Physica A 512 (2018) 1–13

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

A robust semi-parametric approach for measuring income inequality in Malaysia

Muhammad Aslam Mohd Safari *, Nurulkamal Masseran, Kamarulzaman Ibrahim

School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

HIGHLIGHTS

- A robust semi-parametric approach provides good results for measuring income inequality.
- PITSE method provides good parameter estimation of Pareto model in the presence of outliers.
- Various inequality indices based on semi-parametric model are provided.

ARTICLE INFO

Article history:
Received 26 March 2018
Received in revised form 27 July 2018
Available online 7 August 2018

Keywords:
Income distribution
Pareto distribution
Power-law
Probability integral transform statistics
Semi-parametric model
Survey samples

ABSTRACT

In this study, a robust semi-parametric approach which involves combining the empirical distribution and Pareto model is applied for measuring the income distribution in Malaysia based on the survey data of household incomes for the years of 2007, 2009, 2012 and 2014. From our analysis, it is found that the Pareto model is well fitted to the top household incomes for all years considered, suggesting that this model is a suitable parametric distribution to describe the top household incomes in Malaysia. Based on the semi-parametric Gini, generalized entropy and Atkinson indices, it is found that the income inequality shows a decreasing trend over the period, indicating an improvement of income distribution in Malaysia. Moreover, the fitted semi-parametric Lorenz curves show that the economic pie of the high income group reduced slightly after being taken up by the low income group while there is no clear change for the middle household income group.

1. Introduction

The level of income inequality is a matter of concern since it provides the condition of the economic well-being of any particular society. The high level of income inequality which is caused by a substantial spread in the income distribution contributes to a slow economic growth of a country [1–4]. In addition, according to study conducted in Latin America and western societies, Graham and Felton [5] and Ferrer-i-Carbonell and Ramos [6] found that income inequality correlates negatively with the degree of happiness. Besides, studies on the relationship between income inequality and crime have found that income inequality is positively associated with crime, in particular, violent crimes such as burglary and robbery [7–10]. Considering the importance of income inequality, it is certainly of interest of many researchers to investigate methods on determining its measure.

[*] Corresponding author.
E-mail address: aslammohdsafari@gmail.com (M.A.M. Safari).

https://doi.org/10.1016/j.physa.2018.08.029

To determine the measure of income inequality, the parametric method is often applied. In some previous works, Pareto[1] model is found to be particularly suitable to describe the top income data which exhibit heavy-tailed behavior [14–22]. To overcome the problem of poor reporting of top income earners, as explained in [18,23,24], application of a parametric distribution such as Pareto is preferable than non-parametric method since the former approach manages to capture the heavy tail property in the upper tail data. However, some authors have considered semi-parametric approach which involves combining the empirical distribution and Pareto model for describing the lower and upper parts of the income distribution respectively [14,16,18,23].

In many works on Pareto tail modeling, to take into account of the presence of extreme outliers, robust estimators have been proposed as an alternative approach to the classical estimator such as maximum likelihood estimator (MLE) for estimating the shape[2] parameter. Note that the extreme outliers that exist in the upper tail of income distribution are not necessarily to be removed, but should be considered in the analysis of income inequality since they may contain relevant information about the income distribution [16,25]. In the presence of extreme outliers, the usage of MLE should be avoided since it provides no protection against extreme outliers [26,27]. Therefore, a robust estimation of Pareto tail index should be applied in order to obtain a more accurate estimation of income inequality measures for the upper tail of income distribution.

Various methods have been proposed in the literatures on the choice of the threshold for Pareto tail modeling. The choice of optimal threshold is important due to several reasons. If the chosen threshold is low, the estimated shape parameter is biased. On the other hand, if the chosen threshold is over estimated, the sample size is reduced and, consequently, the variance of the parameter estimates increases [13,28]. Probably the most simple and widely used method for determining the threshold is by choosing a fixed proportion of top income distribution, say, 10%, 5% or 1% and study the goodness of fits of the Pareto model to the data [18,19,23,29]. Besides, the graphical tools such as the Zipf plot, Pareto quantile plot and mean excess function plot can be utilized to determine the threshold for Pareto modeling by choosing the leftmost point of the fitted line that shows a positive linear trend [30–32]. However, the choice made based on these two methods are rather subjective, and therefore, the threshold value determined is not optimal. It has been argued that the best technique for determining the optimal threshold is by choosing the value of the threshold that minimizes the empirical distribution function (EDF) statistics such as Kolmogorov–Smirnov (KS) or Anderson–Darling (AD) statistics [13,33].

This study attempts to provide the analysis of income inequality in Malaysia involving Lorenz curve and three inequality indices, namely, Gini index, generalized entropy (GE) index and Atkinson index using a robust semi-parametric approach based on the data for the years of 2007, 2009, 2012 and 2014 which are available from the sample surveys conducted by Malaysia Department of Statistics (DOSM). In the analysis, sample weights are considered in order to allow for the inclusion probabilities of the observations in the population so that the true distribution of the population is accurately reflected [16,34,35]. The rest of the paper is organized as follows. Section 2 focuses on the data sources and the sampling methods used. Section 3 provides the methodologies of this research. Section 4 presents the results of the analysis. Finally, Section 5 concludes the paper.

2. Data

The data set considered consists of household income data[3] which are obtained from the Household Income Surveys (HIS). This official survey was first carried out in 1973 and has been conducted twice in every five years. The main objective of HIS is to measure the economic well-being of the Malaysian population by collecting information on the household incomes and socio-economic background. The statistics found are used for proposing policy and development of economic plan for Malaysia, particularly in terms of eradicating of poverty and developing strategies for fair income distribution.

In the HIS, two-stage stratified sampling design was carried out involving the survey frame which consists of Enumeration Blocks[4] (EB) [36]. In the sampling design, the strata for the first stage consist of the administrative districts for every state in Malaysia while for the second stage is either urban or rural areas. The EB is a sampling frame, consisting of about 80 to 120 living quarters (LQ) that is classified as either urban or rural areas. For each selected EB, a systematic sampling method is applied to ensure that every LQ have an equal probability to be selected in the sample. This procedure is performed in order to produce an unbiased sample, which is representative of the entire population of households in Malaysia.

In this study, the Malaysian household monthly gross incomes are applied for estimating the income inequality in Malaysia from 2007 to 2014. Moreover, in the analysis, the inflation effects are also being taken into account by multiplying all the data for each year by a factor that includes the value of inflation using 2007 as the reference year. For example, as given in Table 1, for the year 2009, all the household income data is multiplied by a factor of $1/1.06 = 0.9434$ since the inflation rate is 6%. By considering the inflation effects, it is fair to compare income inequality over the years. Table 1 shows the reported sample sizes for HIS, total number of households in Malaysia, level of income poverty line of income, annual Consumer Price Index (CPI) and inflation rates for the years 2007 to 2014 with 2007 as the reference year.

[1] In the literature, the Pareto distribution is also often known as power-law distribution and Zipf's law [11–13].

[2] The shape parameter of Pareto distribution (α) is also known as the Pareto tail index, Pareto exponent or Pareto coefficient.

[3] The currency of Malaysia is Ringgit Malaysia (RM). For the year 2007 to 2014, the exchange rate of RM1 is in the range of 0.28 to 0.34 USD.

[4] According to DOSM [36], the EB are geographical contiguous areas of land, identifiable by boundaries which are created for the purpose of survey operation.

Table 1
Sample sizes for HIS, total number of households in Malaysia, level of income poverty line, annual CPI and inflation rate for the years 2007 to 2014 with 2007 as the reference year.
Source: Economic Planning Unit (2016) [37] and Department of Statistics Malaysia (2016) [38].

Year	Sample size, *n*	Total number of household	Poverty line (RM)	Annual CPI	Inflation rate (%) (2007 as ref. year)
2007	12,136	6,195,682	750	92.7	–
2009	12,908	6,557,880	800	98.3	6.00
2012	13,232	6,943,203	860	104.9	13.16
2014	24,463	7,108,210	950	110.5	19.20

3. Methodologies

3.1. Semi-parametric model

Assume that P_{tail} denotes the proportion of top income earners in the population. In order to employ the semi-parametric approach, a Pareto model is fitted to $100P_{tail}$% of the upper tail data. The cumulative distribution function (CDF) of the Pareto distribution is given by

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

where x_0 is the scale or threshold parameter of Pareto distribution and $\alpha > 0$ is the shape parameter. The parameter α measures the heaviness of Pareto tail whereby smaller value of α indicates a heavier tail [39]. In our analysis, the 100(1-P_{tail}) % lower tail data of income distribution is modeled by an empirical distribution. The weighted empirical distribution of an ordered household incomes denoted as $x_{(1)}, x_{(2)},\dots, x_{(n-k)}$ is given by

$$F_{n-k}(x) = \frac{\sum_{i=1}^{n-k} w_i I(x_{(i)} \leq x)}{\sum_{i=1}^{n-k} w_i}, \tag{2}$$

where k is the number of observations in the upper tail of the income distribution, $I(\cdot)$ is the indicator function and w_i is the sample weight of the ith observed household income data. According to Cowell and Victoria-Feser [14], the full semi-parametric distribution $\overline{F}_{x_0;\alpha}(x)$ of the household income X can be written as

$$\overline{F}_{x_0;\alpha}(x) = \begin{cases} F_{n-k}(x), & x < x_0 \\ 1 - P_{tail}\left[\frac{x_0}{x}\right]^{\alpha}, & x \geq x_0 \end{cases} \tag{3}$$

Then, the quantile function of the semi-parametric distribution (3) can be given by

$$\overline{Q}_{x_0;\alpha}(u) = \begin{cases} Q_{n-k}(u), & u < 1 - P_{tail} \\ x_0\left[\frac{P_{tail}}{1-u}\right]^{1/\alpha}, & u \geq 1 - P_{tail} \end{cases}, \tag{4}$$

where $u \in [0, 1]$.

3.2. Pareto quantile plot

The Pareto quantile plot can be used to check whether the upper tail data can be adequately modeled by Pareto distribution [30]. If the tail of a data set is found to follow a Pareto distribution, the observations in the Pareto quantile plot will appear to form a straight line. Any values starting from the leftmost data point of the fitted line can be chosen as the threshold.

As suggested by Beirlant et al. [30], the Pareto quantile plot can be constructed by plotting the logarithms of the income data, i.e. $\log(x_i)$, for $i = 1, 2,\dots, n$ against the theoretical quantiles of the standard exponential distribution. Following Alfons et al. [16], and when the sample weights are considered, the theoretical quantiles of the standard exponential distribution can be written as

$$-\log\left(1 - \frac{\sum_{j=1}^{i} w_j}{\sum_{j=1}^{n} w_j}\frac{n}{n+1}\right), \quad i = 1, 2, \dots, n, \tag{5}$$

where w_j is the sample weights of the jth observed the household income data. Another advantage of the Pareto quantile plot is that one can identify the presence of extreme outliers by noticing points which fall away from the fitted Pareto model [16]. Accordingly, the Pareto quantile plot can be a useful graphical tool for examining the presence of extreme outliers in the upper tail of the income distribution.

3.3. The threshold selection and robust estimator for the shape parameter of Pareto distribution

The optimal threshold for Pareto tail modeling can be determined by using KS statistic [13,40]. The idea is to choose the value of threshold x_0 that minimizes KS statistic. The KS statistic (D) can be written as

$$D = \max_{x \geq x_0} |F_k(x) - \hat{F}_{x_0;\alpha}(x)|, \tag{6}$$

where $F_k(x)$ is the weighted empirical cumulative distribution function for household incomes given the threshold x_0 and $\hat{F}_{x_0;\alpha}(x)$ is the CDF of the fitted Pareto model. Once the threshold has been estimated, denoted as $\hat{x}_0$, the proportion of top household incomes that can be fitted by Pareto model is given by

$$P_{tail} = \frac{\sum_{i=1}^{k} w_{n-k+i}}{\sum_{i=1}^{n} w_i}, \tag{7}$$

where k is the number of top household incomes can also be determined.

In the literature, there are several robust estimators for the Pareto tail index which have been proposed and applied for Pareto tail modeling of top incomes. For example, Cowell and Victoria-Feser [14] have used optimal B-robust estimators (OBRE) for estimating the shape parameter of Pareto model. Apart from that, Alfons et al. [16] have considered two robust estimators, which are integrated squared error (ISE) and partial density component (PDC), for Pareto tail modeling of EUSLIC data. In our analysis, a robust estimator for the shape parameter of Pareto distribution called probability integral transform statistic estimator (PITSE) which is proposed by Finkelstein et al. [26] that have the advantages of being both conceptually and computationally simpler than other robust estimators is utilized. Brzezinski [27] has suggested that PITSE gives a more powerful protection against outliers as compared to MLE and some robust estimators such as PDC and weighted MLE. When the sample weights are taken into account, PITSE can be defined as

$$G_{k,t}(\alpha) = \frac{\sum_{i=1}^{k} w_{n-k+i} \left(\frac{x_0}{x_{n-k+i}} \right)^{\alpha t}}{\sum_{i=1}^{k} w_{n-k+i}}, \tag{8}$$

where $t > 0$ is a tuning parameter that will be used to adjust the balance between efficiency and robustness. By assuming that $\frac{x_0}{x_{n-k+i}}$ can be represented by a uniform random variable and applying the method of moment approach for estimating parameter, the estimate of α can be found by solving the equation

$$\frac{\sum_{i=1}^{k} w_{n-k+i} \left(\frac{x_0}{x_{n-k+i}} \right)^{\alpha t}}{\sum_{i=1}^{k} w_{n-k+i}} = \frac{1}{t+1} \tag{9}$$

using a numerical method such as secant. Note that, for any higher value of t, PITSE gains robustness but losses its relative efficiency. To get a reasonable protection against outliers, Finkelstein et al. [26] suggested to use $t = 0.531$ and 0.883 that correspond to 88% and 78% asymptotic relative efficiency (ARE) of $\hat{\alpha}$.

To evaluate the goodness of fit for fitted Pareto model to the top household incomes, the R^2 coefficient is utilized. The coefficient of R^2 quantifies the correlation between the observed probabilities and the predicted probabilities when a particular distribution is assumed. A high value of R^2 indicates that the Pareto model can adequately explain the top household income data. The R^2 coefficient can be evaluated by

$$R^2 = \frac{\sum_{i=1}^{k} \left[\hat{F}_{x_0;\alpha}(x_{n-k+i}) - \overline{F}_{x_0;\alpha}(x) \right]^2}{\sum_{i=1}^{k} \left[\hat{F}_{x_0;\alpha}(x_{n-k+i}) - \overline{F}_{x_0;\alpha}(x) \right]^2 + \sum_{i=1}^{k} \left[F_k(x_{n-k+i}) - \hat{F}_{x_0;\alpha}(x_{n-k+i}) \right]^2}, \tag{10}$$

where for $i = 1, 2, \ldots, k, F_k(x_{n-k+i})$ is the weighted empirical cumulative probabilities, $\hat{F}_{x_0;\alpha}(x_{n-k+i})$ is the estimated cumulative probabilities under the Pareto model and $\overline{F}_{x_0;\alpha}(x)$ is the weighted average of $\hat{F}_{x_0;\alpha}(x_{n-k+i})$.

3.4. Semi-parametric Lorenz curve and Gini index

Lorenz curve is a graphical representation of income inequality that was developed by Lorenz [41]. Fig. 1 shows an example of a Lorenz curve which shows the actual distribution of income while the line of equality represents an evenly distributed income. The more unequal the distribution of income is, the more the Lorenz curve deviates from the line of equality [39,41].

Based on the semi-parametric Lorenz curve which has been developed by Cowell and Victoria-Feser [14], we defined the semi-parametric Lorenz curve for the household incomes as

$$L(u) = \int_{\underline{x}}^{\overline{Q}_{x_0;\alpha}(u)} \frac{x}{\mu} d\overline{F}_{x_0;\alpha}(x). \tag{11}$$

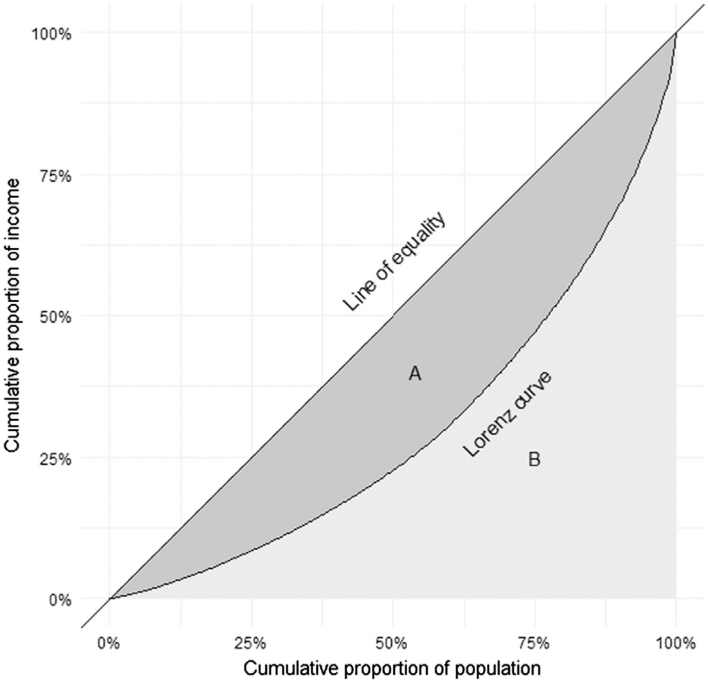


Fig. 1. Example of a Lorenz Curve.

Based on Eq. (11), it can be shown that

$$L(u) = \begin{cases} \frac{1}{\mu} \int_{\underline{x}}^{Q_{n-k}(u)} x dF_{n-k}(x), & u < 1 - P_{tail} \\ \frac{1}{\mu} \left[\int_{\underline{x}}^{Q_{n-k}(1-P_{tail})} x dF_{n-k}(x) + P_{tail} \frac{\alpha x_0}{1-\alpha} \left[\left(\frac{1-u}{P_{tail}} \right)^{(\alpha-1)/\alpha} - 1 \right] \right], & u \geq 1 - P_{tail} \end{cases}, \tag{12}$$

where $\underline{x}$ is the minimum value of household incomes and $\hat{\mu}$ is the weighted semi-parametric mean. The parameter of μ can be estimated by

$$\hat{\mu} = \frac{\sum_{i=1}^{n-k} w_i x_{(i)}}{\sum_{i=1}^{n} w_i} + P_{tail} \frac{\alpha x_0}{\alpha - 1}, \quad \alpha > 1, \tag{13}$$

where $x_{(i)}$ is the ordered household incomes for $i = 1, 2, \ldots, n$ and w_i is the sample weight attached to $x_{(i)}$. The formula proposed by Cowell and Flachaire [42] is applied to estimate the non-parametric part of the Lorenz curve in Eq. (12), and after allowing for the sample weights, we have

$$\frac{1}{\hat{\mu}} \frac{\sum_{i=1}^{G(u)} w_i x_{(i)}}{\sum_{i=1}^{n} w_i}, \tag{14}$$

where $G(u) = \lfloor i - u + 1 \rfloor$ is the largest integer less than $i - u + 1$. Based on the semi-parametric Lorenz curve of Eq. (12), the semi-parametric Gini index, denoted as *Gini* where *Gini*∈ [0,1], can be estimated by

$$Gini = 1 - 2 \int_0^1 L(u) du. \tag{15}$$

Note that in Fig. 1, the value of Gini index is twice the area of *A*. Thus, *Gini* = 2*A* = 1–2*B*. A Gini index of 0 represents a perfect income equality, while a Gini index of 1 represents a perfect income inequality where only one person or household is earning 100 percent of the total income [43,44]. To calculate the Gini index in Eq. (15), one may apply the trapezoidal rule to find the area under the Lorenz curve given by $\int_0^1 L(u) du$.

3.5. Semi-parametric generalized entropy and Atkinson index

The GE index is an inequality measures that incorporates a sensitivity parameter ε which represents the weights given to differences between incomes in different parts of the income distribution [45]. The lower or higher values of ε indicate that GE indices are more sensitive to either lower or upper tail of the income distribution respectively. The semi-parametric GE index can be computed by using the moments of the semi-parametric distribution. In the case when sample weights are considered, the semi-parametric GE index, denoted as $GE(\varepsilon)$ where $GE(\varepsilon) \in [0,\infty)$, is given by

$$GE(\varepsilon) = \frac{1}{\varepsilon^2 - \varepsilon}\left[\frac{v_\varepsilon^*}{\mu^\varepsilon} - 1\right], \quad \varepsilon \neq 0, 1, \tag{16}$$

where μ is the semi-parametric mean and v_ε^* can be given by

$$v_\varepsilon^* = \frac{\sum_{i=1}^{n-k} w_i x_{(i)}^\varepsilon}{\sum_{i=1}^{n} w_i} + P_{tail}\frac{\alpha x_0^\varepsilon}{\alpha - \varepsilon}. \tag{17}$$

The two special cases of the GE index where $\varepsilon = 1$ and $\varepsilon = 0$, known as Theil index and mean logarithmic deviation index respectively, can be determined using

$$GE(1) = \frac{v_1^*}{\mu} - \log(\hat{\mu}) \tag{18}$$

and

$$GE(0) = \log(\mu) - v_0^*, \tag{19}$$

where v_1^* and v_0^* can be given as follows

$$v_1^* = \frac{\sum_{i=1}^{n-k} w_i \left(x_{(i)} \log x_{(i)}\right)}{\sum_{i=1}^{n} w_i} + P_{tail}\frac{\alpha x_0}{\alpha - 1}\left[\log x_0 + \frac{1}{\alpha - 1}\right] \tag{20}$$

and

$$v_0^* = \frac{\sum_{i=1}^{n-k} w_i \log x_{(i)}}{\sum_{i=1}^{n} w_i} + P_{tail}\left[\log x_0 + \frac{1}{\alpha}\right]. \tag{21}$$

The Atkinson index is an inequality measure which is based explicitly on a social welfare evaluation of income distribution where it captures a greater equality in income distribution as higher social welfare. The Atkinson index incorporates a sensitivity parameter ξ which represents the weight given to differences in incomes in different parts of the income distribution [46]. As the parameter ξ increase, more weight is attached to income transfers at the bottom of the distribution. Stern [47] reviews the literature on elasticity of marginal utility of income and suggests that the value of ξ to lie between 1.5 and 2.5. However, in practice, the values of $\xi = 0.5$, 1, 1.5 or 2 are usually used to measure inequality [48]. The semi-parametric Atkinson index can also be computed by applying the moments of the semi-parametric distribution. In the case when sample weights are considered, the semi-parametric Atkinson index, denoted as $At(\xi)$ where $At(\xi) \in [0,1]$, can be written as

$$At(\xi) = 1 - \frac{\left(v_\xi^*\right)^{1/(1-\xi)}}{\mu}, \quad \xi \neq 1, \tag{22}$$

$$At(1) = 1 - \exp[-GE(0)] = 1 - \frac{\exp[v_0^*]}{\mu}, \tag{23}$$

where v_0^* can be computed by using the formula in Eq. (21) and v_ξ^* can be calculated by

$$v_\xi^* = \frac{\sum_{i=1}^{n-k} w_i x_{(i)}^{1-\xi}}{\sum_{i=1}^{n} w_i} + P_{tail}\frac{\alpha x_0^{(1-\xi)}}{\alpha - (1-\xi)}. \tag{24}$$

According to Atkinson [46], $At(\xi)$ can also be determined by considering the equally distributed equivalent income (y_{EDE}) which is the level of income that, if received by every unit or individual, produces the same social welfare as the actual distribution. This measure is given by

$$At(\xi) = 1 - \frac{y_{EDE}}{\mu}. \tag{25}$$

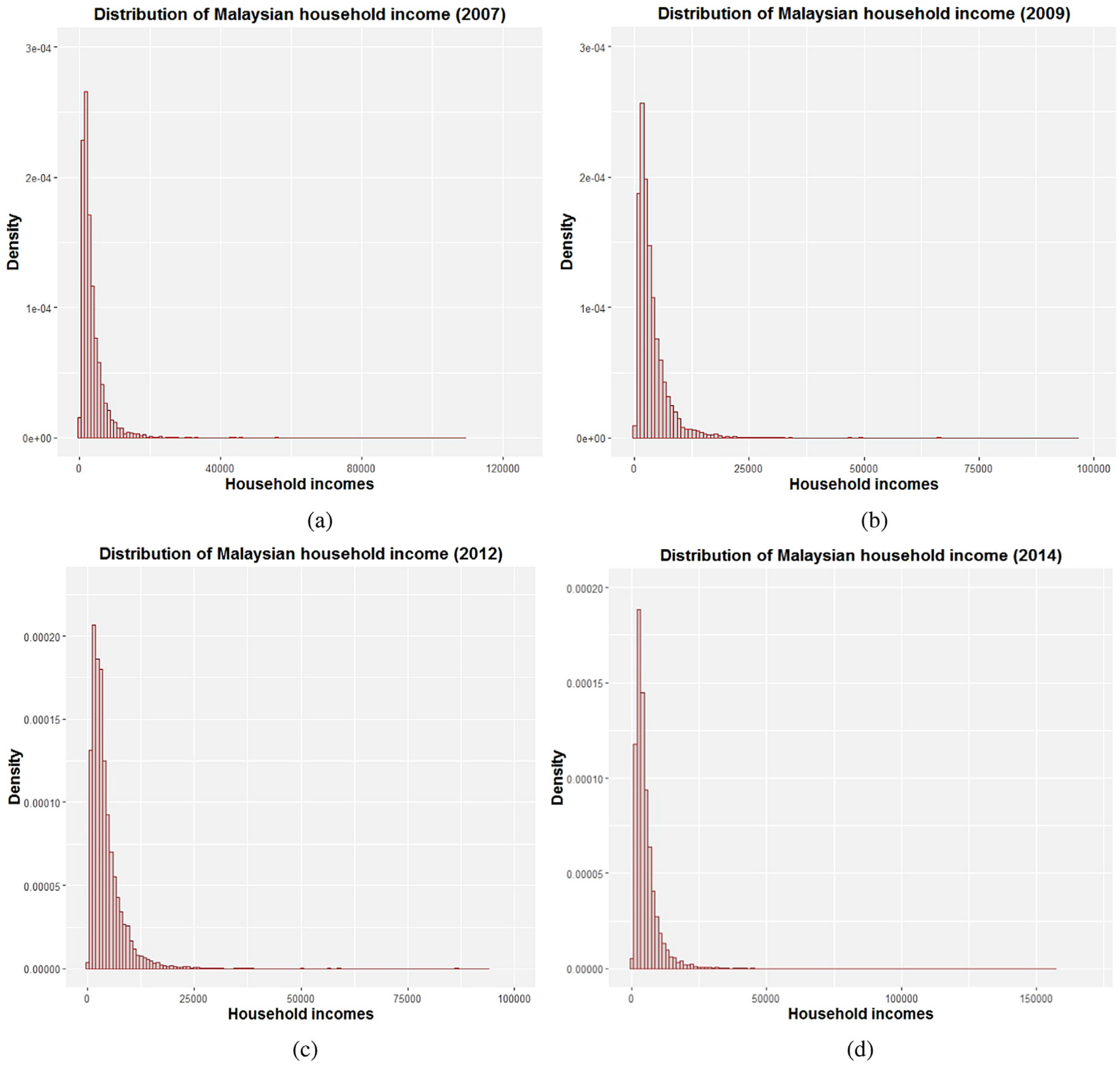
4. Data analysis and results

4.1. Descriptive statistics

Table 2 shows the descriptive statistics which involves mean, median, variance, maximum, minimum and coefficient of skewness for the household income after taking into account of inflation rates from 2007 to 2014. Based on the reported mean and median values for the different years, it is clear that household incomes show an increasing trend, where the highest

Table 2
The descriptive statistics of Malaysian household incomes for the years 2007 to 2014.

Year	Mean (RM)	Median (RM)	Variance	Minimum (RM)	Maximum (RM)	Coefficient of skewness
2007	3588.82	2450.00	16004620	59.17	109036.00	6.1508
2009	3781.70	2681.80	15324824	94.34	96305.03	5.0548
2012	4401.75	3187.27	20994733	132.56	93635.56	5.7244
2014	5259.22	3865.77	27814145	178.69	156788.60	6.1625



(a)

(b)

(c)

(d)

Fig. 2. The distribution of Malaysian household income data for the years of (a) 2007, (b) 2009, (c) 2012 and (d) 2014.

values are observed for the year 2014. The variance of household incomes is also high for each year, indicating that the data is wide spread about the mean. Based on the calculated coefficient of skewness, it can be observed that all the coefficients are positive, indicating that the distribution of Malaysian household incomes does not follow a normal distribution and skews to the right. As shown in Fig. 2, the household incomes appear to be skewed to the right, which explains why the mean is always greater than the median. In addition, it can also be observed that the upper tail of household income distribution for each year exhibits a heavy-tailed behavior. The ranges of minimum and maximum values for all years are between 59.17 to 178.69 and 93635.56 to 156788.60 respectively. From these values, it is shown that the ranges of Malaysian household incomes for the different years are extremely large, which indicates a large dispersion in the data.

Table 3
The estimated threshold levels ($\hat{x}_0$), the estimated shape parameters of Pareto model ($\hat{\alpha}$), R^2 values, number of top household income data (n_{tail}), proportion of top household incomes (P_{tail}) and semi-parametric mean for the different years ($\hat{\mu}$).

Year	$\hat{x}_0$	$\hat{\alpha}$	R^2	$n_{tail}(x \geq \hat{x}_0)$	P_{tail}	$\hat{\mu}$
2007	5871.67	2.1568	0.9987	1499	0.15	3697.64
2009	7227.20	2.3114	0.9980	1183	0.11	3872.32
2012	8700.07	2.6039	0.9990	1065	0.10	4408.21
2014	8494.13	2.2256	0.9984	2903	0.14	5391.11

4.2. Fitting the Pareto distribution to top of household incomes

All methods presented in this paper are analyzed by using R [49]. Fig. 3 shows the Pareto quantile plot of Malaysian household income for the years of 2007, 2009, 2012 and 2014. As shown in Fig. 3, some data points which represent the top incomes appear to deviate from the Pareto model. Thus, it is certain that there exist some extreme outliers in the upper tail of the household income distribution. In addition, it could be seen that the upper part of the Pareto quantile plots appear to form almost a straight line. This indicates that the upper tail of the household income distribution holds the Pareto assumption. Determining the threshold using the leftmost data point of the fitted lines is quite subjective. However, as indicated previously, the optimal threshold of Pareto model can be determined by choosing the value of x_0 that minimizes KS statistic. In Fig. 3, the intersection point of the Pareto quantile plot and the solid line indicates the estimated optimal threshold $\hat{x}_0$ and the values of $\hat{x}_0$ for all years are given in Table 3. Given that the optimal thresholds levels for all the years considered have been estimated based on the Pareto model, the proportion of households that belongs to the Pareto region, denoted as P_{tail}, can be calculated using Eq. (7) and the values are given in Table 3. Based on several studies, such as [20] and [40], in certain country, the proportion of households that belongs to the Pareto region are recognized as the proportion of high income group.

After having identified the existence of extreme outliers, the PITSE[5] (78% ARE) is applied for estimating the shape parameter α. In the presence of extreme outliers, given this particular choice of robust estimate of α, the value of x_0 that minimizes KS statistic is determined. The shape parameter α is a useful measure of income inequality where the smaller value of α, contribute to a more uneven income distribution [50,51]. Since the value of $\hat{\alpha}$ is found lowest for the year 2007, as demonstrated in Table 3, we could say that the earners of the top household income in 2007 had the most unequal income distribution as compared to the other years.

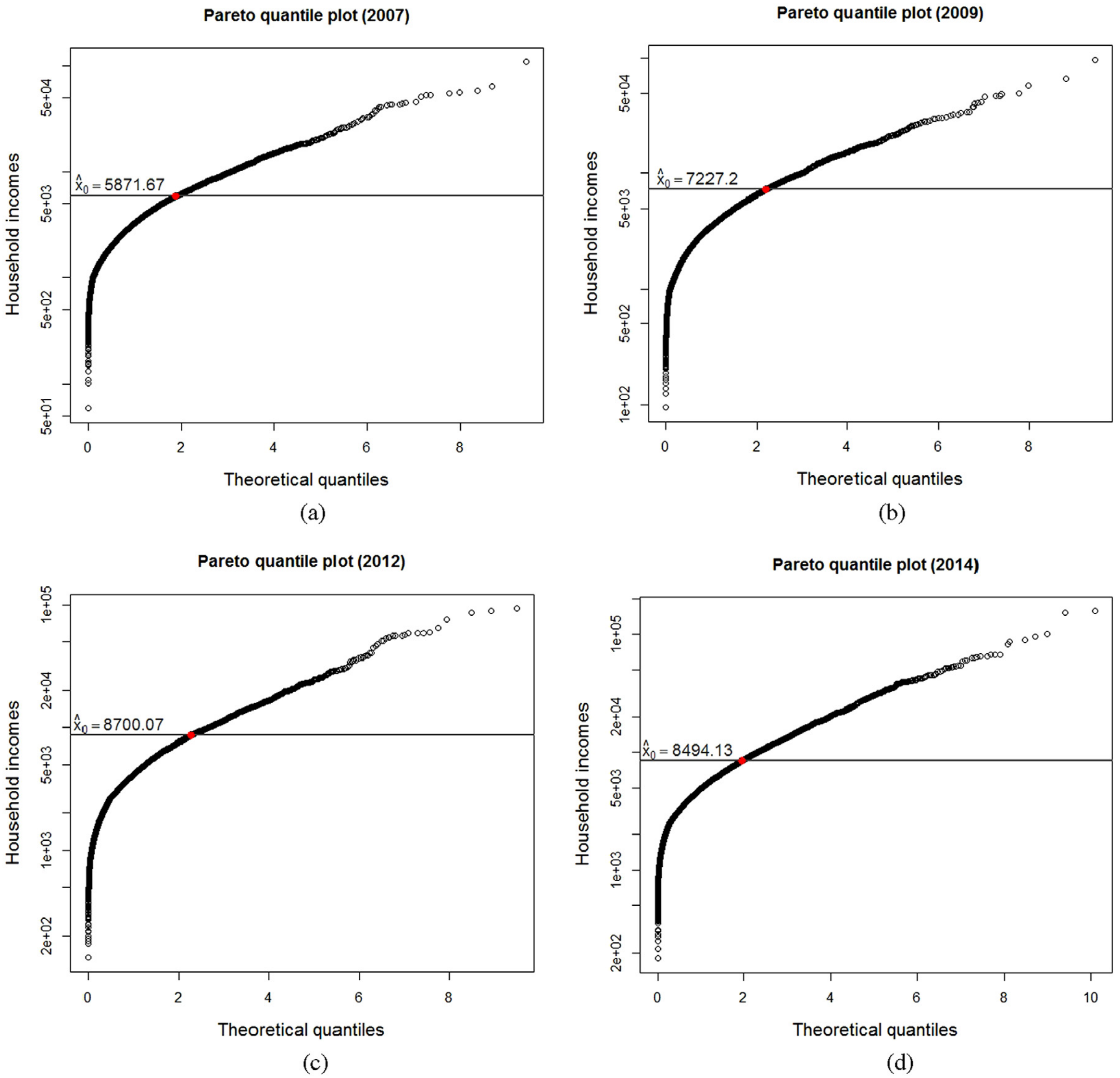
Fig. 4 displays the fitted Pareto model to the top household income data for the years of 2007, 2009, 2012 and 2014. As shown in Table 3, it can be seen that the Pareto distribution is well fitted to the top household income data for each year since all the values of $R^2$are greater than 0.99. This suggests that about 99% of the variation in the data are can be explained by the Pareto model and about 1% of the variation is attributed to error, which cannot be explained by the model. In general, as observed in Table 3, the estimated semi-parametric mean $\hat{\mu}$, shows an increasing trend from 2007 to 2014, indicating that on the average, the household incomes for these years had continued to grow.

4.3. Income inequality in Malaysia based on semi-parametric model

The fitted semi-parametric Lorenz curves that are shown in Fig. 5 describe how the incomes of different household income groups are distributed over the time period. According Eighth Malaysia Plan [52], with respect to the income distribution, there are three income groups: bottom 40% (low), middle 40% (middle), top 20% (high). Table 4 summarizes the proportion of total income shared by each income group from 2007 to 2014. It can be observed over the period that the low income household group only earned between 13.82% and 15.88% of the total household income, indicating a small gradual increase in these proportions. On the other hand, the percentages of the earnings of the middle income household group fluctuated around from 34.42% to 36.60%. There was no clear change in the proportion of total household income earned by this group. However, for the high income household group, it can be observed a slight decreased in the proportion of income earned, reducing from 51.76% in 2007 to 48.60% in 2014. Based on the increase and decrease of income shares for low and high income household groups respectively, we could say that the household income distribution had improved over the period. Nevertheless, it could be seen that the huge difference between the proportions of the household income attained by the high and low income household groups has lead to the income inequality in Malaysia. Since it is observed that the Lorenz curves of Malaysian household incomes cross each other, we cannot rank the distribution of income for the different years in terms of inequality.

Table 5 shows the values of several semi-parametric income inequality indices considered in this study. From this table, the Gini index shows a decreasing trend from 0.4627 in 2007 to 0.4235 in 2014. These values indicate that for the year 2007, only 53.73% of the households shared the total household income and the other 46.27% gained nothing while for the year 2014, the measure of income inequality has reduced, resulting in 57.65% of the households shared the total household

[5] The PITSE (78% ARE) is numerically solved by using secant method.



(a)

(b)

(c)

(d)

Fig. 3. Pareto quantile plots of Malaysian household income data in (a) 2007, (b) 2009, (c) 2012 and (d) 2014.

Table 4
Income shares of income groups located in the bottom 40%, middle 40% and top 20%.

Cumulative percent of household (%)	Percent of household (%)	Percent of income (%)				Cumulative percent of income (%)			
		2007	2009	2012	2014	2007	2009	2012	2014
40	40 (Low)	13.82	14.01	14.80	15.88	13.82	14.01	14.80	15.88
80	40 (Middle)	34.42	35.47	36.60	35.52	48.24	49.48	51.40	51.40
100	20 (High)	51.76	50.52	48.60	48.60	100.00	100.00	100.00	100.00

income and the other 42.35% gained nothing. These results further support the claim that the distribution of incomes among households continued to improve over the years.

In addition to the Gini index, GE and Atkinson indices with different values of sensitivity parameters, denoted by ε and ξ respectively, are also applied to measure income inequality in Malaysia. The GE index with three different values of ε, i.e. ε = 0, 1, and 2, are used to allow for the sensitivity of the GE indices to income differences at the different position of the income

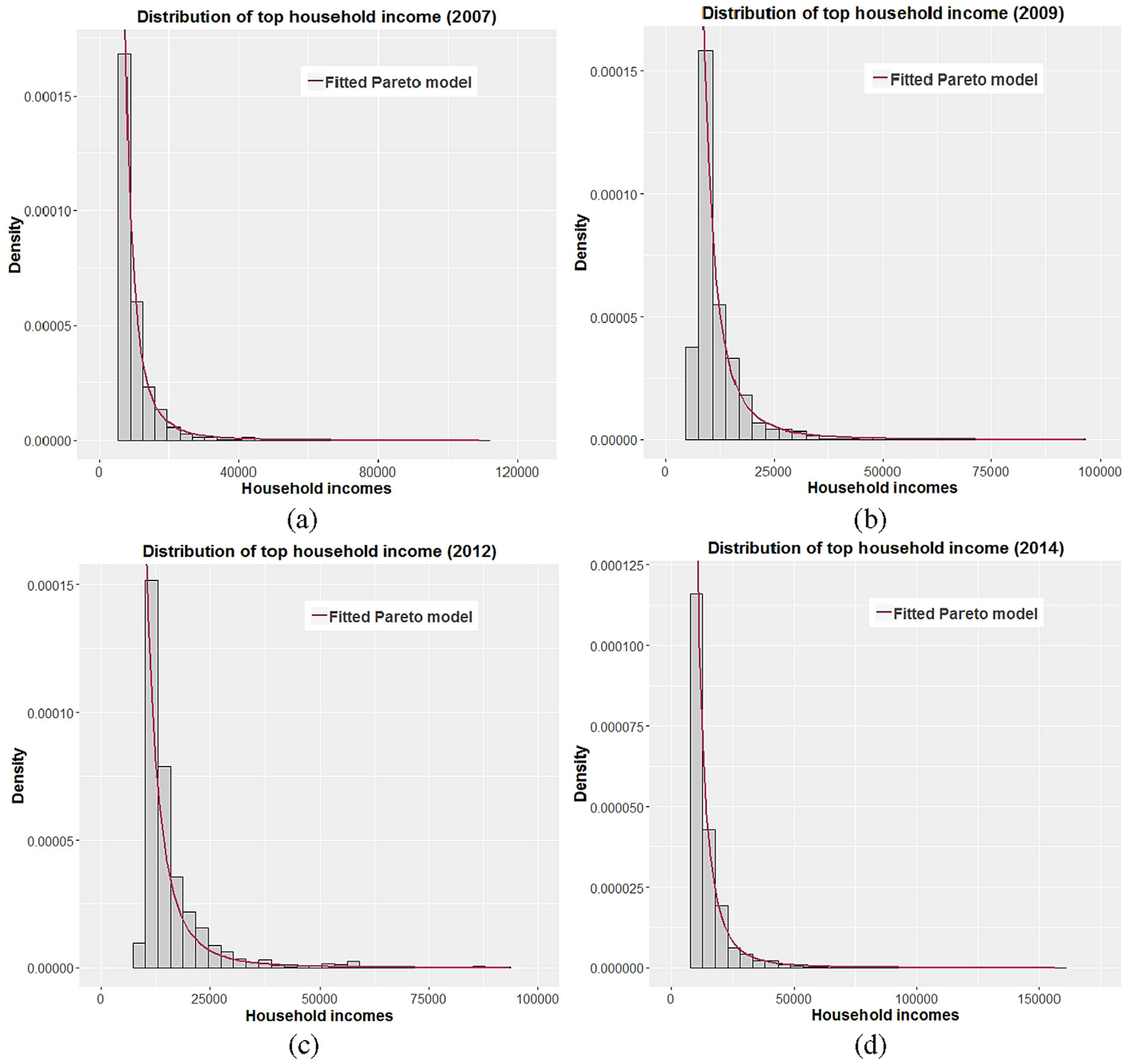


Fig. 4. The fitted Pareto distribution to the top household income data for the years of (a) 2007, (b) 2009, (c) 2012 and (d) 2014.

Table 5
The estimated semi-parametric income inequality measures based on Gini, GE and Atkinson indices with respect to several values of sensitivity parameters.

Year	Gini	$GE(0)$	$GE(1)$	$GE(2)$	$At(0.5)$	$At(1)$	$At(2)$
2007	0.4627	0.3700	0.4395	2.3432	0.1796 (3033.54)	0.3093 (2553.96)	0.4955 (1865.46)
2009	0.4511	0.3525	0.3996	1.2325	0.1680 (3221.77)	0.2971 (2722.24)	0.4863 (1989.21)
2012	0.4299	0.3217	0.3435	0.6868	0.1520 (3738.16)	0.2751 (3195.51)	0.4661 (2353.54)
2014	0.4235	0.3082	0.3580	1.4753	0.1521 (4571.12)	0.2653 (3960.85)	0.4358 (3041.10)

distribution. As shown in Table 5, the Theil index or $GE(1)$ and $GE(2)$ are found to reduce for the period from 2007 to 2012, but slightly increase in 2014. On the other hand, the decline in the values of bottom-sensitive index, i.e. $GE(0)$, suggests that there has been some small change of income distribution for the lower income group, attributed to a slight increase in the income shares attained. The results of $GE(0)$ and $GE(2)$ are further supported by the fitted semi-parametric Lorenz curves, as shown in Table 6. It is clear from Table 6 that the income share attained by the lower income group has slightly

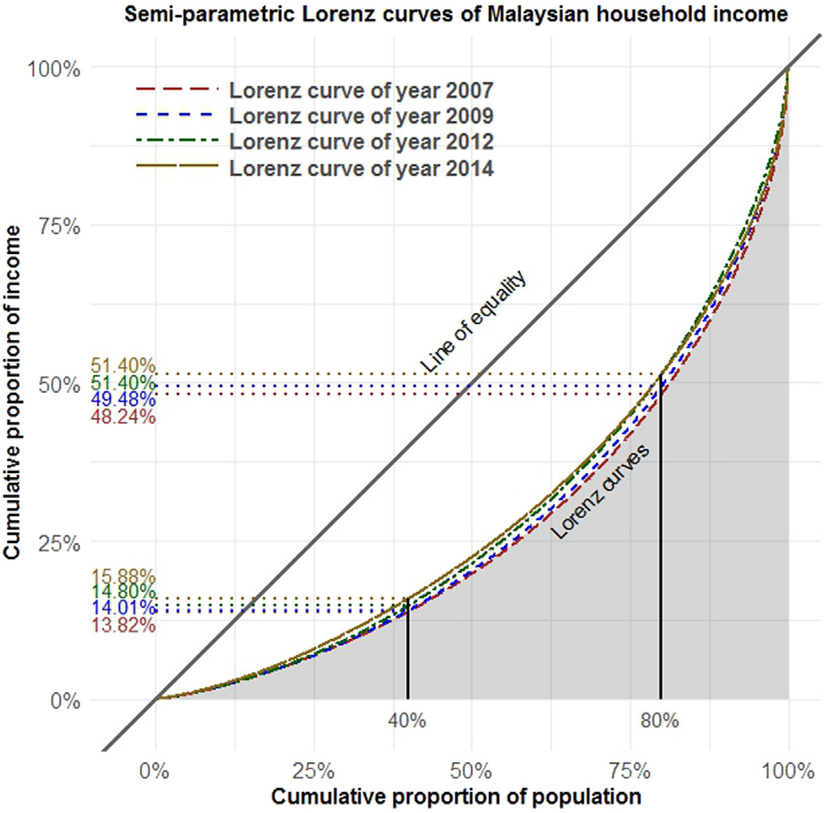


Fig. 5. The fitted semi-parametric Lorenz curves of Malaysian household income for the years of 2007, 2009, 2012 and 2014.

Table 6
Income shares at the very bottom and top of the Malaysian income distribution based on the fitted semi-parametric Lorenz curves.

Year	Bottom 5%	Bottom 10%	Top 5%	Top1%
2007	0.73%	1.90%	24.63%	10.33%
2009	0.74%	1.93%	23.13%	9.20%
2012	0.77%	1.93%	20.91%	7.71%
2014	0.84%	2.14%	22.71%	9.35%

increased and vice-versa for the high income group. This would contribute to the improvement of the income distribution for the population.

The Atkinson inequality index and the equally distributed equivalent income, i.e. y_{EDE} for three different values of the sensitivity parameter $\xi = 0.5, 1$, and 2, are presented in Table 5. For data of 2007, when ξ = 0.5 , y_{EDE} = 3033.54, meaning that if the income is equally distributed, it only requires the income of RM 3033.54 per household per month to achieve the same level of social welfare as the existing distribution with an estimated mean income of RM 3697.64 per household per month as given in Table 3. Thus a proportionate income loss of $(\mu - y_{EDE})/\mu$ = 17.96% arises from the inequality in the distribution of income, which gives a value of 0.1796 for $At(0.5)$. In other words, the same level of social welfare could be reached with only 82.04% (1–0.1796) of the existing total income while the potential welfare gain from redistribution is 17.96% of the existing income distribution in 2007. As the inequality aversion ξ parameter increases, y_{EDE} decreases and the corresponding values of inequality indices $At(\xi)$ also increase, thus indicating larger losses of welfare due to inequalities in the distribution of income. Based on Table 5, it could be seen that the Atkinson inequality index replicated the Gini index trend from 2007 to 2014 for all cases, indicating that the distribution of household incomes has improved from 2007 to 2014. Apart from that, the condition of social welfare has also been improved as the value of y_{EDE} increases for all cases over the period.

The statistical correlations between several inequality indices used are given in Fig. 6. From Fig. 6, it is clear that almost all correlations between the indices are close to 1, indicating a strong positive relationship among the indices. However, the degree of correlation between $GE(2)$ and $At(2)$ is found to be lowest, i.e. 0.4, due to the difference in the level of sensitivity of both indices for the different parts of income distribution, where it can be observed that $GE(2)$ is more sensitive to the upper parts while $At(2)$ is more sensitive to the lower parts.

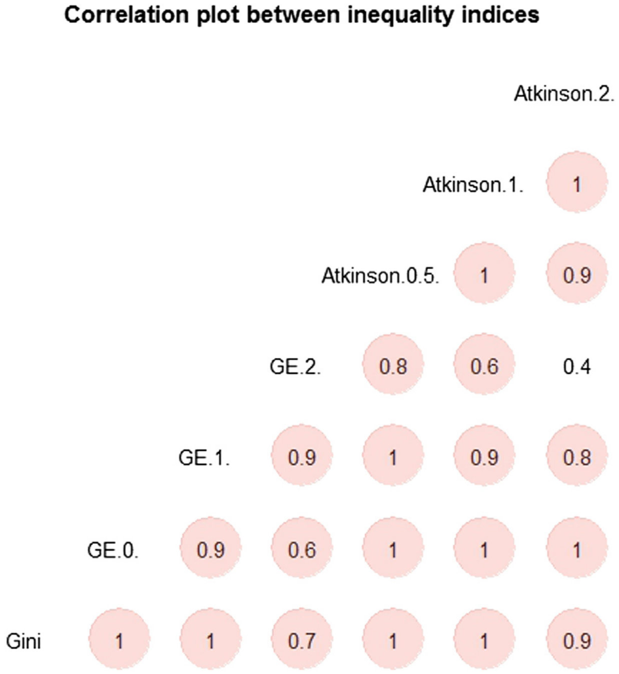


Fig. 6. The correlation plot between several inequality indices for the year of 2007, 2009, 2012 and 2014.

5. Conclusion

This paper attempts to provide the empirical analysis of the income inequality among Malaysian households using samples from the surveys carried out by DOSM over the period from 2007 to 2014. Poor reporting of top incomes based on the survey data is a problem which is often encountered, and need to be addressed so that a more reliable estimate of income inequality measures could be obtained. In this study, this issue is tackled by combining the empirical distribution and Pareto model for describing the lower and upper parts of the income distribution respectively.

Since the data used in this study has been obtained from an official survey known as HIS, it is pertinent to consider sample weights in the analysis; thus, the sample could accurately represent the population. In the analysis, the inflation effects for the different years are also being taken into account by multiplying all the data for each year by a factor that includes the value of inflation using 2007 as reference year. In order to utilize the semi-parametric model, firstly, the $100P_{tail}$ % of the household incomes in the upper part of the data distribution are fitted using Pareto model. After identifying the existence of extreme outliers in the top income data, a robust method called the PITSE is applied for estimating the shape parameter and the optimal threshold parameter is determined using the KS statistic. Finally, to measure the income inequality in Malaysia, the Lorenz curve and three inequality indices namely Gini index, GE index and Atkinson index are considered and the semi-parametric form of these indices are derived after taking sample weights into account.

From our analysis, the Pareto model is adequate for describing the top household incomes for the years of 2007, 2009, 2012 and 2014 since all values of R^2 coefficient are found to be close to one. Based on the fitted semi-parametric Lorenz curves, it was found that, over the period from 2007 to 2014, the proportions of total household income earned by the low household income group had slightly increased while for the high household income group, these proportions had slightly decreased. However, there was no clear change in the proportion of total household income earned by the middle household income group. Based on all the inequality indices found, Malaysia had experienced a decreasing trend in income inequality. This decreasing trend indicates that the overall income distributions from 2007 to 2014 had improved. However, based on the top-sensitive index, which is $GE(2)$, it is found that the income inequality for the top household incomes shows a decreasing trend from 2007 to 2012, but this measure slightly increase in the year 2014, indicating that income gap among the rich had slightly reduced during the first five years and may seem to increase later.

Acknowledgments

The authors are indebted to the Department of Statistics Malaysia and Bank Data of UKM for providing the Household Income Data (HIS). This research would not be possible without the sponsorship from Universiti Kebangsaan Malaysia (grant number GUP-2017-116 and Zamalah sponsorship, Malaysia).

References

[1] T. Persson, G. Tabellini, Is inequality harmful for growth? Amer. Econ. Rev. 84 (3) (1994) 600–621.
[2] D. Qin, M.A. Cagas, G. Ducanes, X. He, R. Liu, S. Liu, Effects of income inequality on China's economic growth, J. Pol. Model. 31 (1) (2009) 69–86.
[3] A. Castelló-Climent, Inequality and growth in advanced economies: an empirical investigation, J. Econ. Ineq. 8 (3) (2010) 293–321.
[4] V. Amarante, Revisiting inequality and growth: evidence for developing countries, Growth Chang. 45 (4) (2014) 571–589.
[5] C. Graham, A. Felton, Inequality and happiness: insights from Latin America, J. Econ. Ineq. 4 (1) (2006) 107–122.
[6] A. Ferrer-i Carbonell, X. Ramos, Inequality and happiness, J. Econ. Surv. 28 (5) (2014) 1016–1027.
[7] E.B. Patterson, Poverty, income inequality, and community crime rates, Criminology 29 (4) (1991) 755–776.
[8] M. Kelly, Inequality and crime, Rev. Econ. Stat. 82 (4) (2000) 530–539.
[9] J. Brush, Does income inequality lead to more crime? A comparison of cross-sectional and time-series analyses of United States counties, Econom. Lett. 96 (2) (2007) 264–268.
[10] J. Choe, Income inequality and crime in the United States, Econom. Lett. 101 (1) (2008) 31–33.
[11] W.J. Reed, The Pareto, Zipf and other power laws, Econom. Lett. 74 (1) (2001) 15–19.
[12] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. 46 (5) (2005) 323–351.
[13] A. Clauset, C.R. Shalizi, M.E.J Newman, Power-law distributions in empirical data, SIAM Rev. 51 (4) (2009) 661–703.
[14] F.A. Cowell, M.P. Victoria-Feser, Robust stochastic dominance: A semi-parametric approach, J. Econ. Ineq. 5 (1) (2007) 21–37.
[15] T. Ogwang, Power laws in top wealth distributions: evidence from Canada, Empir. Econ. 41 (2) (2011) 473–486.
[16] A. Alfons, M. Templ, P. Filzmoser, Robust estimation of economic indicators from survey samples based on Pareto tail modeling, J. R. Stat. Soc. C. Appl. 62 (2) (2013) 271–286.
[17] M. Brzezinski, Do wealth distributions follow power laws? Evidence from 'rich lists', Physica A 406 (2014) 155–162.
[18] N. Ruiz, N. Woloszko, What do household surveys suggest about the top 1% incomes and inequality in OECD countries? Economics Department Working Paper (1265) OECD, Paris, 2016.
[19] A.B. Atkinson, Pareto and the upper tail of the income distribution in the UK: 1799 to the present, Economica 84 (334) (2017) 129–156.
[20] P. Soriano-Hernández, M. del Castillo-Mussot, O. Córdoba-Rodríguez, R. Mansilla-Corona, Non-stationary individual and household income of poor, rich and middle classes in Mexico, Physica A 465 (2017) 403–413.
[21] B. Oancea, T. Andrei, D. Pirjol, Income inequality in Romania: the exponential-Pareto distribution, Physica A 469 (2017) 486–498.
[22] M. Jagielski, K. Czyzewski, R. Kutner, H.E. Stanley, Income and wealth distribution of the richest Norwegian individuals: an inequality analysis, Physica A 474 (2017) 330–333.
[23] V. Hlasny, P. Verme, Top incomes and the measurement of inequality in Egypt, World Bank Econ. Rev. (2016) 1–35, lhw031.
[24] G.M. Giorgi, C. Gigliarano, The gini concentration index: a review of the inference literature, J. Econ. Surv. 31 (4) (2017) 1130–1148.
[25] F.A. Cowell, E. Flachaire, Income distribution and inequality measurement: the problem of extreme values, J. Econom. 141 (2) (2007) 1044–1072.
[26] M. Finkelstein, H.G. Tucker, J. Alan Veeh, Pareto tail index estimation revisited, N. Am. Actuar. J. 10 (1) (2006) 1–10.
[27] M. Brzezinski, Robust estimation of the Pareto tail index: a Monte Carlo analysis, Empir. Econ. 51 (1) (2016) 1–30.
[28] D.J. Dupuis, M.-P. Victoria-Feser, A robust prediction error criterion for Pareto modelling of upper tails, Canad. J. Statist. 34 (4) (2006) 639–658.
[29] F. Clementi, M. Gallegati, Pareto's law of income distribution: evidence for Germany, the United Kingdom, and the United States, in: A. Chatterjee, S. Yarlagadda, B.K. Chakrabarti (Eds.), Econophysics of Wealth Distributions: Econophys-Kolkata I, Springer-Verlag, Mailand, 2005, pp. 3–14.
[30] J. Beirlant, P. Vynckier, J.L. Teugels, Tail index estimation, Pareto quantile plots, and regression diagnostics, J. Am. Stat. Ass. 91 (436) (1996) 1659–1667.
[31] P. Embrechts, C. Klüppelberg, T. Mikosch, Modelling Extremal Events, Springer-Verlag, Berlin, 1997.
[32] P. Cirillo, Are your data really Pareto distributed? Physica A 392 (23) (2013) 5947–5962.
[33] H.F. Coronel-Brizio, A.R. Hernandez-Montoya, On fitting the Pareto–Levy distribution to stock market index data: Selecting a suitable cutoff value, Physica A 354 (2005) 437–449.
[34] R.L. Chambers, Introduction to continuous and general response data, in: R.L. Chambers, C.J. Skinner (Eds.), Analysis of Survey Data, John Wiley & Sons, West Sussex, 2003, pp. 125–131.
[35] R. Valliant, J.A. Dever, F. Kreuter, Practical Tools for Designing and Weighting Survey Samples, Springer, New York, 2013.
[36] Department of Statistics Malaysia, Household Income and Basic Amenities Survey Report 2012, 2013, https://newss.statistics.gov.my/newss-portalx/ep/epProductFreeDownloadSearch.seam.
[37] Economic Planning Unit, Brief household income & poverty statistics newsletter 2007–2014, 2016, http://www.epu.gov.my/en/socio-economic/household-income-poverty.
[38] Department of Statistics Malaysia, Population of Malaysia 1895–2015, 2016, http://www.data.gov.my/data/ms_MY/dataset/population-and-demographic-statistics-malaysia.
[39] C. Kleiber, S. Kotz, Statistical Size Distributions in Economics and Actuarial Sciences, Wiley, New York, 2003.
[40] M.A.M. Safari, N. Masseran, K. Ibrahim, Optimal threshold for Pareto tail modelling in the presence of outliers, Physica A 509 (2018) 169–180.
[41] M.O. Lorenz, Methods of measuring the concentration of wealth, J. Amer. Statist. Assoc. 9 (70) (1905) 209–219.
[42] F.A. Cowell, E. Flachaire, Statistical methods for distributional analysis, in: A.B Atkinson, F. Bourguignon (Eds.), Handbook of Income Distribution, vol. 2, Elsevier Science, New York, 2015, pp. 359–465.
[43] D.G. Champernowne, F.A. Cowell, Economic Inequality and Income Distribution, Cambridge University Press, Cambridge, 1998.
[44] A. Sen, On Economic Inequality, Clarendon Press, Oxford, 1973.
[45] A.F. Shorrocks, The class of additively decomposable inequality measures, Econometrica 48 (3) (1980) 613–625.
[46] A.B. Atkinson, On the measurement of inequality, J. Econ. Theory (1970) 244–263.
[47] N.H. Stern, Welfare weights and the elasticity of the marginal valuation of income, in: M. Artis, R. Nobay (Eds.), Studies in Modern Economic Analysis: Proceedings of Annual Conference of the Association of University Teachers of Economics, Basil Blackwell, Oxford, 1977, pp. 209–257.
[48] F.G. De Maio, Income inequality measures, J. Epidemiol. Community Health 61 (10) (2007) 849–852.
[49] Development Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2017.
[50] D.G. Champernowne, A model of income distribution, Econ. J. 63 (250) (1953) 318–351.
[51] B.C. Arnold, Univariate and multivariate Pareto models, J. Stat. Distrib. Appl. 1 (2014) 11.
[52] Malaysia, Eighth Malaysia Plan 2001–2005, National Printers Malaysia Bhd, Kuala Lumpur, 2001.