Modeling COVID-19 Pandemic Data with New Pareto Model

.

ABSTRACT

This paper aims to find a statistical model for the COVID-19. We used an efficient and superior model for fitting the COVID 19 mortality. A new Extended-Pareto distribution (NE-P) is found. The maximum likelihood method is utilized to obtain the estimator of the parameters. A simulation was carried out using different sample sizes and different values of the parameters. In addition, the goodness of fit test statistics was calculated for proposed model compared with the baseline model to find out that our new model is the best for modeling data COVID-19.

Keywords: *A new Extended-Pareto distribution, COVID 19 mortality, The maximum likelihood method, Goodness of fit.*

1 Introduction

Probability distributions are often used in reliability tests for data modeling which is an important aspect of statistical work in a wide range of scientific and technological fields. Over the last decades, a considerable amount of research was devoted to the creation of lifetime models and several generalization or transformation methods have been proposed. These generalizations are obtained either by adding one or more shape parameters or changing the functional form of distributions to make them more flexible for modeling many lifetime data. One of the most important of these distributions is Pareto distribution (PD) which has attracted the attention of researchers for example, Jayakumar et al. [1] presented a new four parameter distribution called New

Generalized Pareto distribution, which was a generalization of the classical Pareto distribution. Boumaraf et al. [2] used nonlinear optimization methods to find the estimators of beta Pareto distribution.

The probability density function (PDF) of Pareto distribution is given by:

$$g(x) = \frac{\beta}{x^{\beta+1}} \;;\; x \geq 1 \;; \beta > 0 \,. \tag{1}$$

and the cumulative distribution function (CDF) is:

$$G(x) = 1 - x^{-\beta} \;;\; x \geq 1 \,,\; \beta > 0. \tag{2}$$

where β is the scale parameter.

Furthermore, there were many authors who were interested to introduce new models: Ibrahim et al. [3], Bantan et al. [4], in [5], the authors studied a new extended (NE-X) family of distributions. The PDF of NE-X distribution is given by:

$$f(x) = \frac{2\theta^2 g(x) G(x) \left(1-G^2(x)\right)^{\theta-1}}{\left(1-(1-\theta)G^2(x)\right)^{\theta+1}};\; \theta > 0 \;;\; x \in R \,. \tag{3}$$

The CDF of NE-X distribution is:

$$F(x) = 1 - \left(\frac{1-G^2(x)}{1-(1-\theta)G^2(x)}\right)^{\theta} \;;\; \theta > 0;\; x \in R. \tag{4}$$

Depending on Equations (1) , (2) and the family in [5], we deduced a new distribution called it a New Extended-Pareto Distribution (NE-P). The PDF and CDF of NE-P distribution with two parameters(θ, β) is obtained respectively as:

$$f(x) = \frac{2\theta^2 (\beta x^{-(\beta+1)})(1-x^{-\beta})\left(1-(1-x^{-\beta})^2\right)^{\theta-1}}{\left(1-(1-\theta)(1-x^{-\beta})^2\right)^{\theta+1}};\; \theta > 0,\; \beta > 0 \;;\; x \in R. \tag{5}$$

And

$$\mathrm{F}\,(x) = 1 - \left(\frac{1-(1-x^{-\beta})^2}{1-(1-\theta)(1-x^{-\beta})^2}\right)^{\theta} \;;\; \theta > 0,\; \beta > 0 \;;\; x \in R \tag{6}$$

2 Maximum Likelihood Estimation Method

This section presents the maximum likelihood estimator of the NE-P distribution parameters (θ, β). If $x_1, x_2, \ldots, x_n$ is a random sample from NE-P distribution, the log-likelihood function is $L(\underline{x})$ can be obtained as:

$$L(\underline{x}) = (2\theta^2\beta)^n \prod_{i=1}^{n} x_i^{-(\beta+1)} \prod_{i=1}^{n}\left(1 - x_i^{-\beta}\right) \prod_{i=1}^{n}\left(1-\left(1-x_i^{-\beta}\right)^2\right)^{\theta-1}$$
$$\times \prod_{i=1}^{n}\left(1-(1-\theta)(1-x_i^{-\beta})^2\right)^{-(\theta+1)}. \qquad (7)$$

And the log-likelihood function is given as follows

$$l = \log(L) = nlog(2) + n\log(\theta^2) + n\log(\beta) - (\beta+1)\sum_{i=1}^{n}\log x_i$$
$$+\sum_{i=1}^{n}\log\left(1-x_i^{-\beta}\right) + (\theta-1)\sum_{i=1}^{n}\log\left(1-\left(1-x_i^{-\beta}\right)^2\right)$$
$$-\ (\theta+1)\sum_{i=1}^{n}\log\left(1-(1-\theta)\left(1-x_i^{-\beta}\right)^2\right). \qquad (8)$$

Differentiating (8) with respect to each of the parameters θ and β gives

$$\frac{\partial\, log\, L}{\partial\theta} = \frac{2n}{\theta} + \sum_{i=1}^{n}\log\left(1-\left(1-x_i^{-\beta}\right)^2\right) - (\theta+1)$$
$$\times\sum_{i=1}^{n}\frac{\left(1-x_i^{-\beta}\right)^2}{\left(1-(1-\theta)\left(1-x_i^{-\beta}\right)^2\right)} - \sum_{i=1}^{n}\log\left(1-(1-\theta)\left(1-x_i^{-\beta}\right)^2\right), \qquad (9)$$

$$\frac{\partial\, log\, L}{\partial\beta} = \frac{n}{\beta} - \sum_{i=1}^{n}\log x_i + \sum_{i=1}^{n}\frac{x_i^{-\beta}\ln x_i}{1-x_i^{-\beta}} - 2(\theta-1)$$
$$\times\sum_{i=0}^{n}\frac{x_i^{-\beta}ln(x_i)(1-x_i^{-\beta})}{1-\left(1-x_i^{-\beta}\right)^2} + \ 2(1-\theta^2)\sum_{i=0}^{n}\frac{x_i^{-\beta}ln(x_i)(1-x_i^{-\beta})}{\left(1-(1-\theta)\left(1-x_i^{-\beta}\right)^2\right)}. \qquad (10)$$

There isn't a closed form answer to this equations for (θ, β) . As a result, the equations can be solved numerically using a method like the Newton-Raphson method to determine the Maximum Likelihood Estimate $\hat{\theta}_{MLE}$ and $\hat{\beta}_{MLE}$.

3 Simulation Study

In this section, the simulation result for the ML method is given when two parameters are unknown based on complete samples. for various sample sizes and parameter values, the parameter values are selected as β = 2.8,θ=1.5and n = 25, 100, 250, 450, and 1000 This process is repeated N = 500 times . Furthermore, performance of different estimators is considered in terms of their biases and mean square errors (MSEs) that given, respectively, by

$$Bias(\hat{\lambda}) = E(\hat{\lambda}) - \lambda \quad \text{and} \quad \text{MSE}(\hat{\lambda}) = \text{E}(\hat{\lambda} - \lambda)^2.$$

Table 1: Mean, MSEs and Bias for the parameter estimates when β_0 = 2.8, θ_0=1.5.

N		MLE	MSE	BIAS
25	$\hat{\beta}$	3.37488	0.76493	0.574879
	$\hat{\theta}$	1.67533	0.627229	0.17533
100	$\hat{\beta}$	3.00401	0.638085	0.204005
	$\hat{\theta}$	1.67228	0.622801	0.17228
250	$\hat{\beta}$	3.00453	0.27286	0.204525
	$\hat{\theta}$	1.579	0.185657	0.0790023
450	$\hat{\beta}$	2.94519	0.339871	0.145187
	$\hat{\theta}$	1.55915	0.151468	0.059146
1000	$\hat{\beta}$	2.89381	0.069541	0.0938141
	$\hat{\theta}$	1.53192	0.05152	0.0319226

We can see when sample size increases, the mean squared error and bias decrease. Therefore, the maximum likelihood method works very well to estimate the parameters of NE-P distribution.

4 Real Data Applications

In this section, we provide the application with real data sets to assess the flexibility of NE-P distribution comparing with the base line Pareto distribution. The parameters are estimated using maximum likelihood method. Mathematica (V.11.0) is used for computation. Moreover, we consider the model selection criteria, including Akaike information criterion (AIC), Bayesian information criterion (BIC), consistent Akaike information criterion (CAIC), and Hannan-Quinn information criterion (HQIC). They are defined as follows:

$$AIC = -2l\left(\hat{\theta}\right) + 2k$$

$$CAIC = AIC + \frac{2k(k+1)}{n-k-1}$$

$$HQIC = -2logl\left(\hat{\theta}\right) + 2k\log(\log(n))$$

(See Whittaker and Furlow [6]).

5 Daily mortality cases of COVID-19:

In this section, we will study the data number of daily mortality COVID-19 cases will be compared with (PD).

5.1 Data set 1:

The first data represents a COVID-19 mortality rates data belongs to Italy of 59 days, that is recorded from 27 February to 27 April 2020. The data is taken from (Almongy et al. [7]) as follows:

4.571, 7.201, 3.606, 8.479, 11.410 ,8.961, 10.919, 10.908, 6.503 ,18.474, 11.010 ,17.337, 16.561, 13.226, 15.137 ,8.697 ,15.787 ,13.333 ,11.822 ,14.242 ,11.273, 14.330, 16.046, 11.950, 10.282, 11.775 ,10.138 ,9.037 ,12.396 ,10.644, 8.646 ,8.905, 8.906, 7.407, 7.445, 7.214, 6.194, 4.640 ,5.452 ,5.073, 4.416, 4.859 ,4.408 ,4.639 ,3.148 ,4.040 ,4.253 ,4.011 ,3.564, 3.827, 3.134, 2.780, 2.881, 3.341, 2.686, 2.814, 2.508, 2.450 ,1.518.

Table 2: Descriptive statistics for data set 1.

n	Min	Q_1	Median	Mean	Q_3	Max	Skewness	Kurtosis
59	1.518	4.04	7.44	8.15	11.41	18.474	0.45	2.12

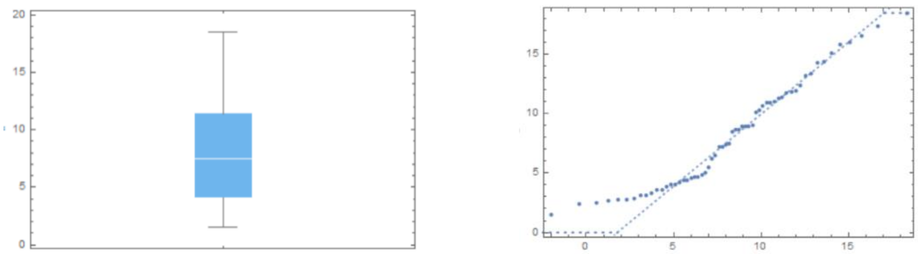


Figure 1. PP plot of the NE-P distribution and the box plot for data set 1.

Table 3. Parameter estimation for various distributions depending on data set 1.

Model	Parameters		LL	AIC	CAIC	HQIC
	$\hat{\theta}$	$\hat{\beta}$				
NE-P	0.31	2.31	-190.47	384.94	385.15	386.56
PD		0.52	-211.25	424.50	424.57	425.31

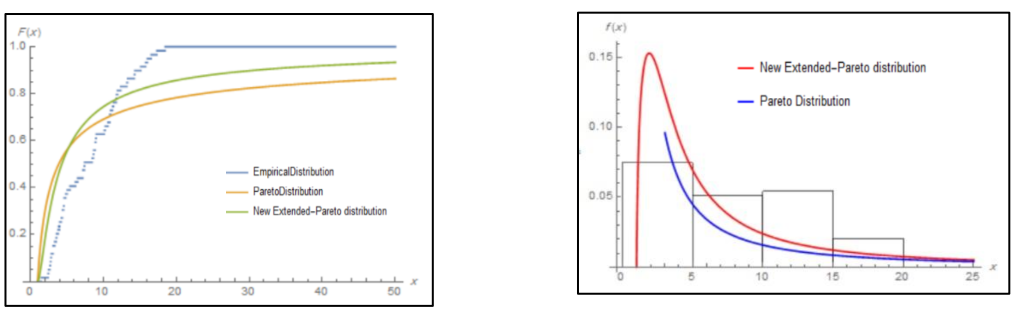


Figure 2. Plots of the fitted CDF (left) and the histogram with fitted PDF (right) of the NE-P model for data set 1.

5.2 Data set 2:

The second data set of data of COVID-19 mortality numbers in Mexico of 108 days, that is recorded from 4 March to 20 July 2020. The data is taken from Nagy et al. The data is taken from (Almongy et al. [7]) as follows:

8.826, 6.105, 10.383, 7.267, 13.220, 6.015, 10.855, 6.122, 10.685, 10.035, 5.242, 7.630, 14.604, 7.903, 6.327, 9.391, 14.962, 4.730, 3.215, 16.498, 11.665, 9.284, 12.878, 6.656, 3.440, 5.854, 8.813, 10.043, 7.260, 5.985, 4.424, 4.344, 5.143, 9.935, 7.840, 9.550, 6.968, 6.370, 3.537, 3.286, 10.158, 8.108, 6.697, 7.151, 6.560, 2.988, 3.336, 6.814, 8.325, 7.854, 8.551, 3.228, 3.499, 3.751, 7.486, 6.625, 6.140, 4.909, 4.661, 1.867, 2.838, 5.392, 12.042, 8.696, 6.412, 3.395, 1.815, 3.327, 5.406, 6.182, 4.949, 4.089, 3.359, 2.070, 3.298, 5.317, 5.442, 4.557, 4.292, 2.500, 6.535, 4.648, 4.697, 5.459, 4.120, 3.922, 3.219, 1.402, 2.438, 3.257, 3.632, 3.233, 3.027, 2.352, 1.205, 2.077, 3.778, 3.218, 2.926, 2.601, 2.065, 1.041, 1.800, 3.029, 2.058, 2.326, 2.506, 1.923.

Table 4: Descriptive statistics for data set 2.

n	Min	Q_1	Median	Mean	Q_3	Max	Skewness	Kurtosis
108	1.041	3.23	5.19	5.75	7.48	16.498	0.98	3.68

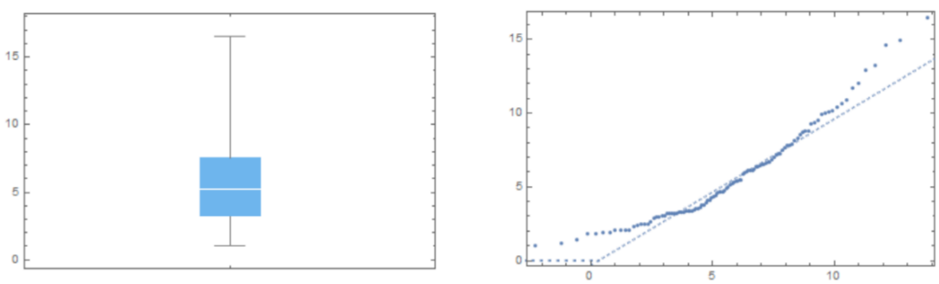


Figure 3. PP plot of the NE-P distribution and the box plot for data set 2.

Table 5. Parameter estimation for various distributions depending on data set 2**.**

Model	Parameters		LL	AIC	CAIC	HQIC
	$\hat{\theta}$	$\hat{\beta}$				
NE-P	0.33	2.65	-296.84	597.69	597.81	599.87
PD		0.63	-329.82	661.64	661.68	662.73

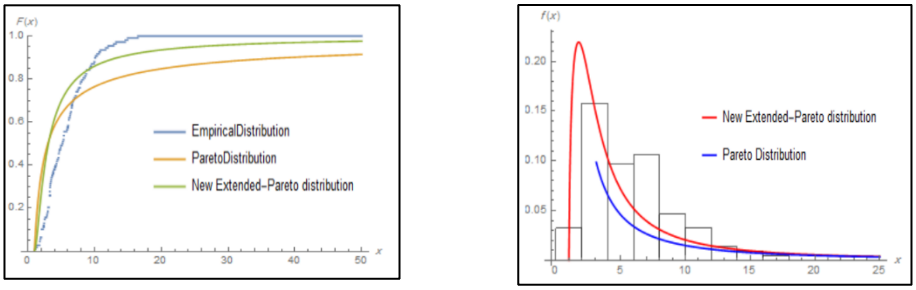


Figure 4. Plots of the fitted CDF (left) and the histogram with fitted PDF (right) of the NE-P model for data set 2.

5.3 Data set 3:

The third data set of a COVID-19 data belonging to the Netherlands of 30 days, that is recorded from 31 March to 30 April 2020. This data formed of rough mortality rate. (see Almongy et al. [7]) The data are as follows: 14.918, 10.656, 12.274 ,10.289, 10.832, 7.099, 5.928, 13.211, 7.968, 7.584, 5.555, 6.027, 4.097, 3.611, 4.960, 7.498, 6.940, 5.307, 5.048, 2.857, 2.254, 5.431, 4.462, 3.883, 3.461, 3.647, 1.974, 1.273, 1.416, 4.235.

Table 6: Descriptive statistics for data set 3.

n	Min	Q_1	Median	Mean	Q_3	Max	Skewness	Kurtosis
30	1.273	3.64	5.36	6.15	7.58	14.918	0.83	2.95

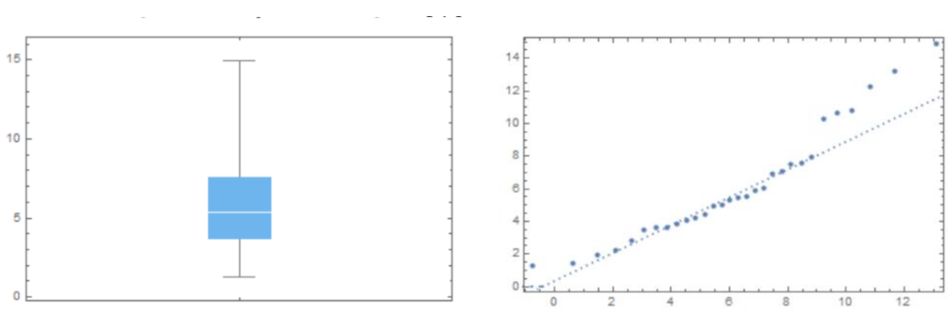


Figure 5. PP plot of the NE-P distribution and the box plot for data set 3.

Table 7. Parameter estimation for various distributions depending on data set 3 .

Model	Parameters		LL	AIC	CAIC	HQIC
	$\hat{\theta}$	$\hat{\beta}$				
NE-P	0.37	2.32	-85.46	174.93	175.38	175.83
PD		0.60	-94.38	190.76	190.91	191.21

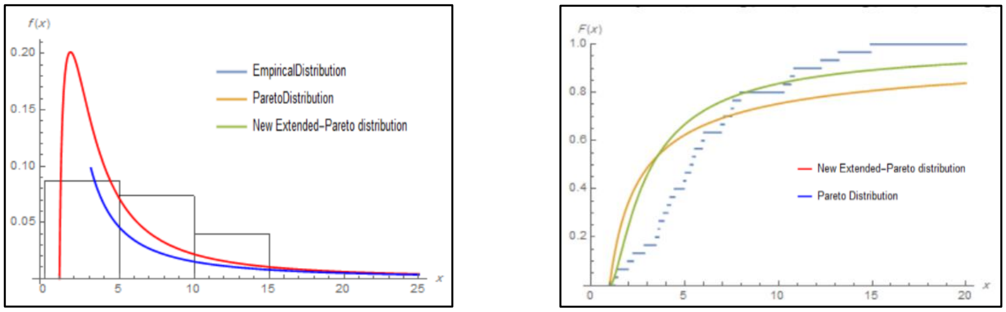


Figure 6. Plots of the fitted CDF (left) and the histogram with fitted PDF (right) of the NE-P model for data set 3.

5.4 Data set 4:

The fourth data set represents a COVID-19 data belong to Canada of 36 days, from 10 April to 15 May 2020 (see Almetwally et al. [8]). These data formed of mortality rate. The data are as follows:
3.1091, 3.3825, 3.1444, 3.2135, 2.4946, 3.5146, 4.9274, 3.3769, 6.8686, 3.0914, 4.9378, 3.1091, 3.2823, 3.8594, 4.0480, 4.1685, 3.6426, 3.2110, 2.8636, 3.2218, 2.9078, 3.6346, 2.7957, 4.2781, 4.2202, 1.5157, 2.6029, 3.3592, 2.8349 ,3.1348, 2.5261, 1.5806, 2.7704, 2.1901, 2.4141, 1.9048.

Table 8: Descriptive statistics for data set 4.

n	Min	Q_1	Median	Mean	Q_3	Max	Skewness	Kurtosis
36	1.5171	2.77	3.17	3.28	3.63	6.8686	1.21	6.15

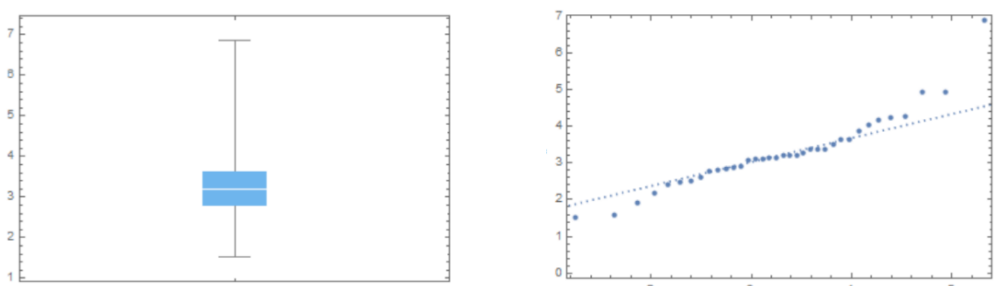
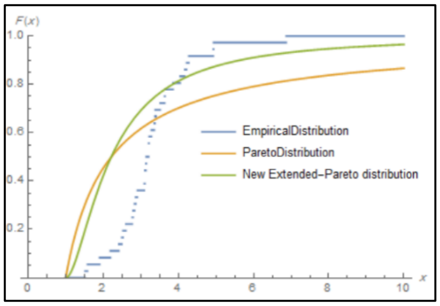


Figure 7. PP plot of the NE-P distribution and the box plot for data set 4.

Table 9. Parameter estimation for various distributions depending on data set 4**.**

Model	Parameters		LL	AIC	CAIC	HQIC
	$\hat{\theta}$	$\hat{\beta}$				
NE-P	0.32	3.75	-68.20	140.41	140.77	141.51
PD		0.87	-82.13	166.27	166.39	166.82



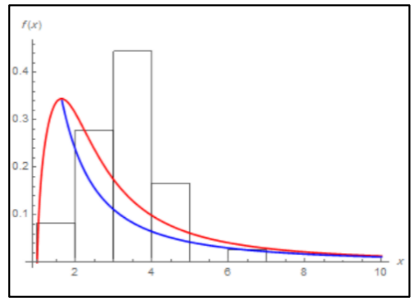


Figure 8. Plots of the fitted CDF (left) and the histogram with fitted PDF (right) of the NE-P model for data set 4.

From Tables 3, 5, 7 and 9, the values of log-likelihood (LL), AIC, CAIC and HQIC are minimum and favorable of NE-P distribution than PD distribution, which indicate that our new model is the best comparing with the competing model.

6 Conclusion

We derive a two parameter New Extended -Pareto distribution. The parameters are estimated by method of maximum likelihood. Performances of MLE are tested through simulation study. Finally, four real data applications of COVID-19 are analyzed in to assess the flexibility of new model. The NE-P distribution is better than the PD in modeling data based on the goodness of fit measures.

7 Refrences:

[1] Jayakumar, K., Krishnan, B. and Hamedani, G. G. (2018). On New Generalization of Pareto Distribution and its Applications. *Communications in Statistics-Simulation and Computation*, *49*(5), 1264-1284.

[2] Boumaraf, B., Seddik-Ameur, N. and Barbu, V. S. (2020). Estimation of Beta-Pareto Distribution Based on Several Optimization Methods. *Mathematics*, *8*(7), 1-12.

[3] Sule, I., Sani, I. D., Audu, I., & Jibril, H. M. (2020). On the Topp Leone exponentiated-G family of distributions: properties and applications.

[4] Bantan, R. A., Jamal, F., Chesneau, C., & Elgarhy, M. (2020). Type II Power Topp-Leone generated family of distributions with statistical inference and applications. *Symmetry*, *12*(1), 75.

[5] Zichuan, M., Hussain, S., Iftikhar, A., Ilyas, M., Ahmad, Z., Khan, D. M. and Manzoor, S. (2020). A new extended family of distributions:

properties and applications. *Computational and mathematical methods in medicine*, 2020, 1-13.

[6] Whittaker, T. A., & Furlow, C. F. (2009). The comparison of model selection criteria when selecting among competing hierarchical linear models. *Journal of Modern Applied Statistical Methods*, *8*(1), 15.

[7] Almongy, H. M., Almetwally, E. M., Aljohani, H. M., Alghamdi, A. S., & Hafez, E. H. (2021). A new extended Rayleigh distribution with applications of COVID-19 data. *Results in Physics*, *23*, 104012.

[8] Almetwally, E. M., Alharbi, R., Alnagar, D., & Hafez, E. H. (2021). A new inverted topp-leone distribution: applications to the COVID-19 mortality rate in two different countries. *Axioms*, *10*(1), 25.