Analysis of Case Fatality Rate Related to Coronavirus Disease 2019

71671086

University of Pennsylvania, USA

Abstract. In this article, we provides a basic idea to classify the Case Fatality Rate (CFR) of Coronavirus Disease 2019 (COVID-19) in terms of region. We propose two different Bayesian mixture models. The first model assumes the low CFR group following truncated normal distribution and the high CFR group following normal distribution, whereas the second model assumes Gamma distribution and Beta distribution respectively. We implement the mixture model by the application of EM algorithm. The simulation results shows that the truncated normal-normal distribution mixture model has a greater predictive power and the results is robust to all level of CFR; and the Gamma-Beta mixture model can only accurately classify the extreme cases of CFR level.

Keywords: COVID-19 · case fatality rate · Bayesian mixture model · EM algorithm

1 Introduction

Coronavirus disease 2019 (COVID-19) is caused by virus known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The disease was first diagnosed in Wuhan, the capital of Hubei province of China, and rapidly spread all over the world, causing an epidemic threat to global health [4].

According to World Health Organization (WHO), the severity of COVID-19 varies [6]. There are mild cases reported having no symptoms whereas a few severe or critical cases died within eight weeks after symptom. The case fatality rate (CFR) in different locations are reported to be very heterogeneous [5]. For example, in 17 April 2020, the CFR in Italy is 13.12%; while the CFR in Iceland is 0.46% [2]. In such a case, it is important for us to analysis the difference of CFR in different regions all around the world so that we are able to provide a general impression of such difference in order to account the location-specific effects of CFR in the future such as the economy, policy, and weather.

In this article, we introduce our data and use the Bayesian approach to classify the CFR of different regions into two groups with relatively high CFR and relatively low CFR using Bayesian mixture model and EM algorithm in Section 2. The data, numerical realization, and results analysis are described in Section 3.

2 Methodology

In this section, we introduce the theoretical Bayesian mixture model to classify the CFR of regions into two groups, which have relatively high CFR and relatively low CFR respectively.

2.1 Basic Model

For simplicity, we first assume the low CFR group follows truncated normal distribution with mode 0 and variance σ^2; and the high CFR group follows the normal distribution with mean μ and same variance as the low CFR group. The assumption of truncated normal distribution is reasonable because the CFR cannot be negative number and we need the mode of lower CFR group to be small. Moreover, for the assumption of higher CFR group, although the assumption implies that CFR is possible to be a negative number, the probability of negative CFR is extremely small. So it is proper for us to assume the distribution of high CFR group to be normal. We use EM algorithm to model this Bayesian mixture model. We denote the CFR of region i as r_i,

$$r_i \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & \alpha \\ Truncated\ Normal\,(0, \sigma^2) & 1-\alpha \end{cases}$$

where α is the probability of region i's CFR belonging to the high CFR group. We denote an indicator variable I_i as follows,

$$I_i = \begin{cases} 1 & region\ i \in high\ CFR\ group \\ 0 & region\ i \in low\ CFR\ group \end{cases}$$

The joint probability density function is

$$\begin{aligned} Pr(\boldsymbol{r}, \boldsymbol{I}|\mu, \sigma^2, \alpha) &= \prod_{i=1}^{n} \Big(\alpha\, \mathcal{N}((\mu, \sigma^2))\Big)^{I_i} \Big((1-\alpha)\; Truncated\; Normal\; ((0, \sigma^2))\Big)^{(1-I_i)} \\ &= \prod_{i=1}^{n} \Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}\Big)^{I_i} \Big(\frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} e^{-\frac{r_i^2}{2\sigma^2}}\Big)^{1-I_i} \alpha^{I_i}(1-\alpha)^{1-I_i}. \end{aligned}$$

Now, we are able to derive the complete log-likelihood function,

$$\begin{aligned} l(\mu, \sigma^2, \alpha|\boldsymbol{r}, \boldsymbol{I}) &= \log \mathcal{L}(\mu, \sigma^2, \alpha|\boldsymbol{r}, \boldsymbol{I}) \\ &= \log \Big\{ \prod_{i=1}^{n} \Big(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r_i-\mu)^2}{2\sigma^2}}\Big)^{I_i} \Big(\frac{\sqrt{2}}{\sqrt{\pi}\sigma} e^{-\frac{1}{2}\frac{r_i^2}{\sigma^2}}\Big)^{1-I_i} \alpha^{I_i}(1-\alpha)^{1-I_i} \Big\} \\ &= k + \sum_{i=1}^{n} \Big\{ -\frac{1}{2}\log\sigma^2 - \frac{r_i^2}{2\sigma^2} + \frac{I_i\mu}{2\sigma^2}(2r_i - \mu) + I_i \log\alpha + (1-I_i)\log(1-\alpha) \Big\} \end{aligned}$$

where k is a constant. Now, we are able to derive the EM algorithm for our mixture model.

E-Step:
First, we take the expectation with respect to I_i.

$$\begin{aligned} \hat{I}_i &= E(I_i|\sigma^2, \mu, r_i) = P(I_i = 1|\sigma^2, \mu, r_i) \\ &= \frac{\alpha\, \mathcal{N}(\mu, \sigma^2)}{\alpha\, \mathcal{N}(\mu, \sigma^2) + (1-\alpha)\, Truncated\; Normal\; (0, \sigma^2)}. \end{aligned}$$

M-Step:
We plug in the expected value of I_i into the complete data log likelihood and then find maximum likelihood estimators for μ, σ^2, α. We substitute the $\hat{I}_i$ for I_i, then we are able to obtain

$$l(\mu, \sigma^2, \alpha|\boldsymbol{y}, \hat{\boldsymbol{I}}) = k + \sum_{i=1}^{n} \Big\{ -\frac{1}{2}\log\sigma^2 - \frac{r_i^2}{2\sigma^2} + \frac{\hat{I}_i\mu}{2\sigma^2}(2r_i - \mu) + \hat{I}_i \log\alpha + (1-\hat{I}_i)\log(1-\alpha) \Big\}.$$

Now, we take partial derivative to the μ, σ^2, α and set them to 0, then we are able to obtain the maximum estimators

$$\begin{cases} \hat{\mu} = \frac{\sum_{i=1}^{n} \hat{I}_i r_i}{\sum_{i=1}^{n} \hat{I}_i} \\ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n} + \frac{\mu^2 \sum_{i=1}^{n} \hat{I}_i}{n} - \frac{2\mu \sum_{i=1}^{n} \hat{I}_i r_i}{n} \\ \hat{\alpha} = \frac{\sum_{i=1}^{n} \hat{I}_i}{n}. \end{cases}$$

2.2 Developed Model

Although it would not cause any serious issue for us to make assumption about normally distributed CFR for simplicity, theoretically it still has possibility that the CFR is out of the defined range (i.e. $r_i \notin [0, 1]$). So, we modify our assumptions that now we are assuming that CFR in high CFR group follows beta distribution as beta distribution is particular statistically suitable for modeling the random behavior of proportions [3]. Moreover, instead of assuming the low CFR group follows the normal distribution truncated at 0 on the left, we assume the low CFR group follows Gamma distribution. With our new assumptions, we have

$$r_i \sim \begin{cases} Beta(\alpha_0, \beta_0), & \omega \\ Gamma(\alpha_1, \beta_1), & 1-\omega. \end{cases}$$

where ω is the probability of region i's CFR belonging to the high CFR group. We again, denote an indicator variable I_i as follows,

$$I_i = \begin{cases} 1 & region\; i \in high\; CFR\; group \\ 0 & region\; i \in low\; CFR\; group \end{cases}$$

With the assumption of non-informative prior distributions for α_0, β_0, α_1, β_1 ($Pr(\alpha_0, \beta_0) \propto 1, Pr(\alpha_1, \beta_1) \propto 1$), the posterior joint probability density function is

$$Pr(\boldsymbol{r}, \boldsymbol{I}|\mu_0, \mu_1, \sigma^2, \omega) = \prod_{i=1}^{n} \Big(\omega\ Beta(\alpha_0, \beta_0)\Big)^{I_i} \Big((1-\omega) Gamma(\alpha_1, \beta_1)\Big)^{1-I_i}.$$

So the complete log-likelihood function is

$$l(\boldsymbol{r}, \boldsymbol{I}|\mu_0, \mu_1, \sigma^2, \omega) = \sum_{i=1}^{n} \Big\{ I_i \log\big(\omega Beta(\alpha_0, \beta_0)\big) + (1-I_i) \log\big((1-\omega) Gamma(\alpha_1, \beta_1)\big) \Big\}$$

Now, we are able to derive the EM algorithm of our mixture model.

E-Step:
First, we take the expectation with respect to I_i.

$$\begin{aligned} \hat{I}_i &= E(I_i|\sigma^2, \mu_0, \mu_1, r_i) = P(I_i = 1|\sigma^2, \mu_0, \mu_1, r_i) \\ &= \frac{\omega\ Beta(\alpha_0, \beta_0)}{\omega\ Beta(\alpha_0, \beta_0) + (1-\omega)\ Gamma(\alpha_1, \beta_1)}. \end{aligned}$$

M-Step:
There is no analytic solution to the maximization step. Thus, we use optimization algorithm to implement M-Step numerically.

3 Results and Analysis

3.1 Data Pre-Processing

We analyze using the CFR of COVID-19 data published by Our World in Data Organization (ourworldindata.org), which updated daily including data on confirmed cases, death, and testing for such disease [2]. The raw data contains the number of cumulative confirmed cases and cumulative deaths from 31 December 2019 to 17 April 2020 in 211 regions around the world. Total CFR for each region is simply obtained by dividing the number of cumulative death at 17 April 2020 by the number of cumulative confirmed cases at the same day in such region. Next, to exclude the bias, we only select the regions with relatively heavy disease where the total number of confirmed cases is greater than or equal to 1000. After selection, there are 92 regions left.

3.2 Simulation and Results Analysis

We implement the EM algorithm for basic model we derived before and use 6 sets of different initial values to see the convergence of our algorithm. We set the stopping criteria to be the maximum absolute value of difference for two consecutive iteration among all three parameters (μ,σ^2,α) to be smaller than 10^{-12} or the number of iteration exceed 1000. The EM algorithm using all sets of initial values converges within 300 iteration. We plot the convergence of the parameters of interest and the absolute difference between two consecutive iterations for two different sets of initial values in Fig. 1. The result of our mixture model is that we classify the regions into two groups, and one has relative high CFR which is normally distributed with mean CFR 0.105 and variance 0.00182 whereas the other one group has relatively low CFR which assumed to follow normal distribution left truncated at 0 with the same variance as the high CFR group. According to our model, there are approximately 13% regions with confirmed cases more than or equal to 1000 belonging to the high CFR group. Fig. 2 provides a histogram of CFR with curves for the fitted densities of the mixture components.

Next, we implement our developed model and follows the same analysis procedure as the basic model. Our EM algorithm converges to two different sets of parameters and same result is obtained using six different sets of initial values. Specifically, we choose two different sets of initial values and plot the convergence in Fig. 3. The two different sets of result parameters are displayed in the Table 1. Theoretically, we are supposed to select the set with the greater log-likelihood value. However, since Set 2 implies that the mean of low CFR group is higher than the mean of high CFR group which is unrealistic, we conclude that the Set 1 is the final results of our developed model. Specifically, the high CFR group follows Beta distribution with two shape parameters $\alpha_0 = 2.3412$, $\beta_0 = 36.454$, and the low CFR group follows Gamma distribution with shape parameter $\alpha_1 = 1.3533$ and rate parameter $\beta_1 = 47.569$, and 50.017% regions belong to the high CFR group. Fig. 4 provides a histogram of CFR with curves for the fitted densities of the mixture components for the developed model. Our histogram of CFR with fitted densities of the mixture components for developed model are shown in Fig. 4.

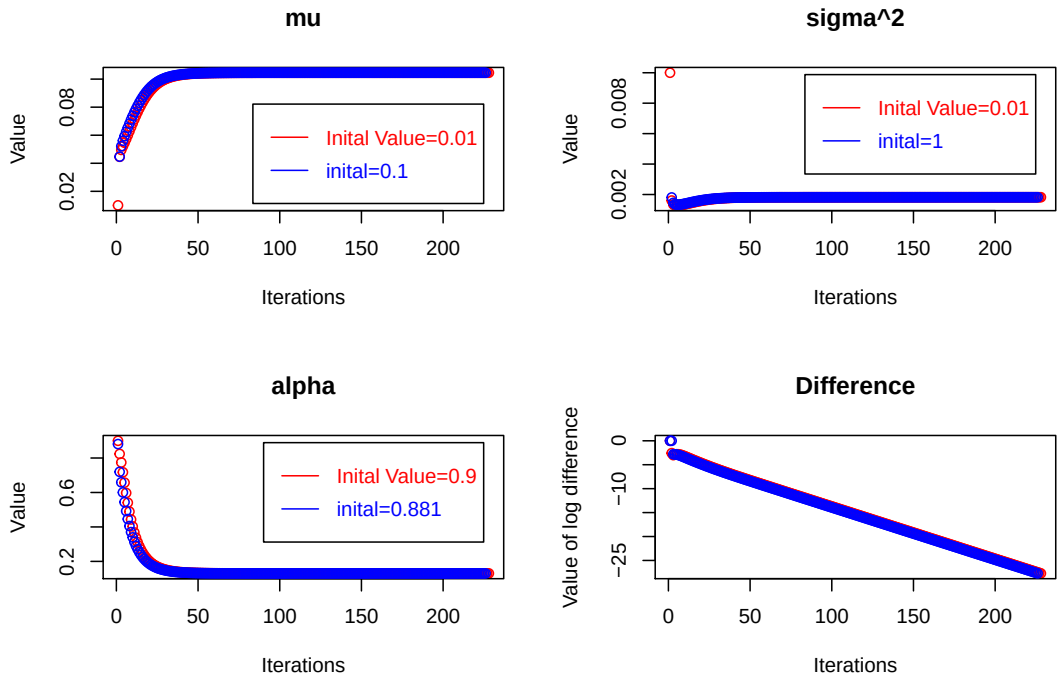

Fig. 1. Convergence of Parameters of Interest and Log Absolute Difference. Four sub-figures plot the convergence of EM algorithm for two different initial value sets. Specifically, the top-left, top-right, and bottom-left sub-figures describe the convergence of parameter μ, σ^2, and α against the number of iterations respectively. The bottom-right sub-figure displays the difference of two consecutive iterations against the number of iterations on the log scale.

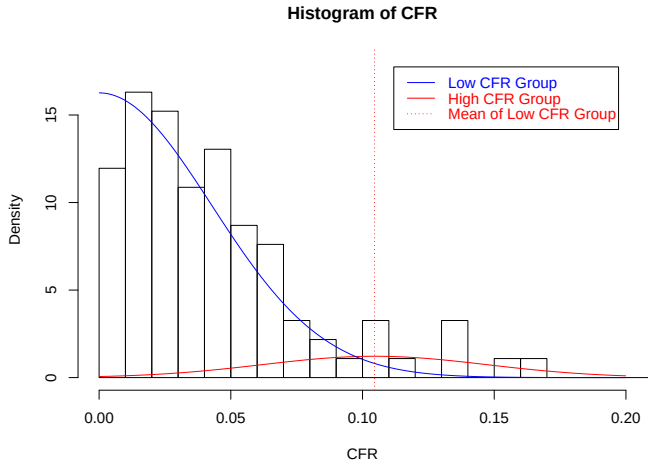

Fig. 2. Histogram of CFR with Fitted Densities of the Mixture Components. The plot shows the histogram of CFR for 92 target regions. The blue line is the distribution of low CFR group and the red line is the distribution of the high CFR group. Moreover, the dashed red line shows the mean of the high CFR group.

We use our fitted distribution to classify the a few regions, estimating the probability of each region belonging to the high CFR group. We select 6 representative regions and the classification results are shown in the Table 2. As we can observe from the predictive results table, both basic and developed model show a quite good performance in terms of the extremely low CFR (Australia, Russia) and extremely high CFR (Italy, United Kingdom). The basic model has strong predictive power for all CFR values. However, for the developed model, the regions with CFR values relatively in the middle level (Japan, United States) cannot be classified accurately. For example, the CFR in United States is much lower than the CFR in United Kingdom, but the developed model concludes that United States has more chance to be classified into the high CFR group than the United Kingdom. The reason of such problem may caused by the underestimate of the proportion of low CFR region in the low CFR group, and the overestimate of the low CFR region in the high CFR group. Therefore, the basic model is more accurate to classify the level of CFR.

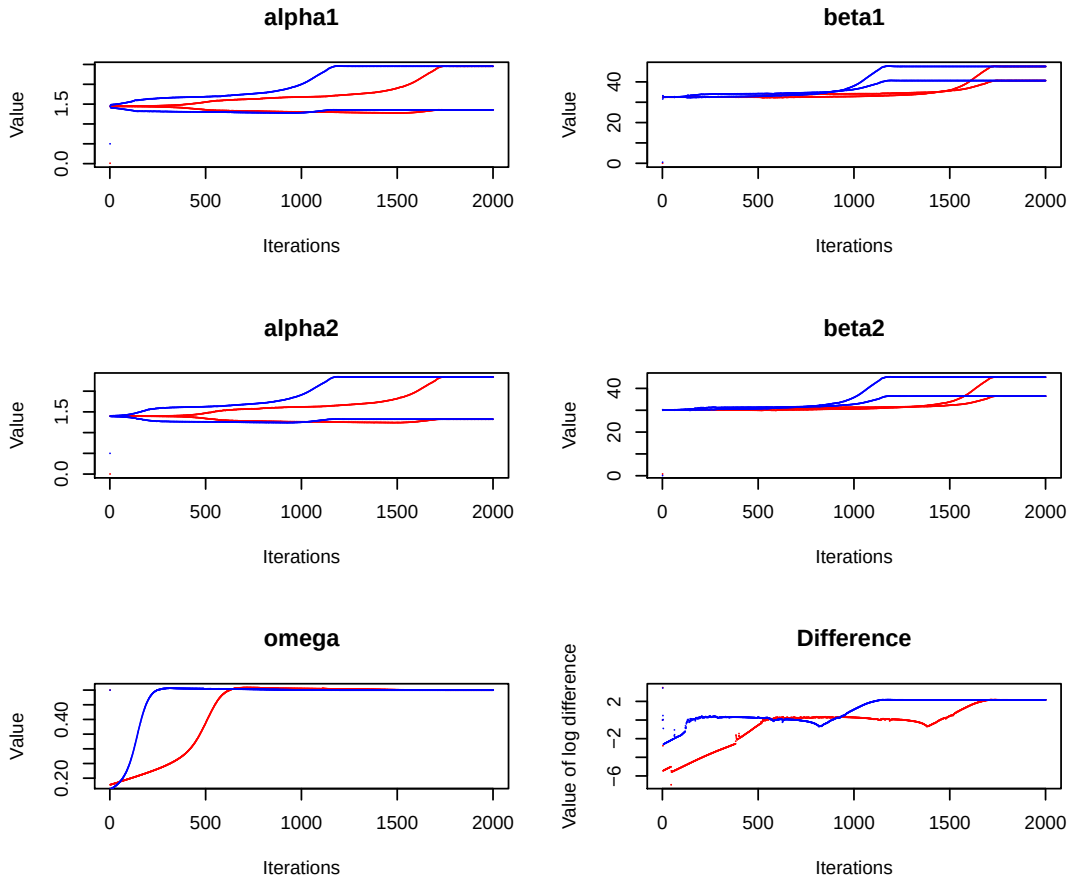

Fig. 3. Convergence of Developed Model. The first five plots show the convergence of parameters $\alpha_1, \beta_1, \alpha_2, \beta_2, \omega$ respectively. The sixth plot shows the maximum of the difference between two consecutive parameters.

Table 1. Results of Developed Model. All six sets of initial values converge to both sets of the results. Five significant digits are reserved for all the numbers in this table.

Set Number	α_1	β_1	α_0	β_0	ω
Set 1	1.3533	47.569	2.3412	36.454	0.50017
Set 2	2.4544	40.631	1.3233	45.161	0.49968

4 Conclusions

This article provides a basic idea to classify the CFR of COVID-19 all around the world. We provide an elementary step for researchers to analyze the different region-related factors for the difference of CFR such as the different level of medical resources, different policy, different climate among different regions. The Bayesian approach is used to classify the CFR level. A mixture model solving by EM algorithm with low CFR group to be normally distributed truncated at 0 on the left with mode 0 and variance 0.00182 and high CFR group to be normally distributed with mean 0.105 and same variance is obtained and analyzed to have a great power of prediction.

Although the mixture model with assumption of two normally distributed groups has a great power of estimation, the model with assumption of Beta and Gamma distributed groups shows poor predictive power especially for the regions with CFR values that are in the middle level. In the future, in order to improve the efficiency of developed model, we may try to add proper constrains to the two CFR groups to force almost all the proportion of the low CFR region to be in the low CFR group, and the high CFR group contribute a relatively low proportion. Furthermore, another approach to improve our model is to add more covariates to our model. With the consideration of different covariates such as level of medical resources, policy related to the public health. We are able to analyze the critical factors for the high level of CFR. In

such circumstance, we are able to provide suggestions in terms of reduce the CFR in certain region.

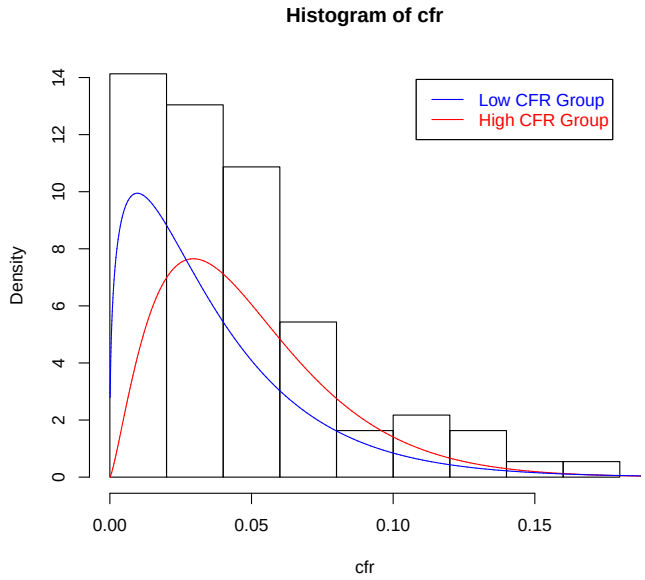

Fig. 4. Histogram of CFR with Fitted Densities of the Mixture Components. The plot shows the histogram of CFR for 92 target regions. The blue line is the distribution of low CFR group and the red line is the distribution of the high CFR group.

Table 2. Predictive Results of Both Basic Model and Developed Model. The first column shows the select six regions; the second and third columns are the number of death and confirmed case in each region; the "Basic Model" column shows the predictive results with application of basic model; and the "Developed Model" column shows the predictive results of the developed model.

Region	Death	Case	CFR	Basic Model	Developed Model
Australia	63	6497	0.009696783	0.006457261	0.2967943
Italy	22172	168941	0.131241084	0.874849171	0.5888049
Japan	148	9167	0.016144867	0.009324500	0.3983288
Russia	232	27938	0.008304102	0.005963851	0.2680625
United Kingdom	13729	103093	0.133171020	0.886489414	0.5855477
United States	33284	671331	0.049579120	0.060339819	0.5957721

Bibliography

[1] Coronavirus disease 2019 (covid-19) — symptoms and causes. Mayo Clinic (2020)
[2] in Data, O.W.: Case fatality rate of covid-19. Our World in Data (2020), https://ourworldindata.org/grapher/coronavirus-cfr?year=2020-04-17
[3] Fleming, P.J., Wallace, J.J.: How not to lie with statistics: the correct way to summarize benchmark results. Communications of the ACM **29**, 218–221 (1986)
[4] Hui, D.S., Azhar], E.I., Madani, T.A., Ntoumi, F., Kock, R., Dar, O., Ippolito, G., Mchugh, T.D., Memish, Z.A., Drosten, C., Zumla, A., Petersen, E.: The continuing 2019-ncov epidemic threat of novel coronaviruses to global health — the latest 2019 novel coronavirus outbreak in wuhan, china. International Journal of Infectious Diseases **91**, 264 – 266 (2020). https://doi.org/https://doi.org/10.1016/j.ijid.2020.01.009, http://www.sciencedirect.com/science/article/pii/S1201971220300114
[5] Lazzerini, Marzia, P.G.: Covid-19 in italy: momentous decisions and many uncertainties. The Lancet Global Health (2020). https://doi.org/10.1016/S2214-109X(20)30110-8, https://doi.org/10.1016/S2214-109X(20)30110-8
[6] WHO: Report of the who-china joint mission on coronavirus disease 2019 (covid-19) (2020), https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf