Generalizations of pareto distribution with applications to lifetime data

H Haj Ahmad¹, E Almetwally²

¹Department of Basic Science, Preparatory Year Deanship, King Faisal University, Hofuf, Al-Ahsa, 31982, Saudi Arabia,
²Faculty of Business Administration, Delta University of Science and Technology, Egypt.

Corresponding author: hhajahmed@kfu.edu.sa

Abstract. In this paper, we consider five different generalizations of the well-known Pareto distribution. In literature, different generalizations were used so that the newly generated lifetime distributions were more flexible and can be used to model skewed, symmetric, and monotone data. We consider some important characteristics of the generalized Pareto distributions, such as monotonicity and hazard rate function. Comparisons are performed between the different generalizations of Pareto distribution. Modelling real data examples are done using the different goodness of fit tests. Numerical methods are used to conduct the suggested tests, and resulted with the models that are most suitable for describing the behaviour for these data. Finally, the conclusion is given to illustrate the purpose of this work.

1. Introduction

The amount of data in real life are growing increasingly, requiring new statistical distributions that enable us to better describe each phenomenon or experiment under study. Defining these new distributions is a very significant problem in statistics, when a researcher aims to predict more accurate future behavior of the data based on an observed set of data he needs a statistical model that better fit these data. Many attempts have been made by several authors to define new distributions or new families of distributions to provide more flexibility for modeling data under investigation.

In this paper, we consider a famous and continuous life time model function that was first introduced by Vilfredo Pareto (1898), it is called Pareto distribution. We deal with two parameters Pareto Type-I distribution which has a probability density function (pdf) as

\[ f_1(x; \theta, \beta) = \frac{\theta \beta^\theta}{x^{\theta+1}}, \quad x \geq \beta, \]

where \( \theta > 0 \) is the shape parameter and \( \beta > 0 \) is the scale parameter. The survival function of Pareto Type-I distribution is \( \tilde{F}(x; \theta, \beta) = \left( \frac{\beta}{x} \right)^\theta \), Where \( \tilde{F}(.) = 1 - F(.) \) and \( F(.) \) is the cumulative distribution function (cdf).

Many forms of Pareto distribution appeared in the literature, and they were used in a wide range of scientific applications. For instance, it was found that they are compliant in lifetime models such as actuarial sciences, finance, economic, life testing, and climatology, where they usually describe the occurrence of extreme events. The areas of applications are successfully mentioned in several books,
such as those by [1] and [2]. Marshall-Olkin Pareto and Marshall-Olkin semi-Pareto distributions were considered by [3], and time series models with modification structure were studied. In [4] Marshall-Olkin of Pareto distribution was considered, and some of its statistical properties and its hazard rate were obtained. It was also showed that the limiting distributions of the sample extremes were of exponential and Fréchet type. [5] Discussed different point and interval estimation for Marshall-Olkin Pareto distribution.

Since several generalizations appeared in the literature, there is a need to compare between the efficiency of these generalized models and apply them to the different lifetime data that appear in many scientific aspects like medical, engineering, economic, weather phenomena, and others.

In this paper, we aim to study some important and newly developed generalizations of Pareto distribution and compare between their statistical properties and characteristics. Another comparison is conducted by testing the goodness of fit of these generalizations to some real life data examples. Different goodness of fit tests are utilized in order to decide which are more suitable for modelling data and having better forecast for future events.

This paper is organized as follow: In Section 2 we introduce different generalization of Pareto distribution, statistical properties are discussed in Section 3. Real data examples are analyzed in Section 4, while conclusions are given in Section 5.

2. Some generalized model

Several generalized forms of Pareto distribution were discussed in the literature. Generalized Pareto (GP) distribution was first studied by [6], and then it was studied by many authors like [7] and [8]. The GP distribution has two parameters and it was used as a model for excesses over thresholds. Since it generalizes Pareto distribution and others like uniform and exponential, it has wide applications including environmental extreme events, ozone levels in the upper atmosphere, large insurance claims or large fluctuation in financial data, and reliability studies. Its areas of applications are successfully addressed in several books, such as those by [1], [2], and [9]. The pdf of GP distribution is

\[ f_2(x; \alpha, \beta) = \frac{1}{\beta} \left(1 + \frac{\beta}{\alpha} \right)^{-\frac{1}{\beta}} \left(1 + \frac{\beta}{\alpha} x^{-\frac{1}{\beta}} \right)^{-\frac{1}{\beta}}, \quad x \geq 0, \]  

where \(-\infty < \theta < \infty\) is the shape parameter and \(\beta > 0\) is the scale parameter. The survival function of GP distribution is

\[ F_2(x; \alpha, \beta) = \left(1 + \frac{\beta}{\alpha} x^{-\frac{1}{\beta}} \right)^{-1}. \]

Later in (1997) a new method was proposed by [10], their idea of obtaining a new distribution depends on adding a parameter (\(\alpha\)) to the original distribution. The new family of distributions includes the original distributions as special cases, and it gives more flexibility to the original models.

A new generalization of Pareto distribution was considered using the generator Marshall-Olkin, and some of its statistical properties and its hazard rate were discussed, see [4].

He also showed that the limiting distributions of the sample extremes were of exponential and Fréchet type. Recently [5] studied Marshall-Olkin Pareto Type I distribution (MOP) which has three parameters and generalizes Pareto distribution, it acts much better than other distributions when the goodness of tests are conducted for real life data. The pdf of MOP distribution is given by:

\[ f_3(x; \alpha, \beta, \gamma) = \frac{\alpha \beta \gamma x^{-(\alpha+1)}}{\left(1 - \frac{\beta}{\alpha} \left(\frac{\beta}{\gamma} x^{\frac{\alpha}{\gamma}}\right)^{\alpha \gamma} \right)^\alpha}, \quad x \geq \beta, \]

where \(\bar{\alpha} = 1 - \alpha, \beta > 0\) is the shape parameter, \(\beta > 0\) is the scale parameter and \(\alpha > 0\) is an additional Marshall-Olkin parameter. When \(\alpha = 1\) the distribution will reduce to original Pareto distribution. The survival function of MOP distribution is given by

\[ F_3(x; \alpha, \beta, \gamma) = \frac{\alpha \left(\frac{\beta}{\gamma} x^{\frac{\alpha}{\gamma}}\right)^\alpha}{1 - \bar{\alpha} \left(\frac{\beta}{\gamma} x^{\frac{\alpha}{\gamma}}\right)^{\alpha}}, \quad x \geq \beta. \]
Another new Pareto-type distribution (NP) was obtained by [11] and discussed many interesting properties of this new model, the flexibility of this model proved that it is applicable in reliability analysis and income data. NP distribution has two parameters and its pdf is

\[ f_1(x; \theta, \beta) = \frac{2 \theta^\beta x^{\theta-1}}{(x^\theta + \beta^\theta)^2}, \quad x \geq \beta, \]  

(4)

where \( \theta > 0 \) is the shape parameter and \( \beta > 0 \) is the scale parameter, and its survival function is given by \( F_1(x; \theta, \beta) = \frac{2 \theta^\beta x^\theta}{x^\theta + \beta^\theta} \). In [12] a new generalization of (MOP) was developed which was called Marshall-Olkin generalized Pareto denoted by (MOGP), this distribution has three parameters and it is a generalization of many other distributions such as: (GP), exponential, uniform, Pareto Type-I and Marshall-Olkin exponential distribution. The pdf and survival function are respectively written as

\[ f_5(x; \alpha, \theta, \beta) = \frac{a}{\beta} \left(1 + \frac{\theta x^\beta}{\beta}ight)^{-\frac{1}{\beta}-1} \frac{1}{\alpha-1} \left(1 - \frac{\theta x^\beta}{\beta}ight)^{1-\frac{1}{\beta}-1} (1 - \frac{\theta x^\beta}{\beta}), \quad x \geq 0, \]  

(5)

where \( \alpha > 0, \theta > 0 \) and \( \beta > 0 \), and \( F_5(x; \alpha, \theta, \beta) = \frac{a(1 + \frac{\theta x^\beta}{\beta})^{-\frac{1}{\beta}}}{1 - \frac{a(1 + \frac{\theta x^\beta}{\beta})^{-\frac{1}{\beta}}}{\alpha-1}} \).

Another new method of obtaining a generalized model is the alpha power transformation family, see [13]. Later on, [14] used this method to obtain an efficient and attractive generalization of Pareto distribution namely, Marshall-Olkin Alpha Power Pareto (MOAPP) distribution. This model owns many nice properties and proved to be a better fit for medical and biological examples. The pdf and survival function of MOAPP distribution are given by

\[ f_6(x; \alpha, \theta, \beta) = \frac{\theta^\beta (\log a)^{\alpha-1-x^{-\beta}}}{(a-1)x^\beta (1 + \theta(\alpha-1)(1-x^{-\beta})^{-1})^2}, \quad x \geq 1 \]  

(6)

where \( \alpha > 0, \theta > 0 \) and \( \beta > 0 \), and \( F_6(x; \alpha, \theta, \beta) = 1 - \frac{\alpha^{1-x^{-\beta}-1}}{(a-1)[\theta + (1 - \theta)(a-1)^{-1}(a^{1-x^{-\beta}} - 1)]}, \)

respectively.

Readers can find many other generalized forms of Pareto distribution, but we consider only the above five generalizations in this research study. Several statistical functions, density properties and some figures are introduced and illustrated in Section 3. Among these are monotonicity of the pdf its hazard rate function.

3. Characteristics of new models

Pareto Type-I distribution and the five generalizations that we introduced in Section 2 have many characteristics related to the monotonicity of their pdf and the behaviour of the hazard functions. For detailed review of Pareto distribution and related topics, one may refer to [15] and [6, 7, 11, 12, 14], and we summarize these properties in sections 3.1 and 3.2 for better comparison between the distributions.

3.1. Monotonicity

Consider the pdfs in Equations (1) to (6), using some algebraic manipulations, one can prove the monotonicity behaviour the density functions of Pareto, GP, MOP, NP, MOGP, and MOAAP, see [6, 7, 11, 12, 14]. With the help of graphics command in Mathematica 9, we can plot the density function for the six distributions, the behaviour of their curves indicates which kind of data can be modelled by these distributions. Figure 1 (left side) illustrates the pdf curves for Pareto, GP, MOP, NP, MOGP, MOAAP distributions respectively.

We may classify these density monotonicity as:
a) Decreasing: Pareto Type-I, GP, and NP distributions have decreasing pdf for all values of its parameters. Hence these distributions are recommended for modelling decreasing data.

b) Non-monotone: MOP, MOGP, and MOAPP have non-monotone pdfs with unimodal, and right skewness or symmetry curves. For some special values of their parameters also they can be decreasing. Hence these distributions are recommended for modelling right skewed or symmetrical data.

3.2. Hazard Rate function

Hazard rate function or failure rate is an important statistical function that is mainly used in survival analysis and reliability theory. The behaviour and the shape of its function curve play a basic role in lifetime failures and real data experiments. The hazard rate function is defined as $h(t) = \frac{f(t)}{1-F(t)}$. For the Pareto and its generalizations, we can easily compute their hazard rate functions and hence can show their behaviour with respect to different parameter values. Using plot command is much helpful to get a clear view on the monotonicity of $h(t)$. See Figure 1 (right side) for the hazard rate plot.
Figure 1. Plots of density functions (left) and hazard function (right) of GP, MOP, NP, MOGP, and MOAPP distributions under different values of the parameters.

We may classify the monotonicity of hazard rate curve as:

a) Decreasing: This is the case for Pareto and GP distributions, where the curve of hazard rate is decreasing for all values of their parameters. Usually, decreasing failure rates occur during the early life or wear out of population units.

b) Non-monotone: For MOP, NP, MOGP, MOAPP, the hazard rate curve is upside down bathtub with unimodal for some value of parameters, and it can be decreasing for specific parameters values. Non-monotone failure rate usually appears in the aging process.

4. Data analysis

In this section, we provide several examples to illustrate the flexibility of the Pareto distribution and its generalizations for data modeling purposes. Modeling is of great importance in many fields, its basic usefulness appears in determining the trend of data in future time periods. This leads to better prediction and forecasting. The goodness of fit tests is a powerful tool used to check the suitability of the distribution to a certain given real data, with the null hypothesis indicating that the data follows $f_i(x; \Phi)$, with $i=1,\ldots,6$, and a significant level of 0.05. The goodness of tests used in this context are:

- Kolmogorov-Smirnov distance (KS) between the fitted and the empirical distribution functions and the corresponding p-values.
- Akaike information criterion (AIC) such that $AIC = -2L + 2q$, where $q$ is the number of parameters in the model and $L$ is the maximized value of the likelihood function for the model.
- Bayesian information criterion (BIC) is also used for comparison between models where BIC can be defined as $BIC = -2L + q\ln(n)$, where $n$ is the sample size.
- Consistent Akaike information criteria (CAIC), defined by $CAIC = -2L + q(\ln(n) + 1)$.
- Hannan-Quinn information criterion (HQIC) such that $HQIC = -2L + 2qln(\ln(n))$.

As a model selection criterion, the researcher should select the model with the minimum KS, AIC, BIC, CAIC, and HQIC. Also, the value of $p$ will give a good answer by which we decide to select a certain model. The following examples are taken from different scientific fields. Hence we try to consider the most important and newly available real data such as, the index of income for the last two years in Brazil, some medical data, the latest records of active cases of COVID-19, annual rainfall in Belgium, and finally an engineering example about failure times. Data analysis is performed using R code, see [16].
4.1. Income data

Index of income is an important economic indicator, it is used to measure the growth estimates for the agricultural, industrial, and service sectors. See [17]. Modelling the index of income plays an important role in predicting and estimating the future growth of economic sectors. The following data are recorded from the central bank of Brazil, and the index of income was monthly observed from January 2019 to June 2020, these data are available at [18].

Table 1 below shows the index values, and Table 2 describes the goodness of fit test of the six models to this data.

| Table 1. Index of income from Jan 2019 to Jun 2020. |
|---------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|
| Jan 2019 | Feb 2019 | Mar 2019 | Apr 2019 | May 2019 | Jun 2019 |
| 133.87   | 133.51   | 138.54   | 138.85   | 135.15   | 142.9    | 142.24   | 138.76   |

| Table 2. Goodness of fit criterion for index of income data |
|---------------------------------|-----------|-----------|-----------|-----------|
|                                   | KS        | P-Value   | AIC       | BIC       |
| P                                 | 0.4245    | 0.0027    | 135.7126  | 137.3790  |
| MOP                               | 0.1633    | 0.6962    | 120.9147  | 123.4144  |
| GP                                | 0.7606    | 0.0000    | 233.3090  | 234.9754  |
| MOGP                              | 0.1593    | 0.7241    | 113.4613  | 115.9610  |
| MOAPP                             | 0.5448    | 0.0000    | 206.5806  | 209.0803  |
| NP                                | 10.1713   | 0.0000    | 131.3795  | 133.0459  |

From Table 2 we can realize the efficiency of using MOGP to model the index of income, the reason is that the minimum values of KS, AIC, BIC, CAIC, and HQIC are achieved under MOGP distribution, also the p-value is 0.7241, which is greater than 0.05 hence the null hypotheses is not rejected which indicated the suitability of MOGP to model this data. We may also notice a second preferable model which is MOP since its p-value is also greater than 0.05. The idea of modelling Index of income by the goodness of fit test of Pareto and its generalizations is new in this area, and its results may lead to better prediction and forecast for the future growth of income.

4.2. Medical data

In this example, we consider a medical data set that contains the relief of 20 patients receiving analgesic, the data set can be found in [19].

| Table 3. Goodness of fit criterion for patients’ relief |
|---------------------------------|-----------|-----------|-----------|-----------|
|                                   | KS        | P-Value   | AIC       | BIC       |
| P                                 | 0.2223    | 0.3047    | 33.5667   | 35.4556   |
| MOP                               | 0.1008    | 0.9904    | 31.7814   | 34.6147   |
| GP                                | 0.3215    | 0.0394    | 57.3696   | 59.2585   |
| MOGP                              | 0.1186    | 0.9521    | 37.6340   | 40.4673   |
| MOAPP                             | 0.1124    | 0.9699    | 34.2692   | 37.1025   |
| NP                                | 2.3433    | 0.0000    | 41.0752   | 42.9641   |

From Table 3 we see that MOP is the most suitable distribution amongst other distributions, it has the minimum value of all goodness of fit test methods, also its p-value support our choice of MOP distribution to model patients’ relief. MOGP and MOAPP are good alternatives for modelling purpose as well, since their p-values are greater than 0.05.
4.3. COVID-19 statistics
Active cases are recorded for COVID-19 in China starting from January 2020 until July 2020, we consider the extreme values of active cases obtained weekly from [20]. Our sample consists of 30 extreme value, and Table 4 summarizes the results of the goodness of fit test procedures in order to select the best distribution that can model these data.

Table 4. Goodness of fit criterion for active cases of COVID-19

|       | KS   | P-Value | AIC     | BIC     | HQIC   | CAIC   |
|-------|------|---------|---------|---------|--------|--------|
| P     | 0.1463 | 0.5174  | 541.3296 | 544.0642 | 542.1860 | 541.7911 |
| MOP   | 0.1003 | 0.9042  | 541.2203 | 543.9322 | 541.5050 | 541.1803 |
| GP    | 0.9052 | 0.0000  | 689.3466 | 692.0812 | 690.2030 | 689.8081 |
| MOGP  | 0.0949 | 0.9345  | 539.7892 | 543.8911 | 541.0738 | 540.7492 |
| MOAPP | 0.0945 | 0.9392  | 539.6083 | 543.7102 | 540.8930 | 540.5683 |
| NP    | 0.5713 | 0.0000  | 529.3981 | 532.1327 | 530.2545 | 529.8596 |

From results in Table 4, we conclude that MOAPP is the best fit for modelling COVID-19 active cases in China, while the second and third candidates are MOGP and MOP, respectively.

4.4. Annual rain fall
In this example, we consider the annual maximum daily rainfall for Uccle (Brussels, Belgium) for the 50-year period from 1934 to 1983, see [21]. Goodness of tests are performed to check the efficiency of our models. From Table 5 we conclude that MOP distribution has the minimum values of KS, AIC, BIC, HQIC, and CAIC, with p-value of 0.9856 which is greater than 0.05. In this case we said that MOP is better to model these data. MOGP and MOAPP are of second priority for modelling rainfall data.

Table 5. Goodness of fit criterion for annual maximum rain fall

|       | KS   | P-Value | AIC     | BIC     | HQIC   | CAIC   |
|-------|------|---------|---------|---------|--------|--------|
| P     | 0.2374 | 0.0071  | 411.8670 | 415.6910 | 413.3232 | 412.1223 |
| MOP   | 0.0644 | 0.9856  | 388.5188 | 394.2548 | 390.7031 | 389.0405 |
| GP    | 0.3394 | 0.0000  | 437.8937 | 441.7178 | 439.3499 | 438.1490 |
| MOGP  | 0.0871 | 0.8426  | 395.5361 | 401.2722 | 397.7204 | 396.0579 |
| MOAPP | 0.0831 | 0.8798  | 393.8740 | 399.6101 | 396.0583 | 394.3957 |
| NP    | 1.4411 | 0.0000  | 401.0533 | 404.8773 | 402.5095 | 401.3086 |

4.5. Failure time data
This real data are mentioned by [22], which represents the data set of failure times, they refer to the fatigue times of 6061-T6 Aluminium coupons. The data set consists of 101 observations with maximum stress per cycle 26,000 psi. In Table 6 the values of KS statistic with p-value, AIC, CAIC, BIC and HQIC are reported. When comparing these values between Pareto and the generalized forms, we obtain the minimum KS, AIC, BIC, HQIC and CAIC for MOGP and MOP. The related p-values of MOGP and MOP also show that our decision is not to reject the null hypothesis. Therefore, this indicates that the MOGP and MOP distribution fits the data set well and better than other distributions. This is also proves the needs of new distributions in managing some sets of data. So in general we can tell that the generalized distributions is superior according to other sub models. This coincide with the monotonicity of MOP and MOGP so that these distributions are better fit for non-monotone data models.
Table 6. Goodness of fit criterion for failure time data

|      | KS    | P-Value | AIC    | BIC    | HQIC   | CAIC   |
|------|-------|---------|--------|--------|--------|--------|
| P    | 0.63304 | 2.20e-16 | 1625.145 | 1630.375 | 1627.262 | 1625.267 |
| MOP  | 0.0572  | 0.896   | 1504.308 | 1507.484 | 1504.556 | 1512.154 |
| GP   | 0.9816  | 2.20e-16 | 2720.843 | 2726.073 | 2722.961 | 2720.966 |
| MOGP | 0.054   | 0.927   | 1500.158 | 1508.003 | 1503.334 | 1500.405 |

5. Conclusion
In this paper, we considered five new generalizations of Pareto distribution. The characteristics and statistical properties of these generalized forms act well and can be used to model different kinds of real data. The monotonicity and hazard rate function curves showed that the generalized forms of Pareto are more flexible and can model skewed, symmetry, and monotone data. Hence we used these generalizations to model different real data examples, this was performed using different goodness of test methods. It was concluded that MOGP is the best distribution to model the index of income in Brazil and modelling the fatigue times of Aluminium coupons, while MOP is preferable to model patients’ relief time and annual rainfall. It was also concluded that MOAPP is the best fit for modeling COVID-19 active cases in China. Modeling different real data example may lead to better forecast and predict either missing values or future data values. Therefore we recommend further work on creating new generalized forms of Pareto and other well-known distributions and obtain a better fit for modeling real data examples.

Acknowledgments
The authors would like to thank the editor and the referees for their careful reading and for their comments which greatly improved the paper.

References
[1] Castillo E, Hadi A S, Balakrishnan N and Sarabia J M 2005 Extreme Value and Related Models with Applications in Engineering and Science (Hoboken, NJ: Wiley) MR2191401
[2] Kotz S and Nadarajah S 2000 Extreme Value Distributions: Theory and Applications (London: Imperial College Press)
[3] Alice T and Jose K K 2004 IAPQR Trans 29 1-9.
[4] Ghitany M E 2005 J. of Appl. Math. 18 1 17-31
[5] Bdair O and Haj Ahmad H 2019 Revista Investigacion Operacional Forth Comming Paper #60E05 18-01.
[6] Pickands J 1975 Ann. of Stat. 3 119—131.
[7] Gupta R C, Gupta R D and Gupta P L 1998 27 887-904.
[8] Hogg R V, McKean J W and Craig A T 2005 Introduction to Mathematical Statistics, 6th ed. (New Jersey: Pearson Prentice-Hall)
[9] Ahsanullah M 1992 Inference and prediction problems of the generalized Pareto distribution based on record values, In: P.K. Sen and L.A. Salama (Eds.), Order Statistics and Nonparametrics: Theory and Applications, (Amsterdam: Elsevier Science Publishers) 47-57
[10] Marshall A W and Olkin I 1997 Biometrika 84 (3) 641-652.
[11] Bourguignon M, Saulo H and Fernandez R, 2016 Physica A, DOI: http://dx.doi.org/10.1016/j.physa.2016.03.043
[12] Haj Ahmad H and Almetwally E M 2020 Pakistan J. of Stat. and Oper. Res. 16 1 21-33
[13] Nassar M, Kumar, D, Sanku Dey S, Gauss M, Cordeiro G, Afify A Z, 2019 J. of Comput. and Appl. Math. 351 41–53
[14] Almetwally, E. and Haj Ahmad H 2020 Statistics in Transition New Series 21 5 61-84 DOI: https://doi.org/10.21307/stattrans-2020-055
[15] Arnold B C 2015 Pareto Distribution, 2 nd edition (John Wiley & Sons, Ltd.)
[16] http://www.R-project.org/
[17] https://www.mql5.com/en/economic-calendar/brazil/bcb-ibc-br-economic-activity
[18] https://www3.bcb.gov.br/
[19] Nadarajah S, Bakouch H S and Tahmasbi R 2011 A generalized Lindley distribution (UK: Technical Report, School of Mathematics, University of Manchester)
[20] http://www.worldometers.info
[21] Demaree G 1985 Intensity-Duration-Frequency Relationship of Point * Precipitation at Uccle, Reference Period 1934–1983 (Institut Royal Meteorologique de Belgique, Publications Series A) 116 1–52
[22] Birnbaum Z W, Saunders S C 1969 Journal of Appl. Probab. 6 2, 328-347