income distribution obey the power-law suggesting that it is naturally fractal. Majid and Ibrahim (2021) employed a composite model to fit the entire Malaysian income distribution which leveraged the Pareto distribution for the upper tail. The Pareto principle, nevertheless, also applies to the low-end observations (Luckstead, Devadoss & Danforth 2017; Masseran et al. 2019). This makes the inverse Pareto distribution capable of visualizing the power law properties for the lower income data (Masseran et al. 2019). As the name suggests, it is an inverse version of the typical Pareto distribution (Kleiber 2003).

The lower quantile of the income distribution can be depicted as the reflection of the upper quantile income distribution. Reed (2003) suggested that the lower tail of the income distribution likewise displays power-law behaviour, indicating that the density of incomes at low-level is proportional to raise to a positive power. As a result of this occurrence, the income distribution shifts from the lower, rapidly growing portion to the middle, which gains influence. Notably, poverty and penury can cause a skewed lower end in income distribution. Prior research showed that the inverse Pareto distribution can fit the lower tail of the income distribution, demonstrating that this model is can provide a sufficient explanation for the lower tail data (Masseran et al. 2019; Safari et al. 2021, 2019).

The income of the poor-income group was comparatively lower than that of other groups, causing it to be at the lower end of the Malaysian income distribution. The presence of households with extremely low incomes indicates extreme observations in the lower tail of the income distribution, which in turn leads to the heavy-tailed property in the lower tail. The heavy-tailed distributions are frequently linked to power-law behaviour (Clauset, Shalizi & Newman 2009). Cowell and Flachaire (2007) stated that the assessment of income disparity will be impacted if the heavy-tailed property is included in the income distribution. A more stable measurement of income disparity can be attained by modelling the income distribution using the parametric distribution to solve this problem.

There are many types of statistical distribution that can fit the lower tail distribution. These include the exponential, stretched exponential, and log-normal distributions (Brzezinski 2015, 2014; Klaus, Yu & Plenz 2011; Laherrere & Sornette 1998). However, the study by Safari et al. (2020) indicates that the inverse Pareto distribution outperformed all these models in terms of fitting the left-tail data of Malaysian household incomes. This finding motivates us to leverage the inverse Pareto distribution to model Malaysian household incomes. The probability density function (PDF), cumulative distribution function (CDF), and quantile function of inverse Pareto distribution are given accordingly as follows.

$$f\left(x;\alpha,x_0\right)=\frac{\alpha x^{\alpha-1}}{x_0^{\alpha}}, \quad 0<x<x_0, \tag{1}$$

$$F\left(x;\alpha,x_0\right)=\left(\frac{x}{x_0}\right)^{\alpha}, \quad 0<x<x_0, \tag{2}$$

$$Q\left(y;a;x_0\right)=x_0 y^{\frac{1}{\alpha}}, \quad 0<y<1 \tag{3}$$

where $\alpha > 0$ is the shape parameter and x_0 represents the scale parameter. Note that a decrease in the value of a leads to a more substantial tail in the inverse Pareto distribution (Safari et al. 2020).

The method most frequently used to estimate the parameter α is the maximum likelihood estimator (MLE). It has been demonstrated that the MLE is relatively effective for standard parametric models, particularly for large sample sizes. Nevertheless, when contamination is present in the data, the MLE is unreliable due to its strong bias. This shortcoming is due to the diverged influence function (IF) of MLE which causes inaccurate estimation of shape parameters. When the outlier is present, finding an alternative, more reliable estimator than MLE is required.

Therefore, we present an outlier-robust estimator method as an alternative to the MLE for better estimating of the shape parameter called the repeated median estimator (RME). We demonstrate the RME's capability to resist the outlier-contamination effects via a high breakdown point and constrained IF. We conducted a Monte Carlo simulation study to assess the performance of our proposed method against existing techniques, including the MLE, method of product spacing (MPS), and method of moments (MOM), in both outlier-free and outlier-influenced scenarios. Afterwards, we employed the proposed method to estimate the inverse Pareto distribution. The estimated inverse Pareto distribution is then, utilized to analyze the lower-income data from Malaysia.

THE PROPOSED REPEATED MEDIAN ESTIMATOR

This section begins with the formulation of RME. Then, we evaluate the estimator's efficiency by employing the concept of asymptotic relative efficiency (ARE), which serves as a criterion for comparing its performance against MLE. Moving forward, we determine the robustness of RME via two key robustness metrics: the breakdown value and the IF.

FORMULATION OF THE REPEATED MEDIAN ESTIMATOR

Suppose a random variable, X, has an inverse Pareto distribution, then, the log(X) quantile plot can be expressed as:

$$G^{-1}(p,\alpha,x_0)=\frac{1}{\alpha}\log(p)+\log\left(x_0\right), \qquad 0<p<1 \tag{4}$$

The quantiles for both the empirical and generic log-inverse Pareto distributions are proportionally related, provided that $Z(x) = G(x;1,1)$. As a result, we may write Eq. (4) as:

$$G^{-1}(p,\alpha,x_0) = \frac{1}{\alpha}Z^{-1}(p) + \log(x_0) \quad , \qquad 0 < p < 1 \quad (5)$$

where p is the percentile of the random variable X.

Let us consider that the log-inverse Pareto model, denoted by $G(x; \alpha, x_0)$, form a location-scale family, where the scale parameter is $\sigma = 1/x_0$, and the location parameter is $\mu = \log(x_0)$:

$$G(\log x, \alpha, x_0) = F(x;\alpha,x_0) = Z\left(\frac{\log x - \mu}{\sigma}\right), \qquad 0 < x < x_0 \quad (6)$$

The quantiles of the standard log-inverse Pareto model, with intercept $\beta_0 = log\ (x_0)$ and slope $\beta_1 = 1/\alpha$, are linearly related to the quantiles of the general log-inverse Pareto distribution, as shown in Equation (5). Therefore, a robust regression approach can be used to estimate the shape parameter of the inverse Pareto distribution. Specifically, by regressing the logarithms of the observed quantiles on the corresponding theoretical quantiles from the standard log-inverse Pareto model, one can obtain parameter estimates. This method provides an alternative estimation procedure that is less sensitive to outliers and may yield a better fit to the data. Inserting empirical values, $G^{-1}(-1)\ (p,\alpha,x_0)=\hat{q}_{i/(n+1)}$ and theoretical quantiles, $p = \frac{i}{n+1}$ into Equation (5) gives us a straight-line regression formula:

$$y_i = \beta_0 + \beta_1 z_i + \epsilon_i, \quad (7)$$

where $y_i = log\ \hat{q}_{i/(n+1)}$, $\beta_0 = log\ (x_0)$, $\beta_1 = 1/\alpha$, $z_i = Z^{-1}\left(\frac{i}{n+1}\right)$ and ϵ_i are the error terms for $i = 1, \ldots, n$.

Keep in mind that this equation's error terms do not all have the same independent distribution. The estimator suggested in this section is based only on the estimated quantiles expressed as a straight-line regression in Equation (7) provided that the error is negligible. The inverse Pareto shape parameter can be estimated as $\hat{\alpha} = 1/\hat{\beta}_1$. Taking into account the repeated median estimation approach put forth by Siegel (1982). Let $med_j\ (z_i) = median(z_1, \ldots, z_n)$. The repeated median gradient estimate is then, written as

$$\hat{\beta}_1 = med_j med_{i \neq j} \frac{y_j - y_i}{z_j - z_i} \quad (8)$$

The fact that, $\frac{y_j - y_i}{z_j - z_i} \geq 0$, implies $\hat{\beta}_1 \geq 0$.

ASYMPTOTIC RELATIVE EFFICIENCY

MLE is well-regarded for its efficiency, serving as a useful benchmark for evaluating statistical measures. Specifically, comparing the asymptotic variances of MLE and RME allows us to assess the relative efficiency of the estimator $\hat{a}$ in relation to a_{MLE}. This concept captures how efficiently an estimator performs compared to MLE. Given that $\hat{a}_n$ is the estimated shape parameter value from any estimation method and $\hat{a}_{MLE}$ is the shape parameter value estimated by MLE, the equation below can be used to evaluate the ARE of an estimator in comparison to MLE (Safari & Masseran 2024).

$$ARE(\hat{\alpha}_n) = \lim_{n\to\infty} \frac{Var(\hat{\alpha}_{MLE})}{Var(\hat{\alpha}_n)} \quad (9)$$

We estimate the ARE of the RME using the simulation of 5,000 generated sample sets with size, n = 10,000. The ratio value, $Var(\alpha_{MLE})/\ Var\ (\alpha_{RME})$, represents the ARE of RME. We discovered that the ARE of RME with respect to MLE is approximately 74.09%.

BREAKDOWN VALUE

Theoretically, the repeated median procedure has an asymptotic breakdown value of 50%. The estimated shape parameter, $\hat{\alpha} = 1/\hat{\beta}_1$ may diverge to infinity when $\hat{\beta}_1$ tends toward zero. As z_i are fixed, this situation may only occur considering that half of the y_i data points overlap. This indicates that 50% corrupted data points are required. Thus, $\hat{a}_{RME}$ retains the 50% breakdown characteristic of repeated median regression estimators. Additionally, when more than $(n + k - 1)/2$ data points are kept fixed while the others are allowed to vary, the RME remains stable. This can be demonstrated using the lemma 1 below which suggested by Siegel (1982).

Lemma 1: Let $\tilde{\theta}_s(i_1,\ldots,i_k)$ be a class of functions, where $i_j \in [1,n]$ are different values of integers and distinct values of s corresponds to different datasets. Suppose $F \subset \{1, \ldots, n\}$ has more than $\frac{1}{2}(n + k - 1)$ elements and that the $\tilde{\theta}_s(i_k,\ldots,i_k)$ are bounded as α varies whenever $i_1,\ldots,i_k \in F$. Then, the repeated median values $\hat{\theta}_s = median^k\left\{\tilde{\theta}_s(i_1,\ldots,i_k)\right\}$ are also bounded.

Proof: Using the induction on k, when k = 1, the breakdown value is equal to the breakdown of the univariate median. Assuming the hypothesis from the lemma is true and computing the first median, we see that $\tilde{\theta}_s(i_1,\ldots,i_{k-1},.) = median\left\{\tilde{\theta}_s(i_1,\ldots,i_k)\right\}$ are bounded provided that $i_1, \cdots i_{k-1} \in F$, because the median has n - k terms, of which more than half are bounded. By the induction hypothesis, $median^{k-1}\left\{\tilde{\theta}_s(i_1,\ldots,i_{k-1},.)\right\}$ is bounded; because this is $\hat{\theta}_s$, the proof is complete.

INFLUENCE FUNCTION

Hampel, Ronchetti and Rousseeuw (1986) introduced an additional technique for evaluating the robustness of a statistical methodology: the influence function (IF). An estimator is considered resilient to outliers when its IF is bounded. As explained by Maronna et al. (2019), the IF provides an asymptotic representation of the estimator's sensitivity curve (SC). The SC measures an estimator's sensitivity to the position of an outlier x_ϵ in a random

sample. In this work, we use the corresponding SCs of each estimator to simplify the construction of the IF for the RME. One can define the estimator's SC as a function of the outlier's location, x_ϵ given a random variable $X_1, \cdots X_n$. The SC of an estimator is given as follows,

$$SC(x_\epsilon) = \hat{a}_n(x_1, x_2, \cdots, x_n, x_\epsilon) - \hat{a}_n(x_1, x_2, \cdots, x_n) \qquad (10)$$

where $\hat{a}_n(x_1, x_2, \cdots, x_n)$ is the estimate based on the sample, $X_1, \cdots X_n$. We obtain the RME's SC by simulating random data with a sample of size $n = 50$ that follows inverse Pareto distribution with $x_0 = 100$ and $\alpha = 1,2,3,4,5$. This sample size is sufficient according to Safari and Masseran (2024). After that, we add a lower outlier data point, x_ϵ to the generated data and decrease it by 0.5, adjusted from 100 to 1. The SC value is computed using Equation (10). As displayed in Figure 1, the sensitivity curve will reach a certain limit as the value of $1/x_\epsilon$ increases. This also suggests that the SC of RME is bounded with certain values as x_ϵ tends to decrease. The visualisation of SC affirms the resistance of RME against single outliers due to its bounded IF.

MONTE CARLO SIMULATION DESIGN

The effectiveness of RME in estimating the shape parameter of the inverse Pareto model is assessed by the Monte Carlo simulation, for both cases - with outliers and no outliers. This method identifies which estimation techniques hold up well when dealing with anomalous data. Furthermore, the Monte Carlo simulation can also compare the performance of RME along with other existing estimators such as MLE, MOM, and MPS. The details on these existing estimators are provided in Appendix A. In this simulation study, the RME is imitated from the inverse Pareto model for a predefined threshold level ($x_0 = 100$). The true values of shape parameter for each case are defined initially as α = 2,3,4,5. For small-sample scenarios, sample sizes of n = 30, 50, 70 are considered, whereas larger-sample scenarios use n = 200, 500, 800. The number of outliers is set to 0, 1, 3, 5, and 7 for small samples, and to proportions of 0%, 2%, 5%, 7%, and 10% for large samples. Lower-tail outliers are generated by dividing randomly selected observations from the datasets by 100. The R programming language (R Core Team 2021) run this simulation 2000 times for 2000 produced datasets. The percentage relative root mean square error (RRMSE) assesses the estimators' performance. Given that α is the exact value of the shape parameter for the inverse Pareto distribution, the RRMSE is defined as

$$RRMSE = \frac{100}{\alpha}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_i - \alpha)^2}, \qquad (11)$$

where N is the number of simulations and $\hat{a}_i$ is the estimated shape parameter for the simulated data set i = 1,2, ... n. An estimator with a lower RRMSE value is thought to be more accurate. Therefore, the estimator that reduces the RRMSE is considered as optimal for estimating the shape parameter α when outliers are present.

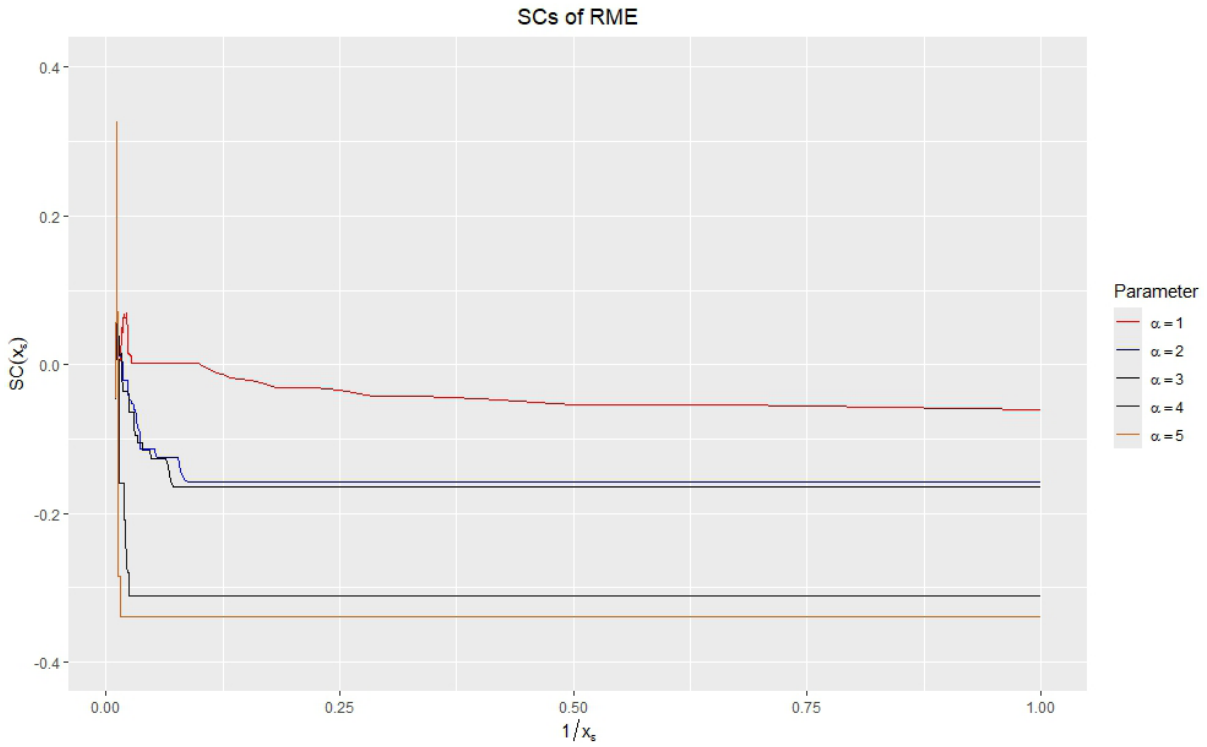

FIGURE 1. Sensitivity curve of RME for α = 1, 2, 3, 4, 5

SIMULATION RESULTS

The simulation findings for the small and large sample sizes are illustrated respectively in Figures 2 and 3. For the most part, all estimators considered generally demonstrates a decline in RRMSE values as sample sizes increase, suggesting the rise in their performance for growing-size samples.

MLE, MPS, and MOM perform about equally well for small sample sets, exhibiting RRMSE values that are relatively similar across various conditions. However, as the analysis progresses, it becomes clear that RME outperforms the other estimators, particularly in scenarios where the number of outliers increases. This trend highlights RME's robustness and underscores the challenges faced by the other methods when dealing with lower extreme values. As the shape parameter value increases, this discrepancy becomes more evident across all small sample sizes, illustrating the importance of considering the nature of the data when choosing an estimator.

For instance, when the shape parameter, α is set to 2, MOM demonstrates commendable performance across all outlier settings, providing reliable estimates even with moderate number of outliers. However, as the value of α increases beyond this point, the performance of MOM begins to decline significantly, leading to less accurate estimates. This deterioration in MOM's effectiveness highlights a critical point: While it may be a suitable in certain contexts, its reliability diminishes under specific conditions.

In contrast, MLE and MPS consistently rank as the least desirable options as the number of outliers rises. Their vulnerability to the influence of outliers makes them less effective in scenarios where data may not conform to ideal assumptions. This analysis ultimately emphasizes the necessity of carefully selecting estimation methods based on the underlying characteristics of the data and the presence of outliers, ensuring more accurate and reliable results in statistical modelling.

In the case of large sample sizes, RME's performance is almost as effective as the other estimators for a no-outlier case. With the increasing percentage of outliers, RME outperforms other estimators with lower RRMSE values. The RRMSE of MOM and MPS begin to increase significantly even with only 2% outliers, indicating that these estimators are unreliable for handling outlier contamination. Even though MOM's performance for α = 2 is roughly on par with RME's, it appears to be losing ground as the value of the shape parameter increases. Unlike the other estimators, RME's performance does not deteriorate with the increase in the value of the shape parameter's value. Its RRMSE is capped at most 25 for all large sample cases.

Overall, the simulation studies suggest that RME is better than MLE, MOM, and MPS in shape parameter estimation of the inverse Pareto distribution for outlier-contaminated datasets, supporting the statement that RME is a robust estimation approach.

DATA APPLICATION

In this study, we utilized data extracted from the monthly gross household earnings provided by the Department of Statistics Malaysia (DOSM) for the years 2012, 2014, 2016, 2019, and 2022. The data is based on the Household Income Survey (HIS), carried out twice over a five-year period. The HIS is used to obtain information on the distribution of household income, compiling data on households living in poverty, and figuring out how easily accessible basic facilities are to homes. This data also helps the government reducing income inequality and eradicates poverty. Table 1 summarize the statistical summary of the Malaysian household income datasets.

It's crucial to emphasize that the study is centred solely on the gross household incomes of low-income earners in Malaysia. The low-income earners in Malaysia are known as B40 which means households within 40% and below out of the total income distribution. These groups are classified into absolute, hard-core, and relative poverty (DOSM 2024). When a household's gross monthly income falls below the poverty line income (PLI), they are considered to be living in absolute poverty. Meanwhile, a household earning less than the food poverty line income (Food PLI) per month is considered to be in hard-core poverty. In addition, a household earning less than half of the national median income each month is in relative poverty. A detailed study on the classification of poverty in Malaysia can be referred in Abu, Hamdan and Sani (2020). Table 2 provides the threshold income for absolute poverty, relative poverty, and hard-core poverty provided by DOSM (2024).

The transformation method can be employed to demonstrates that the reverse exponential distribution, or the mirror image of the exponential distribution across the y-axis, followed by the logarithms of the inverse Pareto random variable, X. A graphical approach for validating the assumption of an inverse Pareto model is proposed: the inverse Pareto quantile plot. This technique involves the log-transformed random variable. By plotting the logarithms of the actual data points, or $log\ (x_i)$ where i = 1, 2, ... , n, versus the theoretical quantiles of the inverted exponential distribution, one can obtain the inverse Pareto quantile plot. The classical inverted exponential distribution's theoretical quantiles are provided by:

$$\log\left(\frac{i}{n+1}\right) \quad , \quad i = 1, 2, \dots n \tag{12}$$

The data points on the inverse Pareto quantile plot will appear to align closely along a straight line if the dataset is consistent with an inverse Pareto distribution. The threshold can be any value, starting from x_0, which is the farthest data point on the fitted line. Based on the quantile plot depicted in Figure 4, we selected the threshold value 1st to the 50th percentile of the Malaysian income dataset.

This study uses the scale parameter that minimizes the Kolmogorov-Smirnov (KS) statistics. This approach of selecting the scale parameter has been used by lots of previous studies. The KS statistic is given as:

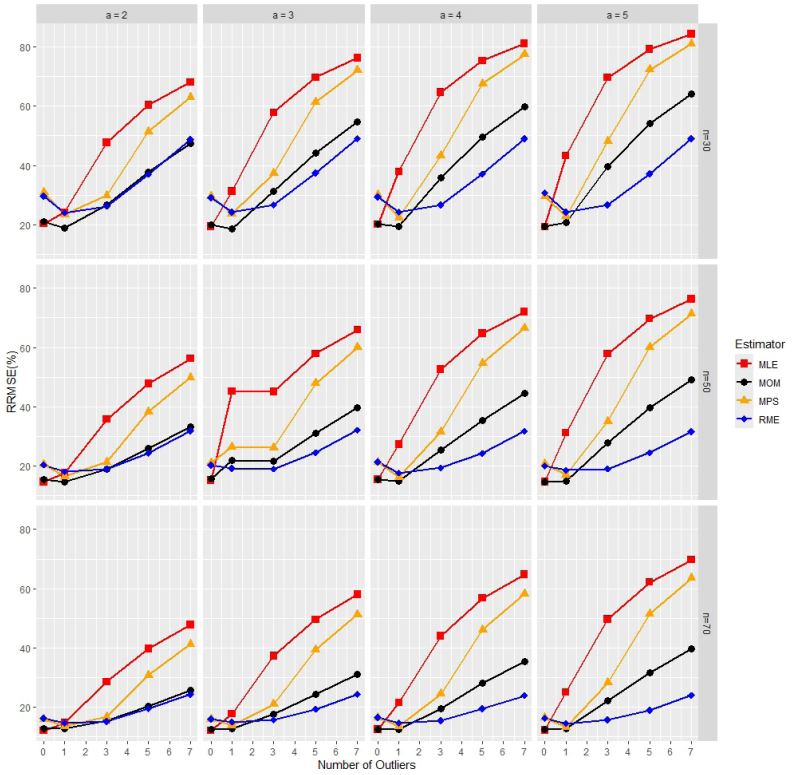

FIGURE 2. RRMSE of the inverse Pareto shape parameter estimations $\alpha = 2$, 3, 4, 5 for $n = 30$, 50, 70, and 0, 1, 3, 5, 7 as the number of outliers

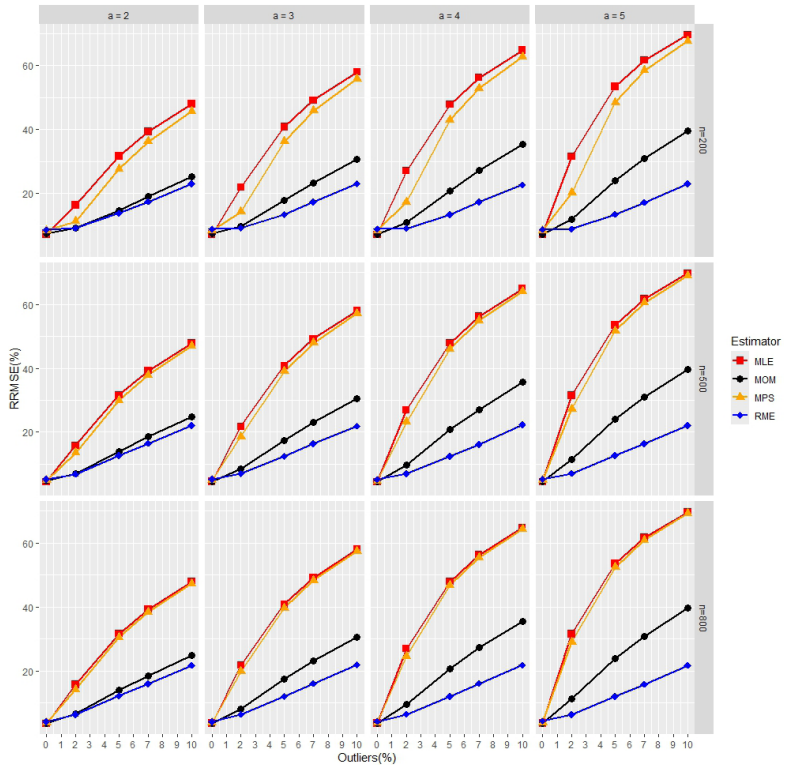

FIGURE 3. RRMSE of the inverse Pareto shape parameter estimations $\alpha = 2$, 3, 4, 5 with $n = 200$, 500, 800 and outlier's percentage = 0%, 2%, 5%, 7%, 10%

TABLE 1. The statistical summary of the Malaysian household income data

Year	Mean (RM)	Median (RM)	Min (RM)	Max (RM)	Variance	Skewness	Kurtosis
2012	4480.00	3221.00	150.00	105958.00	23448300	6.192304	76.0361
2014	5746.80	4251.50	212.50	186892.00	33825399	6.924341	116.9348
2016	6298.20	4701.10	269.60	368585.00	43319581	13.92394	543.8493
2019	6979.50	5142.80	318.20	882163.80	89727701	43.84595	3519656
2022	7549.90	5638.40	451.20	303150.70	58253325	9.726517	241.7010

TABLE 2. The baseline income for absolute poverty, relative poverty, and hard-core poverty

Year	Absolute poverty (RM)	Relative poverty (RM)	Hard-core poverty (RM)
2012	666.63	1573.70	564.07
2014	699.23	2091.50	497.07
2016	754.31	2319.28	583.45
2019	1793.72	2639.14	919.19
2022	1983.33	2870.89	947.71

$$D = \max_{x<x_0} |F_n(x) - F(x)| \tag{13}$$

where $F_n(x)$ represents the empirical distribution function and $F(x)$ is the CDF of the inverse Pareto model in Equation (2) for the lower tail income data.

To evaluate how well the inverse Pareto model fits the data, two measures are used: the KS test and the coefficient of determination, R^2. The model's adequacy is evaluated by contrasting the computed KS statistic, as defined in Equation (13), with the threshold value. Suppose the p-value exceeds the threshold for statistical significance or the KS statistic is lower than the reference value. In that case, it is concluded that the lower household income data in Malaysia adheres to the inverse Pareto distribution.

The relationship between the actual and estimated values of a given distribution is quantified by the R^2coefficient. A higher R^2 value suggests that the theoretical distribution fits the real data more closely. One can compute the coefficient of R^2 using the formula

$$R^2 = \frac{\sum_{i=1}^{n}\left[\hat{F}(x_i;\alpha,x_0) - \bar{F}(x;\alpha,x_0)\right]^2}{\sum_{i=1}^{n}\left[\hat{F}(x_i;\alpha,x_0) - \bar{F}(x;\alpha,x_0)\right]^2 + \sum_{i=1}^{n}\left[F_n(x_i) - \hat{F}(x_i;\alpha,x_0)\right]^2}. \tag{14}$$

where $\hat{F}(x_i)$ represents the estimated cumulative probabilities for the i_{th} income value in the lower end of the distribution according to the inverse Pareto model, $F(x)$ denotes the mean of $\hat{F}(x_i)$, and $F_n(x_i)$ refers to the observed cumulative probabilities for the i_{th} income value in the lower portion of the distribution.

Table 3 provides the result of the goodness-of-fit measure of the estimated inverse Pareto shape parameter as well as the scale parameter value. The fitted inverse Pareto models provide p-values that surpass the significance level of 0.05, suggesting that the inverse Pareto model is a fitting model for the lower tail data pertaining to household incomes in Malaysia for the five selected years. For the 2012 dataset, we discovered that the inverse Pareto model estimated by RME, managed to capture the lowest tail data points with a 779 sample size compared with the other estimators. Based on the KS-test and R^2 coefficient, RME offer the best fitting for the 2012 dataset. For 2019 and 2022 datasets, we discovered that RME provides the best fitting with the highest p-value of KS statistic. It also has R^2 coefficient nearest to 1 for 2022 datasets compared with the other considered estimators. Figure 5 displays the best fit estimated inverse Pareto distribution for the lower household income in Malaysia for the years 2012, 2014, 2016, 2019, and 2022. We discovered that the best fit of the estimated inverse Pareto distribution is sufficient to cover all the hard-core poverty groups for all 5 selected years and absolute poverty groups for 2012, 2014, and 2016. However, it cannot fit all the B40 groups and relative poverty groups.

The Gini index was computed using the fitted inverse Pareto model, as reported in Table 4. The estimated Gini coefficient is slightly below 0.2, indicating a relatively low level of income inequality among the hard-core poverty group in Malaysia. The corresponding Lorenz curve is presented in Figure 6. The curve shows that approximately 74.70% to 76.96% of the total household income is received by the bottom 80% of the population within the

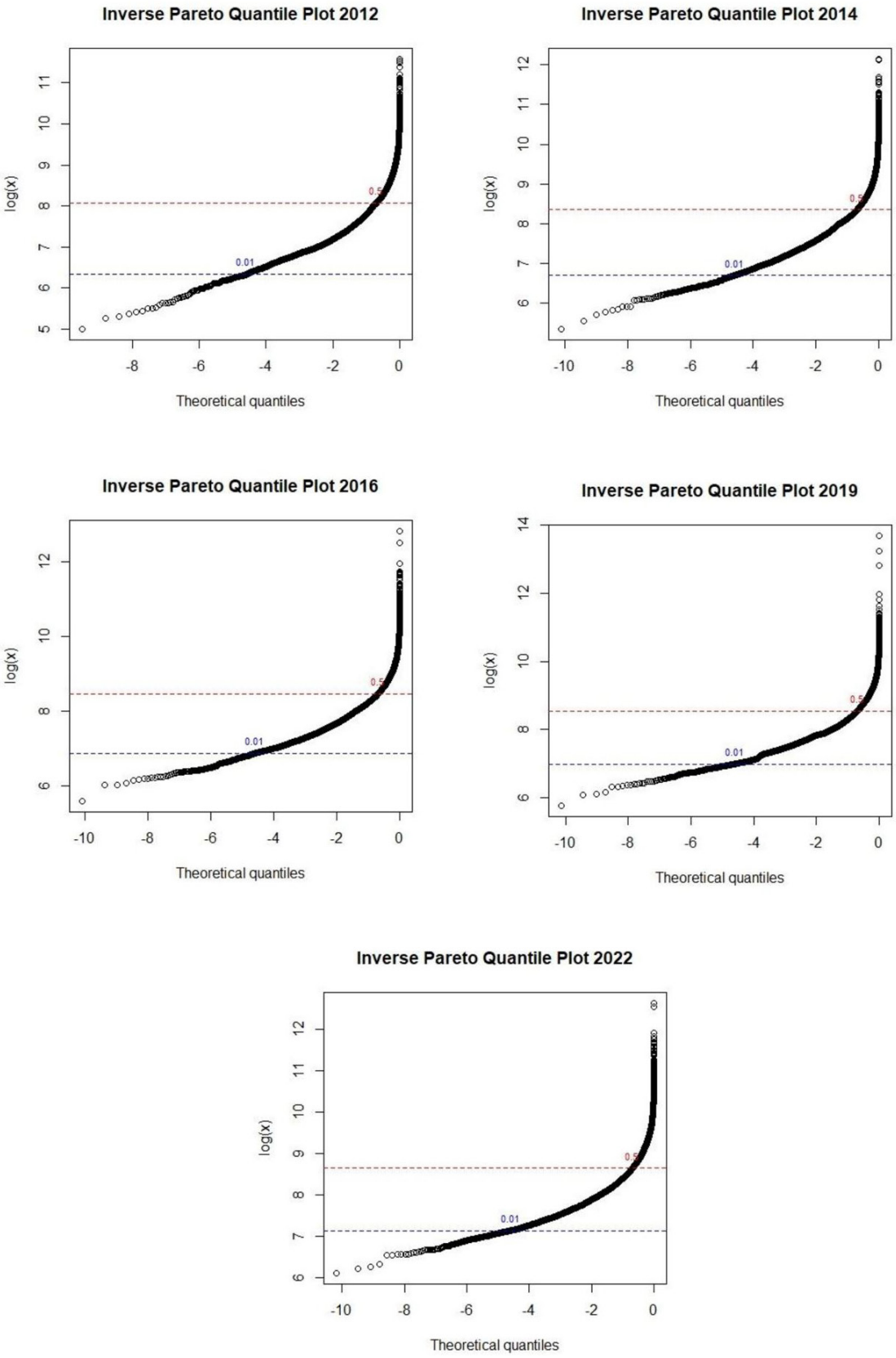

FIGURE 4. Quantile plot of Malaysian household income for the years 2012, 2014, 2016, 2019, and 2022

TABLE 3. Inverse Pareto distribution fit for Malaysian lower-tail income data

Year	Estimator	n_{tail}	$\hat{x}_0$	$\hat{a}$	KS statistic (p-value)	R^2
2012	MLE	769	983.00	3.279393	0.215380 (0.8678807)	0.9990619
	MOM	769	983.00	3.248608	0.018142 (0.9619385)	0.9993385
	MPS	754	978.00	3.188262	0.019182 (0.9442360)	0.9993747
	RME	779	984.83	3.256431	0.016803 (0.9804187)	0.9994285
2014	MLE	888	1179.17	3.599257	0.035115 (0.2235448)	0.9958368
	MOM	875	1173.08	3.531441	0.027110 (0.5409965)	0.9976958
	MPS	842	1155.67	3.545935	0.023434 (0.7442403)	0.9989168
	RME	840	115.33	3.541232	0.023630 (0.7362441)	0.9987985
2016	MLE	548	1162.50	4.286436	0.022214 (0.9496787)	0.9985115
	MOM	547	1161.67	4.248240	0.020142 (0.9795097)	0.9988841
	MPS	537	1157.50	4.194894	0.021786 (0.9607594)	0.9986054
	RME	545	1159.83	4.304627	0.022831 (0.9388524)	0.9987142
2019	MLE	409	1182.67	5.391208	0.017413 (0.9996598)	0.9994427
	MOM	399	1176.17	5.422609	0.017191 (0.9997913)	0.9994568
	MPS	390	1170.89	5.447250	0.061421 (0.9999379)	0.9994289
	RME	390	1170.89	5.461424	0.015488 (0.9999846)	0.9994238
2022	MLE	351	1316.67	5.764246	0.017893 (0.9998723)	0.9992548
	MOM	351	1316.67	5.760188	0.017963 (0.9998615)	0.9992486
	MPS	320	1295.83	5.757015	0.019474 (0.9997232)	0.9991936
	RME	351	1316.67	5.766338	0.017857 (0.9998776)	0.9992577

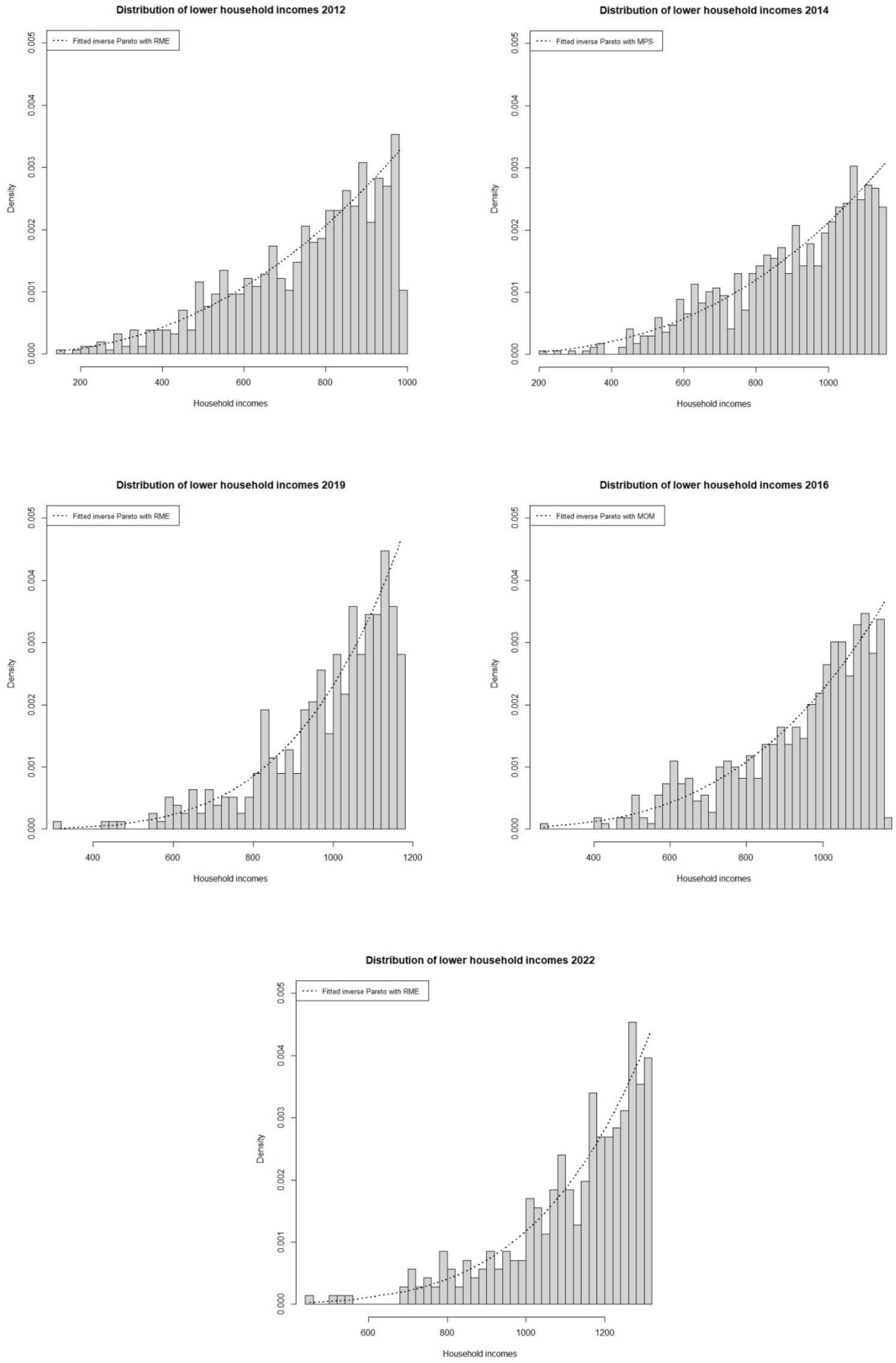

FIGURE 5. Best-fitted inverse Pareto distribution of Malaysian lower-class household income data for the years 2012, 2014, 2016, 2019, and 2022

TABLE 4. The estimated Gini index for Malaysian lower-income group based on the estimated inverse Pareto model

Year	$\hat{a}$	Gini
2012	3.256431	0.1331051
2014	3.545935	0.1235808
2016	4.248240	0.1053022
2019	5.461424	0.0838726
2022	5.766338	0.0797914

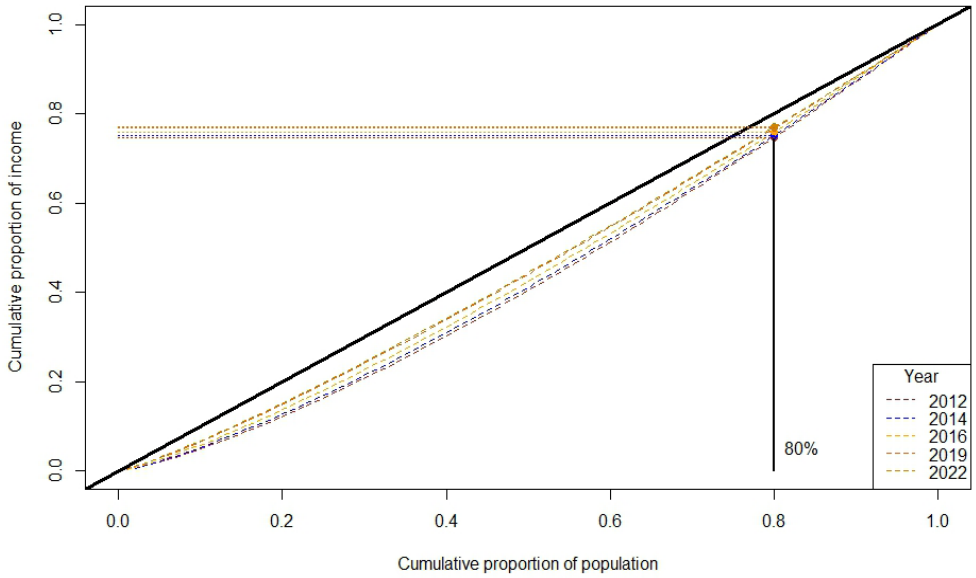

FIGURE 6. The fitted Lorenz curve based on inverse Pareto model for lower-class earners in Malaysia

hard-core poverty group, whereas the top 20% accounts for only about 23.04% to 25.30% of total household income. These findings closely resemble the 80/20 Pareto principle in the context of this low-income group and suggest a relatively equitable income distribution, where the majority of households share a substantial portion of total income.

CONCLUSION

In this study, we have proposed a new robust estimator method for the inverse Pareto distribution, the RME. We have shown that our proposed estimator is efficient due to its high value of ARE. We also demonstrated that RME is robust with acceptable breakdown value and constrained IF. The Monte Carlo studies suggest that RME works well for no-outlier cases and outliers-contaminated cases, surpassing the performance of MLE, MOM, and MPS especially for large sample sizes.

We then proceed with the data application towards the lower income of Malaysian household income datasets for 2012, 2014, 2016, 2019, and 2022. The inverse Pareto threshold parameter is determined by choosing the data point within the 1% to 50% percentile that minimizes the KS statistics. We discovered that RME is the best fit for the 2012, 2014, and 2022 datasets with the highest KS $p-$value near 1. Notably, the estimated inverse Pareto best fit can capture the whole hard-core poverty group for all five selected years. This analysis reaffirmed that the inverse Pareto model is a good fit for the lower-tail data of Malaysian household income, suggesting its effectiveness in explaining the income dynamics of the low earners in the country. Future research can work on extending the present work by exploring alternative models and broader datasets, while relevant authorities and agencies are encouraged to use the study's findings to inform policy development, methodological guidelines, and evidence-based decision-making.

ACKNOWLEDGMENTS

We gratefully acknowledged the financial support from the 'Ministry of Higher Education Malaysia for the Fundamental Research Grant Scheme with Project Code: FRGS/1/2023/STG06/UM/02/11'. We would also would like to express our sincere gratitude to the DOSM for providing the household income datasets that were instrumental in conducting this study.

REFERENCES

Abd Raof, A.S., Haron, M.A., Safari, M.A.M. & Siri, Z. 2022. Modeling the incomes of the upper-class group in Malaysia using new Pareto-type distribution. *Sains Malaysiana* 51(10): 3437-3448.

Abu, A., Hamdan, R. & Sani, N. 2020. Ensemble learning for multidimensional poverty classification. *Sains Malaysiana* 49(2): 447-459.

Alfons, A., Templ, M. & Filzmoser, P. 2013. Robust estimation of economic indicators from survey samples based on Pareto tail modelling. *Journal of the Royal Statistical Society. Series C: Applied Statistics* 62(2): 271-286.

Brzezinski, M. 2015. Power laws in citation distributions: Evidence from scopus. *Sciento-Metrics* 103: 213-228.

Brzezinski, M. 2014. Do wealth distributions follow power laws? evidence from 'rich lists'. *Physica A: Statistical Mechanics and its Applications* 406: 155-162.

Clauset, A., Shalizi, C.R. & Newman, M.E. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4): 661-703.

Cowell, F.A. & Flachaire, E. 2007. Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics* 141(2): 1044-1072.

DOSM: OpenDOSM. 2024. https://open.dosm.gov.my/data-catalogue/hh poverty Accessed on November 4, 2024.

Filimonov, V. & Sornette, D. 2015. Power law scaling and "dragon-kings" in distributions of intraday financial drawdowns. *Chaos, Solitons & Fractals* 74: 27-45.

Hampel, F.R., Ronchetti, E.M. & Rousseeuw, P.J. 1986. *Robust Statistics*. New Jersey: John Wiley & Sons, Inc.

Klaus, A., Yu, S. & Plenz, D. 2011. Statistical analyses support power law distributions found in neuronal avalanches. *PLoS ONE* 6(5): 19779.

Kleiber, C. 2003. *Statistical Size Distributions in Economics and Actuarial Sciences*. New Jersey: John Wiley & Sons, Inc.

Laherrere, J. & Sornette, D. 1988. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems* 2: 525-539.

Luckstead, J., Devadoss, S. & Danforth, D. 2017. The size distributions of all Indian cities. *Physica A: Statistical Mechanics and its Applications* 474: 237-249.

Majid, M.H.A. & Ibrahim, K. 2021. Composite pareto distributions for modelling household income distribution in Malaysia. *Sains Malaysiana* 50(7): 2047-2058.

Maronna, R.A., Martin, R.D., Yohai, V.J. & Salibi´an-Barrera, M. 2019. *Robust Statistics: Theory and Methods (with R).* 2nd ed. New Jersey: John Wiley & Sons. p. 380.

Masseran, N., Yee, L.H., Safari, M.A.M. & Ibrahim, K. 2019. Power law behavior and tail modeling on low income distribution. *Mathematics and Statistics* 7(3): 70-77.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing,* Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org

Razak, F.A. & Shahabuddin, F.A. 2018. Malaysian household income distribution: A fractal point of view. *Sains Malaysiana* 47(9): 2187-2194.

Reed, W.J. 2003. The pareto law of incomes - An explanation and an extension. *Physica A: Statistical Mechanics and its Applications* 319: 469-486.

Safari, M.A.M. & Masseran, N. 2024. Robust estimation techniques for the tail index of the new pareto-type distribution. *Empirical Economics* 66(3): 1161-1189.

Safari, M.A.M., Masseran, N. & Haron, M.A. 2024. Examining tail index estimators in newpareto distribution: Monte Carlo simulations and income data applications. *Sains Malaysiana* 53(2): 461-476.

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018. Optimal threshold for pareto tail modelling in the presence of outliers. *Physica A: Statistical Mechanics and its Applications* 509: 169-180.

Safari, M.A.M., Masseran, N., Ibrahim, K. & Hussain, S.I. 2021. Measuring income inequality: A robust semi-parametric approach. *Physica A: Statistical Mechanics and Its Applications* 562: 125359.

Safari, M.A.M., Masseran, N., Ibrahim, K. & AL-Dhurafi, N.A. 2020. The power-law distribution for the income of poor households. *Physica A: Statistical Mechanics and its Applications* 557: 124893.

Safari, M.A.M., Masseran, N., Ibrahim, K. & Hussain, S.I. 2019. A robust and efficient estimator for the tail index of inverse pareto distribution. *Physica A: Statistical Mechanics and its Applications* 517: 431-439.

Siegel, A.F. 1982. Robust regression using repeated medians. *Biometrika* 69(1): 242-244.

*Corresponding author; email: azmiharon@um.edu.my

APPENDIX A

Here are the details on the alternative estimator methods used in this study.

MLE

The MLE for the inverse Pareto distribution is determined by optimizing the log-likelihood function concerning the shape parameter, α. Consequently, the MLE is given as the following equation.

$$\alpha_{MLE} = \frac{n}{n \log x_0 - \sum_{i=1}^{n} \log x_i}$$

MOM

The MOM estimator is derived by equating the empirical moment from the data and inverse Pareto theoretical moment. It can be defined as follows.

$$\alpha_{MOM} = \frac{\frac{1}{N}\sum_{i=1}^{n} x_i}{x_0 - \frac{1}{N}\sum_{i=1}^{n} x_i}$$

MPS

The approach of maximum product spacings highlights the differences in CDF values for adjacent data points, focusing on the maximization of their geometric mean which can be written as the follows.

$$S = \left[\prod_{i=1}^{n+1} D_i\right]^{\frac{1}{n+1}}$$

where $D_i = F(x_i;\alpha,x_0) - F(x_{i-1};\alpha,x_0)$, for x_i is an ordered sample observation for i = 1, 2, ... , n. Note that, $F(x_0;\alpha,x_0)=0$, $F(x_{n+1};\alpha,x_0)=1$. Similarly, the MPS also can be obtained by maximising the function

$$\log(S) = \frac{1}{1+n}\sum_{i=1}^{n+1} \log(D_i)$$

The reasoning behind maximizing S or log (S) is that the upper limit is constrained by the requirement $\Sigma D_i = 1$. This maximum value is reached only when all D_i values are identical.