http://dx.doi.org/10.18623/rvd.v23.5385

EXTREME VALUE THEORY TO COVID-19 DATA: PROBABILISTIC ANALYSIS OF DAILY NEW CASES AND DEATHS IN BRAZIL

TEORIA DOS VALORES EXTREMOS APLICADA A DADOS DA COVID-19: ANÁLISE PROBABILÍSTICA DE NOVOS CASOS E ÓBITOS DIÁRIOS NO BRASIL

Article received on: 12/10/2025
Article accepted on: 3/13/2026

Ana Carolina Matiussi*
*Escola Superior de Agricultura Luiz de Queiroz (ESALQ-USP), Piracicaba, São Paulo, Brasil
Lattes: http://lattes.cnpq.br/1739047183434576
Orcid: https://orcid.org/0000-0002-6795-703X
anamatiussi@gmail.com

Gilberto Rodrigues Liska**
**Universidade Federal de São Carlos (UFSCAR), Araras, São Paulo, Brasil
Lattes: https://lattes.cnpq.br/2217949943647601
Orcid: https://orcid.org/0000-0002-5108-377X
gilbertoliska@ufscar.br

Luiz Alberto Beijo**
**Universidade Federal de São Carlos (UFSCAR), Araras, São Paulo, Brasil
Lattes: https://lattes.cnpq.br/8194104388434526
Orcid: https://orcid.org/0000-0002-3286-5602
prof.beijo@gmail.com

Thales Rangel Ferreira*
*Escola Superior de Agricultura Luiz de Queiroz (ESALQ-USP), Piracicaba, São Paulo, Brasil
Lattes: https://lattes.cnpq.br/0524689191887659
Orcid: https://orcid.org/0000-0002-5775-3121
thales.rangel8@gmail.com

Daniel Augusto de Melo Pedro***
***Universidade Federal de Alfenas (UNIFAL-MG), Alfenas, Minas Gerais, Brasil
Lattes: https://lattes.cnpq.br/6080002082844426
Orcid: https://orcid.org/0009-0007-2000-9124
danielmelo0416@gmail.com

The authors declare that there is no conflict of interest

Abstract

The COVID-19 pandemic, caused by the novel coronavirus (SARS-CoV-2), has strongly impacted global public health and healthcare systems. In this context, mathematical modeling is essential for understanding extreme events related to the disease, thereby contributing to the prevention and mitigation of its most severe impacts. This study applied Extreme Value Theory (EVT) to the analysis of new COVID-19 cases and deaths in Brazil, estimating the probability of future extreme events and their expected maximum values over monthly and biweekly intervals. Official daily data from Our World in Data, affiliated with the University of Oxford, were used, covering the period from 2020 to 2022. Among the methods evaluated, the

Resumo

A pandemia de COVID-19, causada pelo novo coronavírus (SARS-CoV-2), teve um grande impacto na saúde pública global e nos sistemas de saúde. Nesse contexto, a modelagem matemática é essencial para compreender eventos extremos relacionados à doença, contribuindo, assim, para a prevenção e a mitigação de seus impactos mais severos. Este estudo aplicou a Teoria dos Valores Extremos (TVE) à análise de novos casos e de óbitos por COVID-19 no Brasil, estimando a probabilidade de eventos extremos futuros e seus valores máximos esperados em intervalos mensais e quinzenais. Foram utilizados dados diários oficiais do Our World in Data, vinculado à Universidade de Oxford, que cobrem o período

Block Maxima approach proved to be the most suitable, providing more accurate estimates of the observed extreme values. The results showed that EVT was effective in identifying extreme patterns in emerging diseases, enabling accurate estimates to support public health planning.

Keywords: COVID-19. Extreme Value Theory. Block Maxima Method. Public Health. Probabilistic Analysis.

de 2020 a 2022. Entre os métodos avaliados, a abordagem de Máximos em Blocos mostrou-se a mais adequada, fornecendo estimativas mais precisas dos valores extremos observados. Os resultados mostraram que a TVE foi eficaz na identificação de padrões extremos em doenças emergentes, possibilitando estimativas precisas para subsidiar o planejamento em saúde pública.

Palavras-chave: *COVID-19. Teoria dos Valores Extremos. Método de Máximos em Blocos. Saúde Pública. Análise Probabilística.*

1 INTRODUCTION

In late 2019, the first cases of COVID-19 were reported in China. The disease is an acute respiratory infection caused by the novel coronavirus SARS-CoV-2, characterized by high transmissibility and rapid global spread (OPAS, 2021). On March 11, 2020, the World Health Organization declared a pandemic due to the high number of deaths and the rapid increase in daily cases (WHO, 2020). In Brazil, the first confirmed case occurred on February 26, followed by the first death on March 17, with 121 reported cases (Barbosa *et al*., 2020). Since then, the disease has grown significantly across the country, establishing a public health emergency (Cavalcante; Abreu, 2020).

The rise in cases and deaths put a lot of pressure on the healthcare system, causing crowded hospitals, a lack of supplies and medicines, and problems with services (Bezerra *et al*., 2020b). Although such events cannot be avoided, their occurrence can be anticipated using statistical methods that estimate the probability of extreme situations (Antunes; Cardoso, 2015; Bezerra *et al*., 2020a; Nascimento *et al*., 2020; Vasconcelos; Moura, 2020). Extreme Value Theory (EVT) is one such statistical approach, designed to analyze the probability of rare events across fields such as epidemiology, economics, and climatology (Liska *et al*., 2013; Martins *et al*., 2020). Despite its relevance, few studies have applied EVT in public health, limiting the ability to predict emerging epidemiological scenarios (Guillou *et al*., 2014; Thomas *et al*., 2016; Zhu; Chen, 2021).

This study proposes applying EVT to the evolution of the COVID-19 pandemic in Brazil. The aim was to analyze data on new cases and deaths, estimate the probability

of extreme events, measure expected maximum values over time, and evaluate which probability distributions best fit the data.

2 MATERIALS AND METHODS

2.1 Data

The daily COVID-19 data were obtained from the Our World in Data database, linked to the University of Oxford, covering Brazil from 2020 to 2022. The database is updated daily (Mathieu *et al*., 2020). The data were organized into fortnightly and monthly periods, separated into two intervals. The period from February 26, 2020, to March 31, 2021, was used as the training series, while the period from April 1, 2021, to January 31, 2022, was used for comparison. The variables considered were the number of new cases and the number of new deaths, with the maximum value for each period.

2.2 Methodology and probability distributions

In this section, we briefly describe the statistical methods used to analyze extreme events in COVID-19 cases and deaths, based on Extreme Value Theory.

2.2.1 Block Maxima Method (BM) and Generalized Extreme Value Distribution (GEV)

To analyze the maximum events in COVID-19 records, two approaches from Extreme Value Theory were used. The first approach, called Block Maxima (BM), describes the behavior of the maximum value that occurs within periods. According to Coles *et al*. (2001), EVT states that under suitable conditions, if there exist sequences of constants $a_n > 0$ and $b_n \in \mathbb{R}$, then.

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) \to f(x) \quad n \to \infty \qquad (1)$$

Where

M_n is the maximum of a sample of size n. This means that the normalized sequence of maxima converges in distribution to a non-degenerate asymptotic distribution.

The methodology includes three extreme value distributions: Gumbel ($\xi \to 0$), Fréchet ($\xi > 0$) and Weibull ($\xi < 0$), which can be written using a single common parameterization, called the Generalized Extreme Value (GEV) distribution, where the parameters satisfy $-\infty < \mu < \infty, \sigma > 0\ e\ -\infty < \xi < \infty$, given by:

$$F(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} \quad (2)$$

The model has three parameters: the location parameter μ, which shifts the distribution along the x-axis; the scale parameter σ, which controls the spread; and the shape parameter ξ, which defines the type of distribution. Specifically, Gumbel corresponds to $\xi \to 0$, Fréchet to $\xi > 0$, and Weibull to $\xi < 0$.

Using this combined framework of the three distributions, the statistical implementation through the inference of (ξ) becomes simplified. This model was used to determine which of the three distributions best fit the data and to identify the distribution represented by the dataset. Additionally, the Gumbel distribution was applied in the BM analysis. In the case where ($\xi \to 0$), the cumulative distribution function of the Gumbel distribution, defined in the range $-\infty < x < \infty$ and using the same parameterization as the previous equation with $\mu > 0$, is given by:

$$F(x) = \exp\left[-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right] \quad (3)$$

A random variable x follows a Gumbel distribution when its probability density function (PDF) is given by:

$$f(x) = \frac{1}{\sigma} exp\left\{-\left(\frac{x-\mu}{\sigma}\right) - exp\left[\left(-\frac{x-\mu}{\sigma}\right)\right]\right\} \quad (4)$$

From the derivation of equation (2) with respect to x, it is possible to obtain the PDF of the GEV distribution, which is expressed as:

$$f(x) = \frac{1}{\sigma}\left\{\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\left(\frac{1+\xi}{\xi}\right)} exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}\right\} \quad (5)$$

defined in $-\infty < x < \mu - \sigma/\xi$ for $\xi < 0$ and $\mu - \sigma/\xi < x < \infty$ for $\xi > 0$ (Bautista *et al.*, 2004; Coles *et al*., 2001).

2.2.2 Peaks Over Threshold Method (POT) and Generalized Pareto distribution (GP)

For a more detailed understanding of data variation and to analyze rare events, a second methodology was applied, considering and estimating values exceeding a given high threshold. In this case, the values were estimated using the Peaks Over Threshold (POT) method and analyzed using the Generalized Pareto distribution (GP), whose cumulative distribution function is defined as:

$$F(y) = \begin{cases} 1 - \left(1 + \xi\dfrac{(y-\mu)}{\sigma}\right)^{-\frac{1}{\xi}}, & \xi \neq 0 \\ 1 - exp\left(-\dfrac{y-\mu}{\sigma}\right), & \xi \to 0 \end{cases} \quad (6)$$

From the distribution function in equation (5), we can derive the probability density function (PDF) of the GP, given by:

$$f(y) = \begin{cases} \dfrac{1}{\sigma}\left(1 + \xi\left(\dfrac{y-\mu}{\sigma}\right)\right)^{-\left(\frac{1+1}{\xi}\right)}, & \xi \neq 0 \\ \dfrac{1}{\sigma} exp\left\{-\dfrac{y-\mu}{\sigma}\right\}, & \xi \to 0 \end{cases} \quad (7)$$

This definition applies for $0 \leq y$ when $\xi \neq 0$ and for $0 \leq y \leq 1/\xi$ when $\xi = 0$ (Ma *et al*., 2021). It is important to note that, as in the first methodology, the GP can also be divided into three types of distributions: Exponential ($\xi \to 0$), Pareto ($\xi > 0$), and Beta ($\xi < 0$). To define an appropriate threshold value u in the POT methodology, graphical analyses were used. The mean residual life plot, which is approximately linear, was used. However, its interpretation is subjective. Therefore, the threshold choice plot was also used, as demonstrated by Coles *et al*. (2001).

2.2.3 Maximum Likelihood Estimation

Since the model parameters are unknown, it was necessary to use statistical estimation methods to obtain the estimators and fit the models. For this purpose, we used the Maximum Likelihood Estimation (MLE) method. The statistical model is defined as:

$$L(\theta;\, x_1, \ldots, x_n) = f(x_1;\theta) \times \ldots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta) \qquad (8)$$

This approach consists of estimating the parameters that maximize the likelihood function, i.e., the values that provide the highest probability for the observed x, as described by Assis *et al*. (2021).

2.2.4 Hypothesis testing

To assess the dataset's assumptions, three hypothesis tests were applied. The Ljung-Box test was used to assess the independence of the time series, and the Mann-Kendall test was used to assess the presence of a significant trend in the data (Albuquerque, 2018). The null hypothesis H_0 states that there is no significant trend, while the alternative H_1 states that a trend exists. Both were tested at the 1% significance level ($\alpha = 0.01$). The same logic applies to the Ljung-Box test. The Kolmogorov-Smirnov test was also applied to assess whether the biweekly and monthly COVID-19 maxima follow

a theoretical distribution and to evaluate how well the model fits (Cotta *et al*., 2016). The test statistic is defined as:

$$D = max|F(X_d) - R(X_d)| \quad (9)$$

Where $F(X_d)$ is the empirical distribution function and $R(X_d)$ is the theoretical distribution. The test checks if the data ($d = 1, 2, 3, \ldots, n$) are close to the theoretical model, based on the maximum distance between the two distributions.

To evaluate which extreme value distribution fits the data best, we applied the Likelihood Ratio Test (LRT) to test whether the parameter ξ is statistically equal to zero when comparing the Gumbel and GEV models. The test statistic is given by:

$$T_{RV} = -2\left[l\left(\widehat{\theta_G}\right) - l\left(\widehat{\theta_{GEV}}\right)\right] = 2\left[l\left(\widehat{\theta_{GEV}}\right) - l\left(\widehat{\theta_G}\right)\right] \quad (10)$$

where $l\left(\widehat{\theta_G}\right)$ and $l\left(\widehat{\theta_{GEV}}\right)$ are the maximum log-likelihoods, and $\widehat{\theta_G} = (\mu, \sigma)$ and $\widehat{\theta_{GEV}} = (\mu, \sigma, \xi)$ are the vectors of estimated parameters. For the comparison between the Exponential and GPD models, the likelihood ratio statistic follows the same logic:

$$T_{RV} = -2\left[l\left(\widehat{\theta_E}\right) - l\left(\widehat{\theta_{GPD}}\right)\right] = 2\left[l\left(\widehat{\theta_{GPD}}\right) - l\left(\widehat{\theta_E}\right)\right] \quad (11)$$

The null hypothesis is H_0: $\xi = 0$, while the alternative is H_1: $\xi \neq 0$. At the 1% significance level ($\alpha = 0.01$), H_0 is rejected if the test statistic is larger than the chi-squared critical value with one degree of freedom, or if the p-value is smaller than α (Bautista *et al*., 2004).

2.2.5 Probability of exceedances

To calculate the probability of events exceeding the number of COVID-19 cases and deaths, the following formula was applied:

$$P(X > x) = 1 - F(x) = 1 - exp\left\{-\left[1 + \hat{\xi}\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)\right]^{-\frac{1}{\hat{\xi}}}\right\} \quad (12)$$

For the case where $\xi \to 0$ the formula is defined as:

$$P(X > x) = 1 - exp\left\{-exp\left[-\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right)\right]\right\} \quad (13)$$

Here, x represents the maximum event and $0 \leq x < \infty$, as described by Almeida (2018).

2.2.6 Return level estimation

The return period represents the estimated time interval in which a specific event is expected to occur. It is defined as the inverse of the probability that an event is equal to or exceeds a given value. The return period t is expressed as:

$$t = \frac{1}{1 - F(x)} \quad (14)$$

where $F(x) = p(X \leq x)$. The return level (x_p) associated with a return period is obtained from the expression:

$$F(x_p) = \int_{-\infty}^{x_p} f(x)dx = 1 - p \quad (15)$$

for $p = 1/t$. Using the estimated parameters of the GEV and Gumbel distributions, it is possible to estimate the maximum probable number of COVID-19 cases and deaths. The return level is calculated as the inverse of $F(x_p)$. For the GEV distribution, the return level is given by:

$$\hat{x}_p = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\{1 - [-\ln(1-p)]^{-\hat{\xi}}\}, \qquad \xi \neq 0 \qquad (16)$$

For the Gumbel distribution, the return level is given by:

$$\hat{x}_p = \hat{\mu} - \hat{\sigma}\,\{\ln[-\ln(1-p)]\}, \qquad \xi \to 0 \qquad (17)$$

Return level estimates ($\hat{x}_p$) associated with the return period $t = 1/p$ were obtained using maximum likelihood estimators (MLE), as described by Almeida (2018). For the GP, return levels can be estimated similarly. According to Bautista *et al*. (2004) the confidence interval for $\hat{x}_p$ at a $(1-\alpha)100\%$ confidence level is given by:

$$IC(x_p) = \hat{x}_p \pm z_{\frac{\alpha}{2}}\sqrt{Var(\hat{x}_p)} \qquad (18)$$

where α is the significance level, $z_{\frac{\alpha}{2}}$ is the critical value from the standard normal distribution, and $Var(\hat{x}_p)$ is the variance associated with the estimated return level x_p.

2.2.7 Goodness of fit criteria

After all analyses, statistical methods were applied to evaluate and select the models that best fit the COVID-19 data. Based on these measures, the distribution and methodology that performed best were identified. The quality-of-fit criteria are calculated using the following equations (Ananias *et al*., 2021):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{i(obs)} - \hat{x}_{i(pred)}\right)^2} \qquad (19)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_{i(obs)} - \hat{x}_{i(pred)}}{x_{i(obs)}}\right| \times 100 \qquad (20)$$

$$md = 1 - \frac{\sum_{i=1}^{n}\left|x_{i(obs)} - \hat{x}_{i(pred)}\right|}{\sum_{i=1}^{n}\left(\left|x_{i(obs)} - \bar{x}_{(mean)}\right| + \left|\hat{x}_{i(pred)} - \bar{x}_{(mean)}\right|\right)} \qquad (21)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{i(obs)} - \hat{x}_{i(pred)}\right| \qquad (22)$$

$$NSE = 1 - \frac{\sum_{i=1}^{n}\left(x_{i(obs)} - \hat{x}_{i(pred)}\right)^2}{\sum_{i=1}^{n}\left(x_{i(obs)} - \bar{x}_{(mean)}\right)^2} \qquad (23)$$

Based on goodness-of-fit criteria, it is possible to evaluate and select the statistical models that best fit the COVID-19 data (Abreu *et al*., 2018).

2.2.8 Computational resources

All analyses were conducted in R (R Core Team, 2020), using the packages evd (Stephenson, 2002) for the estimation of extreme value distributions, extRemes (Gilleland; Katz, 2016) for hypothesis testing and model fitting, and hydroGOF (Zambrano-Bigiarini, 2017) for the evaluation of predictive performance.

3 RESULTS AND DISCUSSION

According to the likelihood ratio test (LRT), at the 1% significance level ($\alpha = 0.01$), the pvalue was greater than the tabulated value, meaning that H_0 could not be rejected. Therefore, the parameter ξ was considered statistically equal to zero under both methodologies. Consequently, the Gumbel and Exponential distributions were identified as providing the best fit and were used to analyze the three observed series (Tables 1 and 4).

Similar findings were reported by Lim *et al*. (2020), who also observed that the Gumbel and Exponential distributions provided the best fit for dengue maxima. However, unlike the present study, their model selection relied on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), whereas here it was based on the LRT. This highlights that model selection can be supported by different approaches, as

discussed by the authors. In a related context, Chen *et al*. (2015) applied Extreme Value Theory (EVT) in epidemiology using both methodologies to estimate the probability of highly pathogenic diseases in public health, though in a distinct scenario from the present study.

Table 1

Parameter estimates of the GEV, Gumbel, Exponential, and GPD distributions and their respective hypothesis tests for the number of new COVID-19 cases.

Method	Period	Distribution	μ	σ	ξ	LRT	MK	LB	KS
BM	1st Fortnight	GEV	40513.09	31671.80	-0.5960	0.03055	0.9514	0.4512	0.8705
		Gumbel	40513.10	31671.81	-				0.6007
	2nd Fortnight	GEV	40561.40	32830.39	-0.4682	0.05379	0.3244	0.01422	0.6254
		Gumbel	40561.40	32830.40	-				0.2574
	Monthly	GEV	46447.39	35381.02	-0.6076	0.01336	0.6614	0.01116	0.9043
		Gumbel	46447.40	35381.03	-				0.5787
POT	1st Fortnight	Exponential	75000	8922.38	-	0.0646	0.917	0.3912	0.2528
		GP	75000	12574.36	-0.4346				0.0919
	2nd Fortnight	Exponential	75000	8922.38	-	0.0646	0.5371	0.8223	0.0040
		GP	75000	12574.36	-0.4346				0.0219
	Monthly	Exponential	75000	8922.38	-	0.0646	0.0086	0.1139	0.3052
		GP	75000	12574.36	-0.4347				0.7276

The Mann-Kendall test, applied under the BM methodology for cases, indicated no significant trend, suggesting that the series was randomly distributed (p-value > 0.01). Similar results were observed in the first and second fortnights under the POT methodology, although the monthly analysis showed a slight trend. These results differ from those of Lim *et al*. (2020), who identified trends in dengue maxima series.

The Ljung-Box test indicated independence across all case series. In addition, the Kolmogorov-Smirnov test confirmed that the Gumbel and GEV distributions provided adequate fits in all series. Under the POT methodology, only two series were fitted by the Exponential and GP distributions, with satisfactory adjustment by the GP observed only in the second fortnight (Table 1).

Table 2

Probabilities (%) of new COVID-19 cases and expected maximum values under the BM method.

BM Method											
		Probabilities					Expected (months)				
Periods	Distribution	>10.000	>20.000	>30.000	>50.000	>75.000	2	3	4	5	6
1st Fortnight	GEV	88.25	82.26	74.17	51.27	15.85	50941	62624	68363	71916	74382
	Gumbel	92.72	85.21	75.18	52.34	28.58	52121	69104	79973	88019	94418
2nd Fortnight	GEV	88.53	82.30	74.06	52.02	21.04	51618	64730	71550	75938	79073
	Gumbel	92.09	84.60	74.83	52.77	29.55	52594	70198	81464	89805	96438
Monthly	GEV	89.20	84.31	77.82	59.41	28.02	58072	71031	77363	81270	83974
	Gumbel	93.93	87.90	79.64	59.52	35.99	59415	78387	90529	99517	106665

Table 3

Probabilities (%) of new COVID-19 cases and expected maximum values under the POT method.

POT Method											
		Probabilities					Expected (months)				
Periods	Distribution	>10.000	>20.000	>30.000	>50.000	>75.000	2	3	4	5	6
1st Fortnight	Exponential	32.60	10.63	3.47	0.37	0.02	77062	80679	83246	85237	86864
	GP	37.69	6.69	0.00	0.00	0.00	77764	81992	84570	86360	87698
2nd Fortnight	Exponential	32.60	10.63	3.47	0.37	0.02	77062	80679	83246	85237	86864
	GP	37.69	6.69	0.00	0.00	0.00	77764	81992	84570	86360	87698
Monthly	Exponential	32.60	10.63	3.47	0.37	0.02	76165	78244	79713	80851	81781
	GP	37.69	6.69	0.00	0.00	0.00	71823	74477	75924	76837	77447

Tables 2 and 3 present the probabilities of new COVID-19 cases at different levels, along with expected maximum values. Under the BM methodology with the Gumbel distribution, there was a 93.93% probability of exceeding 10.000 cases and a 35.99% probability of exceeding 75.000. Using the POT methodology with the Exponential distribution, the probabilities were 32.60% and 0.02% for the same thresholds. Regarding return periods, the maximum number of new cases increased over time, reaching 106.665 (BM) and 81.781 (POT) in the sixth month. These results correspond to the monthly analysis.

Lim *et al*. (2020) proposed a different methodology, based on Inhomogeneous Point Processes (IPP) combined with EVT, to estimate the probability of dengue cases exceeding observed levels. They showed that within ten years, dengue cases are expected to surpass historical maxima in more than half of Thailand's provinces. In another example, Thomas & Rootzén (2022) developed a real-time forecasting method for

influenza epidemics in France using the multivariate Generalized Pareto Distribution. Their results estimated a 10% probability that the epidemic would exceed 9.385 cases at least once over the next 10 years. However, the method is applicable only to diseases with historical data, making it less suitable for emerging diseases such as COVID-19.

Table 4

Parameter estimates of the GEV, Gumbel, Exponential, and GP distributions and their respective hypothesis tests for the number of new COVID-19 deaths.

Method	Period	Distribution	μ	σ	ξ	LRT	MK	LB	KS
BM	1st Fortnight	GEV	1106.33	523.93	-0.3105	0.1199	0.1148	0.8284	0.8549
		Gumbel	1057.11	522.59	-				0.7131
	2nd Fortnight	GEV	938.04	628.36	0.0166	0.9091	0.8548	0.02799	0.5906
		Gumbel	938.04	628.36	-				0.5879
	Monthly	GEV	1064.78	679.60	-0.0407	0.7903	0.583	0.01118	0.5841
		Gumbel	1057.39	690.22	-				0.5682
POT	1st Fortnight	Exponential	1600	839.96	-	0.2685	0.1074	0.0571	0.0269
		GP	1600	914.96	-0.2065				0.0264
	2nd Fortnight	Exponential	2800	577.71	-	0.1191	0.0162	0.08962	0.6596
		GP	2800	726.71	-0.5963				0.4912
	Monthly	Exponential	2800	577.71	-	0.1191	0.0162	0.0896	0.6596
		GP	2800	726.71	-0.5963				0.4912

Similar to the case results, the Mann-Kendall test (MK) for deaths indicated no significant trend, suggesting a random distribution of the series. The Ljung-Box test (LB) confirmed data independence, and the Kolmogorov-Smirnov test (KS) showed that all distributions fitted the death series adequately under both methodologies (Table 4).

Table 5

Probabilities (%) of new COVID-19 deaths and expected maximum values under the BM method.

POT Method											
		Probabilities					Expected (months)				
Periods	Distribution	>1.000	>2.000	>3.000	>4.000	>5.000	2	3	4	5	6
1st	GEV	70.40	8.44	0.00	0.00	0.00	1288	1519	1648	1735	1799
Fortnight	Gumbel	67.22	15.18	2.40	0.36	0.05	1249	1529	1708	1841	1947
2nd	GEV	59.59	17.21	4.01	0.92	0.22	1169	1509	1729	1892	2022
Fortnight	Gumbel	59.59	16.85	3.69	0.76	0.16	1168	1505	1720	1880	2007

Monthly	GEV	66.71	21.54	4.73	0.86	0.14	1312	1667	1890	2053	2182
	Gumbel	66.27	22.52	5.82	1.40	0.33	1310	1680	1917	2093	2232

Table 6

Probabilities (%) of new COVID-19 deaths and expected maximum values under the POT method.

BM Method											
		Probabilities					Expected (months)				
Periods	Distribution	>1.000	>2.000	>3.000	>4.000	>5.000	2	3	4	5	6
1st Fortnight	Exponential	30.41	9.25	2.81	0.85	0.26	1984	2324	2566	2753	2906
	GP	28.98	5.46	0.42	0.00	0.00	1999	2323	2537	2694	2817
2nd Fortnight	Exponential	17.71	3.14	0.56	0.10	0.02	2328	2562	2728	2857	2963
	GP	5.61	0.00	0.00	0.00	0.00	2035	2461	2707	2870	2988
Monthly	Exponential	17.71	3.14	0.56	0.10	0.02	2328	2562	2728	2857	2963
	GP	5.61	0.00	0.00	0.00	0.00	2035	2461	2707	2870	2988

Tables 5 and 6 show the probabilities of new COVID-19 deaths and their expected maximum values. Under the BM methodology with the Gumbel distribution, the probability of more than 1.000 deaths in a month was 66.27%, while the probability of exceeding 5.000 was only 0.33%. Under the POT methodology with the Exponential distribution, the probabilities were approximately 18% and 0.02%, respectively. Regarding return periods, the maximum expected deaths increased over time, reaching 2.232 (BM) and 2.963 (POT) in the sixth month, according to the monthly analysis.

Previous studies reinforce the use of EVT for mortality. Thomas *et al*. (2016) and Chiu *et al*. (2018) applied mathematical approaches to estimate mortality events, although their results differed from ours. Both investigated the probability of deaths exceeding specific return levels. In a related context, Campolieti (2021) estimated influenza mortality risks and their extreme values. The results indicated return levels ranging from 122.77 to 232.87 per 100.000 inhabitants over the next 50 years, with mortality rates varying from 3.8% to 7.2%. However, the author emphasized the limitation of the small number of observations, a constraint also noted by Thomas; Rootzén (2022).

Table 7

Goodness-of-fit criteria for probability distributions under different methodologies considering return periods of 2 to 11 months, for the number of cases.

Periods	Method	Distribution	RMSE	MAPE (%)	md	MAE	AMI	NSE
1st Fortnight	BM	Gumbel	93523.14	175.1278	0.10	79411.18	62750.06	-27.90
		GEV	98542.00	138.3594	0.05	77258.99	26917.71	-101.63
	POT	Exponential	94477.31	164.2682	0.04	75135.70	9729.87	-307.78
		GP	95557.84	153.3021	0.02	74383.47	6327.859	-810.17
2nd Fortnight	BM	Gumbel	92382.07	191.656	0.12	80097.77	70382.05	-21.75
		GEV	96786.06	150.5158	0.05	77688.86	29110.81	-79.47
	POT	Exponential	94477.31	164.2682	0.04	75135.70	9729.87	-307.78
		GP	95557.84	153.3021	0.02	74383.47	6327.859	-810.17
Monthly	BM	Gumbel	91939.02	204.9791	0.14	80535.85	74347.76	-18.87
		GEV	96175.14	155.5169	0.05	77614.49	26951.93	-84.35
	POT	Exponential	94477.31	164.2682	0.04	75135.70	9729.87	-307.78
		GP	95557.84	153.3021	0.02	74383.47	6327.859	-810.17

Table 8

Goodness-of-fit criteria for probability distributions under different methodologies considering return periods of 2 to 11 months, for the number of deaths.

Periods	Method	Distribution	RMSE	MAPE (%)	md	MAE	AMI	NSE
1st Fortnight	BM	Gumbel	1241.38	139.114	0.01	1156.65	929.8376	-16.76
		GEV	1147.83	125.8712	0.01	1065.05	682.3222	-25.28
	POT	Exponential	2038.63	252.2347	0.06	1851.14	942.5786	-18.11
		GP	1956.88	242.4996	0.04	1779.89	1741.693	-26.49
2nd Fortnight	BM	Gumbel	1347.06	153.3154	0.01	1255.62	1125.959	-12.24
		GEV	1351.25	153.1154	0.01	1259.34	1200.998	-11.51
	POT	Exponential	1941.84	240.9584	0.05	1749.97	412.7978	-34.91
		GP	2067.34	253.693	0.04	1909.12	4946.539	-11.30
Monthly	BM	Gumbel	1386.68	160.4486	0.02	1291.53	1129.949	-12.56
		GEV	1389.04	160.624	0.02	1293.86	1181.121	-12.33
	POT	Exponential	1941.84	240.9584	0.05	1749.97	412.7978	-34.91
		GP	2067.34	253.693	0.04	1909.12	3660.622	-11.30

In addition to the goodness-of-fit tests, which evaluate the adequacy of distributions to the dataset, it is essential to assess the uncertainties associated with model predictions. This step complements the initial evaluation by measuring model fit using various statistics (Tables 7 and 8). Among the four probability distributions analyzed, the Generalized Pareto (GP) and the Generalized Extreme Value (GEV) distributions performed best for case series, with lower estimation errors across all three periods.

For the death series, the goodness-of-fit measures indicated different outcomes. The criteria showed that GP and GEV were more suitable for the first fortnight when estimated by Maximum Likelihood Estimation (MLE). For the second fortnight and

monthly data, the Gumbel and Exponential distributions performed better when estimated with the L-moments method, which proved more efficient than MLE in these cases. Similar findings were reported by Wong & Collins (2020), who concluded that coronavirus superspreading follows a heavy-tailed distribution, specifically a Fréchet distribution. This interpretation is consistent with Cirillo & Taleb (2020), who also argued that pandemics are heavy-tailed phenomena and correspond to specific cases of the GEV distribution.

Table 9

Comparison of observed and estimated values for the 11-month return period, using the methodology and distribution that provided the most accurate estimates.

Period	Method	Distribution	Observed	Estimated
Number of Cases				
1st Fortnight	BM	GEV	260806*	80559
2nd Fortnight	BM	GEV	260806	87350
Monthly	BM	GEV	260806	90716
Number of Deaths				
1st Fortnight	BM	GEV	780*	1980
2nd Fortnight	BM	Gumbel	780	2415
Monthly	BM	Gumbel	780	2679

*The observed values refer to the maximum peak in January 2022.

Table 9 compares the observed and estimated values. In some cases, the predicted values were consistently lower than the observed ones, indicating underestimation. For deaths, the opposite occurred: predictions exceeded observations, suggesting overestimation. This discrepancy may be explained by external factors such as vaccination, mandatory mask use, physical distancing, and social isolation. These measures helped reduce cases and deaths during the pandemic (Brasil, 2020). Therefore, since the pandemic and preventive strategies were still in effect during the study period, occasional discrepancies between estimates and observed values were expected.

4 CONCLUSIONS

Given the number of cases and deaths, the Block Maxima methodology proved most suitable, providing more consistent forecasts of COVID-19 extreme values. For cases, the monthly period yielded more accurate estimates, while for deaths, the first

fortnight period was the most satisfactory. The applied theory and methods proved functional, enabling the calculation of the probability of an emerging disease and the estimation of its expected maximum value within a defined time frame, even with a limited number of observations. As highlighted in previous studies, this methodology can be extended to other public health contexts, supporting the prediction of epidemiological scenarios and providing relevant information for healthcare systems. It is important to emphasize that forecasts do not represent absolute certainties but rather the probability of future extreme values occurring in situations such as emerging pandemics. Thus, this study offers a methodology based on probability distributions that produces more accurate estimates with lower prediction errors than other approaches.

REFERENCES

ABREU, Marcel Carvalho *et al.* Critérios para escolha de distribuições de probabilidades em estudos de eventos extremos de precipitação. **Revista Brasileira de Meteorologia**, v. 33, p. 601-613, 2018.

ALBUQUERQUE, R. C. **Modelagem em séries temporais**: aplicação em dados de precipitação na região do sertão de Pernambuco-Brasil. Dissertação de Mestrado. Universidade Federal Rural de Pernambuco. 2018.

ALMEIDA, G. C. **Uma abordagem bayesiana para a modelagem dos ventos máximos de Sorocaba-SP e Bauru-SP**. Dissertação de Mestrado. Universidade Federal de Alfenas. 2018.

ANANIAS, Denis Rafael Silveira *et al.* The assessment of annual rainfall field by applying different interpolation methods in the state of Rio Grande do Sul, Brazil. **SN Applied Sciences**, v. 3, n. 7, p. 687, 2021.

ANTUNES, José Leopoldo Ferreira; CARDOSO, Maria Regina Alves. Uso da análise de séries temporais em estudos epidemiológicos. **Epidemiologia e Serviços de Saúde**, v. 24, p. 565-576, 2015.

ASSIS, J. P. *et al.* **Estimação Estatística**. Pantanal Editora, 2021.

BARBOSA, Isabelle Ribeiro *et al.* Incidence of and mortality from COVID-19 in the older Brazilian population and its relationship with contextual indicators: an ecological study. **Revista Brasileira de Geriatria e Gerontologia**, v. 23, n. 01, p. e200171, 2020.

BAUTISTA, Ezequiel Abraham López; ZOCCHI, Silvio Sandoval; ANGELOCCI, Luiz Roberto. A distribuição Generalizada de Valores Extremos aplicada ao ajuste dos

dados de velocidade máxima do vento em Piracicaba, São Paulo, Brasil. **Revista de Matemática e Estatística**. 2004.

BEZERRA, Anselmo César Vasconcelos *et al.* Factors associated with people's behavior in social isolation during the COVID-19 pandemic. **Ciencia & Saude Coletiva**, v. 25, p. 2411-2421, 2020.

BEZERRA, Évilly Carine Dias *et al.* Spatial analysis of Brazil's COVID-19 response capacity: a proposal for a Healthcare Infrastructure Index. **Ciência & Saúde Coletiva**, v. 25, p. 4957-4967, 2020.

BRASIL. **Painel Coronavírus Ministério da Saúde**. Brasília, 2020.

CAMPOLIETI, Michele. Tail risks and infectious disease: Influenza mortality in the US, 1900–2018. **Infectious Disease Modelling**, v. 6, p. 1135-1143, 2021.

CAVALCANTE, João Roberto; ABREU, Ariane de Jesus Lopes de. COVID-19 no município do Rio de Janeiro: análise espacial da ocorrência dos primeiros casos e óbitos confirmados. **Epidemiologia e Serviços de Saúde**, v. 29, p. e2020204, 2020.

CHEN, Jiangpeng *et al.* Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang, China. **PloS one**, v. 10, n. 2, p. e0118521, 2015.

CHIU, Y. *et al.* Mortality and morbidity peaks modeling: An extreme value theory approach. **Statistical Methods in Medical Research**, v. 27, n. 5, p. 1498-1512, 2018.

CIRILLO, Pasquale; TALEB, Nassim Nicholas. Tail risk of contagious diseases. **Nature Physics**, v. 16, n. 6, p. 606-613, 2020.

COLES, Stuart *et al.* **An introduction to statistical modeling of extreme values**. London: Springer, 2001.

COTTA, Higor Henrique Aranda; CORRÊA, Wesley de Souza Campos; ALMEIDA ALBUQUERQUE, Taciana Toledo. Aplicação da distribuição de Gumbel para valores extremos de precipitação no município de Vitória-ES. **Revista Brasileira de Climatologia**, v. 19, 2016.

GILLELAND, Eric; KATZ, Richard W. extRemes 2.0: an extreme value analysis package in R. **Journal of Statistical Software**, v. 72, p. 1-39, 2016.

GUILLOU, Armelle; KRATZ, Marie; STRAT, Y. Le. An extreme value theory approach for the early detection of time clusters. A simulation-based assessment and an illustration to the surveillance of Salmonella. **Statistics in Medicine**, v. 33, n. 28, p. 5015-5027, 2014.

LIM, Jue Tao; DICKENS, Borame Sue Lee; COOK, Alex R. Modelling the epidemic extremities of dengue transmissions in Thailand. **Epidemics**, v. 33, p. 100402, 2020.

LISKA, Gilberto Rodrigues *et al.* Estimativas de velocidade máxima de vento em Piracicaba-SP via Séries Temporais e Teoria de Valores Extremos. **Revista Brasileira de Biometria**, v. 2, p. 295-309, 2013.

MA, Ning; BAI, Yanbing; MENG, Shengwang. Return period evaluation of the largest possible earthquake magnitudes in mainland China based on extreme value theory. **Sensors**, v. 21, n. 10, p. 3519, 2021.

MARTINS, Amanda Larissa Alves *et al.* Generalized Pareto distribution applied to the analysis of maximum rainfall events in Uruguaiana, RS, Brazil. **SN Applied Sciences**, v. 2, n. 9, p. 1479, 2020.

MATHIEU, Edouard *et al.* **COVID-19 pandemic**. Our World in Data, 2020.

NASCIMENTO, Igor Ferreira; NASCIMENTO, Alex Rodrigues; YAOHAO, Peng. Uma análise estatística comparativa das evidências de subnotificação da COVID-19 no Brasil. **Revista Eletrônica Gestão e Saúde**, v. 11, n. 3, p. 261-280, 2020.

OPAS. **Doença causada pelo novo coronavírus (COVID-19) OPAS/OMS**. 2021.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing. 2020.

STEPHENSON, Alec G. *et al.* evd: Extreme value distributions. **R News**, v. 2, n. 2, p. 31-32, 2002.

THOMAS, Maud *et al.* Applications of extreme value theory in public health. **PloS one**, v. 11, n. 7, p. e0159312, 2016.

THOMAS, Maud; ROOTZÉN, Holger. Real-time prediction of severe influenza epidemics using extreme value statistics. **Journal of the Royal Statistical Society Series C: Applied Statistics**, v. 71, n. 2, p. 376-394, 2022.

VASCONCELOS, Fernando Freire; MOURA, Heber José de. Elaboração de uma metodologia baseada em estatística para encaminhamento dos casos da COVID-19. **Revista de Administração Pública**, v. 54, p. 1417-1428, 2020.

WHO. **Brazil - WHO Coronavirus (COVID-19) Dashboard**. World Health Organisation 2020.

WONG, Felix; COLLINS, James J. Evidence that coronavirus superspreading is fat-tailed. **Proceedings of the National Academy of Sciences**, v. 117, n. 47, p. 29416-29418, 2020.

ZAMBRANO-BIGIARINI, Mauricio. **Goodness-of-fit functions for comparison of simulated and observed hydrological time series**. R Package Version 0.3-8, 2017.

ZHU, Yifan; CHEN, Ying Qing. On a statistical transmission model in analysis of the early phase of COVID-19 outbreak. **Statistics in Biosciences**, v. 13, n. 1, p. 1-17, 2021.

Authors' Contribution

All authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA)

Matiussi, A. C., Liska, G. R., Beijo, L. A., Ferreira, T. R., & Pedro, D. A. de M. (2026). EXTREME VALUE THEORY TO COVID-19 DATA: PROBABILISTIC ANALYSIS OF DAILY NEW CASES AND DEATHS IN BRAZIL. *Veredas Do Direito*, *23*(6), e235385. https://doi.org/10.18623/rvd.v23.5385