Sains Malaysiana 51(10)(2022): 3437-3448
http://doi.org/10.17576/jsm-2022-5110-26

Modeling the Incomes of the Upper-Class Group in Malaysia using New Pareto-Type Distribution

(Pemodelan Pendapatan Isi Rumah Kelas Atas di Malaysia menggunakan Taburan Pareto Jenis Baharu)

ANIS SYAZWANI ABD RAOF[1], MOHD AZMI HARON[1,*], MUHAMMAD ASLAM MOHD SAFARI[2] & ZAILAN SIRI[1]

[1]*Institute of Mathematical Sciences, Faculty of Science, Universiti Malaya, 50603 Kuala Lumpur, Federal Territory, Malaysia*
[2]*Department of Mathematics and Statistics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor Darul Ehsan, Malaysia*

Received: 11 September 2021/Accepted: 23 May 2022

ABSTRACT

The new Pareto-type distribution has been previously introduced as an alternative to the conventional Pareto distribution in modeling income distribution. It is claimed to provide better flexibility for mathematical simplicity of probability functions and has a more straightforward mathematical form. In this study, the new Pareto-type distribution is used to model the income of the Malaysian upper-class group. The threshold is determined using the fixed proportion technique and the maximum likelihood estimator method is used to estimate the shape parameter. Then, the goodness-of-fit of the fitted new Pareto model is measured using the coefficient of determination, R^2 and Kolmogorov–Smirnov statistics. We also measure the income inequality among the Malaysian top income earners using the Lorenz curve, Gini and Theil indices based on the fitted new Pareto model. Finally, the new Pareto distribution is compared to alternative distributions to analyze which model can give the best fit for the data. Our analysis shows that the Pareto type-1 and the new Pareto models are well fitted to the top income data for all years considered. However, the new Pareto model provides better flexibility which covering more incomes in the upper tail of the distribution than the Pareto type-1 model.

Keywords: Gini index; income inequality; Lorenz curve; Pareto model; Theil index

ABSTRAK

Taburan Pareto baharu telah diperkenalkan sebagai alternatif kepada taburan Pareto konvensional dalam permodelan taburan pendapatan. Kelebihan menggunakan taburan Pareto baharu dapat dilihat dari segi bentuk fungsinya yang mudah dan lebih fleksibel dalam memodelkan data. Dalam kajian ini, taburan Pareto baharu digunakan untuk memodelkan data pendapatan isi rumah kelas atas di Malaysia. Anggaran nilai ambang dan nilai parameter bentuk bagi taburan Pareto baharu, masing-masing ditentukan menggunakan teknik pernisbahan tetap dan kaedah anggaran kebolehjadian maksimum. Seterusnya, kebagusan penyuaian taburan Pareto baharu terhadap data pendapatan kelas atas dinilai menggunakan pekali penentuan, R^2 dan statistik Kolmogorov–Smirnov. Kajian ini juga mengukur ketaksamaan pendapatan antara golongan atas menggunakan keluk Lorenz, indeks Gini dan indeks Theil berdasarkan taburan Pareto baharu. Akhir sekali, perbandingan antara taburan Pareto baharu dan pelbagai taburan lain dilakukan bagi mengenal pasti taburan yang mampu memberikan penyuaian terbaik dalam menerangkan data pendapatan kelas atas. Hasil kajian mendapati kedua-dua taburan Pareto baharu dan Pareto jenis-1 mampu menerangkan data pendapatan kelas atas. Namun, taburan Pareto baharu memberikan kefleksibelan yang lebih baik dan taburan ini mampu untuk menerangkan data pendapatan yang lebih banyak berbanding taburan Pareto jenis-1.

Kata kunci: Indeks Gini; indeks Theil; keluk Lorenz; ketaksamaan pendapatan; model Pareto

INTRODUCTION

Income distribution within a country remains as one of the major debated issue in previous years (Kulub et al. 2020; Kusnic & Vanzo 1980; Shakil et al. 2015). It is described as unimodal with a heavy right tailed distribution (Fellman 2018). Therefore, different skewed

models, particularly the lognormal and Pareto models are considered as a suitable descriptive models of income distribution (Charpentier & Flachaire 2019). The Pareto model is used to model the upper tail of incomes, whereas the lognormal, gamma, and exponential models model the lower part of the incomes (Safari et al. 2020). However, the income distribution can follow a wealth distribution classified by a two-part function. The low to medium range is fitted to a different but interrelated function with the other function used to fit the upper part of the population (Chami Figueira et al. 2011). For instance, the distribution of income can be expressed as a combination of the Gompertz curve which focusing on the majority population (99%) and the Pareto power-law representing the most affluent minority (1%) (Chami Figueira et al. 2011).

Apart from the argument in assessing the lower tail of income distribution, most researchers agreed that the upper tail of income distribution captures the Pareto behavior following power law (Banerjee et al. 2006; Chami Figueira et al. 2011; Moura & Ribeiro 2009). The power law can be described by $Cx^{-\alpha}$ for some positive values C and α, where it often applies for the income data starting from $x = x_0$ and going up to the maximum value (Abdul Majid & Ibrahim 2021; Clauset et al. 2009). Oancea et al. (2018) examines distributional analysis of the capital income in Romania and the findings obtained showed that the capital income is well described by a Pareto type-1 distribution in its upper tail. Masseran et al. (2020) also conducted a study to analyze the distribution of Malaysian annual gross income among upper class group using Pareto model and evaluate the income inequality based on Gini index. However, subsequent studies found that Pareto type-1 model can only give a significant fit to a small proportions of top income for about 1-3% of the population (Banerjee et al. 2006).

Hence, Bourguignon et al. (2016) had introduced a new Pareto model which can provide a better fit for data sampling than the Pareto type-1 in describing income data. Sarabia et al. (2019) highlighted the relationships between the new Pareto model and other Pareto distributions. For example, the Pareto type-1 model is more practical than the new Pareto model for proving the power-law behavior. However, the new Pareto model is significantly better for evaluating a more comprehensive range of data. Additionally, new Pareto model can be used to represent an upside-down bathtub or a decreasing hazard rate function depending on the values of its parameter whereas the Pareto type-1 model can only represent a decreasing hazard rates. In practice, Pareto modeling describes the upper tail of distributions in economic inequality, economic losses, and insurance. For example, Clementi and Gallegati (2005) analyzed the income inequality for Italian personal incomes using the Gini coefficient and concluded that the increasing level of inequality relates to a sharp decline of the Pareto index.

Income inequality refers to the uneven distribution of income within a population in a country. It is the income gap in a society where a group of the population earns higher incomes than others (Islam et al. 2017). The higher the income gap, the higher the income inequality among society. Income inequality damages economic development, leading to social instability and conflicts (Law & Tan 2009). Therefore, efforts to reduce income inequality are championed by the government in the country. For instance, increasing the minimum wages and improving job qualities among workers, and adjusting educational systems to fit the employment standards (Islam et al. 2017). The level of income inequality can be measured using different methods including the Lorenz curve and Gini coefficient, the most used indicators of income inequality (Csörgo et al. 1998; Giorgi & Crescenzi 2001; Pundir et al. 2005). Other than that, many researchers have used the Zenga curve, Theil index, and Atkinson index to measure a country's income inequality (Chakravarty & Sarkar 2020; Hernández-Ramírez et al. 2021; Razak & Shahabuddin 2018).

This study applies the new Pareto distribution to explain the income of the upper-class group using the Malaysian household incomes dataset from 2012-2019. The threshold is determined using the fixed proportion technique for top 40%, 35%, 30%, 25%, 20%, 15%, 10%, and 5% of the sample datasets yearly. Then, the maximum likelihood estimator (MLE) method is used to estimate the shape parameter. Next, the income inequality among Malaysian top incomes is measured using the Lorenz curve, Gini and Theil indices based on the fitted new Pareto model. Finally, the new Pareto model is compared with alternative models such as the Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential models to analyze which model best fits the household income data.

DATA COLLECTION

The datasets used consist of monthly gross income data of Malaysian households obtained from the Department of Statistics Malaysia (DOSM) for 2012, 2014, 2016, and 2019. The income data are provided from the official

survey known as the Household Income Survey and Basic Amenities (HIS & BA) survey, first conducted in 1973 and taken twice every five years to collect the data on income, poverty, and basic amenities among citizens. Hence, the data and statistics obtained are used to evaluate policies and strategize the economic development plans of Malaysia.

In Malaysia, household incomes are categorized into three groups; B40, M40, and T20. The group names represent the percentage of shared income of the country's population, which are Bottom 40%, Middle 40%, and Top 20%. HIS survey in 2019 reported that the monthly incomes for B40, M40, and T20 ranged from RM4,849 and below, RM4,850 to RM10,959, and RM10,959 and above, respectively (Department of Statistics Malaysia 2020). Thus, T20 is the upper-class income earners in Malaysia. However, the income group definitions are not fixed and varies yearly. Besides, the scope of this study only focuses on the gross household incomes of the top income earners in Malaysia.

NEW PARETO MODEL

This study employs the new Pareto model by Bourguignon et al. (2016) with the assumption that the random variable X follows the new Pareto distribution which is denoted as as $X \sim NP(\alpha, x_0)$. The cumulative distribution function (CDF), probability density function (PDF), and quantile function of the new Pareto distribution are given as follows:

The CDF is,

$$F(x;\alpha,x_0) = 1 - \frac{2{x_0}^{\alpha}}{x^{\alpha}+{x_0}^{\alpha}}, x \geq x_0 \quad (1)$$

where $\alpha > 0$ is the shape parameter and $x_0 > 0$ is the scale parameter or the threshold of the new Pareto model. The PDF is,

$$f(x;\alpha,x_0) = \frac{2\alpha{x_0}^{\alpha}x^{\alpha-1}}{(x^{\alpha}+{x_0}^{\alpha})^2}, x \geq x_0 \quad (2)$$

The quantile function is,

$$Q(u) = F^{-1}(u) = x_0\left(\frac{1+u}{1-u}\right)^{\frac{1}{\alpha}}, 0 < u < 1 \quad (3)$$

Sarabia et al. (2019) established a more straightforward expression for the moments of the new Pareto model in terms of the incomplete beta function. Thus, the r^{th} moment of X is,

$$E(X^r) = 2x_0 B\left(\frac{1}{2}; 1-\frac{r}{\alpha}, 1+\frac{r}{\alpha}\right), \alpha > r \quad (4)$$

where the incomplete beta function denoted by $B(x;p,q)$ is given by,

$$B(x;p,q) = \int_0^x t^{p-1}(1-t)^{q-1}dt, p \text{ and } q > 0,\ 0 < x < 1 \quad (5)$$

Therefore, the mean of new Pareto distribution is,

$$\mu = E(X) = 2x_0 B\left(\frac{1}{2}; 1-\frac{1}{\alpha}, 1+\frac{1}{\alpha}\right), \alpha > 1 \quad (6)$$

THE ESTIMATION OF THRESHOLD AND SHAPE PARAMETER OF NEW PARETO DISTRIBUTION

The threshold, x_0 of the new Pareto model can be determined using various techniques. The optimal threshold can be estimated by minimizing the goodness-of-fit of the empirical distribution function statistics. For example, the KS statistics, Kuiper, Anderson-Darling, and Watson (Brzezinski 2014; Safari et al. 2018b). The other techniques include graphical techniques such as the Zipf plot, Pareto quantile plot, and means excess function plot (Safari et al. 2018b). From these graphical techniques, the observations form a straight line on a log-log plot if the Pareto distribution is well fitted to the data. Nevertheless, these techniques can generate a subjective and not optimal threshold value because of the noise or fluctuation sensitivity in the data (Brzezinski 2014; Safari et al. 2018b).

In this study, the threshold of the new Pareto distribution is determined using a fixed proportion technique where each dataset is divided into six different proportions of income, P_{tail} such as top 40%, 35%, 30%, 25%, 20%, 15%, 10%, and 5%. Then, the value of threshold for each P_{tail} is represented by its quantile. This technique is preferable to find out which proportion can give the best fitted new Pareto distribution of top income data since new Pareto model is claimed to be better than Pareto type-1 in analyzing a wider range of data (Sarabia et al. 2019).

The shape parameter estimate, $\hat{\alpha}$ of the new Pareto distribution are obtained using MLE technique. Additionally, a numerical method which is the bisection method is applied to solve the derivative of the log likelihood function (Burden & Faires 2011). Given a known $\hat{x}_0$, the MLE $\hat{\alpha}$ of α is obtained as the solution to the following equation (Bourguignon et al. 2016),

$$\frac{\partial \ell(\alpha,x_0)}{\partial \alpha} = -2\sum_{i=2}^{n} \frac{\left(\frac{x_0}{x_i}\right)^{\alpha}\log\left(\frac{x_0}{x_i}\right)}{1+\left(\frac{x_0}{x_i}\right)^{\alpha}} + \sum_{i=2}^{n}\log\left(\frac{x_0}{x_i}\right) + \frac{n}{\alpha} = 0 \quad (7)$$

ASSESSMENT GOODNESS-OF-FIT

The goodness-of-fit of the fitted models are measured

using the KS test and R^2. KS test is performed by computing the KS statistics, D which is defined as,

$$D = \max_{x>x_0} |F_n(x) - F_{x_0;\alpha}(x)| \quad (8)$$

where $F(x; n)$ is the empirical cumulative distribution function and $F(x; \alpha, x_0)$ is the CDF of the new Pareto model in Equation (1). The KS test hypotheses are given by,

H_0: The upper tail data of household incomes follow the new Pareto distribution

H_1: The upper tail data of household incomes do not follow the new Pareto distribution

Additionally, the R^2 analysis is performed to support the KS test for determining the best-fitted models by measuring the correlation between observed data and fitted CDF of distribution (Safari et al. 2018a). Therefore, R^2 close to 1 implies that the new Pareto distribution adequately explains the upper tail of the household income data. However, if R^2 is close to 0, the new Pareto distribution gives a poor explanation for the top income data. Following the study conducted by Safari et al. (2020), R^2 is given by,

$$R^2 = \frac{\sum_{i=1}^{n}[\hat{F}(x_i;\alpha,x_0)-\bar{F}(x;\alpha,x_0)]^2}{\sum_{i=1}^{n}[\hat{F}(x_i;\alpha,x_0)-\bar{F}(x;\alpha,x_0)]^2+\sum_{i=1}^{n}[F(x_i;n)-\hat{F}(x_i;\alpha,x_0)]^2} \quad (9)$$

where $\hat{F}(x_i; \alpha, x_0)$ is the estimated CDF of the new Pareto distribution for i^{th} household income data where $x_i > x_0$, $\bar{F}(x; \alpha, x_0)$ is the average for $\hat{F}(x_i; \alpha, x_0)$ and $F(x_i; n)$ is the empirical cumulative distribution function for i^{th} household income data where $x_i > x_0$.

INCOME INEQUALITY MEASURES BASED ON NEW PARETO MODEL

This section presents the techniques used to measure the income inequality which includes Lorenz curve, Gini index, and Theil index. The Lorenz curve is a graphical diagram of income inequality (Safari et al. 2018a). It is the most prominent inequality curve in previous literature. Arcagni and Porro (2014) stated that the Lorenz curve is always characterized as convex with a straight diagonal line whose slope is 1. The straight diagonal line is called the 'line of equality', representing an equally distributed income. The higher the income inequality, the more the Lorenz curve shifts away from the 'line of equality' (Safari et al. 2021). Sarabia et al. (2019) provides the Lorenz curve for the household incomes of the new Pareto distribution as,

$$L(u;\alpha) = 1 - \frac{B\left(\frac{1-u}{2};1-\frac{1}{\alpha},1+\frac{1}{\alpha}\right)}{B\left(\frac{1}{2};1-\frac{1}{\alpha},1+\frac{1}{\alpha}\right)}, 0 \le u \le 1 \quad (10)$$

where $\alpha > 1$ and $B(x; p, q)$ is the incomplete beta function in Equation (5).

After the Lorenz curve has been plotted, the Gini coefficient is obtained by computing the area below the Lorenz curve (Moura & Ribeiro 2009). From Figure 1, the area of A is between the 'line of equality' and the Lorenz curve. The value of the Gini index is double the area of A which is computed using the formula $Gini = 1\text{-}2\int_0^1 L(p)$ dp (Safari et al. 2018a). This can be simply expressed as, $Gini = 2A = 1\text{-}2B$. When the Gini coefficient is 0, the allocation of total income is perfectly equal within the population. Thus, $Gini = 0$ indicates perfect equality, whereas $Gini = 1$ indicates perfect inequality (Safari et al. 2020). The Gini index corresponding to the new Pareto model is,

$$G = 1 - \frac{4B(\frac{1}{2};2-\frac{1}{\alpha},1+\frac{1}{\alpha})}{B(\frac{1}{2};1-\frac{1}{\alpha},1+\frac{1}{\alpha})}, \alpha > 1 \quad (11)$$

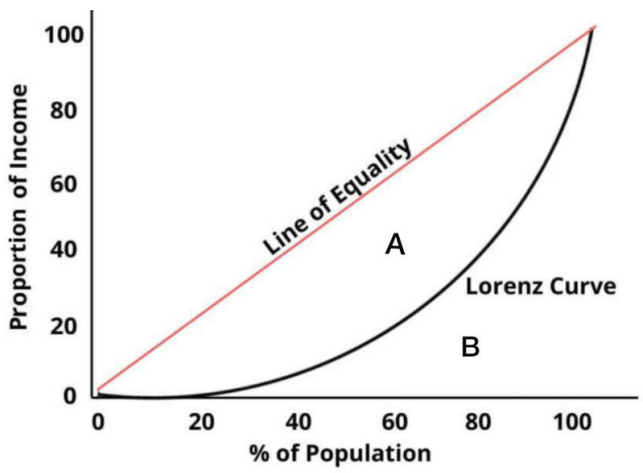

FIGURE 1. An example of a Lorenz curve

Another measure of income inequality is the generalized entropy index denoted by $GE(\varepsilon)$ where ε represents a parameter of the weight assigned to distances between income in different parts of the income distribution. The higher the value of ε, the greater the index sensitivity toward large income data (Safari et al. 2018a). Cowell (2003) stated that the Theil index is a special case of the generalized entropy index when $\varepsilon = 1$. A smaller Theil coefficient implies that income inequality is lower and the incomes become less unequal (Bahari et al. 2015). Based on Safari et al. (2020), the Theil index of the new Pareto distribution is derived as,

$$T = \int_{x_0}^{\infty} \frac{x}{E(X)} \log\left(\frac{x}{E(X)}\right) f(x)\, dx = \frac{v}{E(X)} - \log\,[E(X)] \quad (12)$$

where

$$v = \int_{x_0}^{\infty} (x \log x)\left(\frac{2\alpha x_0{}^{\alpha} x^{\alpha-1}}{(x^{\alpha}+x_0{}^{\alpha})^2}\right) dx,\ x_0 > 0. \quad (13)$$

$$v = \frac{x_0\left(2\alpha^2 \log(x_0)+2\alpha\left(H_{-\frac{1}{2\alpha}} - H_{-\frac{\alpha+1}{2\alpha}}\right)(\log(x_0)+1)+\psi^{(1)}\left(\frac{\alpha-1}{2\alpha}\right)-\psi^{(1)}\left(1-\frac{1}{2\alpha}\right)\right)}{2\alpha^2} \quad (14)$$

where the harmonic number denoted by H_n is,

$$H_n = \int_0^1 \frac{1-x^n}{1-x}\, dx \quad (15)$$

and the polygamma function denoted by $\psi^{(m)}(z)$ is,

$$\psi^{(m)}(z) = (-1)^{m+1} m! \sum_{k=0}^{\infty} \frac{1}{(z+k)^{m+1}} \quad (16)$$

In order to make a valid comparison, a fixed proportion of income which is $P_{tail} = 0.30$ is used to evaluate the income inequality among top income earners considering that the given proportion is adequate in explaining top income data for all years considered.

DESCRIPTIVE STATISTICS

This study performed all methods and data analysis using the R software. Table 1 shows the descriptive statistics of Malaysian household incomes for each year. From the table, the mean and median household incomes show a yearly increasing trend. The variances for the years considered are high, showing the widespread data on the mean. Besides, the yearly coefficients of the skewness of the income data are all positive which means the distribution of Malaysian household incomes is right-skewed instead of following a normal distribution.

TABLE 1. The descriptive statistics of Malaysian household incomes for the years 2012, 2014, 2016, and 2019

Year	Mean	Median	Min	Max	Variance	Coefficient of skewness
2012	4480.00	3221.00	150.00	105958.00	23448300	6.1923
2014	5746.80	4251.50	212.50	186892.00	33825399	6.9243
2016	6298.20	4701.10	269.60	368585.00	43319581	13.9239
2019	6979.50	5142.80	318.20	882163.80	89727701	43.8460

FITTING THE NEW PARETO DISTRIBUTION TO THE UPPER TAIL OF HOUSEHOLD INCOME DATA

Table 2 presents the yearly threshold values and shape parameter estimates of the best-fitted new Pareto distributions for Malaysian household income data. The threshold values of the fitted new Pareto distribution are calculated using the proportion of incomes, P_{tail} of a dataset. For each dataset, six proportions (40%, 35%, 30%, 25%, 20%, 15%, 10%, and 5%) of the upper tail data are analyzed to find the best-fitted new Pareto distribution of the income data. Hence, we observed that the threshold in 2014 ($\hat{x}_0 = 5070.30$) was less than that in 2016 ($\hat{x}_0 = 8740.75$) due to the well-fitness of new Pareto distribution at 40% of upper tail income data in 2014 rather than 20% in 2016. The difference in P_{tail} can be foreseen since the income percentage of the best fitted new Pareto model may fluctuate from year to year. Thus, regardless of the yearly differences in P_{tail}, the new Pareto model is suitable to model the middle to upper parts of the Malaysian income distribution.

Next, we determine the shape parameter value for the fitted new Pareto distribution using MLE. The value of shape parameter indicates the heaviness of a tailed distribution (Safari et al. 2018b). If the value of the shape parameter is lower, the tailed distribution is heavier. It also reflects the income inequality measure. The smaller the value of the shape parameter, the higher is the income inequality for income data. Therefore, the lowest shape parameter for the income distribution in 2014 indicates the heaviest upper tail and the highest income inequality among the top earners. However, the heaviness of tailed distribution and income inequality level may vary due to the proportions of top incomes since the results show different proportions that are best fitted to the new Pareto model for each year.

The KS test and R^2 values assess the adequacy of the new Pareto model in describing the top income data. Table 2 shows that all p-values are higher than the significance level of 0.05 and the R^2 values are greater than 0.99 for each year. These results imply that the new Pareto distribution is significantly describes the income data.

TABLE 2. The estimated threshold levels ($\hat{x}_0$), the estimated shape parameters of the new Pareto model ($\hat{\alpha}$), p-value of KS test, values, the proportion of top household incomes (), the number of top household income data (), KS statistics (D) for the years 2012, 2014, 2016, and 2019

Year	$\hat{x}_0$	$\hat{\alpha}$	p-value (KS test)	R^2	P_{tail}	n_{tail} ($x \geq \hat{x}_0$)	D
2012	5443.79	2.8034	0.8323	0.9997*	0.25	3308	0.0108
2014	5070.30	2.5757	0.9092	0.9999*	0.40	9785	0.0057
2016	8740.75	3.3008	0.1529	0.9990*	0.20	4708	0.0165
2019	7663.92	2.9323	0.6309	0.9998*	0.30	7462	0.0087

*p-value

MEASURING THE INCOME INEQUALITY BASED ON NEW PARETO MODEL

Figure 3 shows the fitted Lorenz curve based on the new Pareto distribution for the top income data for each year. From the figure, there is an overlapping between the fitted Lorenz curve of top income data for 2016 and 2019. Both of them are seen to be the closest curves to the 'line of equality'. Apart from that, the figure shows that the fitted Lorenz curve for 2012 is the farthest curve from the equality line. These indicate that the top income data in 2019 have lower income inequality than the top income data in 2012.

Table 3 also shows a summary of the estimated Gini and Theil coefficients of the Malaysian upper-class for each year. The income inequality among top household incomes had reduced as the estimated Gini coefficient decreased from 0.2760 in 2012 to 0.2452 in 2016. These results are attributed to the increase in the percentage of households, accounting for total top household incomes from 72.40% in 2012 to 75.48% in 2016. Also, the percentage of households who gained nothing decreases from 27.60% in 2012 to 24.52% in 2016. Besides, the estimated Theil coefficient also gives almost similar result to the estimated Gini coefficient with a decreasing trend from 2012 to 2016 and a slight increase from 2016 to 2019. From the table, it can be seen that the estimated Theil coefficient decreases from 0.1683 in 2012 to 0.1291 in 2016, indicating a reduction of income inequality from 2012 to 2016 among the Malaysian upper-class. However, the income inequality had risen modestly from 2016 to 2019.

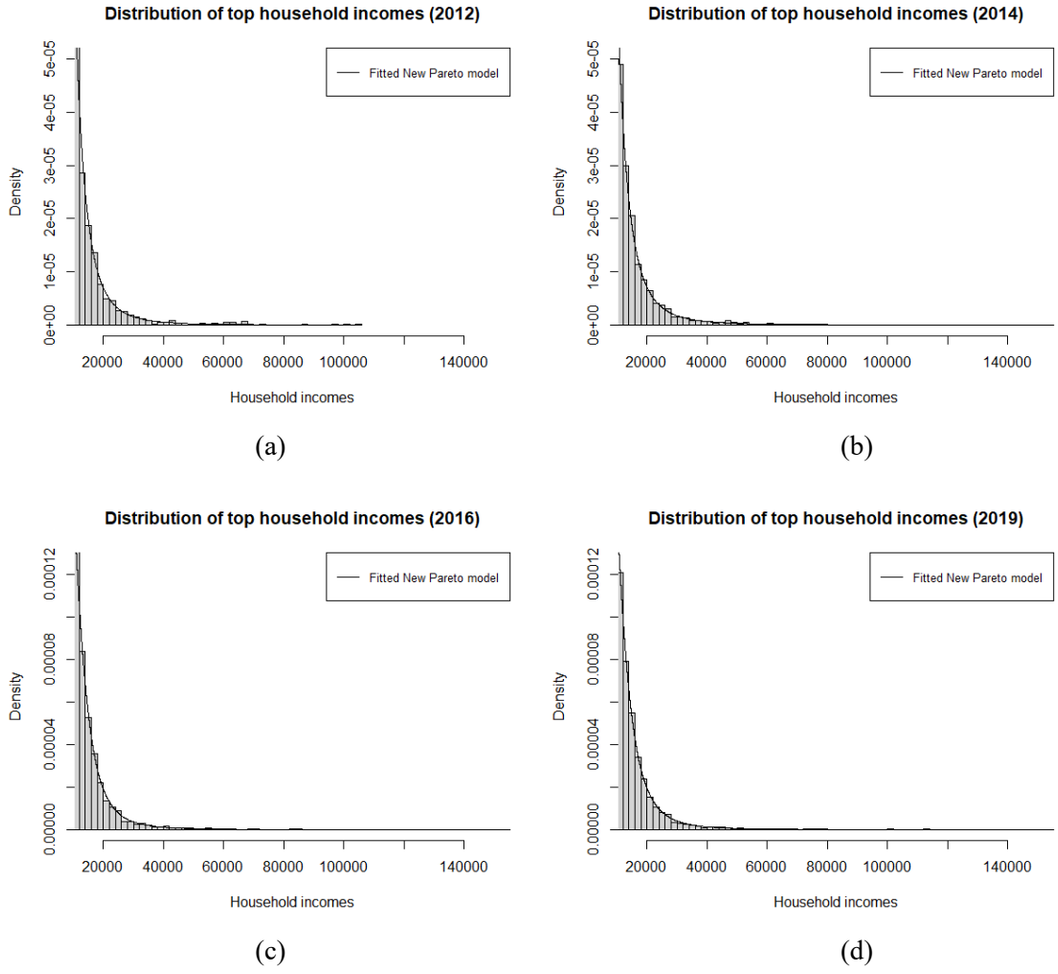

FIGURE 2. Best-fitted new Pareto distribution of the top household income data for the years (a) 2012, (b) 2014, (c) 2016, and (d) 2019

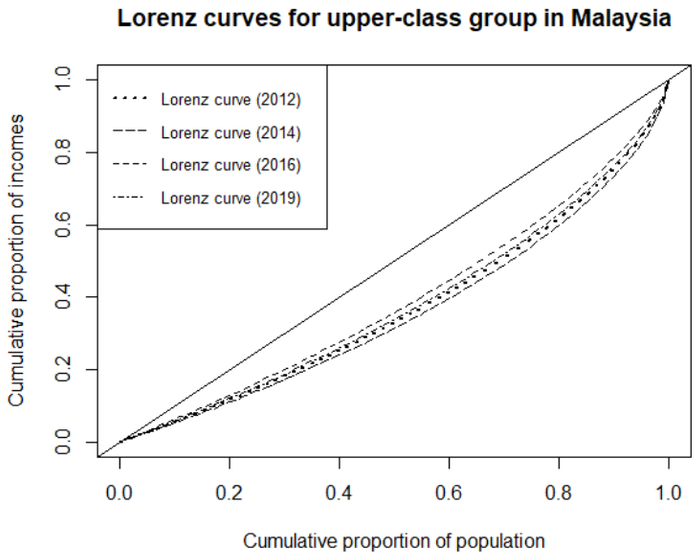

FIGURE 3. Fitted Lorenz curves of Malaysian household incomes based on new Pareto model for the years 2012, 2014, 2016, and 2019

TABLE 3. Income shares for the bottom and top of the Malaysian upper-class based on the fitted Lorenz curve and the estimated Gini and Theil coefficients for the years 2012, 2014, 2016, and 2019

Year	P_{tail}	Bottom 80%	Top 20%	Gini	Theil
2012	0.25	62.00%	38.00%	0.2587	0.1455
2014	0.40	59.89%	40.11%	0.2859	0.1825
2016	0.20	65.39%	34.61%	0.2141	0.0958
2019	0.30	63.02%	36.98%	0.2455	0.1294

COMPARISONS BETWEEN NEW PARETO AND OTHER DISTRIBUTIONS FOR DESCRIBING TOP INCOMES

This study compares the new Pareto distribution with the alternative distributions, particularly the Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential distributions (Banerjee et al. 2006; Clementi & Gallegati 2005; Raqab et al. 2019). Table 4 shows the PDF and CDF of the alternative distributions. This section finds a distribution that better explains the income distribution in Malaysia, especially for the upper tail data lying above the threshold value. For fair comparisons, this study applies the same technique in estimating the parameter for each model using the MLE technique. The MLE functions for the parameters of each distribution are given in Table 5. The scale parameter for the Pareto type-1 distribution is similar to that of the new Pareto distribution and the shifting parameters, for all considered models are assumed to be equal to the scale parameter of the new Pareto distribution for each different proportion (40%, 35%, 30%, 25%, 20%, 15%, 10%, and 5%). Note that the MLE for the shifted stretched exponential distribution is not written in a closed-form expression. Thus, the parameter estimates and are evaluated using the optimization function in R software. Figure 4 shows the graphs of the best-fitted graph of Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential models for the top incomes from 2012 to 2019. From Table 6, we conclude that the shifted exponential distribution gives the worst and inadequate fit for explaining the top income data because its *p*-value (KS test) is less than the significance level, 0.05. Then, followed by the shifted lognormal distribution with poor and inadequate fitting of data for each year. Next, the result shows that the shifted stretched exponential distribution is adequate for describing top income data in 2012 and 2014 (both = 0.05) but inadequate for 2016 and 2019. The Pareto type-1 distribution is adequate for modeling the top income data for each year and gives a good fit from = 0.05 to= 0.10. Finally, the new Pareto distribution is also adequate for explaining the income data for each year by providing a good fit from = 0.20 to = 0.30, except in 2014, where it also fits for = 0.35 to = 0.40. However, it can be concluded that the new Pareto distribution gives best explanation for larger of top income data compared to the Pareto type-1 which models smaller more suitably.

TABLE 4. PDF and CDF of the Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential distributions

Distributions	PDF	CDF
Pareto type-1	$\frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}$,	$1-\left(\frac{x_0}{x}\right)^{\alpha}$,
	$x \geq x_0 > 0$ and $\alpha > 0$	$x \geq x_0 > 0$ and $\alpha > 0$
Shifted lognormal	$\frac{1}{(x-x_a)\sigma\sqrt{2\pi}} e^{\left(-\frac{(\log(x-x_a)-\mu)^2}{2\sigma^2}\right)}$,	$\frac{1}{2}+\frac{1}{2}\text{erf}\left(\frac{\log(x-x_a)-\mu}{\sigma\sqrt{2}}\right)$,
	$x > x_a$, $\sigma > 0$ and $\mu \in (-\infty, +\infty)$	$x > x_a$, $\sigma > 0$ and $\mu \in (-\infty, +\infty)$
Shifted exponential	$\theta e^{-\theta(x-x_a)}, x > x_a$	$1-e^{(-\theta(x-x_a))}$,
	$x > x_a$ and $\theta > 0$	$x > x_a$ and $\theta > 0$
Shifted stretched exponential	$\rho(x-x_a)^{\rho-1}\theta^{-\rho} e^{-\left(\frac{x-x_a}{\theta}\right)^{\rho}}$,	$1-e^{-\left(\frac{x-x_a}{\theta}\right)^{\rho}}$,
	$x > x_a, \rho > 0$ and $\theta > 0$	$x > x_a, \rho > 0$ and $\theta > 0$

TABLE 5. MLE for the parameters of Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential distributions

Distributions	MLE
Pareto type-1	$\hat{a} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{x_0}\right)}$
Shifted lognormal	$\hat{\mu} = \frac{\sum_{i=1}^{n} \log(x_i - x_a)}{n}$, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log(x_i - x_a) - \mu)^2}{n}$
Shifted exponential	$\hat{\theta} = \frac{n}{\sum(x_i - x_a)}$
Shifted stretched exponential	$\log(\theta) + (\rho - 1)\sum_{i-1}^{n} \log(x_i - x_a) - \sum_{i=1}^{n} (\frac{x_i - x_a}{\theta})^{\rho}$, $\theta > 0$ and $\rho > 1$

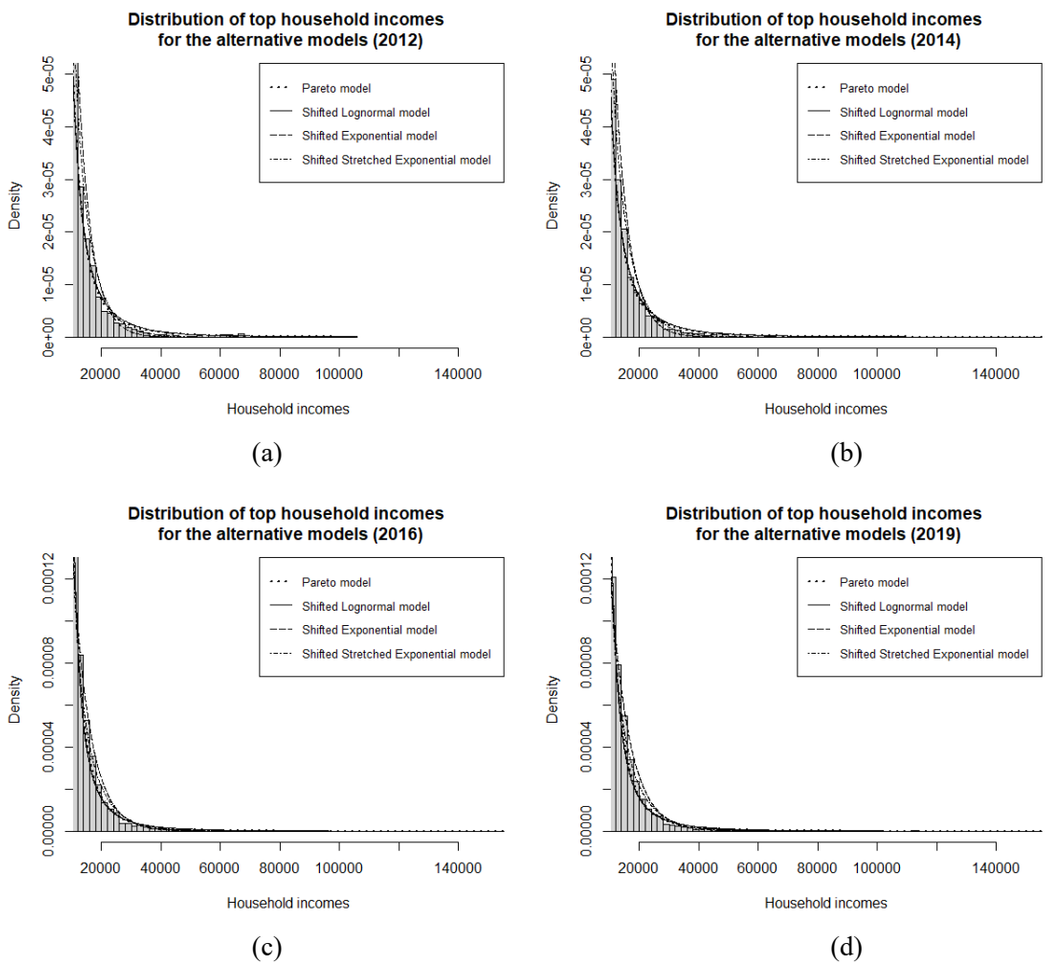

FIGURE 4. Best-fitted graphs of Pareto type-1, shifted lognormal, shifted exponential, and shifted stretched exponential models for the top household income data for the years (a) 2012, (b) 2014, (c) 2016, and (d) 2019

TABLE 6. Parameter estimation and goodness-of-fit of the best-fitted new Pareto and the alternative models based on household income data for 2012, 2014, 2016, and 2019

Year	Distributions	P_{tail}	Estimated Parameters	p-value (KS test)	D	R^2
2012	NP(x_0, α)	0.25	$\hat{\alpha} = 2.8034$	0.8323*	0.0108	0.9998
	P(x_0, α)	0.05	$\hat{\alpha} = 2.5223$	0.4736*	0.0328	0.9981
	SLN (x_α, μ, σ)	0.05	$\hat{\mu} = 8.0991$ $\hat{\sigma} = 1.4257$	0.0051	0.0671	0.9858
	SExp(x_α, θ)	0.40	$\hat{\theta} = 0.000244$	< 0.0001	0.0594	0.9808
	SSExp(x_α, θ, ρ)	0.05	$\hat{\theta} = 6440.2936$ $\hat{\rho} = 0.8010$	0.0642*	0.0510	0.9918
2014	NP(x_0, α)	0.4	$\hat{\alpha} = 2.5757$	0.9092*	0.0057	0.9999
	P(x_0, α)	0.05	$\hat{\alpha} = 2.5393$	0.4286*	0.0250	0.9990
	SLN (x_α, μ, σ)	0.10	$\hat{\mu} = 8.0633$ $\hat{\sigma} = 1.4419$	< 0.0001	0.0526	0.9895
	SExp(x_α, θ)	0.40	$\hat{\theta} = 0.000208$	< 0.0001	0.0746	0.9720
	SSExp(x_α, θ, ρ)	0.05	$\hat{\theta} = 7872.5050$ $\hat{\rho} = 0.8035$	0.1195*	0.0339	0.9967
2016	NP(x_0, α)	0.20	$\hat{\alpha} = 3.3008$	0.1529*	0.0165	0.9990
	P(x_0, α)	0.05	$\hat{\alpha} = 2.8727$	0.4446*	0.0252	0.9986
	SLN (x_α, μ, σ)	0.05	$\hat{\mu} = 8.1843$ $\hat{\sigma} = 1.4586$	0.0140	0.0459	0.9918
	SExp(x_α, θ)	0.40	$\hat{\theta} = 0.000195$	< 0.0001	0.0535	0.9850
	SSExp(x_α, θ, ρ)	0.05	$\hat{\theta} = 7150.5067$ $\hat{\rho} = 0.7573$	0.0061	0.0496	0.9897
2019	NP(x_0, α)	0.30	$\hat{\alpha} = 2.9323$	0.6309*	0.0087	0.9999
	P(x_0, α)	0.10	$\hat{\alpha} = 2.7565$	0.8915*	0.0116	0.9997
	SLN (x_α, μ, σ))	0.05	$\hat{\mu} = 8.3076$ $\hat{\sigma} = 1.4845$	0.0046	0.0494	0.9905
	SExp(x_α, θ)	0.40	$\hat{\theta} = 0.000172$	< 0.0001	0.0545	0.9829
	SSExp(x_α, θ, ρ)	0.05	$\hat{\theta} = 8131.4358$ $\hat{\rho} = 0.7319$	0.0008	0.0561	0.9876

*p-value > $\alpha = 0.05$; the highlighted R^2 values indicate the best-fitted models

CONCLUSION

From the analysis, the new Pareto distribution provides an appropriate explanation for the middle part to the upper part of the top income distribution. On the other hand, the Pareto type-1 can only explain the upper part of the data. Therefore, the new Pareto model allows a bigger scale of data compared to the Pareto type-1 model which only captures a smaller proportion of top income data. These results correspond to the advantages of the new Pareto model mentioned by Sarabia et al. (2019). The new Pareto model is more appropriate in dealing with a wider range of data than the Pareto type-1 model. Finally, we conclude that the new Pareto distribution can describe the incomes of the T20 group and partially that of the M40 group, whereas the Pareto type-1 only covers for some part (upper part) of the incomes of the T20 group.

ACKNOWLEDGEMENTS

The authors thank the Department of Statistics Malaysia for providing the Household Income Survey (HIS) data. This research is also supported and sponsored by Universiti Malaya under the Grant GPF088B-2020 and RF011B 2018FS.

REFERENCES

Abdul Majid, M.H. & Ibrahim, K. 2021. Composite pareto distributions for modelling household income distribution in Malaysia. *Sains Malaysiana* 50(7): 2047-2058. https://doi.org/10.17576/jsm-2021-5007-19

Arcagni, A. & Porro, F. 2014. The graphical representation of inequality. *Revista Colombiana de Estadística* 37(2Spe): 419. https://doi.org/10.15446/rce.v37n2spe.47947

Bahari, Z., Mohd, S. & Hamat, A.F.C. 2015. Relationship between poverty and inequality: A case study of Bumiputera household in the northern states of Malaysia. *Proceedingss of International Conference on Development and Socio Spatial Inequalities* 1: 113-120. http://eprints.usm.my/35089/1/PPIK16.pdf

Banerjee, A., Yakovenko, V.M. & Matteo, T.D. 2006. A study of the personal income distribution in Australia. *Physica A: Statistical Mechanics and Its Applications* 370(1): 54-59. https://doi.org/10.1016/j.physa.2006.04.023

Bourguignon, M., Saulo, H. & Fernandez, R.N. 2016. A new pareto-type distribution with applications in reliability and income data. *Physica A: Statistical Mechanics and Its Applications* 457: 166-175. https://doi.org/10.1016/j.physa.2016.03.043

Brzezinski, M. 2014. Do wealth distributions follow power laws? Evidence from 'Rich Lists'. *Physica A: Statistical Mechanics and Its Applications* 406: 155-162. https://doi.org/10.1016/j.physa.2014.03.052

Burden, R.L. & Faires, J.D. 2011. *Numerical Analysis.* Brooks/Cole, Cengage Learning.

Chakravarty, S.R. & Sarkar, P. 2020. New perspectives on the Gini and Bonferroni indices of inequality. Berlin, Heidelberg: Springer doi: https://doi.org/10.1007/s00355-021-01311-4

Charpentier, A. & Flachaire, E. 2019. *Pareto Models for Top Incomes.* HAL Archieve Ouverte, Id: hal-02145024

Clauset, A., Shalizi, C.R. & Newman, M.E.J. 2009. Power-law distributions in empirical data. *SIAM Review* 51(4): 661-703. https://doi.org/10.1137/070710111

Clementi, F. & Gallegati, M. 2005. Power law tails in the Italian personal income distribution. *Physica A: Statistical Mechanics and Its Applications* 350 (2-4): 427-438. https://doi.org/10.1016/j.physa.2004.11.038

Cowell, F.A. 2003. Theil, inequality and the structure of income distribution. *Distributional Analysis Research Programme* 67(May): 1-19.

Csörgo, M., Gastwirth, J.L. & Zitikis, R. 1998. Asymptotic confidence bands for the Lorenz and Bonferroni curves based on the empirical Lorenz curve. *Journal of Statistical Planning and Inference* 74(1): 65-91. https://doi.org/10.1016/s0378-3758(98)00103-7

Department of Statistics Malaysia. 2020. *Laporan Survei Pendapatan Isi Rumah dan Kemudahan Asas 2019*. pp. 3-57. https://www.dosm.gov.my/v1/index.php?r=column/cthemeByCat&cat=120&bul_id=TU00TmRhQ1N5TUxHVWN0T2VjbXJYZz09&menu_id=amVoWU54UTl0a21NWmdhMjFMMWcyZz09

Fellman, J. 2018. Income inequality measures. *Theoretical Economics Letters* 8(3): 557-574. https://doi.org/10.4236/tel.2018.83039.

Figueira, F.C., Moura, N.J. & Ribeiro, M.B. 2011. The Gompertz-Pareto income distribution. *Physica A: Statistical Mechanics and Its Applications* 390(4): 689-698. https://doi.org/10.1016/j.physa.2010.10.014

Giorgi, G.M. & Crescenzi, M. 2001. A proposal of poverty measures based on the Bonferroni Inequality Index. *Metron* 59(3-4): 3-16.

Hernández-Ramírez, E., Castillo-Mussot, M.D. & Hernández-Casildo, J. 2021. World per capita gross domestic product measured nominally and across countries with purchasing power parity: Stretched exponential or Boltzmann–Gibbs distribution? *Physica A: Statistical Mechanics and Its Applications* 568: 125690. https://doi.org/10.1016/j.physa.2020.125690

Islam, R., Ghani, A.B.A., Abidin, I.Z. & Rayaiappan, J.M. 2017. Impact on poverty and income inequality in Malaysia's economic growth. *Problems and Perspectives in Management* 15(1): 55-62. https://doi.org/10.21511/ppm.15(1).2017.05

Kusnic, M.W. & Vanzo, J. 1980. Income inequality and the definition of income: The case of Malaysia. *The RAND Corporation.* https://www.rand.org/content/dam/rand/pubs/reports/2008/R2416.pdf

Law, S. & Tan, H. 2009. The role of financial development on income inequality in Malaysia. *Journal of Economic Development* 34(2): 153-168. http://www.jed.or.kr/full-text/34-2/8.pdf

Masseran, N., Safari, M.A.M., Lok, C.M. & Hussain, S.I. 2020. Analysis of inequality in the upper-tail of urban household incomes in Malaysia. *AIP Conference Proceedings,* 2266(October). https://doi.org/10.1063/5.0018073

Moura, N.J. & Ribeiro, M.B. 2009. Evidence for the Gompertz curve in the income distribution of Brazil 1978-2005. *European Physical Journal B* 67(1): 101-120. https://doi.org/10.1140/epjb/e2008-00469-1.

Oancea, B., Pirjol, D. & Andrei, T. 2018. A Pareto upper tail for capital income distribution. *Physica A: Statistical Mechanics and Its Applications* 492: 403-417. https://doi.org/10.1016/j.physa.2017.09.034

Pundir, S., Arora, S. & Jain, K. 2005. Bonferroni curve and the related statistical inference. *Statistics and Probability Letters* 75(2): 140-150. https://doi.org/10.1016/j.spl.2005.05.024.

Rashid, N.K.A., Rahizal, N.A. & Possumah, B.T. 2020. Does income difference cause different household expenditure consumption? *International Journal of Innovation, Creativity and Change* 12(12): 1314-1340.

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

Raqab, M.Z., Alkhalfan, L.A., Bdair, O.M. & Balakrishnan, N. 2019. Maximum likelihood prediction of records from 3-parameter Weibull distribution and some approximations. *Journal of Computational and Applied Mathematics* 356: 118-132. https://doi.org/10.1016/j.cam.2019.02.006

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. https://doi.org/10.1016/j.physa.2020.125359

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. https://doi.org/10.1016/j.physa.2020.124893

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018a. A robust semi-parametric approach for measuring income inequality in Malaysia. *Physica A: Statistical Mechanics and Its Applications* 512: 1-13. https://doi.org/10.1016/j.physa.2018.08.029

Safari, M.A.M., Masseran, N. & Ibrahim, K. 2018b. Optimal threshold for Pareto tail modelling in the presence of outliers. *Physica A: Statistical Mechanics and Its Applications* 509: 169-180. https://doi.org/10.1016/j.physa.2018.06.007

Sarabia, J.M., Jordá, V. & Prieto, F. 2019. On a new Pareto-type distribution with applications in the study of income inequality and risk analysis. *Physica A: Statistical Mechanics and Its Applications* 527: 121277. https://doi.org/10.1016/j.physa.2019.121277.

Shakil, N.S.M., Mohd, S., Bahari, Z. & Hamat, A.F.C. 2015. Patterns of income distribution in the northern states of Malaysia: A life cycle approach. *Proceedingss of International Conference on Development and Socio Spatial Inequalities* 1980: 128-134.

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