Data Article

Application of probabilistic models for extreme values to the COVID-2019 epidemic daily dataset

Daniel Canton Enriquez, Jose A. Niembro- Ceceña, Martin Muñoz Mandujano, Daniel Alarcon, Jorge Arcadia Guerrero, Ivan Gonzalez Garcia, Agueda Areli Montes Gutierrez, Alfonso Gutierrez-Lopez

*Facultad de Informática, Universidad Autónoma de Querétaro, Juriquilla, Queretaro 76230, Mexico

bFacultad de Ingeniería, Universidad Autónoma de Querétaro, Centro Universitario, Queretaro 76010, Mexico

cWater Research Center, Centro de Investigaciones del Agua-Queretaro (CIAQ), International Flood Initiative, Latin-American and the Caribbean Region (IFI-LAC), Intergovernmental Hydrological Programme (IHP-UNESCO), Universidad Autónoma de Queretaro, Queretaro 76010, Mexico

Article history:
Received 14 September 2021
Revised 23 December 2021
Accepted 29 December 2021
Available online 1 January 2022

Keywords:
Daily new cases statistical analysis
Coronavirus
Gumbel distribution
Exponential distribution
Probabilistic analysis

Abstract

Worldwide, COVID-19 coronavirus disease is spreading rapidly in a second and third wave of infections. In this context of increasing infections, it is critical to know the probability of a specific number of cases being reported. We collated data on new daily confirmed cases of COVID-19 breakouts in: Argentina, Brazil, China, Colombia, France, Germany, India, Indonesia, Iran, Italy, Mexico, Poland, Russia, Spain, U.K., and the United States, from the 20th of January, 2020 to 28th of August 2021. A selected sample of almost ten thousand data is used to validate the proposed models. Generalized Extreme-Value Distribution Type 1-Gumbel and Exponential (1, 2 parameters) models were introduced to analyze the probability of new daily confirmed cases. The data presented in this document for each country provide the daily probability of rate incidence. In addition, the frequencies of historical events expressed as a return period in days of the complete data set is provided.

* Corresponding author.
E-mail address: alfonso.gutierrez@uaq.mx (A. Gutierrez-Lopez).

https://doi.org/10.1016/j.dib.2021.107783
2352-3409/© 2022 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Specifications Table

| Subject | Data Mining and Statistical Analysis. Infectious Diseases |
|---------|-----------------------------------------------------------|
| Specific subject area | Generalized Extreme-Value Distribution Type 1-Gumbel and Exponential (1, 2 parameters) models applied to characterize probabilistically COVID-19 daily cases |
| Type of data | Table, Graph, Figure |
| How the data were acquired | The data on daily recent confirmed cases of COVID-19 were carefully collected from Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) Database. The data were built as a time-series database by Excel and probabilistic models for extreme values were satisfactorily established for analysis using Matlab. |
| Data format | Analyzed |
| Parameters for data collection | Under the framework of frequency analysis and the Moments estimation parameter method, a probabilistic fitting was carried out to the daily new confirmed Covid cases. Raw data from Argentina, Brazil, China, Colombia, France, Germany, India, Indonesia, Iran, Italy, Mexico, Poland, Russia, Spain, U.K., and the United States, were used. |
| Description of data collection | Daily data on new confirmed cases of COVID-19 outbreaks in Argentina, Brazil, China, Colombia, France, Germany, India, Indonesia, Iran, Italy, Mexico, Poland, Russia, Spain, U.K., and the United States from the 20th of January, 2020 to 28th of August 2021 are available in the Database. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (https://coronavirus.jhu.edu/). In addition, there are no missing values and the Excel file of the daily data is presented in Supplementary Data. This is the data repository for the 2019 Novel Coronavirus Visual Dashboard operated by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). Also, Supported by ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab (JHU APL) [1]. |
| Data source location | Argentina, Brazil, China, Colombia, France, Germany, India, Indonesia, Iran, Italy, Mexico, Poland, Russia, Spain, U.K., and the United States. |
| Data accessibility | The analyzed data is publicly hosted in the mendeley repositories with the following data: Repository name: Frequency analysis of new Covid-19 infections Matlab code: https://github.com/dCantonE/FrequencyAnalysis Supplementary material associated with this article: https://data.mendeley.com/datasets/kvnsn8myhg/3 |

Value of the Data

- Data on daily Covid cases are now easy to obtain. Authorities there are beginning to compile, cross-check and release these data to examine and analyze it. Thus, they are widely available in most countries. However, it is not easy to associate a probability of event occurrence to each daily case report data.
- These data can be updated through official reports and specialized websites. The database presented here is easy to update during the progress of the epidemic (including the third wave in some countries). In data-set of new daily cases are associated with their probability of frequency. They can be wielded to determine the probability of recent infections at specific sites.
- The likelihood of a new outbreak of Covid in any of the countries above can be estimated employing the extreme values probability distribution with the best fit.
• This dataset also supports expanding understanding of the differences in geographic scale in forecasting COVID-19 case counts [2]. Show that statistically significant differences exist based on percentage error metrics when using the same forecasting method at different levels of geographic resolution.

• The probability distributions presented are a complement to a forecasting model. This dataset provides daily probability of rate incidence that could be explored alongside forecasting data to gain further insight into the validity of different forecasts at varied geographic scales as a result of population size differences across countries.

• In order to provide health institutions, research centers and authorities with probabilistic tools to respond to changes in the epidemic. The Matlab code for the systematic of the frequency calculations is included.

1. Data Description

Worldwide, COVID-19 coronavirus disease is spreading rapidly in a second and third wave of infections. In this context of increasing infections, it is critical to know the probability of a specific number of cases being reported [3]. Daily data on new confirmed cases of COVID-19 outbreaks in 16 most affected countries: Argentina, Brazil, China, Colombia, Italy, Spain, France [4], Germany, India [5], Indonesia, Iran, Mexico, Poland, Russia, U.K., and the United States from the 20th of January, 2020 to 28th of August 2021 were collected from COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (https://coronavirus.jhu.edu/). A sample of more than ten thousand daily data is utilized to validate the proposed models. Figs. 1 to 4 shows an example of fit frequency analysis. Comparison between fit proposed models Exp-1P, Exp-2P and Gumbel with daily data on confirmed cases of COVID-19 showed in Figs. 5 to 7.

Very specific studies on COVID-19 forecasting are currently available. It is common to use autoregressive models of the type ARMA(p,q). For example [6], utilize Hidden Markov Chain Models of Moroccan data y [7] using Recurrent Neural Networks; these studies are “forecasting” models. However, there are few studies on the probability of a specific number of infections happening in a day. This is one of the highlights of this dataset. It is proposed to use a frequency analysis to assign a probability of occurrence (infection) of a very particular day in a specific country.

Fig. 1. Comparison between fit proposed models Exp-1P, Exp-2P and Gumbel with daily data on confirmed cases of COVID-19 in Germany.
A theoretical frequency analysis means to fit a series of data to a probability distribution function $P(x)$, which represents the probability of occurrence of a random variable. This procedure must be applied when it is desirable to know an event associated with a return period greater than the maximum length of data record; this is why it is called theoretical because it is not possible to estimate the event using an empirical frequency table. There are several probability distribution functions. Those most successfully used are: normal, log-normal, exponential, gamma, Pearson type III (or three-parameter gamma), log-Pearson type III and those of extreme values types I, II and III; or Gumbel, Frechet and Weibull, respectively. Mixed probability functions are also used, i.e. they can take into account two or three data sets. For daily covid data we propose to use the extreme distributions shown below.
Fig. 4. Comparison between fit proposed models Exp-1P, Exp-2P and Gumbel with daily data on confirmed cases of COVID-19 in Mexico.

Fig. 5. Daily new confirmed cases in Italy, probabilistic characterization with Exp-2P.

1.1. Gumbel distribution

\[ F(x) = e^{-e^{-\alpha (x - \beta)}} \]  

(1)

Where

\[ \alpha = \frac{1.2825}{s}; \beta = \bar{x} - 0.45s \]  

(2)

Where \( s \) is the standard deviation and \( \bar{x} \) is the mean. \( \alpha \) is the scale parameter. \( \beta \) is the shape parameter. Then to equal the probability function of the return period \( P = 1 - \frac{1}{T} \) with the dis-
Fig. 6. Daily new confirmed cases in Argentina, probabilistic characterization with Exp-1P.

Fig. 7. Daily new confirmed cases in Mexico, probabilistic characterization with Gumbel.

distribution function is.

\[ F(X) = \int_0^x f(x) \, dx = P(X \leq x) = 1 - \frac{1}{T} \]  \hspace{1cm} (3)

And solving \( x \)

\[ x = \beta - \frac{1}{\alpha} \left[ \ln \ln \frac{T}{T-1} \right] \]  \hspace{1cm} (4)
1.2. Exponential distribution

\[ F(x) = 1 - e^{-\beta x} \]  
(5)

Where

\[ \beta = \frac{1}{\bar{x}} \]  
(6)

\( \bar{x} \) is the mean, \( \beta \) is the location parameter. According to the return period is:

\[ F(X) = \int_{0}^{X} f(x)dx = P(x \leq X) = 1 - \frac{1}{T} \]  
(7)

And solving \( x \)

\[ x = \frac{1}{\beta} \ln(Tr) \]  
(8)

1.3. Exponential II distribution

\[ F(x) = 1 - e^{-\left(\frac{x-\mu}{\beta}\right)} \]  
(9)

Where

\[ \beta = s; \mu = \bar{x} - s \]  
(10)

Where \( s \) is the standard deviation and \( \bar{x} \) is the mean. \( \beta \) is the scale parameter. \( \mu \) is the shape parameter. According to the return period is:

\[ F(X) = \int_{0}^{X} f(x)dx = P(x \leq X) = 1 - \frac{1}{T} \]  
(11)

And solving \( x \)

\[ x = \mu + \left(-\ln\left(\frac{1}{Tr}\right)\beta\right) \]  
(12)

2. Experimental design, materials and methods

Generalized Extreme-Value Distribution Type-1 (Gumbel) [8] and Exponential models were introduced to analyze the probability of new daily confirmed cases. The data presented in this document for each country provide the daily probability of rate incidence [9]. In addition, the frequencies of historical events expressed as a return period in days of the complete data set is provided. Table 1 shows the estimation of the parameters of the distributions used. This probabilistic analysis comes from the frequency analysis in each of the countries. Only some countries are shown here as examples. The total of the probabilistic analysis can be obtained from the database of this paper. If a series of extreme values is used, the maximum data recorded in each day must be used. This series is used when the design must be based on the most adverse conditions. The empirical return period of this data series is obtained with the following expression proposed by Hosking et al. [10].

\[ T = \frac{(n+1)}{m} \]  
(13)

Where
Table 1
Parameters for fit proposed models Exponential (1, 2 parameters) and Gumbel.

| Country  | Total analyzed data | Gumbel Scale parameter | Gumbel Shape parameter | Exponential 2p Scale parameter | Exponential 2p Shape parameter | Exponential 1p Location parameter |
|----------|---------------------|------------------------|------------------------|--------------------------------|--------------------------------|----------------------------------|
| Argentina | 540                 | 7448.0                 | 4556.5                 | 8284.5                         | 1267.9                         | 0.0001046                        |
| Brazil   | 546                 | 300,052.0              | 13,869.0               | 25,216.4                       | 13,326.7                       | 0.0000259                        |
| China    | 581                 | 126.8                  | 460.6                  | 837.5                          | -674.9                         | 0.0061500                        |
| Colombia | 537                 | 7106.9                 | 4143.5                 | 9114.9                         | 1581.3                         | 0.0001090                        |
| France   | 579                 | 11,938.3               | 5691.2                 | 15,305.6                       | -2732.5                        | 0.0000790                        |
| Germany  | 576                 | 6768.8                 | 2873.0                 | 8771.6                         | -1593.2                        | 0.0001440                        |
| India    | 573                 | 44,240.5               | 45,262.6               | 882,956.6                      | -25,555.0                      | 0.0000176                        |
| Indonesia | 541                 | 7890.7                 | 2856.5                 | 10,116.2                       | -2707.4                        | 0.0000135                        |
| Iran     | 553                 | 6942.9                 | 4595.5                 | 8901.2                         | -300.1                         | 0.0001160                        |
| Italy    | 573                 | 7049.8                 | 3789.4                 | 9041.7                         | 1183.6                         | 0.0001273                        |
| Mexico   | 544                 | 2498.0                 | 4529.4                 | 5174.9                         | 796.3                          | 1867.000                         |
| Poland   | 539                 | 6089.8                 | 1841.6                 | 7810.4                         | 2454.2                         | 0.0001867                        |
| Russia   | 572                 | 6366.8                 | 8022.4                 | 8165.7                         | 3531.2                         | 0.0000855                        |
| Spain    | 571                 | 9843.1                 | 2891.5                 | 12,619.4                       | -4049.1                        | 8570.26                          |
| Turkey   | 532                 | 10,352.3               | 4252.0                 | 13,277.3                       | -3050.5                        | 0.0000978                        |
| U. Kingdom | 572             | 10,913.0               | 5271.3                 | 13,991.1                       | -2423.8                        | 0.0000860                        |
| United States | 580          | 50,807.4               | 36,334.4               | 65,137.7                       | 508.7                          | 65,646.35                        |

$T$ is the empirical return period, in days

$n$ is the total number of data in each country

$m$ is the order-number in a list from high to low value

When historical records of a phenomenon are used, defined as daily data, they should be assigned a return period according to their observed cumulative frequencies (frequencies table). To calculate it, it is assumed that the frequency or recurrence interval of each observed event, allows assigning a return period to each data. This is known as the observed (empirical) return period. Since the return period has a completely probabilistic definition, in its mathematical form $T$ of a daily event $x$, it should be defined as the inverse of the probability $P(x)$ of that event $x$ to occur. This means that the probability of being equalized or exceeded by another event $x$ must be expressed as:

$$T = \frac{1}{P(x)} = \frac{1}{P(X \geq x)}$$

Ethics Statements

The authors paid attention to the ethical rules in the study. There is no violation of ethics. The authors declare that this work does not involve the use of human subjects or experimentation with animals.

CRediT Author Statement

Daniel Canton Enriquez and Alfonso Gutierrez-Lopez: Designed the model and the computational framework. All Authors analyzed the data, carried out the implementation and performed the calculations; Alfonso Gutierrez-Lopez and Martin Muñoz Mandujano: Wrote the manuscript with input from all authors; Ivan Gonzalez Garcia, Jose A. Niembro-Ceceña1 and Jorge Arcadia Guerrero: Were in charge of overall direction and planning.
Funding

This work was financially supported by Consejo Nacional de Ciencia y Tecnología, CONACYT, Mexico.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to the Universidad Autonoma de Queretaro. Corresponding author is grateful to the Risk Management Unit of the UNESCO Regional Office of Science for Latin America and the Caribbean.

Supplementary Materials

– Matlab code: https://github.com/dCantonE/FrequencyAnalysis [11]
– Supplementary material associated with this article: https://data.mendeley.com/datasets/kvnsn8nyhg/3

References

[1] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. 20 (5) (2020), doi:10.1016/s1473-3099(20)30120-1.
[2] C.J. Lynch, R. Gore, Short-range forecasting of COVID-19 during early onset at county, health district, and state geographic levels using seven methods: comparative forecasting study, J. Med. Internet Res. 23 (3) (2021) e24925.
[3] M. Perc, N. Gorišek Miksić, M. Slavinec, A. Stožer, Forecasting COVID-19, Front. Phys. 8 (2020), doi:10.3389/fphy.2020.00127.
[4] Z. Ceylan, Estimation of COVID-19 prevalence in Italy, Spain, and France, Sci. Total Environ. (2020) 138817, doi:10.1016/j.scitotenv.2020.138817.
[5] R. Gupta, S.K. Pal, (2020), Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv. doi:10.1101/2020.03.26.20044511.
[6] A. Marfak, D. Achak, A. Azizi, C. Nejjari, K. Aboudi, E. Saad, I. Youlyouz-Marfak, The hidden Markov chain modelling of the COVID-19 spreading using Moroccan dataset. Data Brief 32 (2020) 106067.
[7] M. Hawas, Generated time-series prediction data of COVID-19’s daily infections in Brazil by using recurrent neural networks. Data Brief 32 (2020) 106175.
[8] J.P. Molina-Aguilar, A. Gutierrez-Lopez, J.A. Raynal-Villaseñor, L.G. Garcia-Valenzuela, Optimization of parameters in the generalized extreme-value distribution type 1 for three populations using harmonic search, Atmosphere 10 (5) (2019) 257, doi:10.3390/atmos10050257.
[9] A. Jalilian, J. Mateu, A hierarchical spatio-temporal model to analyze relative risk variations of COVID-19: a focus on Spain, Italy and Germany, Stoch. Environ. Res. Risk Assess. 35 (4) (2021) 797–812, doi:10.1007/s00477-021-02003-2.
[10] J.R.M. Hosking, J.R. Wallis, E.F. Wood, Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, Technometrics 27 (3) (1985) 251–261.
[11] D. Canton, J. Niembro, M. Muñoz, D. Alarcon, J. Arcadia, I. Gonzalez, A. Montes, A. Gutierrez, (2021). Frequency analysis for confirmed cases of COVID-19 in 17 countries. (Version 1.0.0) [Computer software]. doi:10.13140/RG.2.2.14228.63361