Research Article

# Method for Estimating Time Series Data of COVID-19 Deaths using a Gumbel Model

**Furutani H[1*], Hiroyasu T[1, 2], Okuhara Y[3]**

1AI x Humanity Research Center, Doshisha University, Kyoto, Japan
2Department of Biomedical Information, Doshisha University, Kyoto, Japan
3Medical School, Kochi University, Kochi, Japan

***Corresponding author:** Hiroshi Furutani, AI x Humanity Research Center, Doshisha University, Kyoto, Japan

**Received:** 27 December 2021; **Accepted:** 04 January 2022; **Published:** 11 January 2022

**Citation:** Furutani H, Hiroyasu T, Okuhara Y. Method for Estimating Time Series Data of COVID-19 Deaths using a Gumbel Model. Archives of Clinical and Biomedical Research 6 (2022): 50-64.

## Abstract

The purpose of the present paper is to introduce a method for forecasting the daily number of COVID-19 associated deaths. We apply the Gumbel distribution function for the analysis of time series data of the first-wave outbreak. The Gumbel distribution function $F_G(t)$ has the property of having a mode (peak) point $b$, where $F_G(b) = 1/2.718$. The probability density function of $F_G(t)$ has a single-peaked and right-skewed form. Using this distribution, we can estimate the total number of deaths $N$, and forecast daily numbers in the decreasing phase. We study the early-stage data of Italy, Canada, Belgium, Switzerland, the Netherlands, Sweden, Germany, China, France, and the United Kingdom. The data of New York City and England are also analyzed. In general, the Gumbel distribution reasonably reproduces the time series data of daily counts. However, significant discrepancies between theory and data are observed for China, France, and Sweden. We re-analyze the data of these countries and the Netherlands using different approaches.

**Keywords:** COVID-19; Forecast; Extreme values; Gumbel distribution; Daily number of deaths

## 1. Introduction

Coronaviruses are large, enveloped RNA viruses, the importance of which has been recognized through the identification of a newly emerged coronavirus as the causative agent of severe acute respiratory syndrome (SARS) [1]. The worldwide pandemic called COVID-19 is caused by the most recently discovered coronavirus, SARS-CoV-2, and is ``the third documented spillover of an animal coronavirus to humans in only three decades that has resulted in a major epidemic" [2]. The 2003 outbreak of SARS had a case-fatality rate of approximately 10% (774 deaths), and MERS killed 34% of people with illness between 2012 and 2019 (2,494 deaths). While the case-fatality rate of COVID-19 is a few percent, the associated pandemic has caused many more deaths worldwide [3]. The international research community is currently fighting this global health crisis. There have been numerous efforts to develop mathematical methods in the field of epidemiology [4, 5]. Kermack and McKendrick developed a fundamental theory of epidemics called the SIR model, which is composed of three compartments: Susceptible *S*, Infected *I*, and Recovered/Removed R [6]. Their approach is one of the most frequently used mathematical models for analyzing epidemics. The SIR model is defined by coupled ordinary differential equations including a term of quadratic nonlinearity *SI*. They also showed that the SIR model can be reduced to a logistic model under natural assumptions. This logistic model was applied to the analysis of death counts for the 1905–1906 plague of the island of Bombay, and it was shown that the logistic distribution can fit the weekly number of deaths very well.

The cumulative distribution function (CDF) of the logistic model is defined in the region $-\infty < t < \infty$ as

$$F_L(t) = \{1 + e^{-a(t-b)}\}^{-1}, \tag{1}$$

where $a > 0$. The mean and mode of the distribution are both $b$, and the variance is $\pi^2/(3a^2)$. The probability density function of the logistic model $f_L(t)$ is symmetrical about the $t = b$ axis.

The application of the logistic model to the study of the COVID-19 outbreak has been reported by many researchers. In China, COVID-19 caused a great threat to the public health system from mid-January till mid-March 2020. Analysis with the standard logistic distribution model demonstrated that this simple function can summarize the outbreak dynamic of COVID-19 for 20 provinces in China excluding Hubei [7]. The cumulative number of cases was very well described by a logistic growth model with a coefficient of determination $R^2$ greater than 0.98 for all provinces. The logistic model was also applied to the analysis of the COVID-19 epidemic in Italy [8-11]. The analysis of severely infected people from 2020/2/24 to 2020/4/1 in Italy and its five regions demonstrated that the logistic growth curve reproduces these time series data very well [10]. However, most data after the peak are not included in this analysis. A similar analysis of the daily number of infected people was carried out using the data from 2020/2/27 to 2020/6/3 [11]. After a smoothing procedure, it was shown that the time series data have a right-skewed distribution. Since the logistic distribution is symmetric, the fitting by this model cannot reproduce the data reported for Italy. Therefore, the authors introduced a new method using a logistic equation called the stretched logistic equation with a time-dependent growth rate. This extended logistic function reproduced the data of the entire period very well.

In the first wave of a pandemic in many regions, the daily plot of reported deaths is single-peaked and skewed to the right. Time series data generally have three phases: (A) an exponential increase in the early stage, (B) an intermediate stage changing from increasing to decreasing, and (C) a slowly decreasing final stage. Taking into account the characteristics of COVID-19 data, the present study makes use of extreme value (EV) statistics to estimate the daily numbers for (C) and the total number of deaths from the data of (A) and (B). Extreme value theory provides methods by which to analyze the maximum (or minimum) value samples derived from distributions, such as normal or exponential distributions. Extreme value theory has been used in many fields to estimate the probability of rare events from observed data [14]. The Fisher-Tippett type I distribution [15], i.e., the Gumbel distribution, is a distribution of maximum (or minimum) values from the data of several distributions. The Gumbel function for maximum values has a right-skewed form, and that for minimum values has a left-skewed form.

The present study considers the Gumbel function for maximum values. The CDF of this function is given as

$$F_G(t) = \exp\{-e^{-a(t-b)}\}. \tag{2}$$

A notable property of the Gumbel model is that we can estimate $N$ from the data of $t \leq b$. To this end, we make use of the value of $F_G(t)$ at the mode:

$$F_G(b) = e^{-1} = 1/2.718. \tag{3}$$

Since the partial sum of daily reports $N_1$ in the period $t \leq b$ is

$$N_1 = NF_G(b),$$

$N$ is estimated by

$$N = N_1/F_G(b) = e\,N_1 = 2.718N_1. \tag{4}$$

This means that the total number of deaths can be estimated from the early stage of daily data $t \leq b$. Using the relation $n(b) = Na/e$, where $n(b)$ is the maximum daily number at $t = b$, we can estimate $a$ by

$$a \simeq \frac{e\,n(b)}{N}.$$

The proposed method is applied to the first-wave data of the COVID-19 daily number of deaths in Italy, Canada, Belgium, Switzerland, the Netherlands, Sweden, Germany, China, the United Kingdom, England and three of its regions, and New York City.

## 2. Methods

### 2.1 Dataset

The present study uses four datasets for the analysis of daily COVID-19 deaths.

- Dataset A: For the analysis of data in Canada, France, Belgium, Netherlands, Germany, Italy, Switzerland, and China, we downloaded the dataset from the website of the European Center for Disease Prevention and Control

  https://www.ecdc.europa.eu/en/publications-data/.

  The dataset consists of historical data (to 14 December 2020) for the daily number of COVID-19 cases and deaths by country worldwide.

  The file name is ``COVID-19-geographic-distribution-worldwide-2020-12-14.xlsx"(accessed on 2021/4/21).

- Dataset B: We downloaded the daily number of COVID-19 deaths from the official website of New York City https://www1.nyc.gov/site/doh/covid/covid-19/.

  COVID-19 deaths are categorized as probable or confirmed.

  Confirmed death: Death within 60 days of a positive molecular test.

  Probable death: Cause of death on the health certificate is COVID-19 or similar, but a positive molecular test is not on record.

  The file name is ``deaths-by-day.csv" (accessed on 2021/5/15).
- Dataset C: We used the dataset of the official UK government website for data on coronavirus. https://www.coronavirus.data.gov.uk/details/deaths/ (accessed on 2021/4/22)

  The downloaded files include (1) the number of deaths of people whose death certificate mentioned COVID-19 as one of the causes and (2) the daily number of deaths within 28 days of a positive test.
- Dataset D: The dataset of England NHS is downloaded for the analysis https://www.england.nhs.uk/statistics /statistical work areas/covid-19-daily-deaths/ (accessed on 2021/5/1).

  This website contains the dataset of COVID-19 deaths in England and seven of its regions. The data of London and two regions are selected for the analysis. The present study treats the deaths of patients who have died in hospitals in England and have tested positive for COVID-19.

## 2.2 Gumbel model

The probability density function $f(t)$ of the Gumbel distribution is given by

$$f(t) = ae^{-y(t)}\, F_G(t), \quad y(t) = a(t-b). \tag{5}$$

The mean $\bar{t}$ and variance $V$ of the distribution are

$$\bar{t} = b + 0.5772/a, \quad V = \frac{\pi^2}{6a^2},$$

and the mode, which is the peak position of $f(t)$, is $b$. Useful information for the parameter estimation is obtained at $t = b$:

$$F_G(b) = \frac{1}{e}, \quad f(b) = \frac{a}{e}. \tag{6}$$

For the Gumbel distribution, it is easy to calculate the time $t = q(P)$ that satisfies the relation $F(t) = P$ with $0 < P < 1$. Using eq. (2), we obtain the following quantile:

$$q(P) = b - (\ln\ln\frac{1}{P})/a.$$

Quantiles for $P = 0.5$ and $0.95$ are

$$q(0.5) = b - (\ln\ln 2)/a = b + 0.3665/a, \quad q(0.95) = b + 2.9702/a,$$

where $q(0.5)$ is the median. We define the convergence time $t_c$ measured from the peak position by the quantile of $P = 0.95$ as

$$t_c = 2.9702/a. \tag{7}$$

## 2.3 Estimation of daily and total numbers

This subsection describes the process for estimating the daily data and the total number. The date is indexed as $t = 0$ when the first case was reported. Later, we will redefine the starting date for the convenience of comparing the data of different regions. We use the symmetric 7-day moving average $n(t)$ as the daily data in this analysis:

$$n(t) = (d_{t-3} + d_{t-2} + \cdots + d_{t+3})/7,$$

where $d_t$ denotes the reported daily number. Note that $n(t)$ includes the information of $d_{t+1}$, $d_{t+2}$, and $d_{t+3}$. In the process of calculations, note that, for the data of Switzerland, it is necessary to use the 9-day average because even data for the 7-day average fluctuate considerably around the trend line.

The accumulated number $U(t)$ is calculated using $d_i$:

$$U(t) = \sum_{i=0}^{t} d_i.$$

From $U(t)$, we estimate the distribution function $F_G(t)$.

The first step is the estimation of $N = \sum_t n(t)$. To this end, we used eq. (3). Therefore, it is necessary to estimate mode $b$. We define the estimated mode $t_b$, which takes an integer value recursively defined as

$$t_b = \arg\max_t f(t)\ (t \leq t_b) \quad \text{and} \quad f(t_b) = \max\{f(t_b), f(t_b+1), \ldots f(t_b+5)\}.$$

We define $N_1$ as the partial sum of daily data $\{d_t, t \leq t_b\}$. The estimate $N_e$ can be calculated as

$$N_1 = U(t_b), \quad N_e = eN_1 = 2.718\, N_1. \tag{8}$$

The next step is the estimation of the daily number $n(t)$. The CDF $F_G(t)$ is approximated using $U(t)$ and $N_e$:

$$\tilde{F}_G(t) = U(t)/N_e. \tag{9}$$

The linear function $y(t) = a(t-b)$ is estimated by regression analysis of the data of intermediate estimate $\tilde{y}$:

$$\tilde{y}(t) = -\ln\{-\ln\tilde{F}_G(t)\}.$$

Next, we rewrite the definition of $t$ using $\tilde{F}_G(t)$ of eq. (9). The first time for the analysis $t = 1$ is defined by the condition $\tilde{F}_G(t) > 0.01$.

## 2.4 Forecasting of daily numbers

We calculate the estimate of $y(t)$ from the intermediate estimate $\tilde{y}$ using the regression analysis program. After performing a linear regression of the dependent variable $\tilde{y}$ and the independent variable $t$, we obtain two estimated parameters, $a$ and $b$. The present study uses $t_{max}$ points for the regression:

$$\text{T: } 1, 2, \ldots, t_{max} \quad \text{and} \quad \text{Y: } \tilde{y}(1), \tilde{y}(2), \ldots, \tilde{y}(t_{max}).$$

Here, time (day) starts from $t = 1$. We define the estimate of $y(t)$ as

$$y_e(t) = a(t-b). \tag{10}$$

The goal is to calculate the estimate of $F_G(t)$:

$$F_e(t) = \exp\{-\exp[-y_e(t)]\} = \exp\{-\exp[-a(t-b)]\}. \tag{11}$$

The daily moving average $n(t)$ is estimated by

$$n_e(t) = \{F_e(t) - F_e(t-1)\}N_e.$$

There is another way to estimate $a$. From eq. (6), we obtain an estimate of $a$ as

$$\hat{a} = \frac{e\,n(t_b)}{N_e} = \frac{n(t_b)}{N_1}. \tag{12}$$

## 3. Results

The present study analyzes the daily numbers and the total number of deaths in the first wave of the COVID-19 pandemic. We selected the data of ten countries: Italy, Canada, Belgium, Switzerland, the Netherlands, Sweden, France, Germany, China, and the United Kingdom. One city in the United States, New York City, was also used. We investigate the data of England and three of its regions, including London. All time series plots in the present study are single-peaked and skewed to the right, except for the data for China.

### 3.1 Results using dataset A

Table 1 shows a summary of the analysis using Dataset A.

| | date of t=1 | $t_b$ | $N_e$ | $y_e$ (t) | $t_c$ |
|---|---|---|---|---|---|
| Italy | 2020/3/9 | 23 | 31,508 | 0.06547 (t-23.93) | 45.37 |
| Canada | 2020/4/2 | 33 | 10,009 | 0.04689 (t-32.56) | 63.34 |
| Belgium | 2020/3/21 | 22 | 10,911 | 0.07010 (t-22.33) | 42.37 |
| Switzerland | 2020/3/18 | 21 | 1,587 | 0.07799 (t-20.72) | 38.08 |
| Netherlands | 2020/3/19 | 19 | 4,800 | 0.08214 (t-19.46) | 36.16 |
| Sweden | 2020/3/20 | 26 | 3,961 | 0.05995 (t-27.39) | 49.54 |
| Germany | 2020/3/24 | 26 | 11,172 | 0.06243 (t-25.27) | 47.58 |
| China | 2020/1/26 | 22 | 4,529 | 0.06591 (t-23.12) | 45.06 |
| France | 2020/3/19 | 20 | 24,223 | 0.06633 (t-23.94) | 44.78 |

**Table 1:** Summary of Gumbel analysis using Dataset A.

**3.1.1 Italy, canada, belgium, and switzerland:** The results of the Gumbel analysis for (a) Italy, (b) Canada, (c) Belgium, and (d) Switzerland are shown in Fig. 1. The number of data points used in the regression analysis is **15** for these countries. The government of Italy extended the lockdown imposed on northern regions to the entire country on 2020/3/9, with containment measures including restrictions on travel and public gatherings [10]. The date 2020/3/9 corresponds to t=1 in Figure 1(a). In Canada, the first confirmed death was reported on 2020/3/8 [17]. The border was closed to non-citizens except travelers from the US on 2020/3/16. The border with the US was closed on 2020/3/22. Belgium reported the first confirmed death on 2020/3/10 [17]. The government of Belgium imposed a lockdown on 2020/3/17 (t=-3) with travel and gatherings prohibited and non-essential shops closed. The lockdown was extended until 2020/5/3 (t=44). Thus, the lockdown period of Belgium is shown in Figure 1(c). In Switzerland, the first COVID-19 death was reported on 2020/3/5. On 2020/3/13, the federal government closed all schools until 2020/4/4 and banned gatherings of more than 100 people. On 2020/3/16, the federal government declared the existence of an extraordinary situation. On 2020/3/20, the federal government announced that no complete lockdown would be implemented, but banned gatherings of more than 5 people.

**3.1.2 The netherlands and sweden:** The results of analysis for the Netherlands and Sweden are presented in Figs. 1(e) and 1(f). The number of data points used in the regression analysis is **15** for the Netherlands and **20** for Sweden. The theory estimates lower than reported values of daily death for both countries. In particular, this discrepancy is significant for Sweden. The Netherlands has taken a liberal policy course [18]. A mild lockdown was implemented with individual freedom and responsibility. Sweden has managed the pandemic with no lockdowns, less regulation, and more voluntary action. There has been much discussion on the policy of Sweden regarding the perspective of herd immunity [19].

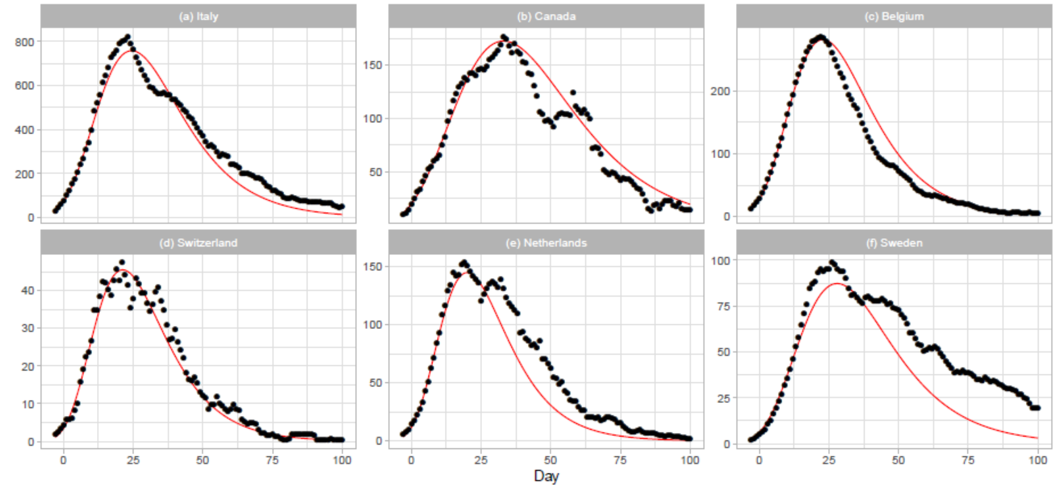

**Figure 1:** Gumbel model estimation based on the time series data for (a) Italy, (b) Canada, (c) Belgium, (d) Switzerland, (e) the Netherlands, and (f) Sweden. Vertical axes show daily death counts. The reported data points are indicated by black points, and the theoretical estimations are indicated by solid red lines. The parameters of the Gumbel distribution are presented in Table 1. The reported data show the 7-day moving averages of daily numbers except for Switzerland. The reported data for Switzerland are the data for a 9-day moving average.

**3.1.3 Germany and china:** In Figure 2, we show the analysis of the daily COVID-19 deaths for Germany and China. The solid red lines show the theoretical results of the Gumbel model. The number of data points used in the regression analysis is **15** for both countries. Germany is one of most affected countries in Europe. In March, Germany established several interventions to contain the virus spread, including the closure of schools and child day care facilities (2020/3/16-18) and a full lockdown with forced social distancing and closures of nonessential services (2020/3/23) [20]. Therefore, most of the data presented in Figure 2(a) were collected during the period of lockdown. In early December 2019, the Covid-19 outbreak began in the city of Wuhan, Hubei Province of China, and spread to other parts of China. On 2020/1/23,Wuhan was locked down. In the daily report of the dataset for China, the number of deaths on 2020/4/17 (t=83) is given as 1,290. We excluded this number from the analysis, considering that it may have been affected by a data gathering problem. In Figure 2(b), the result of logistic model estimation is presented by the dashed line. The parameters of the model are $\boldsymbol{N} = \mathbf{3,345}$, $\boldsymbol{a} = \mathbf{0.1827}$, and $\boldsymbol{b} = \mathbf{21.32}$. The parameters $\boldsymbol{a}$ and $\boldsymbol{b}$ are obtained by a linear regression analysis using the 22 daily numbers of $\mathbf{1} \leq \boldsymbol{t} \leq \mathbf{22}$.

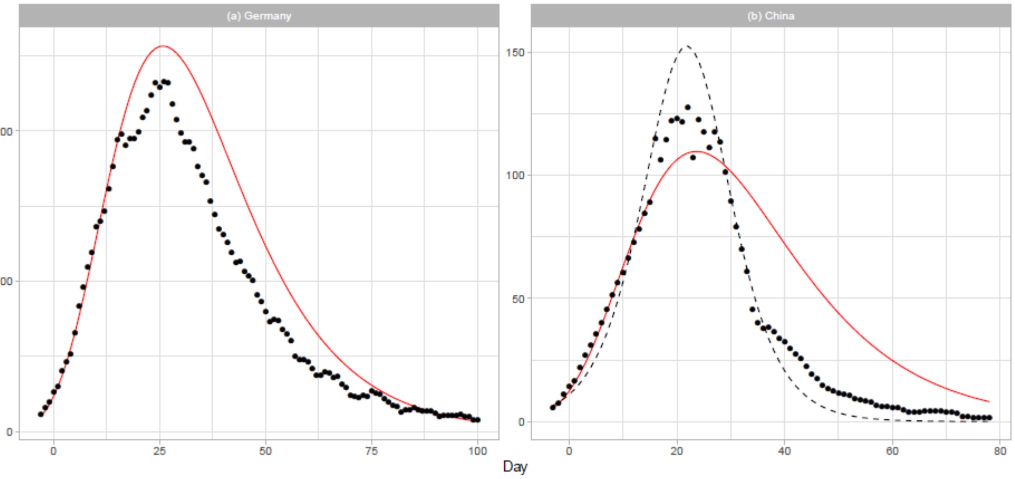

**Figure 2:** Gumbel model estimation based on the time series data for (a) Germany and (b) China. Vertical axes show daily death counts. The reported data are indicated by the points, and the theoretical estimations are indicated by the solid red lines. The points indicate 7-day moving averages of the reported data. The linear functions used for the Gumbel distributions are given in Table 1. For China, the logistic estimation is shown by the dashed line.

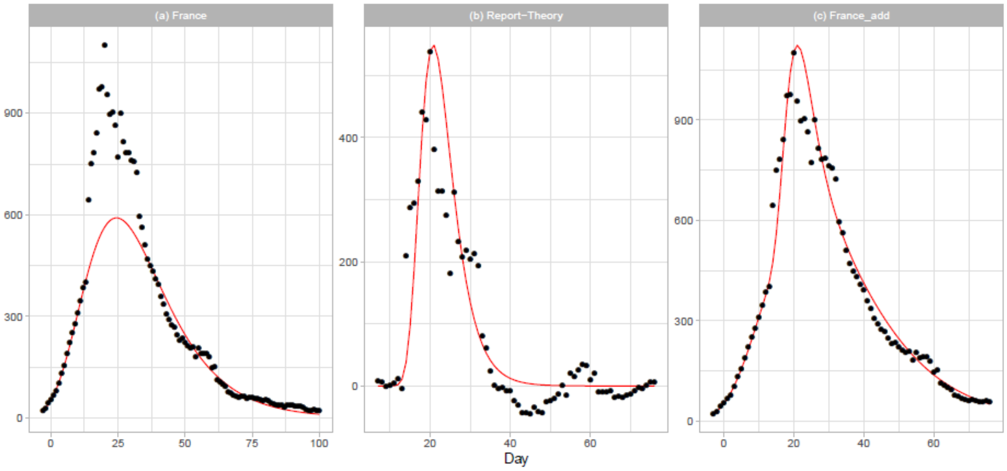

**Figure 3:** Gumbel model estimation based on the time series data for France: (a) the 7-day moving averages of the reported data (black points) and the Gumbel model estimation (solid red line), (b) difference of the reported data and the Gumbel model estimation (black points) and the second Gumbel model estimation (solid red line), and (c) the reported data ( black points) and the sum of the two Gumbel model estimations (solid red line). Vertical axes show

daily death counts.

**3.1.4 France:** In Figure 3, we show the results of Gumbel analysis for France. In Figure 3(a), the solid red line indicates the theoretical estimation of the daily number of deaths with the model parameters presented in Table 1. The number of data used in the analysis is 15. Points show 7-day moving averages of the reported daily number of deaths. The government of France imposed a nationwide lockdown from 2020/3/17 to 2020/5/11 ($-1 \leq t \leq 54$ in Figure 3(a)) [17]. In contradiction with the implementation of severe restrictions, the government decided to allow municipal elections to proceed as scheduled starting on 2020/3/15, with minimal changes to voting procedures. Figure 3(b) shows the difference between the reported data and the theoretical estimation presented in Figure 3(a). The points are $\{\text{reported data} - \text{Gumbel model estimation}\}$, and the solid red line is the Gumbel model fitting to the difference. The fitting parameters are Ne=6,000, a=0.2497, and b=20.24. The parameters $a$ and $b$ are calculated using 20 data points within the region $14 \leq t \leq 33$. Figure 3(c) shows the reported data by points, and the sum of two Gumbel model estimations are shown by the solid red lines.

### 3.2 New york city

New York City (NYC) was an epicenter of the COVID-19 outbreak in the United States during spring 2020 [21]. On 2020/3/7, the governor declared a state of emergency in New York State. Th e mayor of NYC announced that all schools, bars, and restaurants in the city were to be closed from 2020/3/17, except for food takeout and delivery. In Figure 4, we report the results of our analysis of the daily data for NYC in Dataset B. The distribution parameters are presented in Table 2. Parameters $a$ and $b$ are obtained by a linear regression analysis using the 15 daily numbers of $1 \leq t \leq 15$. Figure 4(a) shows the result of analysis of the sum of probable and confirmed deaths, and Figure 4(b) shows the result of analysis of the sum of confirmed deaths.

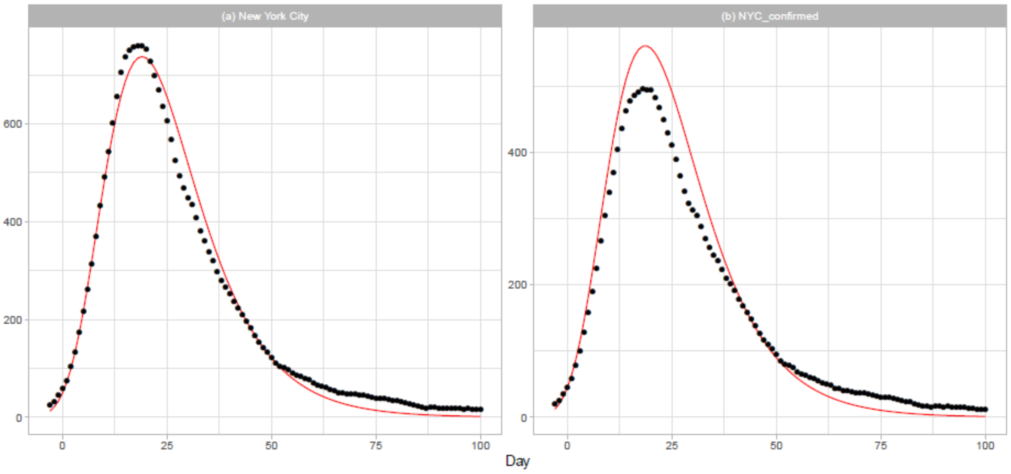

**Figure 4:** Gumbel model estimation based on the time series data for New York City: (a) probable and confirmed deaths, and (b) confirmed deaths. Vertical axes show daily death counts. The 7-day moving averages of daily data are indicated by the points, and the theoretical estimations are indicated by the solid red lines. Parameters for the

estimation are given in Table 2.

| | | date of $t = 1$ | $t_b$ | $N_e$ | $y_e(t)$ |
|---|---|---|---|---|---|
| New York City | (a) reported | 2020/3/22 | 18 | 22,624 | 0.08862 (t-18.51) |
| | (b) confirmed | 2020/3/22 | 18 | 17,280 | 0.08830 (t-18.12) |
| United Kingdom | (a) certificate | 2020/3/20 | 22 | 46,510 | 0.06867 (t-23.46) |
| | (b) 28days | 2020/3/20 | 22 | 36,925 | 0.06852 (t-22.91) |
| England | (c) hospital | 2020/3/19 | 21 | 26,675 | 0.07048 (t-22.42) |
| London | (d) hospital | 2020/3/17 | 21 | 7,127 | 0.07292 (t-21.83) |
| North East | (d) hospital | 2020/3/22 | 17 | 2,582 | 0.09262 (t-17.84) |
| South West | (d) hospital | 2020/3/20 | 20 | 1,166 | 0.07204 (t-21.67) |

**Table 2:** Parameters of Gumbel estimation for analyzing Datasets B, C, and D.

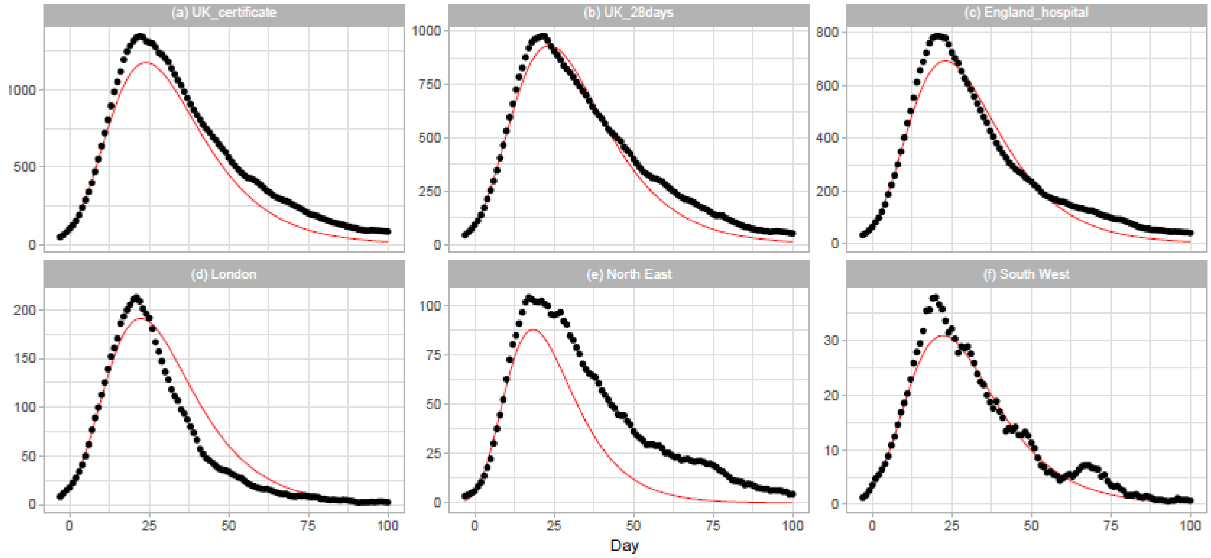

**Figure 5:** Gumbel model estimation based on the time series data for the United Kingdom: (a) the United Kingdom, death certificates, (b) the United Kingdom, deaths within 28 days of the first positive test, (c) England, (d) London, (e) North East and Yorkshire, and (f) South West. Vertical axes show daily death counts. The 7-day moving averages of the reported data are indicated by the points, and the theoretical estimations are indicated by the solid red lines. Parameters for the estimation are given in Table 2.

### 3.3 United kingdom

On 2020/3/23, the United Kingdom Government imposed the instruction to stay at home, and people were allowed to leave their home for very limited purposes. The government closed all shops selling nonessential goods, stopped all gatherings of more than two people in public, and banned all social events. The results of Gumbel analysis for the

United Kingdom are shown in Figure 5. The number of data points used in the regression analysis is 15. The daily numbers of Dataset C are used in Figs. 5(a) and 5(b), and those of Dataset D are used in Figures 5(c), 5(d), 5(e), and 5(f). Figure 5(a) shows the results for daily counts of deaths of people whose death certificate mentioned COVID-19 as one of the causes, and Fig. 5(b) shows the results for daily counts of deaths of people who had had a positive test result for COVID-19 and died within 28 days of the first positive test. We downloaded daily records from the database of the National Health Service (NHS) of England for the analysis of England and three regions of the NHS of England: London, North East and Yorkshire, and South West. The dataset includes information on deaths of patients who have died in hospitals in England.

### 3.4 Different approach for the netherlands and sweden

An optimistic estimation or underestimation may cause the risk of overwhelming health services. The estimations for the Netherlands and Sweden in Figure 1 are such cases. The reported data are far larger than the estimated deaths after $t > t_b$. Therefore, we show an approach to modify estimation parameters after the date of the peak. If reported numbers are larger than those estimated by an epidemiologic model, we have to modify the initial estimation parameters. First, we replace the number of total deaths $N_e$ with $N_e'$. For simplicity, we use three cases of $N_e'$: $1.1N_e$, $1.3N_e$, and $1.5N_e$. For each parameter $N_e'$, we calculate the Gumbel parameters $a$ and $b$ by linear regression. For this analysis, we use the cumulative number $U(t)$ in the period $t_b \leq t \leq t_b + 14$. Thus, the proposed method gives the estimation for $t \geq t_b + 15$. Table 3 presents the renewed Gumbel parameters for two countries. Figure 6 shows the results of new estimation. The result for the Netherlands in Figure 6(a) shows that the estimated number of daily deaths decreases along the line of $1.3N_e$. In Figure 6(b), the reported data for Sweden decrease between the lines of $1.3N_e$ and $1.5N_e$ and finally decrease more slowly than the estimation of $1.5N_e$. The reported data for Sweden indicate an increase in daily cases.

| | $N_e'$ | $y_e$ (t) |
|---|---|---|
| **The Netherlands** | 1.1 $N_e$ | 0.07833 (t-20.30) |
| | 1.3 $N_e$ | 0.06164 (t-22.68) |
| | 1.5 $N_e$ | 0.05228 (t-25.27) |
| **Sweden** | 1.1 $N_e$ | 0.05548 (t-27.32) |
| | 1.3 $N_e$ | 0.04510 (t-30.68) |
| | 1.5 $N_e$ | 0.03891 (t-34.16) |

**Table 3:** Parameters of Gumbel estimation for renewed analysis of the Netherlands and Sweden.

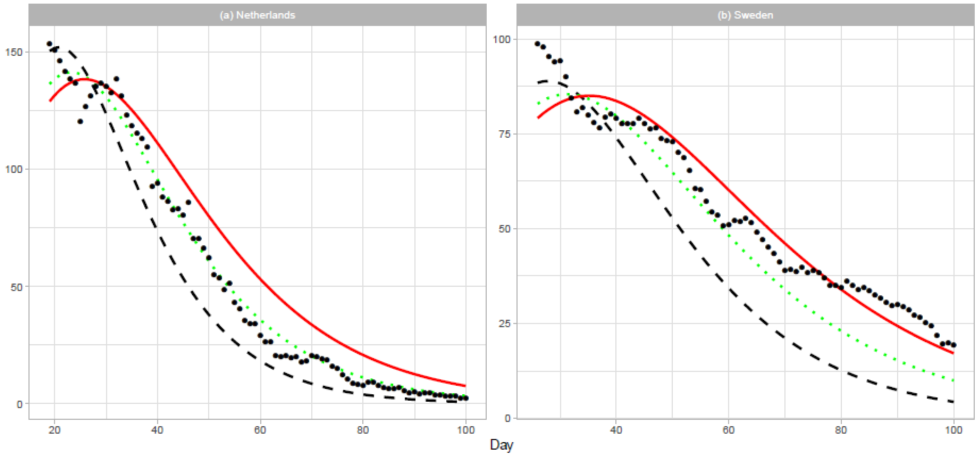

**Figure 6:** Renewed Gumbel model estimation based on the time series data of (a) the Netherlands and (b) Sweden. Vertical axes show daily death counts. The reported data are indicated by the points, and the theoretical estimations are indicated by dashed lines for $N_e' = 1.1N_e$, dotted lines for $1.3N_e$, and solid red lines for $1.5N_e$.

## 4. Discussion

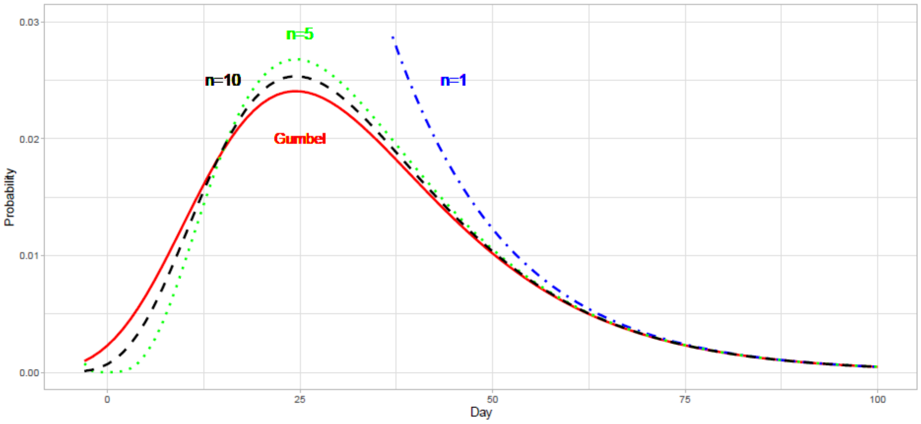

**Figure 7:** Gumbel distribution as a generalization of an exponential distribution.

Some approaches use the phenomenological model, initially assuming a certain type of distribution function in data analysis. The most popular approach is an exponential distribution, the CDF of which is defined as

$$F_X(t) = 1 - e^{-a(t-b)}, \quad \text{where} \quad a > 0, t \geq b, \tag{13}$$

In the epidemiology analysis, many studies suggest that this model can reasonably reproduce the data in the decreasing phase (e.g., [5]). The logistic distribution $F_L(t)$ has been also applied in the field of epidemiology, as shown in the present paper. The important point is that the logistic model is derived from the SIR model with very natural assumptions [22]. Thus, this distribution has the background of a dynamical model of infection. For large $t$, the logistic distribution is approximated by the following exponential distribution:

$$F_L(t) \approx F_X(t).$$

The third candidate is the Gumbel distribution. There are two types of Gumbel functions: the Gumbel function for maximum values and the Gumbel function for minimum values. The present study used the Gumbel function for maximum values. In a similar way as the logistic distribution, the Gumbel distribution approaches the exponential distribution for large $t$

$$F_G(t) \approx F_X(t).$$

Furthermore, the Gumbel distribution is generated from $F_X(t)$, as shown in Chapter 5 of [12]

$$F_n(t) = (1 - e^{-a(t-b)-\ln n})^n = (1 - \frac{1}{n} e^{-a(t-b)})^n \approx F_G(t), \tag{14}$$

for large $n$. Figure 7 shows the Gumbel function and exponential functions $F_n(t)$ for $n = 1$, $n = 5$, and $n = 10$ in eq. (14). Here, $F_1(t) = F_X(t)$. Parameters $a$ and $b$ are those of Italy in Table 1. Gompertz proposed a formula for describing the mortality rates of the elderly. This law states that death rates increase exponentially with age [23]. Gumbel discussed the Gompertz model in one section (6.3.5 Oldest Ages) of his book[12]. He applied the Gumbel model of minimum values for the analysis of mortality. The CDF is given by

$$F_g(t) = 1 - \exp\{-e^{a(t-b)}\}.$$

The survival function of the Gompertz model $S(t)$ is defined for $t \geq 0$. It can be shown [24] that $S(t)$ is approximated as the survival function of the Gumbel distribution for $-\infty < t < \infty$,

$$S(t) \approx 1 - F_g(t) = \exp\{-e^{a(t-b)}\}.$$

There are theoretical efforts to derive the Gumbel distribution of the smallest values to explain the phenomena of exponentially increasing mortality. Abernethy (1979) applied EV theory to this problem [25]. Their study used an analogy between the death of an organism and the failure of a multicomponent system and mathematically derived this type of Gumbel distribution. We hope that this line of approach can explain the applicability of the Gumbel distribution of maximum values to pandemic data.

The present study demonstrates the advantage of the EV theory for investigating time series data of deaths in pandemic outbreaks. We make use of one class of Gumbel distribution for the extremes of maximum values. The Gumbel distribution provides a means of extrapolation in time with theoretical justification by the EV theory and can be used to estimate the data in the decreasing phase from the data in the exponentially increasing phase. The Gumbel distribution reproduces the time series data of daily counts in many regions. However, notable discrepancies between estimation and reported data are observed for China, France, and Sweden. Therefore, we try to analyze the data of these countries under different approaches. In many regions, it is reported that the outbreak of a second wave of the pandemic began before the first wave of COVID-19 converged to a final stage [26]. We carried out calculations for the first wave data of the United States using Dataset A. The estimated total number of deaths $N_e$ is approximately 100,000. However, the second wave of the pandemic began during the decreasing phase of the first wave, so we cannot compare the time series of reported data and theoretical estimation for the entire period. We have designed a tool for analyzing the first and second waves of a pandemic simultaneously and intend to report this tool in the future.

## Author Contributions

The present study was conducted equally by the authors. H.F. wrote the initial manuscript.

## Competing Interests

The authors declare no competing interests.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

## References

1. Masters PS. The molecular biology of coronaviruses. Adv. Virus Res 66 (2006): 193-292.
2. V'kovski P, Kratzel A, Steiner S, Stadler H, Thiel V. Coronavirus biology and replication: implications for SARS-Cov-2. Nat. Rev. Microbiol (2020): 1-16.
3. Mahase E. Coronavirus: covid-19 has killed more people than SARS and MERS combined, despite lower case fatality. BMJ 368 (2020).
4. Brauer F. Compartmental models in epidemiology. In: Mathematical Epidemiology. Lecture Notes in Mathematics 1945 (2008): 19-79.
5. Chowell G, Sattenspiel L, Bansal S, Viboud C. Mathematical models to characterize early epidemic growth: A review. Phys. Life Rev 18 (2016): 66-97.
6. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proc. Roy. Soc. London 115 (1927): 700-721.
7. Zou Y, Pan S, Zhao P, Lei Han, Xiaoxiang Wang, Lia Hemerik, et al. Outbreak analysis with a logistic growth model shows COVID-19 suppression dynamics in China PLoS ONE 15 (2020): e0235247.
8. Postnikov E B. Estimation of COVID-19 dynamics ``on a back-of-envelope'': Does the simplest SIR model provide quantitative parameters and predictions?. Chaos Solitons Fractals 135 (2020): 109841.
9. Efim Pelinovsky, Andrey Kurkin, Oxana Kurkina, Maria Kokoulina, Anastasia Epifanovaa. Logistic equation and COVID-19. Chaos Solitons Fractals 140 (2020): 110241.
10. Gabriele Martelloni, Gianluca Martelloni. Analysis of the evolution of the Sars-Cov-2 in Italy, the role of the asymptomatics and the success of Logistic model. Chaos Solitons Fractals 140 (2020): 110150.
11. Consolini G, Materassi M. A stretched logistic equation for pandemic spreading. Chaos Solitons Fractals 140 (2020): 110113.
12. Gumbel EJ. Statistics of Extremes, Columbia University Press, New York (1958).
13. Coles S. An Introduction to Statistical Modeling of Extreme Values, Springer-Verlag, London (2001).
14. Asadi ZS, Melchers RE. Extreme value statistics for pitting corrosion of old underground cast iron pipes. Reliab. Eng. Syst. Saf. 162 (2017): 64-71.
15. Fisher RA, Tippett LHC. Limiting forms of the frequency distributions of the largest or smallest numbers of

a sample. Proc. Cambridge Phil. Soc. 24 (1928): 180-190.

16. Carter DJT, Challenor PG. Methods of fitting the Fisher-Tippett type 1 extreme value distribution. Ocean Engng 10 (1983): 191-199.
17. Desson Z, Weller E, McMeekin P, Ammi M. An analysis of the policy responses to the COVID-19 pandemic in France, Belgium and Canada. Health Policy Technol 9 (2020): 430-446.
18. Hoekman LM, Smits MMV, Koolman X. The Dutch COVID-19 approach: Regional differences in a small country. Health Policy Technol 9 (2020): 613-622.
19. Jones D, Helmreich S. A history of herd immunity. Lancet 396 (2020): 810-811.
20. Wieland T. A phenomenological approach to assessing the effectiveness of COVID-19 related nonpharmaceutical interventions in Germany. Saf. Sci. 131 (2020): 104924.
21. Ana S Gonzalez-Reiche, Matthew M Hernandez, Mitchell J Sullivan, Brianne Ciferri, Hala Alshammary, Ajay Obla, et al. Introductions and early spread of SARS-Cov-2 in the New York City area. Science 369 (2020): 297-301.
22. Barlow NS, Weinstein SJ. Accurate closed-form solution of the SIR epidemic model. Physica D 408 (2020): 132540.
23. Gompertz B. On the nature of the function expressive of the law of human mortality, and on a new method of determining the value of life contingencies. Phil. Trans. R. Soc. 115 (1825): 513-585.
24. Pflaumer P. Life table forecasting with the Gompertz distribution. JSM Proc. Alexandria, VA (2007): 3564-3571.
25. Abernethy JD. The exponential increase in mortality rate with age attributed to wearing-out of biological components. J. Theor. Biol. 80 (1979): 333-354.
26. Cacciapaglia G, Cot C, Sannino F. Second wave COVID-19 pandemic in Europe: A temporal playbook. Sci. Rep. 10 (2020): 15514.