Bioinformatic Modelling of SARS-CoV-2 Pandemic with a Focus on Country-Specific Dynamics

Jakub Liu
Wroclaw University of Environmental and Life Sciences
Tomasz Suchocki
Wroclaw University of Environmental and Life Sciences
Joanna Szyda (joanna.szyda@upwr.edu.pl)
Wroclaw University of Environmental and Life Sciences

Research Article

Keywords: COVID-19, mixture model, outlier, SIRD

Posted Date: December 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-745759/v2

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Public Health on January 21st, 2023.
See the published version at https://doi.org/10.1186/s12889-023-15092-1.

Abstract

Background: One of the seminal events since 2019 has been the outbreak of the SARS-CoV-2 pandemic. Countries have adopted various policies to deal with it, but they also differ in their socio-geographical characteristics and the public health care facilities. Our study aimed to investigate differences between epidemiological parameters across countries.

Method: The analysed data represents SARS-CoV-2 repository provided by the Johns Hopkins University. Separately for each country, we estimated recovery and mortality rates using the SIRD model applied to the first 30, 60, 150, and 300 days of the pandemic. Moreover, a mixture of normal distributions was fitted to the number of confirmed cases and deaths during the first 300 days. The estimates of peaks' means and variances were used to identify countries with outlying parameters.

Results: For 300 days Belgium, Cyprus, France, the Netherlands, Serbia, and the UK were classified as outliers by all three outlier detection methods. Yemen was classified as an outlier for each of the four considered timeframes, due to high mortality rates. During the first 300 days of the pandemic, the majority of countries underwent three peaks in the number of confirmed cases, except Australia and Kazakhstan with two peaks.

Conclusions: Considering recovery and mortality rates we observed heterogeneity between countries. Liechtenstein was the "positive" outlier with low mortality rates and high recovery rates, at the opposite, Yemen represented a "negative" outlier with high mortality for all four considered periods and low recovery for 30 and 60 days.

Background

One of the most seminal global events since 2019 has been the outbreak of the SARS-CoV-2 (Severe Acute Respiratory Syndrome – Coronavirus - 2) infections in humans, which occurred in December in China [1] and has been following worldwide spread in 2020 (see e.g. WHO Coronavirus disease situation reports at www.who.int/emergencies/diseases/novel-coronavirus-2019/situationreports). It is commonly believed that the current pandemic has had its roots in the seafood market located in Wuhan [2]. Since late 2019 the virus has spread globally and evolved into several strains throughout 2020 and 2021. The two most common means of infection with the SARS-CoV-2 virus are the so-called respiratory droplets and touching an infected surface. When an infected person sneezes or coughs the viral particles become airborne, can travel up to several meters and infect an oral or nasal cavity of a nearby person [3]. SARS-CoV-2 can survive up to several hours on most surfaces and even several days on certain surfaces [4]. When a person touches an infected surface, it is easy to transfer viruses onto one's mucus membranes such as eyes, nose, or mouth. The main location of viral replication is lung epithelial cells where a pathogen uses its S protein to bind with the ACE2 receptor of a host cell [5]. After binding, a virus is ready to infect a cell either via direct entry, where the membrane of the host cell and the viral envelope fuse together, or by endocytosis, where an entire virus is being engulfed by a part of the cellular membrane [6].

Inside a host cell, a virus uses cellular machinery (mainly the endoplasmic reticulum and the Golgi Apparatus) to assemble new virions. The release of newly produced viruses places a lot of stress on human cells and consequently leads to apoptosis [7] and leads to a major immune response, often referred to as a cytokine storm. This strong inflammation leads to an excess build-up of mucus, mostly in the alveoli. Such a condition is responsible for shortness of breath, which is one of the most common symptoms of the SARS-CoV-2 infection.

Various countries have adopted different policies to deal with the dynamics of this pandemic and, atop of those policies, countries also differ in terms of their socio-geographical characteristics (such as climate, total population, and population density, etc.) as well as in the public health care facilities. Our study aimed to fit SIRD models to early (30 days), mid- (60 and 150 days), and long-term (300 days) time-frames of SARS-CoV-2 pandemic, separately to each country, to estimate country-specific pandemic parameters. However, the SIRD model only allows for fitting one peak of infection during the pandemic, which in the case of SARS-CoV-2 results in averaging over multiple peaks. Therefore, in addition to SIRD, linear mixture models were fitted to the same set of data to estimate fluctuations in the dynamics of the pandemic. The final goal of the study was the assessment of (dis)similarities between countries applying outlier detection methodology.

Methods

Study population

The analysed data comprised cumulative daily numbers of: confirmed cases, deaths, and recoveries, reported for 191 countries, beginning from the 21st of January 2020. This data set represents SARS-CoV-2 Data Repository provided and curated by the Center for Systems Science and Engineering at Johns Hopkins University [8] and was obtained via Github (github.com/CSSEGISandData/COVID-19) using custom-written Python scripts. Separately for each country, the pandemic's dynamics was modelled for 30 days (D30), 60 days (D60), 150 days (D150), and 300 days (D300), beginning from the first day in which the number of confirmed cases in a given country exceeded zero, what resulted in various calendar dates – depending on a country.

SIRD model

The SIRD model and the least-squares estimators for its parameters were implemented using custom-written Python scripts following Anastassopoulou et al. [9]. In brief, the model is defined by the following equations:

$$S(t) = S(t-1) - \frac{\alpha}{N} S(t-1) I(t-1)$$

$$I(t) = I(t-1) + \frac{\alpha}{N} S(t-1) I(t-1) - \beta I(t-1) - \gamma I(t-1)$$

$$R(t) = R(t-1) + \beta I(t-1)$$

$$D(t) = D(t-1) + \gamma I(t-1)$$

where, $S(t)$ represents the number of susceptible individuals expressed by the total country population (N) diminished by the number of infected individuals at time t $I(t)$, the numbered of recovered individuals at time t $R(t)$, and the number of dead individuals at time t $D(t)$, obtained from the Github repository [8]. The estimated model parameters comprise infection rate (α), recovery rate (β) and mortality rate (γ). In our study we estimated country-specific β and γ, assuming a time unit represented by a calendar day, following the estimators defined by Anastassopoulou et al. [9], given by $\widehat{\beta} = \left[C\Delta I(t)^T C\Delta I(t)\right]^{-1}\left[C\Delta I(t)^T C\Delta R(t)\right]$ and $\widehat{\gamma} = \left[C\Delta I(t)^T C\Delta I(t)\right]^{-1}\left[C\Delta I(t)^T C\Delta D(t)\right]$, separately for the time-frame of 30-days, 60-days and 150-days, and 300-days, counting from the occurrence of the first infected individual in a given country. $C\Delta I$, $C\Delta R$, and $C\Delta D$ represent the cumulative numbers of infected, recovered and dead individuals respectively within the considered time-frame.

Linear mixture model

Linear mixture models were fitted to the daily numbers of confirmed cases as the daily numbers of deaths during the first 300 days of a pandemic using the *GaussianMixture* module of the Scikit-learn [10] Python library, separately for each country and time period. First, models with three normal components were fitted to each country's data reflecting the underlying assumptions of three peaks of the pandemic in this time period. Further on, two-sample t-tests were calculated for the two consecutive means and standard deviations estimated for each country to determine whether this assumption was appropriate i.e. the difference between the estimated means exists. The difference between the means was determined as nonsignificant based on a P-value threshold above 0.05 after a multiple testing correction based on the Bonferroni approach, meaning that two, instead of three, components were sufficient for a given country during the first 300 pandemic days.

Outlier detection

Detection of outlying countries was based on $\widehat{\beta}$ and $\widehat{\gamma}$ from the SIRD models as well as on the estimated means of normal distributions from the linear mixture models. For this purpose, the country-specific data points given by the abovementioned estimates of model parameters (X) were divided into countries located within the estimated decision boundary (non-outliers) and outside the boundary (outliers). Three following approaches implemented via the Scikit-learn were used for classification. (1) Support Vector Machine classifier, implemented through the *svm.OneClassSVM* function, fitting the sigmoid kernel function $f(X) = tahn\left(\gamma\left(X, X^T\right)\right)$ and allowing for maximally 20% of countries to be located outside of the estimated decision boundary and thus representing outliers. (2) A classifier based on the Local Outlier Factor, implemented through *neighbors.LocalOutlierFactor* function, classifying countries using their Euclidean distance from the decision boundary, which was estimated based on 20 neighbouring countries. Countries located outside a given distance from this boundary were classified as

outlying observations. (3) Density-Based Spatial Clustering approach, implemented through *cluster.DBSCAN* function. In our application, the maximum distance between two samples to form a cluster was set to 0.9, which resulted in forming a single cluster of countries and a set of outlying countries.

Results

Epidemiological parameters

Based on recovery and mortality rates estimated from the SIRD model no marked overlap between outliers detected by the three methods as well as between periods (D30, D60, D150, and D300) was observed (Figure 1 and Table 1). For the longest modelling period of 300 days Belgium, Cyprus, France, the Netherlands, Serbia, and the UK were classified as outliers by all three methods. Among them, the UK and the Netherlands, also based on the first 150 days of the pandemic. Still, a pattern emerges in which Yemen was classified as an outlier for each of the considered timeframes, which was due to mortality rates estimated higher than in the bulk of countries amounting to 0.178, 0.223, 0.282, and 0.288 respectively for D30, D60, D150, and D300. During the beginning of the pandemic, those high mortality rates were accompanied by low recovery rates of 0.034 for D30 and 0.041 for D60, which were however much higher when a longer time span was considered. Comparably low recovery rates, albeit with varying mortality rates, starting from the 60th day of the pandemic were estimated for Netherlands ($\widehat{\beta}_{60} = 0.002$, $\widehat{\beta}_{150} = 0.004$, $\widehat{\beta}_{300} = 0.014$) and the UK ($\widehat{\beta}_{60} = 0.006$, $\widehat{\beta}_{150} = 0.005$, $\widehat{\beta}_{300} = 0.003$), which caused the classification of both countries as outliers.

Table 1

mortality rate (γ) 0-0.30/ recovery rate (β) and 0-1.0

Country	D30 [γ/β]	D60 [γ/β]	D150 [γ/β]	D300 [γ/β]
Algeria	0.138/0.343			
Austria	0.005/0.006			
Belgium			0.158/0.262	0.037/0.045
Belize	0.123/0.039		0.011/0.113	
Bolivia	0.064/0.008			
Botswana			0.002/0.101	
Brunei	0.007/0.673			
Burundi	0.113/0.352			
Chad				0.063/0.853
Chile	0.001/0.012			
Cyprus				0.006/0.156
Ecuador				0.073/0.825
Equatorial Guinea		0.008/0.030		
Finland		0.005/0.012		
France			0.147/0.332	0.035/0.107
Guinea-Bissau		0.005/0.036		
Guyana	0.153/0.154			
Haiti	0.060/0.006			
Honduras	0.065/0.018		0.030/0.122	
Hungary			0.133/0.571	0.025/0.311
Indonesia	0.088/0.051			
Ireland	0.013/0.003			
Italy			0.140/0.556	0.064/0.472
Liberia	0.112/0.076			
Libya			0.020/0.128	

Country	D30 [γ/β]	D60 [γ/β]	D150 [γ/β]	D300 [γ/β]
Liechtenstein	0.010/0.841	0.010/0.841	0.012/0.949	
Lithuania	0.010/0.000			
Maldives		0.003/0.042		
Mauritania	0.169/0.355			
Mexico			0.116/0.762	0.099/0.788
Namibia			0.005/0.133	
Netherlands		0.111/0.002	0.122/0.004	0.021/0.014
Nicaragua	0.302/0.400	0.302/0.400		
Norway	0.004/0.001	0.019/0.004		
San Marino	0.097/0.024			
Serbia				0.010/0.004
South Sudan		0.010/0.007		
Spain			0.113/0.565	0.039/0.172
Sudan	0.175/0.159			0.063/0.530
the United Kingdom		0.052/0.006	0.145/0.005	0.060/0.003
the US		0.018/0.006		
Yemen	0.178/0.034	0.223/0.041	0.282/0.497	0.288/0.615
Zimbabwe	0.159/0.041	0.119/0.207		

Considering the longest time span (D300), for each country, mortality rates were lower than recovery rates. Except for Yemen, which was classified as an outlying country, all the remaining countries revealed mortality rates below 10%. On the other hand, estimated recovery rates varied greatly between countries from as low as less than 1% in Serbia and the UK to over 95% in Bahrain, Djibouti, Ghana, Qatar, and Uzbekistan (Figure 2).

Dynamics of the numbers of confirmed cases across time

During the first 300 days of the pandemic, the vast majority of countries underwent three major peaks in the number of confirmed cases (Figure 3). The two exceptions were Australia with two peaks at day 63 and then the second between 188th and 192nd day of the pandemic and Kazakhstan for which the 1st peak was estimated at day 131 and the 2nd peak at day 263. For the remaining countries, the average timeframe between the two first peaks was 92±50 days, with the shortest interval of only 6 days

estimated for Suriname and the longest for Liechtenstein – 230 days, followed by the average time between the second and the third peak of 91±35 days varying between 4 days for the Diamond Princess and 207 days for SanMarino. Those are four locations with (very) small populations. The earliest peaks were estimated for the Diamond Princess at day 3, 11, and 15 of the pandemic, with small corresponding standard deviations of only 1.7, 2.1. and 7.4 days. Burma “had” its first peak as late as in day 205 while Seychelles had the latest 2nd and 3rd peaks at day 263 and 295 respectively.

The Local Outlier Factor approach identified five countries and the Diamond Princess with outlying patterns of peaks in the number of confirmed cases, which however differ between each other and underlie no clear geographic pattern (Figure 4). For Burma, relatively late peaks were estimated at days 205, 242, and 271. The opposite was observed for the Diamond Princess (3rd, 11th, and 15th day) and Tanzania (32nd, 45th, and 54th day). Quite a similar pattern with all three peaks distributed closely around the middle of the analysed 300-day period was estimated for the Central African Republic (79th, 104th, and 176th day), Singapore (98th, 130th, and 184th day), and Gambia (151st, 162nd, and 194th day).

Dynamics of the numbers of deaths across time

Considering the 99 countries which experienced at least 1000 deaths assigned to the SARS-CoV-2 infection during the first 300 days of the pandemic, we observed that the earliest peaks, expressed by the estimated means of the mixture of three normal distributions were attributable to China (20th, 33rd, and 87th day of pandemic). China is also the only country classified as an outlier, based on the Local Outlier Factor approach applied to the three estimated means of the fitted normal distributions (Figure 5). The latest first peak of the number of deaths appeared in Argentina, at day 159, the latest second peak at day 255 occurred in Germany, while the third peak was the latest in Zimbabwe, by day 294. The latest peak in the number of daily deaths showed the least variation in time and each country (excluding China as the outlier) appeared after the 200th day of the pandemic (Figure 5).

Discussion

Globally, a much greater variation was observed in recovery rates than in mortality rates. Significant sources of heterogeneity in SARS-CoV-2 associated mortality rates across countries are differences in governmental policies related to social aspects of life during the pandemic [11], country globalization level [12], and genetic differences between country- and continent-specific populations [13]. Some of our findings reflect the abovementioned sources of inter-country variation. Australia which was outlying in the numbers of confirmed cases over time by showing two, rather than three peaks over the first 300 days of the pandemic, and China outlying by having all three peaks of confirmed deaths estimated in a short time-span one to another were previously mentioned by Pearce et al. [14] as countries that imposed early restrictions to social contacts. Moreover, high mortality rates estimated by us using the SIRD model aligned with highly globalised countries (Belgium, France, United Kingdom, Italy, Hungary, and the Netherlands for D150). However, neither the patterns of outlying countries nor the sole estimated

mortality rates did align with their geographical locations as it may have been expected providing genetic differences between populations, a phenomenon also observed by Balmford et al. [11].

Conclusions

Considering both major outcomes of the SARS-CoV-2 pandemic, i.e. the recovery and death, the heterogeneity between countries exists. The goal is to be classified as an outlying country based on exceptionally low mortality rates and high recovery rates. Liechtenstein was the closest to this favourable pattern being a “positive” outlier for D30, D60, and D150. On the other end, Yemen represents a “negative” outlier with high mortality rates for all four considered periods and low recovery rates for D30 and D60. We believe that the observed country-specific differences result from a mixture of factors including the biology of the virus strains, different policies adopted by countries to mitigate the spread of infections, but also different accuracy of data reporting.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The input data underlying this article is publicly available from the SARS-CoV-2 Data Repository at https://github.com/CSSEGISandData/COVID-19.

Competing interests

None declared.

Funding

No external funding was involved.

Authors' contributions

JL implemented and performed all the analyses, contributed to the manuscript text. TS wrote Python scripts for the estimation of model parameters, adapted the statistical methodology and contributed to the manuscript text. JS conceived the study and significantly contributed to the manuscript text.

Acknowledgments

Not applicable.

References

1. Anastassopoulou, C., Russo, L., Tsakris, A., Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS One.* **15**, e0230405 (2020). doi:10.1371/journal.pone.0230405
2. Dong, E., Du, H., Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* **20**, 533–534 (2020). doi:10.1016/S1473-3099(20)30120-1
3. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol.* **19**, 155–170 (2021). doi:10.1038/s41579-020-00468-6
4. Shang, J., Wan, Y., Luo, C. et al. Cell entry mechanisms of SARS-CoV-2. *Proc Natl Acad Sci U S A.* **117**, 11727–11734 (2020). doi:10.1073/pnas.2003138117
5. Perrotta, F., Matera, M. G., Cazzola, M., Bianco, A. Severe respiratory SARS-CoV2 infection: Does ACE2 receptor matter? *Respir Med*. **168**, 105996 (2020). doi:10.1016/j.rmed.2020.105996
6. Aboubakr, H.A., Sharafeldin, T.A., Goyal, S.M. Stability of SARS-CoV-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: A review. *Transbound Emerg Dis.* **68**, 296–312 (2021). doi:10.1111/tbed.13707
7. Jayaweera, M., Perera, H., Gunawardana, B., Manatunge, J. Transmission of COVID-19 virus by droplets and aerosols: A critical review on the unresolved dichotomy. *Environ Res.* **188**, 109819 (2020). doi:10.1016/j.envres.2020.109819
8. Zhu, N., Zhang, D., Wang, W., et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med.* **382**, 727–733 (2020). doi:10.1056/nejmoa2001017
9. Li, Q., Guan, X., Wu, P., et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N Engl J Med*. **382**, 1199–1207 (2020). doi:10.1056/nejmoa2001316
10. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* **12**, 2829–2830 (2011).
11. Balmford, B., Annan, J. D., Hargreaves, J. C., Altoè, M., Bateman, I. J. Cross-Country Comparisons of Covid-19: Policy, Politics and the Price of Life. *Environ Resour Econ.* **76**, 525–551 (2020). doi:10.1007/s10640-020-00466-5
12. Farzanegan, M. R., Feizi, M., Gholipour, H. F. Globalization and the Outbreak of COVID-19: An Empirical Analysis. *J Risk Financ Manag.* **14**, 105 (2021). doi:10.3390/jrfm14030105
13. Asselta, R., Paraboschi, E. M., Mantovani, A., Duga, S. ACE2 and TMPRSS2 Variants and Expression as Candidates to Sex and Country Differences in COVID-19 Severity in Italy. *SSRN Electron J*. (2020). doi:10.2139/ssrn.3559608
14. Pearce, N., Lawlor, D. A., Brickley, E. B. Comparisons between countries are essential for the control of COVID-19. *Int J Epidemiol.* **49**, 1059–1062 (2020). doi:10.1093/ije/dyaa108

Figures

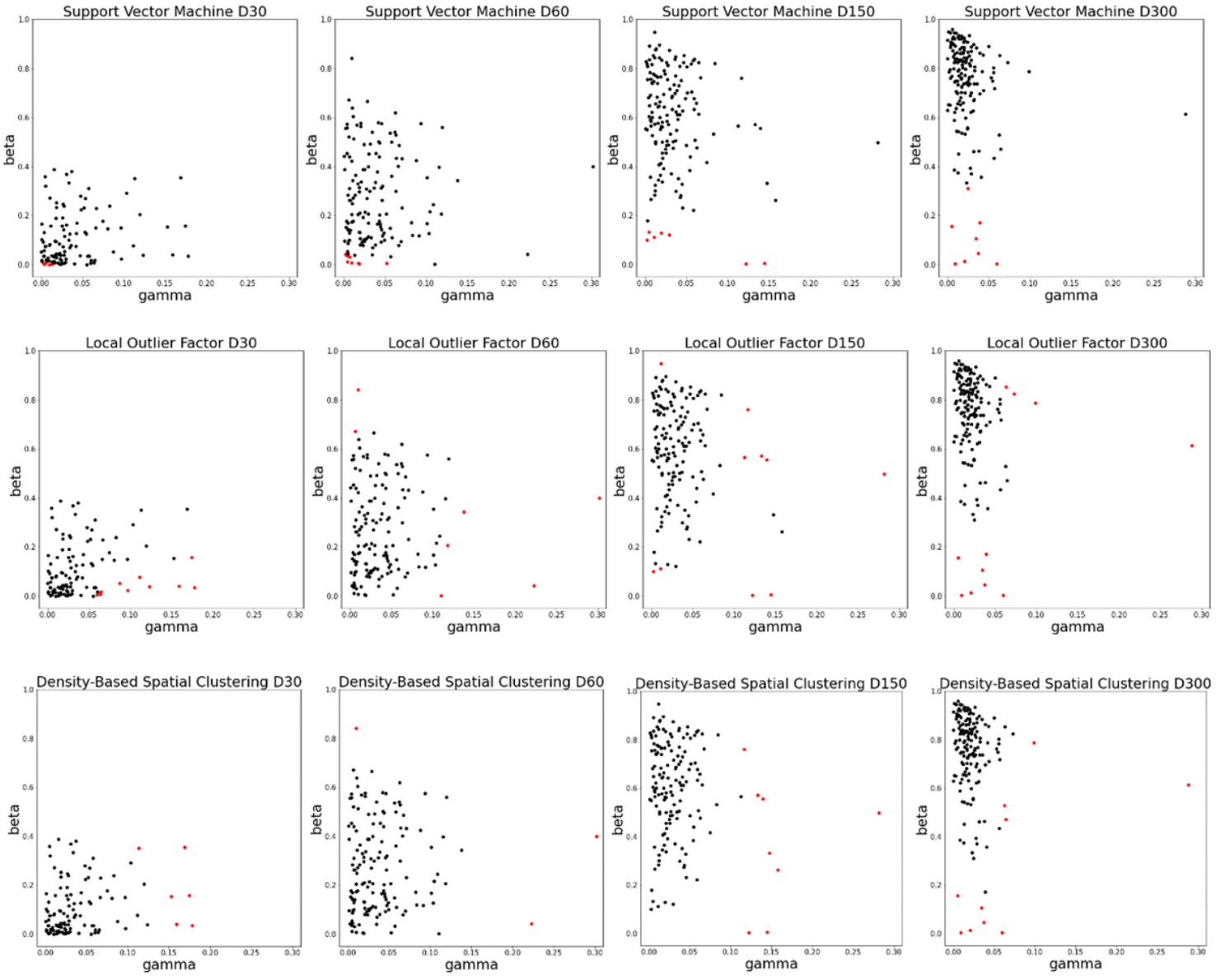

Figure 1 Clustering based on country-specific $\hat{\beta}$ (recovery rate) and $\hat{\gamma}$ (mortality rate) from the SIRD model fitted to the first 30 (D30), 60 (D60), 150 (D150), and 300 (D300) days of the pandemic. Countries classified as outliers are marked in red.

Figure 1

"See image above for figure legend"

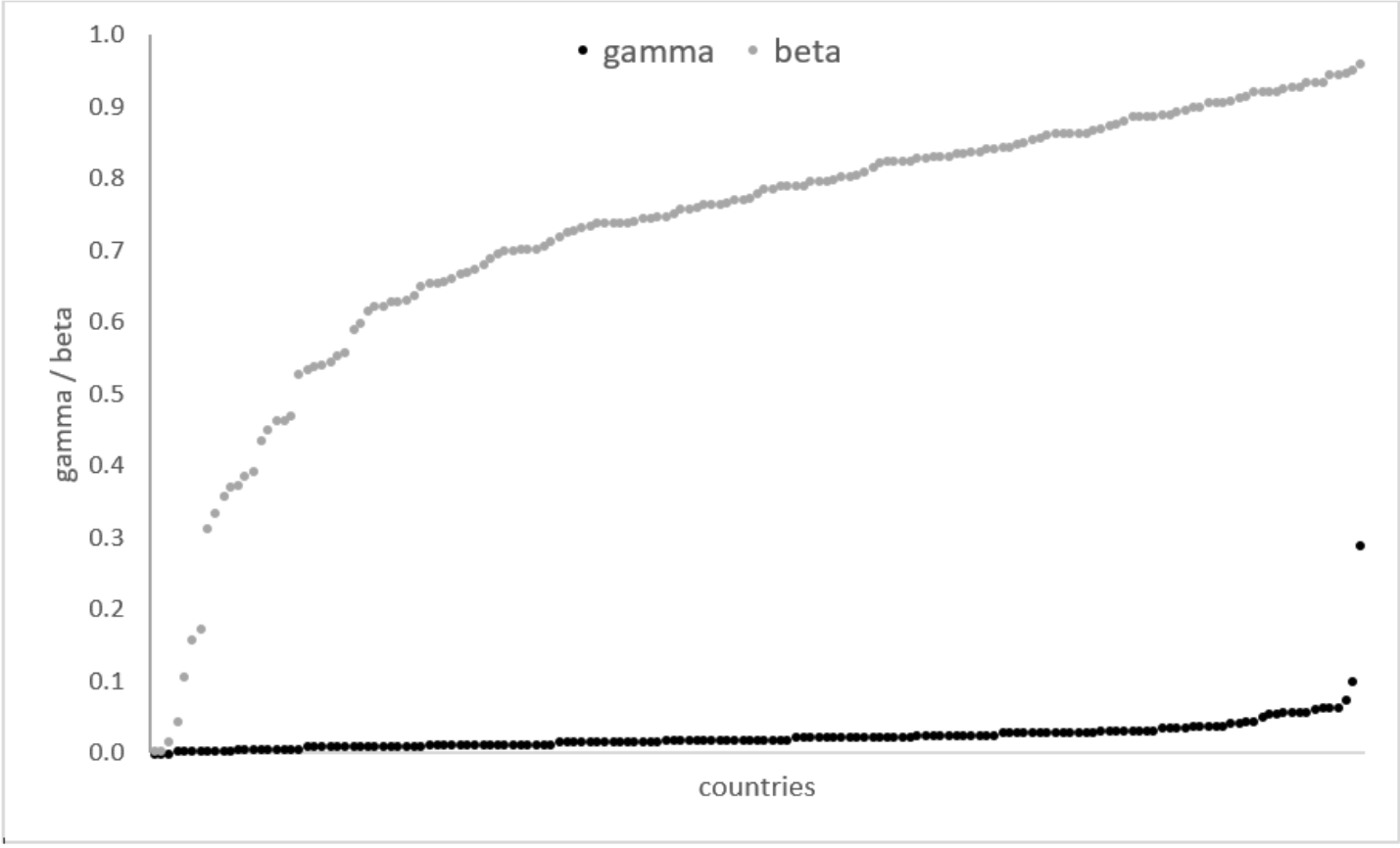

Figure 2 Country-specific $\hat{\beta}$ (recovery rate) and $\hat{\gamma}$ (mortality rate) from the SIRD model fitted to the first 300 (D300) days of the pandemic.

Figure 2

"See image above for figure legend"

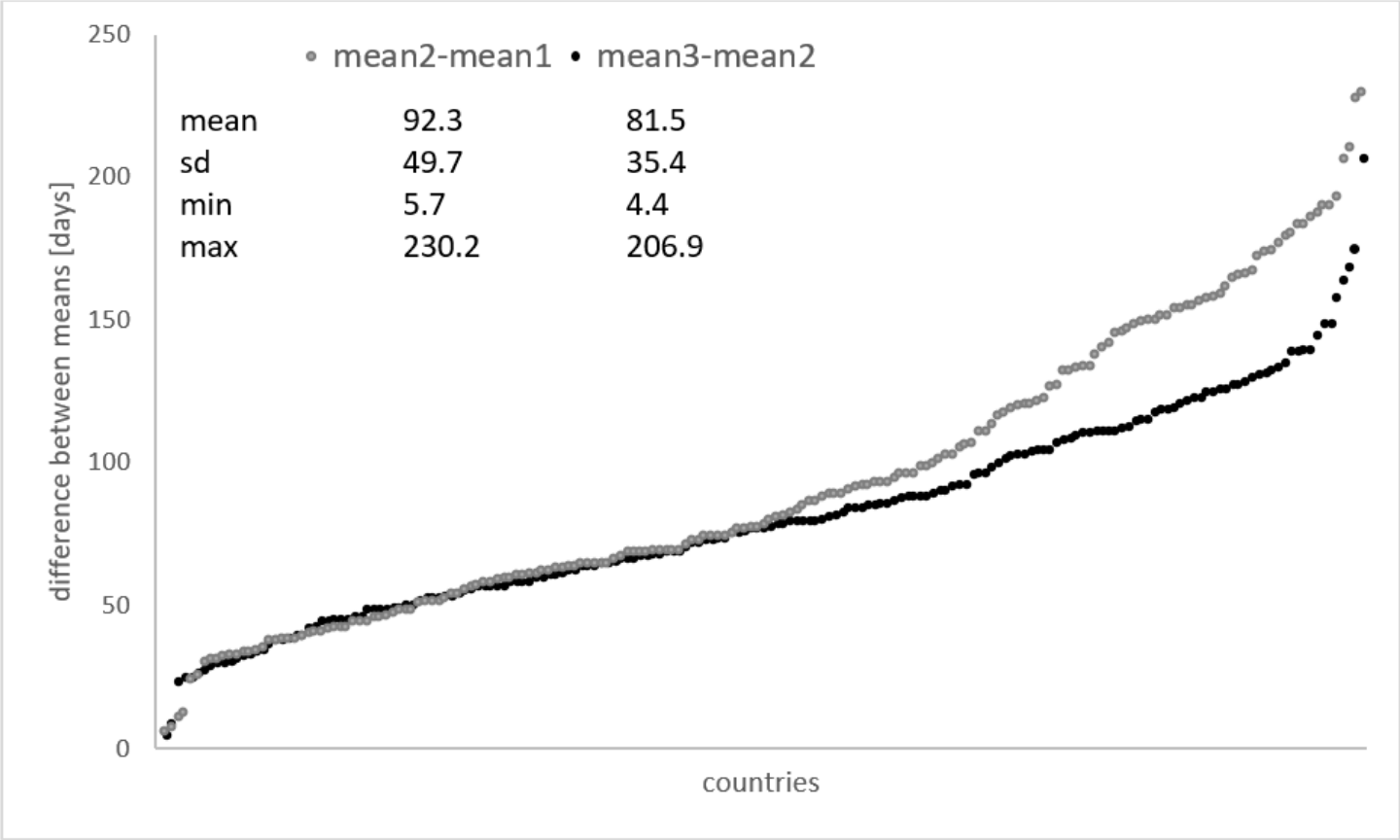

Figure 3

Differences in days between two consecutive peaks of the number of confirmed cases estimated by the linear mixture model for the first 300 days of the pandemic.

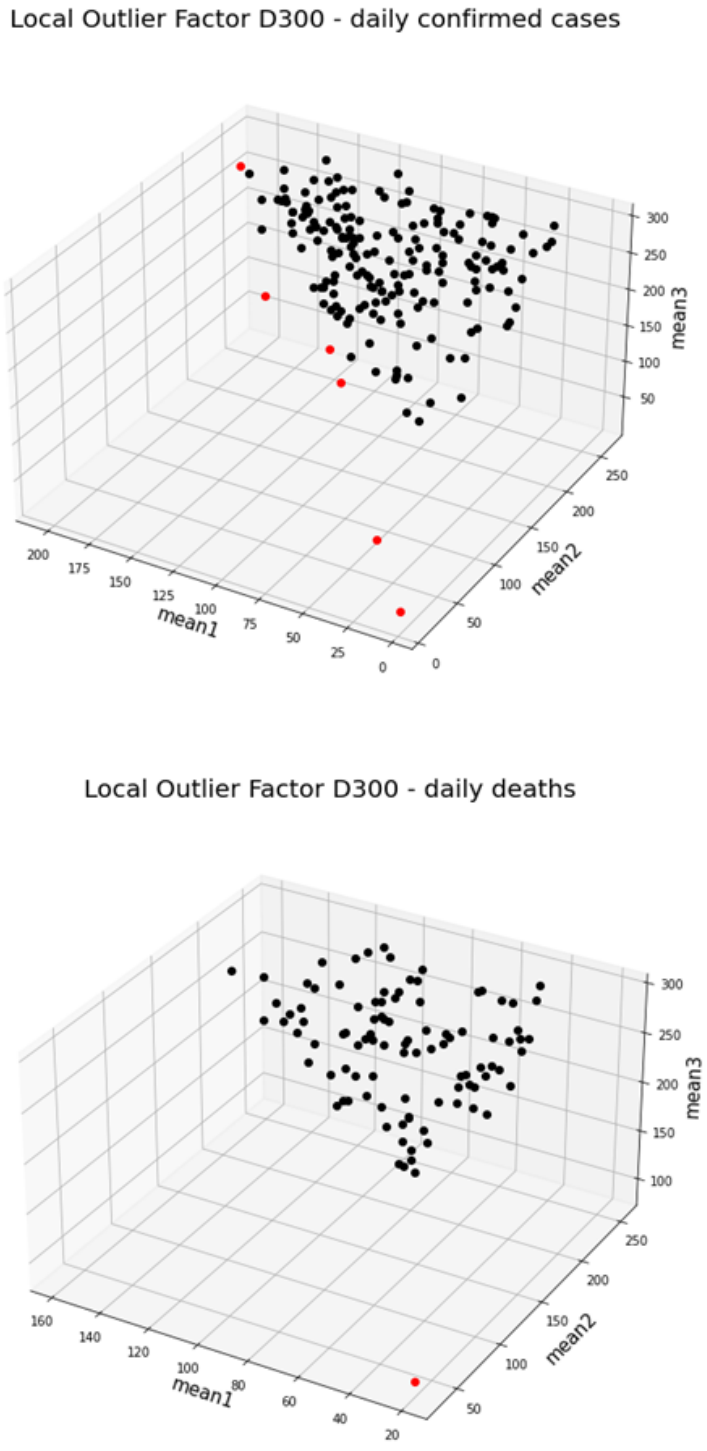

Figure 4

Clustering based on country-specific means of three normal distributions fitted by the linear mixture model to the first 300 (D300) days of the pandemic. Countries classified as outliers are marked in red.

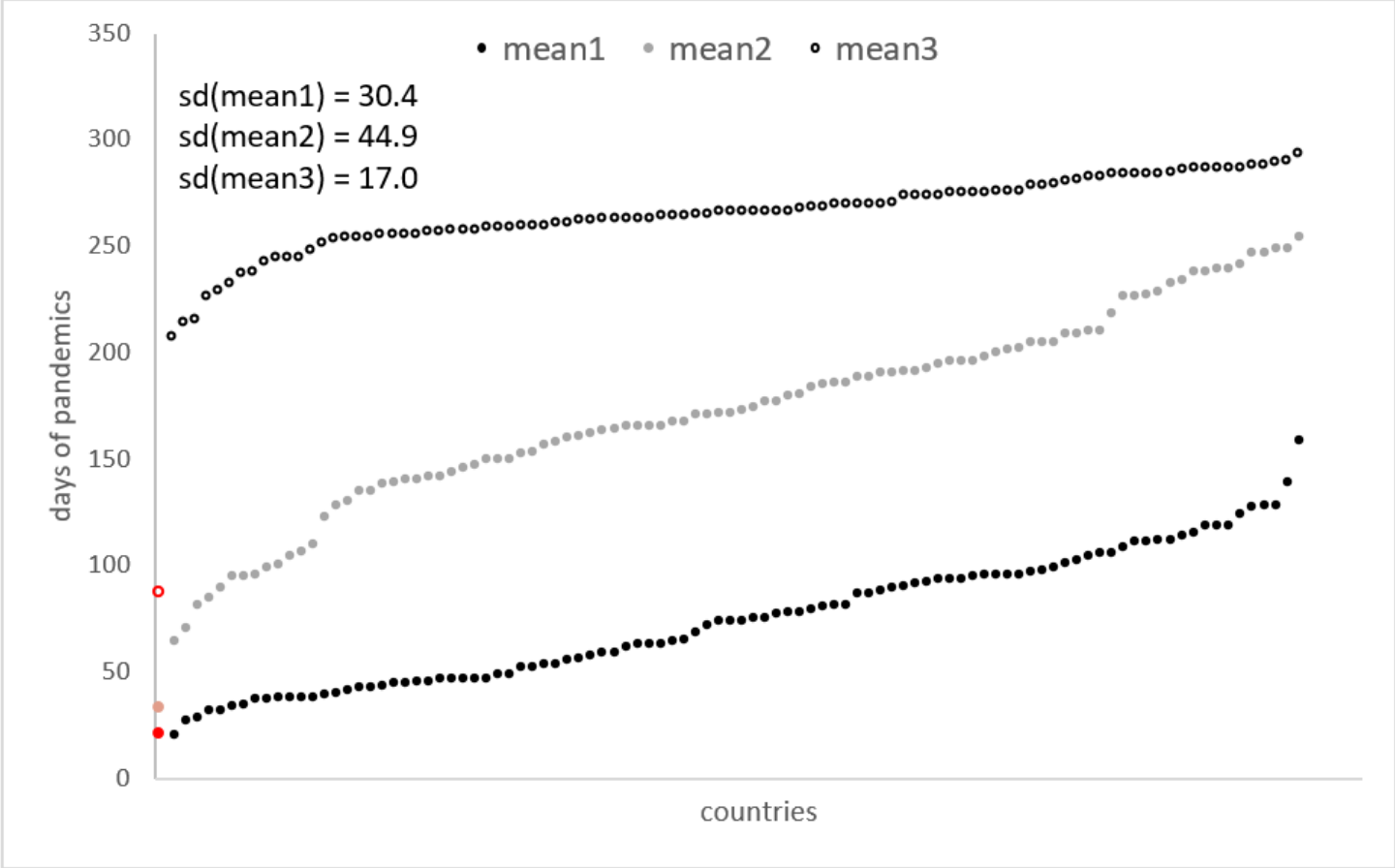

Figure 5

Estimated means of peaks of the number of daily deaths estimated by the linear mixture model for the first 300 days of the pandemic. Means corresponding to China, which was classified as an outlier, are marked in red. Standard deviations were calculated excluding China.