ISSN: 0067-2904

Statistical Analysis of COVID-19 Data in Iraq

Tasnim Hasan Kadhim Albaldawi
Department of Mathematical, College of Science, University of Baghdad, Baghdad, Iraq

Received: 22/8/2022 Accepted: 28/10/2022 Published: 30/6/2023

Abstract

The analysis of COVID-19 data in Iraq is carried out. Data includes daily cases and deaths since the outbreak of the pandemic in Iraq on February 2020 until the 28th of June 2022. This is done by fitting some distributions to the data in order to find out the most appropriate distribution fit to both daily cases and deaths due to the COVID-19 pandemic. The statistical analysis includes estimation of the parameters, the goodness of fit tests and illustrative probability plots. It was found that the generalized extreme value and the generalized Pareto distributions may provide a good fit for the data for both daily cases and deaths. However, they were rejected by the goodness of fit test statistics due to the high variability of the data.

Keywords: statistical modeling, extreme value distributions, Kolmogorov-Smirnov Test, Anderson-Darling Test, chi-Squared Test

التحليل الاحصائي لبيانات كوفيد-19 في العراق

تسنيم حسن كاظم

قسم الرياضيات ، كلية العلوم، جامعة بغداد، بغداد، العراق

الخلاصة

أجري تحليل لبيانات كوفيد-19 في العراق. تضمنت البيانات الاصابات والوفيات اليومية منذ تفشي الجائحة في العراق في شباط 2020 ولغاية 28 من حزيران 2022. تم عمل ذلك بموائمة بعض التوزيعات للبيانات والكشف عن التوزيع الاكثر موائمة لكلا من الاصابات والوفيات اليومية الناتجة عن جائحة كوفيد-19. يشمل التحليل الاحصائي تقدير المعلمات، اختبارات جودة الموائمة و رسوم توضيحية. وجد ان توزيعي القيمة المتطرفة المعممة وباريتو المعمم يمكن ان يزودنا بأفضل موائمة للبيانات لكلا من الاصابات والوفيات اليومية. رغم ذلك، جرى رفضها احصائيا بواسطة اختبارات جودة الموائمة بسبب التشتت العالي في البيانات.

1.Introduction

The spread of the COVID-19 pandemic has greatly affected people's lives all over the world. The outbreak and the spread of the pandemic vary in each country. Fitting a probability model for the rapid spread of the COVID-19 pandemic locally and globally is highly recommended. Therefore, statistical models for each country should be evaluated separately. During the past two years, a great number of researchers were centering their efforts to help control this pandemic. Yonar, H. et al. [1] collected datasets including the number of cases of COVID-19 in eight selected countries. Cases were modeled by using some curve fitting models; namely,

*Email: tasnim.h@sc.uobaghdad.edu.iq

the Box-Jenkins (ARIMA) time series model and forecasted by using the Brown/ Holt linear exponential smoothing method. Zhao, J. et al. [2] compared the COVID-19 pandemic dynamics between two neighboring Asian countries Iran and Pakistan. They developed a new statistical model that provides the best fitting for the COVID-19 daily death data in the two countries. Xia J, Bin Z. and Jinming C., [3] studied the dynamics of the infectious diseases model besides the time series model to detect the trend and provide short-term prediction of the transmission of the COVID-19 pandemic. Woody, S. et al. [4] built a model using a nonparametric technique in regression called locally weighted polynomial regression (LOWESS). They incorporated a set of predictors based on mobile phone social distancing data which provides well-informed predictions on COVID-19 death rates in the United States. Shukur, S., D. and Kadhim, T., H. [5] applied time series analysis to model and forecast COVID-19 daily deaths in Iraq. They found that modeling the Coronavirus deaths series in Iraq was diagnosed as Threshold GARCH (1, 1). Besides, the Holt-winter-additive method of forecasting was the best method amongst the exponential smoothing methods. Further, a number of recent studies on modeling the COVID-19 pandemic data are based on the extreme value theory approach and the machine learning methods [6, 7, 8, 9, 10].

This study aims to provide a statistical model that best fits daily cases and deaths of the COVID-19 pandemic in Iraq. A variety of statistical distributions are selected and fitted to the data. Four distributions are introduced that are well fitted by both the daily cases dataset and the daily deaths dataset.

2. Identification of Probability Models

In this section, we will introduce the probability distributions that are applied in our study that are best-fitted distributions [11, 12].

2.1 Exponential Distribution

The probability density function is

$f(x) = \lambda \exp(-\lambda x),\ x > 0, \lambda > 0$

The cumulative distribution function is

$F(x) = 1 - \exp(-\lambda x)$

2.2 Gamma Distribution

The probability density function is

$f(x) = \frac{x^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} \exp(-x/\beta),\ x > 0, \alpha, \beta > 0$

The cumulative distribution function is

$F(x) = \frac{\Gamma_{x/\beta}(\alpha)}{\Gamma(\alpha)}$, where Γ is the gamma function, and Γ_x is the incomplete gamma function.

2.3 Weibull Distribution

The probability density function is

$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right),\ x > 0, \alpha, \beta > 0$

The cumulative distribution function is

$F(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$

2.4 Generalized Extreme Value Distribution

The probability density function is

$$f(x) = \begin{cases} \frac{1}{\sigma}\ exp\left(-(1+kz)^{-1/k}\right)(1+kz)^{-1-1/k} \\ \frac{1}{\sigma}\exp(-z-\exp(-z)) \end{cases} \begin{matrix} , k \neq 0 \\ , k = 0 \end{matrix}$$

where $z = \frac{x-\mu}{\sigma}$

The cumulative distribution function is

$$F(x) = \begin{cases} exp\left(-(1+kz)^{-1/k}\right), & k \neq 0 \\ \exp(-\exp(-z)), & k = 0 \end{cases}$$

2.5 Generalized Pareto Distribution

The probability density function is

$$f(x) = \begin{cases} \frac{1}{\sigma}\left(1 + k\frac{(x-\mu)}{\sigma}\right)^{-1-1/k}, & k \neq 0 \\ \frac{1}{\sigma} exp\left(-\frac{(x-\mu)}{\sigma}\right), & k = 0 \end{cases}$$

The cumulative distribution function is

$$F(x) = \begin{cases} 1 - \left(1 + k\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{k}}, k \neq 0 \\ 1 - exp\left(-\frac{(x-\mu)}{\sigma}\right), k = 0 \end{cases}$$

3. The Goodness of Fit Tests

The Goodness of fit tests measures the compatibility of a random sample with a theoretical probability distribution [13]. Three types of goodness of fit tests are applied in this study. The common null and the alternative hypotheses of these tests are

H_0: The data follow the specific theoretical distribution

H_1: The data do not follow the specific theoretical distribution

3.1 The Kolmogorov-Smirnov Test

Assume that we have a random sample X_1, ..., X_n from some continuous distribution with CDF *F(x)*. The empirical CDF is denoted by

$$F_n(x) = \frac{1}{n}[Number\ of\ observations \leq x]$$

The Kolmogorov-Smirnov statistic D is based on the largest vertical difference between the CDF *F(x)* and the empirical CDF $F_n(x)$. That is

$$D_n = sup_x\,|F_n(x) - F(x)|$$

The hypothesis is rejected at the selected significance level α if the calculated statistic D exceeds the critical value obtained from a table.

3.2 The Anderson-Darling Test

The Anderson-Darling test gives more weight to the tails than the Kolmogorov-Smirnov test. It is a general test to compare the fit of an observed CDF to an expected CDF.

The Anderson-Darling test statistic A^2 is given as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(x_i) + \ln(1 - F(x_{n-i+1}))]$$

The hypothesis is rejected at the selected significance level (α) if the test statistic, A^2 exceeds the critical value obtained from a table.

3.3 The Chi-Squared Test

The Chi-Squared test is used to determine whether the sample data follow a specified distribution. This test is applied to binned data, so the value of the test statistic depends on how the data werebinned. Please note that this test is available for continuous sample data only.

The Chi-Squared statistic is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where k is the number of bins which is calculated based on a sample size n as $k = 1 + log_2 n$. The observed frequency for bin i is O_i, and the expected frequency for bin i is E_i, where $E_i = F(x_2) - F(x_1)$, and x_1, x_2 are the limits for bin i. The hypothesis is rejected at the selected significance level α if the χ^2 test statistic is greater than the critical value of the chi square distribution $\left(\chi^2{}_{1-\alpha,k-1}\right)$.

4. Data Description

The total number of daily cases and deaths in Iraq were collected according to the WHO statistics [14] from February 2020 until the 28th of June 2022. Tables 1- 2 show summary statistics and percentile values concerning the daily cases and deaths, respectively

Table 1: Descriptive statistics and percentile values of daily cases

Statistic	Duration (Days)	Range	Mean	Variance	Standard Deviation	Coef. of Var.	Std. Error	Skewness	Kurtosis
Value	856	13515	2734.9	7.3028E+6	2702.4	0.98812	92.365	1.1698	1.0748

Percentile	Min	5%	10%	25% (Q1)	50% Median	75% (Q3)	90%	95%	Max
Value	0	31	80.7	410	2038	4268	6778.3	8153.5	13515

Table 2: Descriptive statistics and percentile values of daily deaths

Statistic	Duration (Days)	Range	Mean	Variance	Standard Deviation	Coef. of Var.	Std. Error	Skewness	Kurtosis
Value	856	122	29.478	756.56	27.506	0.9331	0.94012	0.86454	-1.16113

Percentile	Min	5%	10%	25% (Q1)	50% Median	75% (Q3)	90%	95%	Max
Value	0	0	1	5	24	47	72.3	83	122

It is noticed from Tables 1- 2 the high variability in the data, particularly for the number of daily cases. The coefficients of variation have values almost close to one since the values of the means and standard deviations have close levels. This indicates that the data varies exponentially. Plots of COVID-19 daily cases and deaths are presented in Figures 1- 2. From Figure 1 one can notice that the number of daily cases takes the shape of waves of increasing cases over time. Figure 2 reveals that the number of daily deaths is highly increasing at the beginning of the pandemic and deceasing gradually at the end of the period. The reduction in the number of daily deaths is related to the effect of taking the vaccine.

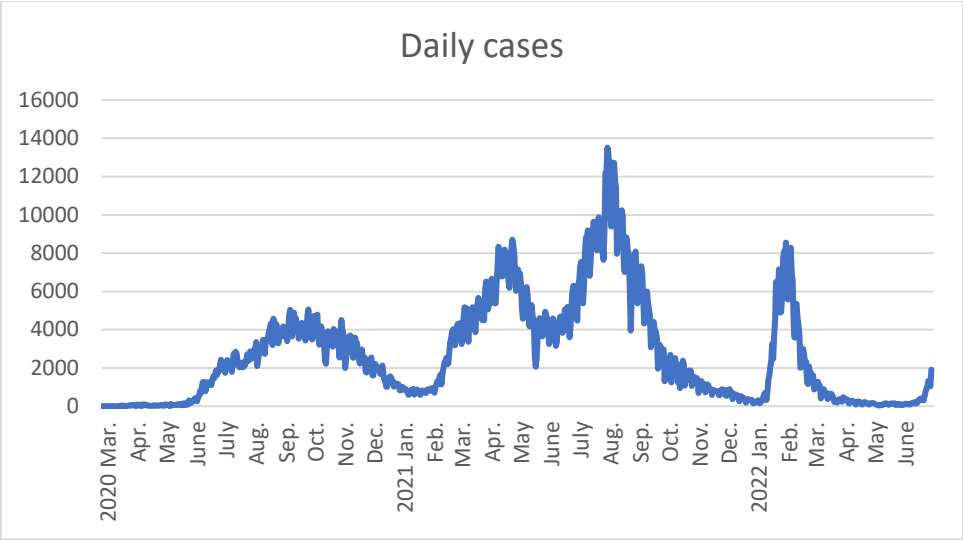

Figure 1: COVID-19 number of daily cases

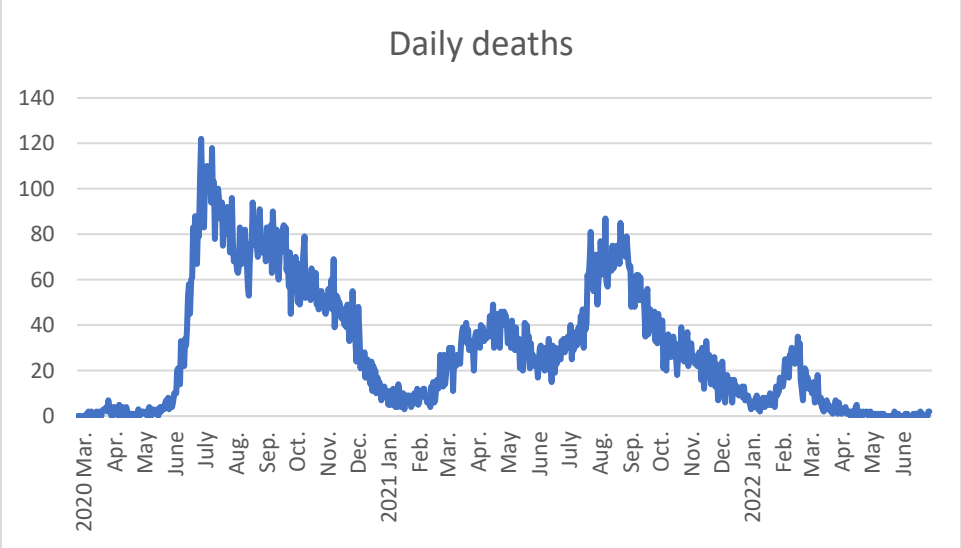

Figure 2: COVID-19 number of daily deaths

5. Data Analysis

In this section, we will try to find the probability distribution that best fits Covid-19 daily cases and deaths in Iraq. First, we choose a selection of some well-known probability distributions, fit our data to these distributions and find out the distributions that get the highest four ranks according to the three types of goodness of fit tests. We will do this with the aid of the well-known Easy-Fit software.

5.1 Analysis of Daily cases

Results of the analysis are presented in Tables 3-4. Maximum likelihood estimates of the parameters are presented in Table 3. The ranking of each distribution according to the three goodness of fit test statistics is presented in Table 4. It is shown that the best-fitted distributions to the daily cases data are the generalized extreme value and the generalized Pareto distributions. Albeit, the best-fitted distributions are all rejected according to the goodness of fit tests. Figure 3, shows the PP-plot for the best-fitted distributions. From Figure 3, we can observe that the data fitted by the generalized extreme value and the generalized Pareto distributions are approximately lined up on the probability plot, which indicates a good fit.

Table 3: Maximum likelihood estimates of parameters

Distribution	Parameters
Exponential	λ=3.6607E-4
Gen. Extreme Value	k=0.15014 σ= 1782.2 μ=1394.9
Gen. Pareto	k=-0.1493 σ= 3578.2 μ=-381.66
Weibull	α=0.5453 β=2770.4

Table 4: Distribution ranking and goodness of fit statistics

Distribution	Kolmogorov Smirnov Critical Value=0.04646 (α=0.05)			Anderson-Darling Critical Value=2.5018 (α=0.05)			Chi-Squared Critical Value=16.919 (α=0.05)		
	Statistic	Reject ?	Rank	Statistic	Reject ?	Rank	Statistic	Reject ?	Rank
Gen. Pareto	0.10078	Yes	2	8.3681	Yes	1	45.803	Yes	1
Gen. Extreme Value	0.09917	Yes	1	15.644	Yes	2	86.381	Yes	2
Weibull	0.12402	Yes	4	41.961	Yes	3	178.88	Yes	3
Exponential	0.12088	Yes	3	49.783	Yes	4	188.81	Yes	4

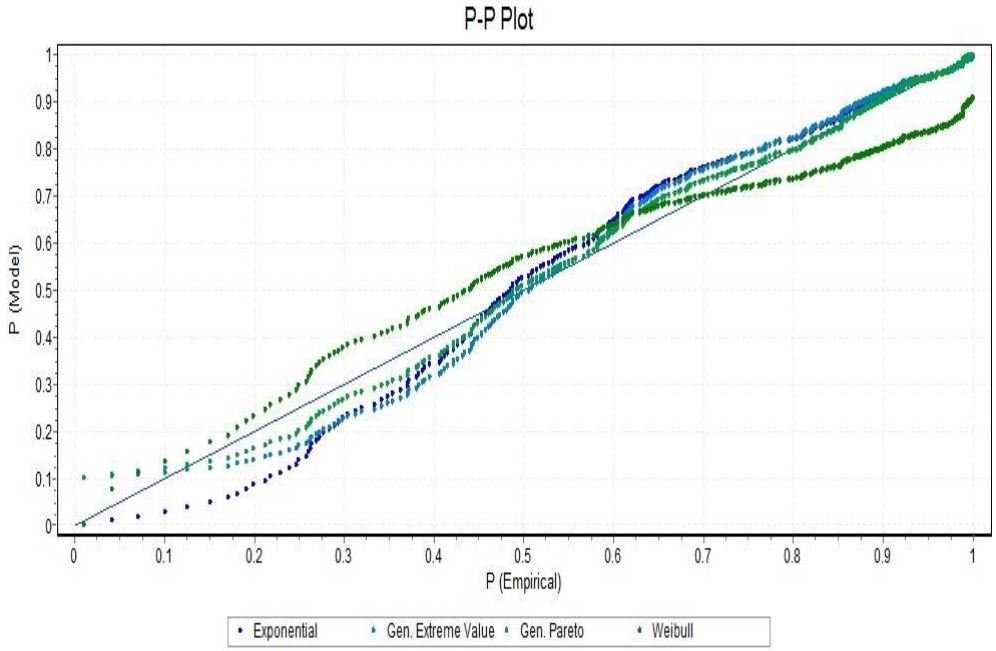

Figure 3: A PP Plot for the best-fitted models of daily cases

5.2 Analysis of Daily Deaths

Results of the analysis are presented in Tables 5-6. Maximum likelihood estimates of the parameters are presented in Table 5. The ranking of each distribution according to the three goodness of fit test statistics is presented in Table 6. It is shown that the best-fitted distributions to the daily deaths data are the same as in the daily cases data: the generalized extreme value and the generalized Pareto distributions. Also, the best-fitted distributions are all rejected according to the goodness of fit tests. Figure 4, shows the PP-plot for the fitted distributions. From Figure 4, we can observe that the data fitted by the generalized extreme value and the generalized Pareto distributions are approximately lined up on the probability plot, which indicates a better fit.

Table 5: Maximum likelihood estimates of parameters

Distribution	Parameters
Exponential	λ=0.03396
Gen. Extreme Value	k=0.09705 σ=19.793 μ=15.93
Gen. Pareto	k=-0.24192 σ=42.131 μ=-4.4804
Gamma	α=1.1457 λ=25.7

Table 6: Distribution ranking and goodness of fit statistics

Distribution	Kolmogorov Smirnov Critical Value=0.04639 (α=0.05)			Anderson-Darling Critical Value=2.5018 (α=0.05)			Chi-Squared Critical Value=16.919 (α=0.05)		
	Statistic	Reject ?	Rank	Statistic	Reject ?	Rank	Statistic	Reject ?	Rank
Gen. Pareto	0.10097	Yes	2	7.6385	Yes	1	39.694	Yes	1
Gen. Extreme Value	0.09985	Yes	1	14.773	Yes	2	106.63	Yes	2
Exponential	0.11869	Yes	3	137.54	Yes	3	181.61	Yes	4
Gamma	0.13825	Yes	4	150.05	Yes	4	176.7	Yes	3

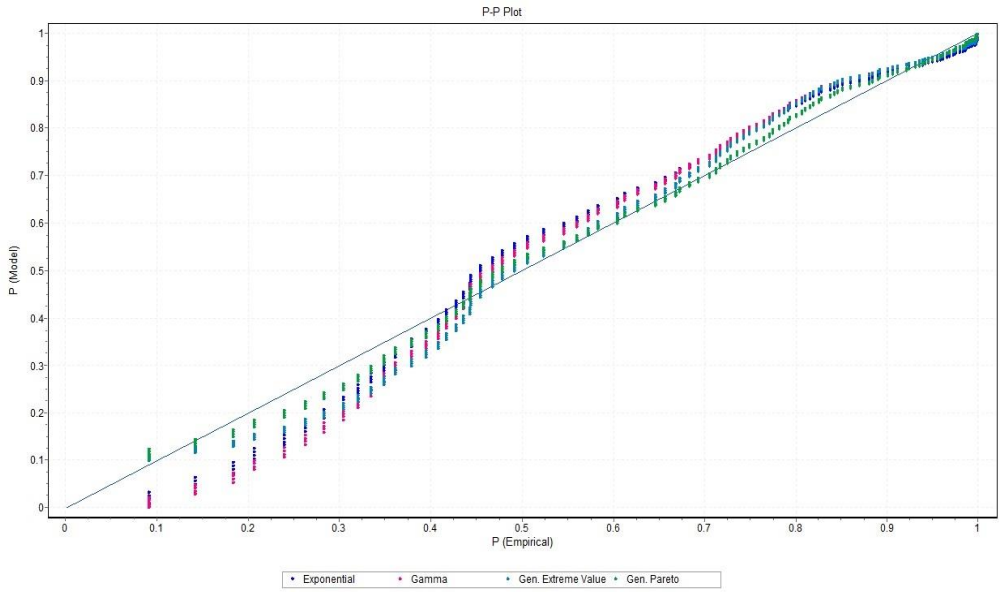

Figure 4: A PP Plot for the best-fitted models of daily deaths

6. Conclusions

The plot of the COVID-19 daily cases presented in Figure 1 reveals that during the period of the study the pandemic went through four waves of which the third is the most severe in the number of infected cases. On the other hand, the plot of the COVID-19 daily deaths presented in Figure 2 revealed a dramatic increase in the number of deaths during the first wave. However, the daily deaths decrease apparently at the end of the period due to the effect of taking the vaccine.

Based on the PP plots presented in Figures 3 and 4 and the goodness of fit tests presented in Table 4 and 6, the fitting results shows that the generalized extreme value distribution and the generalized Pareto distribution were the best-fitted models for both daily cases and deaths. However, the null hypothesis presented in section 3 is rejected for all the suggested distributions by the goodness of fit test statistics.

Accordingly, it is recommended to carry out the statistical analysis with extreme value data rather than the original daily data, since both the generalized extreme value distribution and the generalized Pareto distribution deal with extreme value or block maxima data.

References

[1] H. Yonar , A. Yonar , MA. Tekindal, M. Tekindal," Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods." *EJMO*, Vol.4, no.2, pp.160–165,2020.

[2] J. Zhao, et al.," Modeling COVID-19 Pandemic Dynamics in Two Asian Countries," *Computers, Materials & Continua*, vol. 67, no.1, pp.965-977,2021.

[3] J. Xia, Z. Bin, C. Jinming," Statistical Analysis on COVID-19," *Biomed J Sci & Tech Res*., vol.26, no.2,2020. BJSTR. MS.ID.004310.

[4] S,Woody, et al., " Projections for first-wave COVID-19 deaths across the US using social-distancing measures derived from mobile phones," *MedRxivhttps*://doi.org/10.1101 /2020.04. 16.20068163 (2020).

[5] S. D Shukur, T.H. Kadhim, " Time series analysis of the number of Covid-19 deaths in Iraq," *Int. J. Nonlinear Anal. Appl*. vol.12, no.2, pp.1997-2007,2021.

[6] D.,C. Enriques, et al.," Application of probabilistic models for extreme values to the COVID-2019 epidemic daily dataset," *Data in Brief*, vol. 40,107783, Feb. 2022.

[7] M. Aadhityaa, et al., " A Global Scale Estimate of Novel Coronavirus (COVID-19) 2 Cases Using Extreme Value Distributions," medRxiv preprint 2020. doi: https://doi.org/10.1101/2020.04.17.20069500.

[8] F. Wong, J., J. Collins, "Evidence that coronavirus superspreading is fat-tailed," *Proc. Natl. Acad. Sci. U.S.A.* vol.117, pp.29416–29418, 2020.

[9] S. Ballı, "Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods,' *Chaos Solitons Fractals," 2021,2142:110512.* doi: 10.1016/j.chaos.2020.110512.

[10] Wang, Peipei Zheng, Xinqi Li, Jiayang Zhu, Bangren, " Prediction of epidemic trends in COVID-19 with logistic model and machine learning technics," *Chaos, Solitons & Fractals, Elsevier*, vol. 139(C), 2020.

[11] M. Evans, N. Hastings and B. Peacock, "*Statistical Distributions*" (3rd ed.), *New York: John Wiley*,2000.

[12] S. Coles," *An Introduction to Statistical Modeling of Extreme Values,"* Springer, London,2001.

[13] R. B. D'Agostino and M. A. Stephens, "*Goodness-of-fit Techniques*," *New York: Marcel Dekker*,1986.

[14] WHO-COVID-19-global-data.csv, Coronavirus COVID-19 daily new and cumulative cases and deaths by country. URL: https://covid19.who.int/WHO-COVID-19-global-data.csv.