Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million

Roy Cerqueti[1,2] and Valerio Ficcadenti*[2]

[1]Sapienza University of Rome, Department of Social and Economic Sciences,
Piazzale Aldo Moro, 5, 00185, Rome, Italy
[2]London South Bank University, Business School, Borough Road, 103, SE1 0AA,
London, United Kingdom

March 31, 2022

Abstract

This paper deals with the cluster analysis of selected countries based on COVID-19 new deaths per million data. We implement a statistical procedure that combines a rank-size exploration and a k-means approach for clustering. Specifically, we first carry out a best-fit exercise on a suitable polynomial rank-size law at an individual country level; then, we cluster the considered countries by adopting a k-means clustering procedure based on the calibrated best-fit parameters. The investigated countries are selected considering those with a high value for the Healthcare Access and Quality Index to make a consistent analysis and reduce biases from the data collection phase. Interesting results emerge from the meaningful interpretation of the parameters of the best-fit curves; in particular, we show some relevant properties of the considered countries when dealing with the days with the highest number of new daily deaths per million and waves. Moreover, the exploration of the obtained clusters allows explaining some common countries' features.

Key words— COVID-19; k-means clustering; rank-size analysis

1 Introduction

The spatio-temporal patterns of COVID-19 represent one of the most relevant themes for statistical research nowadays, given the crucial relevance of the pandemic disease in contexts of society such as economics and, of course, health.

This pandemic has heterogeneous implications on countries and regional realities. The most common example we can mention is given by the different applications of the so-called non-pharmaceutical interventions in the preliminary phases (see, e.g. Flaxman

*Corresponding author: e-mail:ficcadv2@lsbu.ac.uk

et al. 2020, Tian et al. 2021). These differences must be included in the premise of an effective exploration of COVID-19 repercussions.

Some authors deal with forecasting exercises of the future evolution of deaths and infections dynamics (see, e.g., Bertozzi et al. 2020, Moein et al. 2021, Nabi 2020, Tang et al. 2021, Prasanth et al. 2021). In this respect, Ioannidis et al. (2020)'s authors state that the reliability of the predictions related to COVID-19 is debatable for several reasons, including the relevant sensitivity of the estimates on the employed methodology.

This paper takes the opposite perspective – hence, overcoming the criticism raised by Ioannidis et al. (2020) – presenting a lookback of the spatio-temporal data related to COVID-19. It is worth mentioning some relevant contributions on the matter. Bartolucci & Farcomeni (2021) propose the study of the cases of COVID-19 infections in the Italian regions by employing a model based on latent variables and estimating it through a Markov chain Monte Carlo (MCMC) algorithm. Still, in the context of MCMC, Lee et al. (2021) discuss the propagation of COVID-19 in Scotland by adopting a Bayesian-type framework. In Schneble et al. (2021), the registered death counts related to COVID-19 are modelled to monitor the dynamic behaviour of the infections on a small-area level in Germany.

Differently from the studies above, we consider a selection of countries and deal with the exploration of their daily data about COVID-19 new deaths per million. We combine a rank-size best-fit exercise – being the size, the considered variable – and a cluster analysis of k-means type. So, it is shown that a rank-size law of third-degree polynomial type provides high-quality goodness-of-fit parameters. The calibrated parameters feed the k-means cluster analysis based on a Euclidean distance, with $k = 3$ (the reasons for this choice are presented in Section 2). In this paper, we do not intend to propose a new method for clustering COVID-19 data in terms of countries as Zubair et al. (2020) have done; instead, we want to consider a statistical clustering technique well known in the literature and widely used in the context of even operations research, and apply it to a novel, relevant problem, with highly informative results. To assure data reliability and to reduce possible sources of biases in the data collection, countries are selected by taking those with a high value of the Healthcare Access and Quality Index (HAQ hereafter, see Barber et al. 2017). Moreover, to avoid distortions in the best-fit procedure, we have removed the outliers at a country level during the data pre-treatment phase. The results interpretation is grounded on the meaningfulness of the calibrated parameters in terms of the polynomial curve shape; the analysis of the obtained clusters allows highlighting regularities and deviations of the considered countries.

Other contributions present a cluster analysis of the data related to COVID-19 and are summarised in Table 1. For instance, James & Menzies (2020), and Rios et al. (2021) respectively employ k-means and hierarchical to analyze public policies along the time (former) and to make forecasting of pandemic waves (latter). In these studies, the time is considered because of the purposes of the researches. Similarly, Li et al. (2021) run health parameters-based classification of the patients in Wuhan covering the beginning of the pandemic spread. Hutagalung et al. (2021) perform a cluster analysis via k-means taking $k = 3$ to group South-East Asian countries. In this case, the time is not considered to

get a more global view of the country's conditions during the pandemic. Similar works can by Abdullah et al. (2021) cover a broader set of countries or in Kumar (2020)'s research, where Indian territories are classified in terms of similar infection propagation, to pursue optimal monitoring strategies. On a different note, certain studies include additional features for investigating the pandemic. For example, Siddiqui et al. (2020) consider the relationship between temperature in Chinese areas and the spreading of the disease to cluster China's territories. A similar exercise is done by Vadyala et al. (2021) where humidity is also considered but for exploring Louisiana's pandemic related data. Kiaghadi et al. (2020) consider a larger number of variables. The authors included in the clustering elements like "Access to Medical Services and Sociodemographic, Behavioral, and Lifestyle Factors" to determine the most vulnerable areas to COVID-19 in Harris County, Texas. Rizvi et al. (2021) clustered 79 countries using socio-economic factors, disease prevalence and health system indicators considering COVID-19 confirmed cases and COVID-19 death cases. Zubair et al. (2020) propose a methodological work where a k-means variation is introduced for the case of COVID-19 data.
Our work aligns with Tuli et al. (2020) for the early steps the authors take, even if they do not explore clusters but focus on forecasting. Tuli et al. (2020) find estimation of new cases distributions. Among others, Weibull and Gaussian distributions on COVID-19 are used. With the modelled distributions, the authors can perform forecasts. We are close to the work by Machado & Lopes (2020) as well. The authors model the infected cases in more than 70 countries and visualize the results via a clustering approach. Machado & Lopes (2020) use daily log changes to create empirical distributions, and then they test multiple families of distributions to model the data. Then, with estimated distributions' parameters, the clusters are found.

Paper	Clustering Method	COVID-19 Data
James & Menzies (2020)	k-means	COVID-19 cases and deaths in multiple countries.
Rios et al. (2021)	hierarchical	COVID-19 cases and deaths in multiple countries.
Zubair et al. (2020)	k-means	COVID-19 cases, deaths and recovery in multiple countries.
Siddiqui et al. (2020)	k-means	COVID-19 confirmed, suspected and death cases in China.
Hutagalung et al. (2021)	k-means	COVID-19 cases and deaths in multiple countries (South-East Asia).
Vadyala et al. (2021)	k-means	COVID-19 cases in Louisana state, USA.
Zhang & Lin (2021)	k-means	COVID-19 new cases in USA.
Abdullah et al. (2021)	k-means	COVID-19 confirmed, death, and recovered cases in Indonesia.
Kiaghadi et al. (2020)	k-means	COVID-19 confirmed cases in Texas, USA.
Machado & Lopes (2020)	hierarchical	COVID-19 cases in in multiple countries.
Kumar (2020)	hierarchical	COVID-19 cases, deaths and recovery in India.
Li et al. (2021)	k-means	COVID-19 cases in China.
Rizvi et al. (2021)	k-means	COVID-19 cases and deaths in multiple countries.

Table 1: Sample of recent studies related to our work for the methods employed and for the data used.

Moreover, some papers treat COVID-19 by adopting a rank-size analysis approach. For example, Kennedy & Yam (2020) use Zipf's law to detect COVID-19 data inconsistencies, while Jiang & de Rijke (2021) employ power-law relationships to explore USA populations, deaths and infections. Vasconcelos et al. (2021) "analyze the rank-frequency distribution of preprints servers, ordered by the number of COVID-19 preprints they host" and Small & Sousa (2021) apply rank-size distributions to model the spatio-temporal evolution of COVID-19 in USA and China, but not directly with data regarding deaths or new cases.

More in general, the usage of rank-size laws is typically driven by robust compliance of the data to theoretical models – see, e.g., the work of Ficcadenti et al. 2019 for text analysis or Ficcadenti & Cerqueti 2017 for earthquakes cost evaluations based on rank-size law – namely, when the best fit is appropriate, the goodness of fit must result excellent. It is the case in this paper, as presented in the section devoted to the results. In studying other researches attempting to model similar information related to COVID-19, one can notice, for example, Table 1 by Tuli et al. (2020) where the R^2s are lower than those usually obtained with rank-size best fits. In addition, Machado & Lopes (2020) write in section "Regression models for describing the spread of COVID-19", that "a single model with a limited number of parameters is not able to fit well the time series [...] for all countries". So, even if they have found some goodness of fit comparable to those expected for rank-size compliance (e.g., they report an $R^2 = 0.99$ for Italian and Chinese data), the issue of identifying a model that works well for all the countries remain open. It also involves some consideration around over-fitting the data with many parameters and the increasing computational complexity in fitting and then clustering. Therefore, another advantage of the approach proposed in the present study is the rank-size relationships' capacity to create a unified environment where comparisons are possible. Namely, we can fit each country's data and compare the results, ensuring that the best fit capacity does not affect the clustering activity.

The rank-size approach has the advantage (in this case) of allowing the analysis without data's temporal feature. Namely, in sorting the observations of new deaths per million and ranking them, the dates in which the causalities occurred are no longer relevant to reach conclusions regarding the countries. In this way, the issues presented by Middelburg & Rosendaal (2020) and Zarikas et al. (2020) do not affect our analysis. Zarikas et al. (2020) "compare the time series of COVID-19 regarding active cases or similar variables" to model the evolution of the pandemic. In our work, we do not make clustering based on different types of time evolution (e.g., strong, medium, mild etc.), but our ranking uses the number of new deaths per million in a certain time period. Furthermore, Zarikas et al. (2020) concern solely the first wave while the present paper mixes different waves to capture information on the phenomenon as a whole. Besides, the study is advancing a unique combination of rank-size and clustering analysis to evaluate past realizations of COVID-19 patterns. This is novel in the literature to the best of our knowledge.

The rest of the paper is organised as follows. Section 2 describes the considered dataset and presents the methodologies employed for the analysis. Section 3 contains the empirical results, along with a discussion of them.

2 Data and methods

The time series of daily new deaths per million by country has been downloaded from Roser et al. (2020). The data source collects a comprehensive set of variables describing many features related to COVID-19, and it has been employed in several authoritative studies (see, e.g. those from Zhao et al. 2020, Hasell et al. 2020, Berg et al. 2020). Each country has a specific reference period, depending on the registered beginning of the pandemic propagation. All the investigated periods ends on April 18th, 2021 – when data have been retrieved – while the starting points are reported in the last column of Table 2.

Countries are rather heterogeneous in terms of health care standards. This might create different reporting best practices, especially at the beginning of the pandemic (see for example McDonell 2020). To overcome this potential bias and obtain a more reliable dataset, we have chosen countries with a high level of HAQ, presented by Barber et al. (2017). Such an indicator is listed for 195 countries. Despite the index published in 2017, the most recent levels are reported for 2015 when Andorra had the highest level with 94.6, and the Central African Republic had the minimum with 28.6. So, we have chosen to keep the 39 countries appearing in the last 20th percentile of the HAQ distribution in 2015; see the first two columns of Table 2 for the details. The same table contains the main descriptive statistics of the considered data at a country level for a more detailed overview.

We have preprocessed the dataset to make the best fit more effective and avoid distortions. First, we have removed the outliers in each analyzed series by applying an interquartile method. Namely, for each series, we have calculated the 75th ($Q3$) and 25th ($Q1$) percentiles; then, we eliminated all the observations being outside the range $[Q1 - 1.5 \times (Q3 - Q1), Q3 + 1.5 \times (Q3 - Q1)]$. In doing so, we have faced the so-called king and vice-roy and queen and harem effect, i.e. the deviations due to outliers at low and high ranks in the rank-size analysis (see, e.g. Ausloos 2014, Cerqueti & Ausloos 2015, Ficcadenti et al. 2020). The country data presents tails on the right side only so that the eliminated observations always sit on the right side of the upper limit. Second, we have removed Andorra, Iceland, New Zealand, and Singapore from the investigated sample since they are not relevant in a rank-size context. Indeed, such countries luckily had just a few days in which deaths were experienced, so they present zero new deaths in most of the days in the period under analysis. Therefore, we have obtained $N = 35$ countries after this preprocessing phase.

The rank-size analysis is implemented at an individual country level. Each country's new daily deaths per million (size) represent the (daily) sizes. The ranks are associated with the daily sizes in decreasing order. Specifically, we have given rank one to the day with the highest level of new deaths per million and the highest rank to the day associated with the smallest number of new daily deaths – possibly, zero.
After many trials of different functional forms such as the Zipf-Mandelbrot by (see Mandelbrot 1961, 1953) and the Universal Law by (see Ausloos & Cerqueti 2016), we have used for each country the following third-degree polynomial relationship which, as

	Locations	Care quality index	Daily deaths per million stats							
	Country	HAQ	Min	Max	μ	m	σ	Skew	Kurt	starting date
0	Andorra	94.6	0.0	77.66	4.05	0.00	9.39	3.19	13.71	2020-03-22
1	Australia	89.8	0.0	2.31	0.09	0.00	0.21	4.95	37.01	2020-03-01
2	Austria	88.2	0.0	24.20	2.73	1.33	3.66	2.01	4.67	2020-03-12
3	Belgium	87.9	0.0	42.80	5.10	2.85	6.50	2.24	6.16	2020-03-11
4	Canada	87.6	0.0	6.46	1.54	0.99	1.47	1.00	0.17	2020-03-09
5	Croatia	81.6	0.0	22.41	4.04	1.22	5.22	1.42	1.11	2020-03-19
6	Cyprus	85.3	0.0	9.13	0.87	0.00	1.58	2.46	6.99	2020-03-22
7	Czechia	84.8	0.0	27.55	6.76	2.15	7.35	0.63	-0.95	2020-03-22
8	Denmark	85.7	0.0	10.36	1.06	0.34	1.50	2.21	5.80	2020-03-14
9	Estonia	81.4	0.0	12.82	2.12	0.75	2.97	1.40	1.05	2020-03-25
10	Finland	89.6	0.0	7.76	0.41	0.00	0.74	4.00	27.54	2020-03-21
11	France	87.9	0.0	21.10	3.46	2.22	4.06	1.75	3.38	2020-02-15
12	Germany	86.4	0.0	20.70	2.35	0.81	3.35	2.05	4.45	2020-03-09
13	Greece	87.0	0.0	11.61	2.25	0.58	2.89	1.27	0.46	2020-03-11
14	Hungary	79.6	0.0	32.19	6.52	1.45	7.94	1.18	0.56	2020-03-15
15	Iceland	93.6	0.0	14.65	0.22	0.00	1.13	7.85	79.29	2020-03-21
16	Ireland	88.4	0.0	44.55	2.43	0.91	4.13	4.12	29.18	2020-03-11
17	Israel	85.5	0.0	11.67	1.85	1.27	1.86	1.83	4.37	2020-03-20
18	Italy	88.7	0.0	16.42	4.57	4.18	4.14	0.53	-0.82	2020-02-21
19	Japan	89.0	0.0	1.96	0.18	0.08	0.23	2.38	9.57	2020-02-13
20	Kuwait	82.0	0.0	3.28	0.89	0.70	0.67	0.86	0.33	2020-04-04
21	Lebanon	80.0	0.0	51.42	2.51	1.03	3.99	5.24	55.34	2020-03-10
22	Luxembourg	89.3	0.0	46.33	3.14	0.00	5.79	3.28	14.76	2020-03-14
23	Malta	85.1	0.0	15.85	2.48	1.13	3.15	1.23	0.91	2020-04-08
24	Montenegro	80.7	0.0	28.66	5.80	3.18	5.99	0.98	0.38	2020-03-23
25	Netherlands	89.5	0.0	13.66	2.45	1.63	2.52	1.29	1.45	2020-03-06
26	New Zealand	86.2	0.0	0.83	0.01	0.00	0.07	6.96	58.30	2020-03-29
27	Norway	90.5	0.0	4.98	0.33	0.00	0.65	3.18	12.64	2020-03-14
28	Poland	79.6	0.0	25.26	4.07	0.77	5.30	1.30	0.69	2020-03-12
29	Portugal	84.5	0.0	29.72	4.18	1.42	6.13	2.37	5.55	2020-03-17
30	Qatar	85.2	0.0	3.47	0.34	0.00	0.54	2.58	8.59	2020-03-28
31	Saudi Arabia	79.4	0.0	1.67	0.50	0.34	0.39	0.79	-0.47	2020-03-24
32	Singapore	86.3	0.0	0.34	0.01	0.00	0.05	4.08	17.55	2020-03-21
33	Slovenia	87.4	0.0	31.75	4.99	1.44	7.09	1.52	1.33	2020-03-14
34	South Korea	85.8	0.0	0.78	0.08	0.04	0.11	2.51	7.73	2020-02-20
35	Spain	89.6	0.0	34.71	4.10	2.11	5.25	1.64	3.36	2020-03-03
36	Sweden	90.5	0.0	46.93	3.43	0.50	6.33	3.44	15.33	2020-03-10
37	Switzerland	91.8	0.0	19.76	2.99	0.92	4.34	1.87	3.00	2020-03-05
38	United Kingdom	84.6	0.0	26.90	4.59	2.22	5.48	1.55	2.05	2020-03-06
39	United States	81.3	0.0	13.52	4.13	3.28	2.95	1.10	0.73	2020-02-29

Table 2: Level of the last twentieth percentile of the HAQ index and the statistical summary of the number of daily deaths per million.

we will see, gives satisfactory best fit outcomes:

$$z = a + b \cdot r + c \cdot r^2 + d \cdot r^3, \tag{1}$$

where z is the size, r represents the rank related to size z and a, b, c and d are real parameters to be calibrated. To implement the best fit procedure, we have used the Scikit-learn Python's library (Pedregosa et al. 2011) which leads to a parameters' estimation "by adding higher-order polynomial terms of existing data features as new features in the dataset" as reported by Bisong (2019).

Once the best fit procedure is performed, each country $i = 1, \ldots, N$ remains associated to four calibrated parameters, collected in a vector $x_i = (\hat{a_i}, \hat{b_i}, \hat{c_i}, \hat{d_i})$.

Such parameters have been used in the countries' clustering procedure by adopting the *"k-means++"* Scikit-learn Python algorithm (see contributions from Pedregosa et al. 2011, Arthur & Vassilvitskii 2006). Anyway, the effect of random initialization have been tested, and they did not impact the results. To implement the k-means++ procedure, the parameters have been standardized over the considered countries. We set $k = 3$ after inspecting different possibilities via the Silhouette, Calinski, Davies and Dunn coefficients. The results of the evaluation are reported in Table 3, as it is possible to note, they straightforwardly suggest $k = 3$ as the best option. Indeed, obtaining adjacent neighbourhoods of clusters representing different countries' structures and regimes is preferred. So, taking $k = 3$ means selecting the condition where the distance between clusters is the minimum. Hence, the groups are close to each other, and the variance in the clusters is maximum, ensuring more comprehensive clusters' perimeters, namely a higher probability of capturing the countries behaving similarly, falling in the same clusters' area.

The proposed clustering algorithm selects the three clusters' centroids that minimize the within-clusters sum-of-squares criterion:

$$\sum_{i=1}^{N} \min_{\mu^{(J)} \in \mathbb{R}^4 : J=0,1,2} (||x_i - \mu^{(J)}||^2) \tag{2}$$

where $||x - \mu||$ is the Euclidean distance between the four-dimensional vectors x and μ.

Let us denote the centroids coming from the optimization procedure in (2) by $\bar{\mu}^{(0)}, \bar{\mu}^{(1)}, \bar{\mu}^{(2)}$; they are associated to clusters labelled with "0", "1" and "2", respectively. Such optimized centroids lead to a classification of countries $1, \ldots, N$, by stating that $i \in J$ if and only if $||x_i - \bar{\mu}^{(J)}||$ is the minimum value of the Euclidean distances between x_i and the centroids $\bar{\mu}^{(0)}, \bar{\mu}^{(1)}, \bar{\mu}^{(2)}$, for $i = 1, \ldots, N$ and $J = 0, 1, 2$.

To summarise the proposed procedure, we report in pseudo-code what described above in Algorithm 1.

Algorithm 1 Summary of the rank-size analysis and k-means clustering.

```
Select the countries and their new daily deaths per millions series;   ▷ in our case they are 35
for $i = [1, \dots, N]$ do   ▷ where N is the number of countries
    Eliminate the outliers by using the interquartile methods;
    Sort in descending order the series of new daily deaths per millions and assign ranks;
    Run a third degree polynomial regression to estimate Eq. (1)'s parameters;
    Save the results;
end for
Determine the optimal number of clusters $k$;   ▷ in our case $k = 3$, see Table 3
Place the initial clusters' centroids according to the "k-means++" variant, see Arthur & Vassilvitskii (2006);
repeat
    for $i = [1, \dots, N]$ do
        find the $x_i$'s nearest centroid using the minimum Euclidean distance: $\min_{\mu_j \in \mathbb{R}^4 : j=[0,1,2]}(||x_i - \mu_j||^2)$;   ▷ note that $x_i$ is the $i^{th}$ quadruple of parameters
        assign the $i^{th}$ data point to the cluster having the closest centroid;
    end for
    Update the centroids with the average of the values belonging to the respective clusters;
until Convergence of centroids reach steady points or until a fixed number of iterations is reached.
```

Clusters # k	2	3	4	5
Silhouette	0.4725	**0.4245**	0.4563	0.4362
Calinski	41.8432	**36.0943**	41.4165	44.1670
Davies	0.7532	**0.8654**	0.6857	0.5328
Dunn	0.1232	**0.0596**	0.0827	0.1272

Table 3: Evaluation of the best k for the k-means cluster analysis. The reference for the indexes used are listed here in the same order they appear in the table Rousseeuw (1987), Davies & Bouldin (1979), Dunn (1974), Caliński & Harabasz (1974)

3 Results

The results of the best fit for Eq. (1) are reported in the first six columns of Table 4. The R^2 and the RSME are outstanding, proving the ability of Eq. (1) to represent the rank-size relationship. The interpretation of the calibrated parameters $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$ leads to relevant comments related to the considered countries. The parameter $\hat{a}$ is the intercept of the best fit curve with the y-axis. Hence, such a parameter is positively

influenced by the highest level of new daily deaths per million experienced by the countries. We observe the maximum value of $\hat{a}$ held by Hungary and the minimum one by Australia.
Differently, $\hat{b}$ is associated with the slope of the decay; thus, not unexpectedly, it is always negative in our case. In particular, at low ranks, a high value of the absolute value of $\hat{b}$ stands for a steep curve, while $\hat{b}$ close to zero means that the curve is rather flat. The difference between such cases is the distance between the sizes at the low ranks, which is large for the steep cases and small for the flat ones. We observe that countries experiencing a single pandemic wave with low daily deaths have similar values. This explains why such countries have parameter $\hat{b}$ much closer to zero – see the case of Australia in Figure 1; on the contrary, the maximum value of the absolute value of $\hat{b}$ is scored by Slovenia.
The concavity is driven by $\hat{c}$, that is positive or negative according to a convex or concave shape of the curve at low ranks, respectively. In our case, such a calibrated parameter is always positive, except for Czechia and Italy. Therefore, for all the other countries, decrements of the low-rank sizes decrease as the rank grows. This means that the highest number of daily deaths per million form a peak in the overall distribution. The flatter shape of the concave curve at low ranks for Italy and Czechia points to more homogeneous values of the daily deaths per million at low ranks. Such behaviours are amplified as the absolute value of $\hat{c}$ increases. We notice that Slovenia scores the maximum value of such a parameter in this respect.
Concluding, an easy computation gives that rank $r = -\frac{\hat{c}}{3 \cdot \hat{d}}$ represents the unique inflection point of the best fit curve, where a change of concavity is observed. This quantity is reported in column "Inflection point" of Table 4. The highest value of the rank associated with the inflection point is scored by Czechia, while the lowest is by Slovenia. Such values further confirm the aforementioned logic, with Czechia having experienced a more recent and prolonged critical situation than Slovenia (see Figure 1).
The standardised parameters $\hat{a}$, $\hat{b}$, $\hat{c}$ and $\hat{d}$ are employed to feed the clustering algorithm. The resulting clusters are reported in the column "Clusters" in Table 4 which, jointly with Figures 1, 2 and 3, further informs about the features captured in fitting the data with Eq. (1). The cluster identified with the colour blue and the number "0" mainly contains countries with a relatively low number of deaths per million; Australia, Japan and South Korea obtain the lowest losses. This is further confirmed by sorting the results by $\hat{a}$. It is relevant to point out that the same countries appear with the lowest values of the calibrated parameter $\hat{b}$. Such a finding makes sense because when a series of deaths gets a shock, it takes time to get back to zero, so this is reflected in a more gentle decay registered in the rank-size relationship. Interestingly, the cluster indicated in orange and identified by the number "1" is characterised by low values of $\hat{b}$ and $\hat{d}$, but high values of $\hat{c}$. For example, the countries with the lowest $\hat{b}$ are Slovenia, Hungary, Poland, Croatia and Belgium. The highest $\hat{c}$ are reported by Poland, Croatia, Luxembourg, Belgium and Slovenia. The lowest $\hat{d}$s are scored by Slovenia, Belgium, Luxembourg, Portugal and Croatia. Despite the outlier removal procedure, these relatively small countries suffered losses with high daily peaks. Finally, the green cluster

identified with the number "2" has countries sitting in the middle of the distribution when values are sorted by $\hat{b}$. Similarly, when they are ordered by $\hat{c}$ except for Czechia and Italy, which present the lowest values of $\hat{c}$ even if they belong to cluster number 2. To justify such a result, we look at the "Inflection point" column in Table 4. Indeed, for Czechia and Italy, the inflection point falls at low ranks, and from Figure 1 we can see that the quoted countries have experienced lengthened periods of high new deaths per million. Montenegro and United States also show similar patterns having long periods of high levels of new daily death per million. However, for them, such a condition is quite evident over the investigated period; therefore, the situation does not allow for a change in concavity to happen early in the rank, even if Montenegro and United States belong to the same cluster of Czechia and Italy.

Moving on to an example, Figure 4 contains a comparison of Italy and the UK. It allows concluding that Italy's change in the concavity at middle ranks lead the country to be in Cluster 2, on the other hand, the UK's number of days with high deaths per million presented at low ranks makes the country more suitable for Cluster 1.

4 Conclusions

This paper aims at providing a unified framework at a country level of the number of deaths for COVID-19 by moving from the daily data and through different waves. To this aim, we implement a rank-size analysis via a four-parameter third-degree polynomial on the series of COVID-19 new deaths per million registered in 35 countries. The statistical soundness of the results allows the identification of different regimes in the data. Specifically, it is possible to make comparisons between countries by using a rank-size approach because the best fit of Eq. (1) are statistically good in all the considered areas (see Table 4). We have provided a reasonable interpretation of the four estimated parameters in Eq. (1), hence capturing insightful information regarding the COVID-19 severity in the countries. In this respect, the clustering activity is grounded on the parameters calibrated from the best-fit exercise, and we group countries according to the phenomenon's features captured by such rank-size function's parameters. The main determinants are given by days with picks of deaths, the steadiness of casualties number, endured COVID-19 waves and other elements that affect the shape of the ranked data. In Figure 3 a visual representation of the cluster profiles is reported, and in Table 4 the clusters are summarized. The clustering exercise leads to relevant information for policymakers. Indeed, Government and supra-national health institutions might carry out common strategies for contrasting the diffusion of COVID-19 and reducing its fatality rate. That can be done by investigating the similarities and divergences among countries described by the clustering procedure's results. Furthermore, countries' policymakers monitoring their own conditions and those of interrelated countries, for example, because of import/export relationships, may benefit from the clusterisation to detect risks and define actions points. Namely, at a given point in time, the cluster to which a country belongs and the ones of its partners/competitors provide proxies to evaluate the exposure to the pandemic, suggesting actions like stock up on essential resources for preserving economic interests or evaluating countries' common features. This is relevant, especially in the case of pandemics, because countries ahead experiencing waves can be seen as flags for interrelated countries (e.g., countries strongly connected via single transportation systems) not yet in the same situation; or because knowing that other countries are in the same cluster provides a view on their managerial abilities and infrastructures conditions. Moreover, the rank-size best-fit curve can effectively describe the pattern of the pandemic in that it provides a clear illustration of the ratios between the consecutive ordered ranked data. Thus, a clustering procedure rank-size based is able to distinguish the countries where the pandemic maintains a generally stable number of fatalities from those with remarkable high peaks of fatalities. So, a policymaker can gain several insights into the effects of countries' policies on the pandemic evolution.

It is important to notice that the rank-size approach allows for evaluations of the overall phenomenon without referring to specific periods and time ranges. In doing so, the proposed approach is free from biases associated with the time inconsistency of the data at country levels. Hence, we do not need here to implement time-based normalising procedures, which would demand an additional transformation of the data as reported by Zarikas et al. (2020), Middelburg & Rosendaal (2020).

	Country	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	R2	RSME	Clusters	Inflection Point	Max Rank
0	Australia	0.07989	-0.00142	0.00001	-1.000000e-08	0.89	0.01	0	205.51	328.0
1	Austria	7.71274	-0.06628	0.00020	-2.000000e-07	0.99	0.25	2	324.46	370.0
2	Belgium	14.71732	-0.14039	0.00051	-6.600000e-07	0.99	0.41	1	259.22	373.0
3	Canada	5.05506	-0.03312	0.00008	-7.000000e-08	1.00	0.07	0	383.40	403.0
4	Croatia	15.38722	-0.14446	0.00046	-4.900000e-07	1.00	0.24	1	311.59	377.0
5	Cyprus	2.74948	-0.03370	0.00013	-1.600000e-07	0.92	0.22	0	266.48	355.0
6	Czechia	22.19969	-0.09063	-0.00010	5.000000e-07	0.98	1.12	2	67.33	393.0
7	Denmark	2.96188	-0.03126	0.00012	-1.500000e-07	0.99	0.08	0	255.95	361.0
8	Estonia	8.98464	-0.08503	0.00026	-2.500000e-07	0.99	0.20	2	340.76	376.0
9	Finland	1.32399	-0.01447	0.00005	-6.000000e-08	0.99	0.04	0	290.43	361.0
10	France	10.77632	-0.07676	0.00020	-1.900000e-07	0.98	0.40	2	353.08	411.0
11	Germany	6.72854	-0.06306	0.00020	-2.200000e-07	1.00	0.10	2	307.50	370.0
12	Greece	9.41917	-0.08441	0.00025	-2.500000e-07	1.00	0.13	2	336.68	392.0
13	Hungary	24.11789	-0.17821	0.00040	-2.500000e-07	0.98	0.91	1	531.04	389.0
14	Ireland	5.81396	-0.06108	0.00023	-3.000000e-07	0.98	0.21	2	258.42	359.0
15	Israel	5.22539	-0.04165	0.00014	-1.800000e-07	0.99	0.11	0	259.90	377.0
16	Italy	13.09504	-0.04930	-0.00000	1.000000e-07	0.99	0.48	2	2.78	423.0
17	Japan	0.56195	-0.00512	0.00002	-2.000000e-08	1.00	0.01	0	288.22	401.0
18	Kuwait	2.10195	-0.01331	0.00004	-6.000000e-08	0.98	0.07	0	232.29	361.0
19	Lebanon	8.66072	-0.08967	0.00032	-3.800000e-07	0.98	0.29	2	280.91	372.0
20	Luxembourg	11.25358	-0.12697	0.00047	-5.500000e-07	0.99	0.36	1	279.46	374.0
21	Malta	10.49816	-0.08532	0.00022	-1.700000e-07	0.96	0.59	2	425.74	375.0
22	Montenegro	19.16544	-0.11586	0.00021	-9.000000e-08	0.99	0.56	2	739.01	389.0
23	Netherlands	7.95225	-0.05571	0.00015	-1.400000e-07	0.99	0.21	2	339.69	400.0
24	Norway	0.95166	-0.01289	0.00006	-8.000000e-08	0.97	0.04	0	243.20	358.0
25	Poland	18.07872	-0.15548	0.00044	-3.900000e-07	0.99	0.43	1	368.18	399.0
26	Portugal	11.46477	-0.11435	0.00041	-4.900000e-07	0.98	0.37	1	274.74	366.0
27	Qatar	0.78369	-0.00656	0.00001	-0.000000e+00	0.90	0.08	0	1005.22	339.0
28	Saudi Arabia	1.42653	-0.00809	0.00002	-1.000000e-08	0.99	0.03	0	493.72	391.0
29	Slovenia	18.51710	-0.20607	0.00076	-9.400000e-07	0.99	0.38	1	272.28	368.0
30	South Korea	0.19842	-0.00168	0.00001	-1.000000e-08	0.98	0.01	0	269.90	385.0
31	Spain	15.86274	-0.12499	0.00032	-2.800000e-07	0.99	0.36	1	390.35	402.0
32	Sweden	10.45896	-0.10913	0.00037	-4.200000e-07	1.00	0.11	1	297.29	373.0
33	Switzerland	9.27636	-0.09700	0.00035	-4.100000e-07	1.00	0.15	1	278.21	372.0
34	United Kingdom	15.66925	-0.12922	0.00038	-3.800000e-07	0.99	0.41	1	327.88	391.0
35	United States	10.15715	-0.07638	0.00029	-4.100000e-07	1.00	0.15	2	235.39	394.0

Table 4: Estimated parameters and clusters per each country. The last two columns respectively represent the rank at which the best fit of Eq. (1) presents a change in concavity and the maximum rank obtained for that country, namely the length of the series.

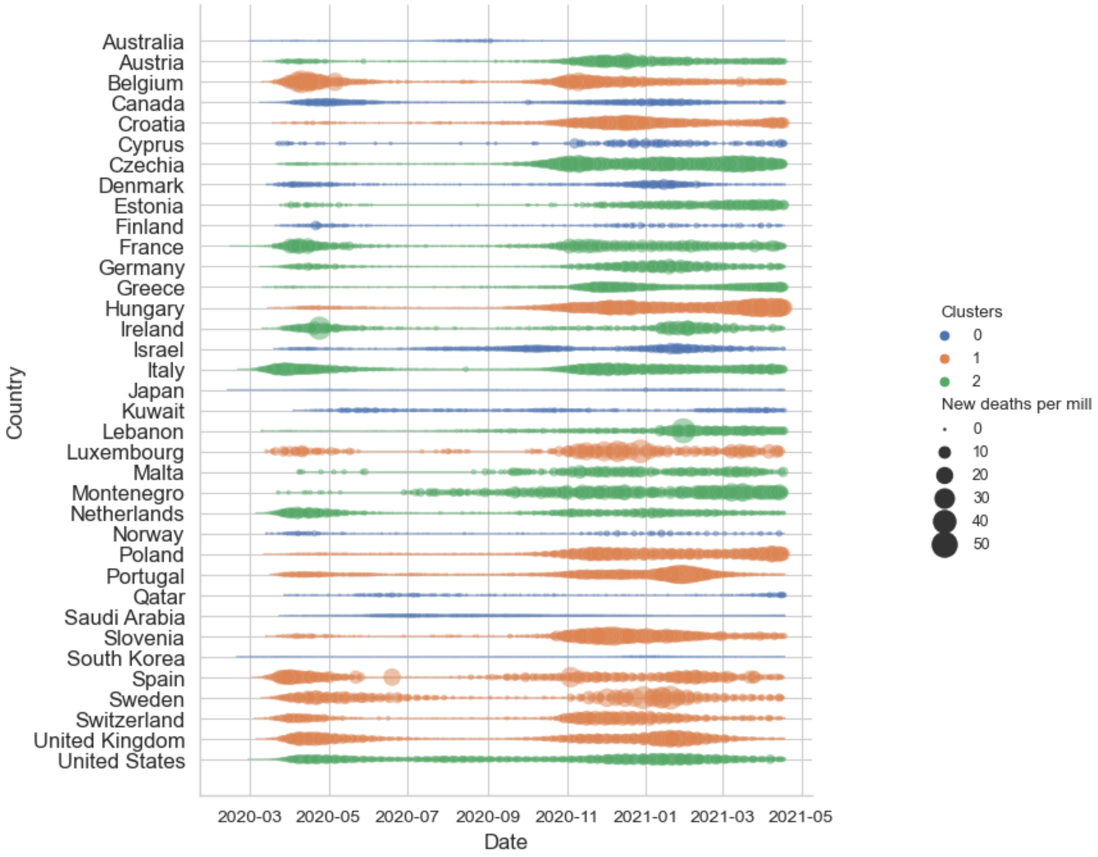

Figure 1: Time series of the daily new deaths per million occurred in each considered country. The colours represent the clusters, and the size of the dots show the level of new daily deaths.

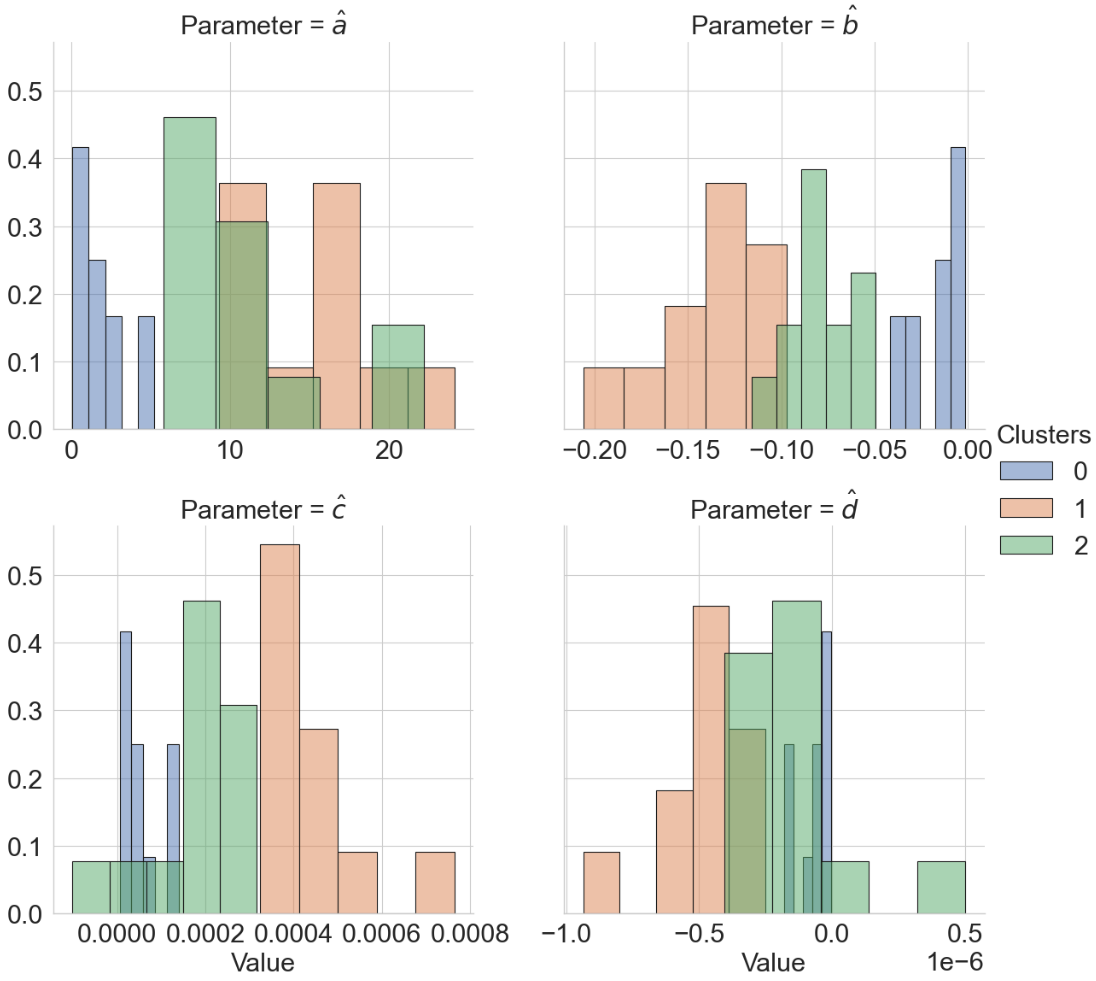

Figure 2: Histograms of the estimated parameters divided by the resulting clusters, five bins per colour, are searched. On the y-axes, there are the relative frequencies that sum to one per cluster.

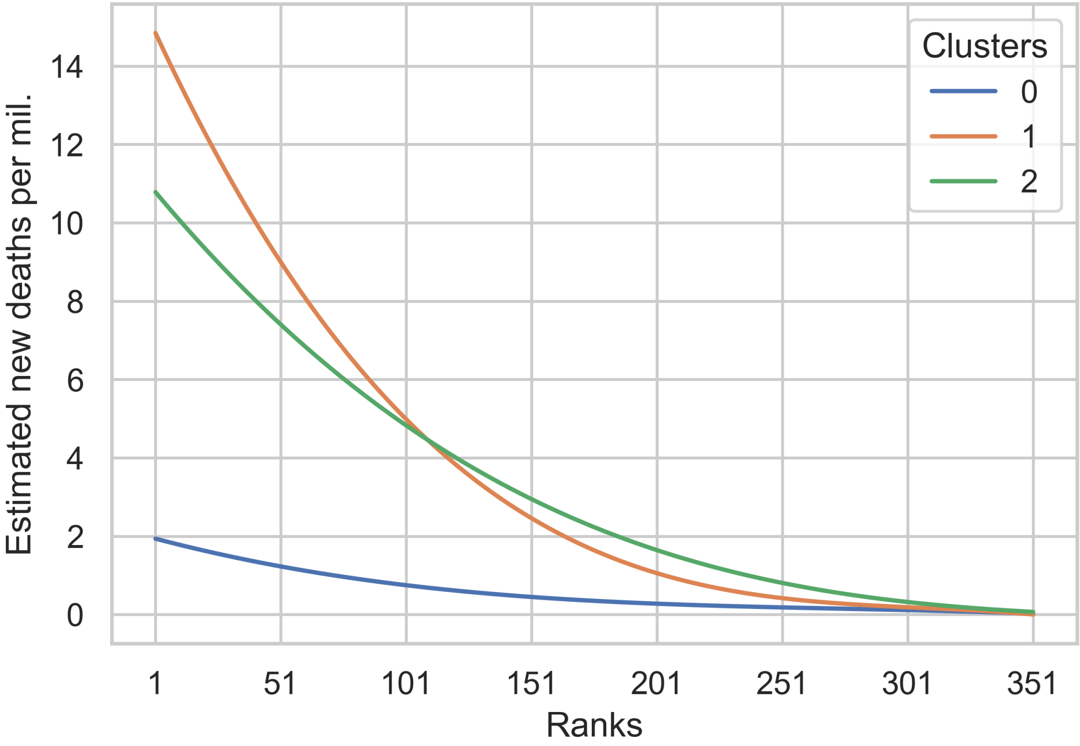

Figure 3: The 3 curves are obtained by plugging into Eq. (1), $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{d}$ corresponding to the centroid of the clusters $\{0,1,2\}$.

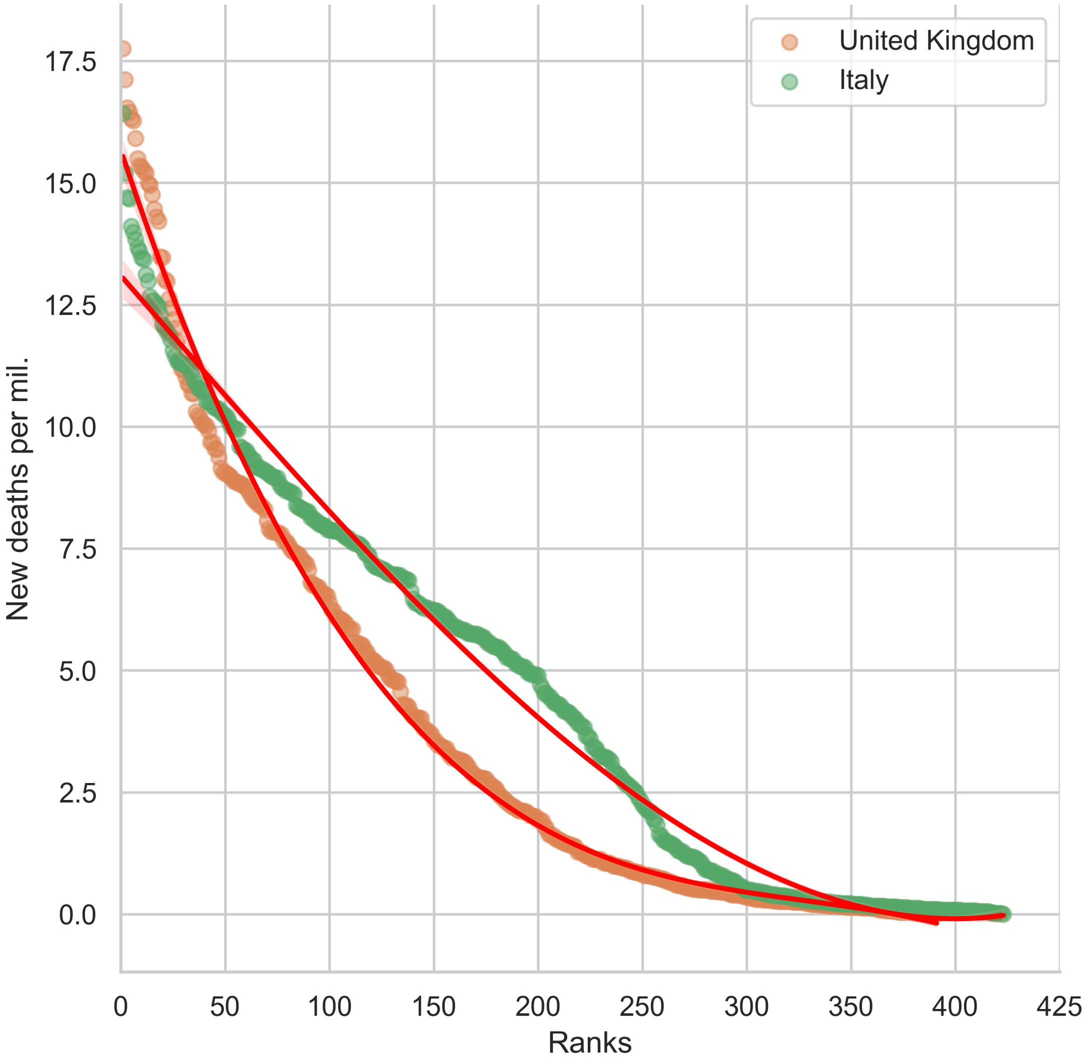

Figure 4: The cases of the United Kingdom and Italy are reported here with their best fits (red lines) obtained through Eq. (1) and the countries' respective parameters $\hat{a}$, $\hat{b}$, $\hat{c}$, and $\hat{d}$ from Table 4. The colours of the dots correspond to the different clusters, 1 for the UK and 2 for Italy.

Country	$\hat{a}$	$\hat{b}$	$\hat{c}$	$\hat{d}$	R^2	RSME
Italy	13.09504	-0.04930	-0.00000	1.000000e-07	0.99	0.48
United Kingdom	15.66925	-0.12922	0.00038	-3.800000e-07	0.99	0.41

References

Abdullah, D., Susilo, S., Ahmar, A. S., Rusli, R. & Hidayat, R. (2021), 'The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data', *Quality & Quantity* pp. 1–9.

Arthur, D. & Vassilvitskii, S. (2006), k-means++: The advantages of careful seeding, Technical Report 2006-13, Stanford InfoLab.
URL: *http://ilpubs.stanford.edu:8090/778/*

Ausloos, M. (2014), 'Zipf–Mandelbrot–Pareto model for co-authorship popularity', *Scientometrics* **101**(3), 1565–1586.

Ausloos, M. & Cerqueti, R. (2016), 'A universal rank-size law', *PloS One* **11**(11), e0166011.

Barber, R. M., Fullman, N., Sorensen, R. J., Bollyky, T., McKee, M., Nolte, E., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M. et al. (2017), 'Healthcare access and quality index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015', *The Lancet* **390**(10091), 231–266.

Bartolucci, F. & Farcomeni, A. (2021), 'A spatio-temporal model based on discrete latent variables for the analysis of COVID-19 incidence', *Spatial Statistics* p. 100504.

Berg, M. K., Yu, Q., Salvador, C. E., Melani, I. & Kitayama, S. (2020), 'Mandated bacillus calmette-guérin (BCG) vaccination predicts flattened curves for the spread of COVID-19', *Science Advances* **6**(32), eabc1463.

Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. (2020), 'The challenges of modeling and forecasting the spread of COVID-19', *Proceedings of the National Academy of Sciences* **117**(29), 16732–16738.

Bisong, E. (2019), *Building Machine Learning and Deep Learning Models on Google Cloud Platform*, Apress, Berkeley, CA, chapter Linear Regression, pp. 231–241.

Caliński, T. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics* **3**(1), 1–27.
URL: *https://www.tandfonline.com/doi/abs/10.1080/03610927408827101*

Cerqueti, R. & Ausloos, M. (2015), 'Evidence of economic regularities and disparities of Italian regions from aggregated tax income size data', *Physica A: Statistical Mechanics and its Applications* **421**, 187–207.

Davies, D. L. & Bouldin, D. W. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-1**(2), 224–227.

Dunn, J. (1974), 'Well-separated clusters and optimal fuzzy partitions', *Journal of Cybernetics* **4**(1), 95–104.
URL: *https://doi.org/10.1080/01969727408546059*

Ficcadenti, V. & Cerqueti, R. (2017), 'Earthquakes economic costs through rank-size laws', *Journal of Statistical Mechanics: Theory and Experiment* **2017**(8), 083401.

Ficcadenti, V., Cerqueti, R. & Ausloos, M. (2019), 'A joint text mining-rank size investigation of the rhetoric structures of the US Presidents' speeches', *Expert Systems with Applications* **123**, 127–142.

Ficcadenti, V., Cerqueti, R., Ausloos, M. & Dhesi, G. (2020), 'Words ranking and Hirsch index for identifying the core of the hapaxes in political texts', *Journal of Informetrics* **14**(3), 101054.

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W. et al. (2020), 'Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe', *Nature* **584**(7820), 257–261.

Hasell, J., Mathieu, E., Beltekian, D., Macdonald, B., Giattino, C., Ortiz-Ospina, E., Roser, M. & Ritchie, H. (2020), 'A cross-country database of COVID-19 testing', *Scientific Data* **7**(1), 1–7.

Hutagalung, J., Ginantra, N. L. W. S. R., Bhawika, G. W., Parwita, W. G. S., Wanto, A. & Panjaitan, P. D. (2021), COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm, *in* 'Journal of Physics: Conference Series', Vol. 1783, IOP Publishing, p. 012027.
URL: *https://doi.org/10.1088/1742-6596/1783/1/012027*

Ioannidis, J. P., Cripps, S. & Tanner, M. A. (2020), 'Forecasting for COVID-19 has failed', *International Journal of Forecasting* . doi: 10.1016/j.ijforecast.2020.08.004.
URL: *https://www.sciencedirect.com/science/article/pii/S0169207020301199*

James, N. & Menzies, M. (2020), 'Cluster-based dual evolution for multivariate time series: Analyzing COVID-19', *Chaos: An Interdisciplinary Journal of Nonlinear Science* **30**(6), 061108.

Jiang, B. & de Rijke, C. (2021), 'A power-law-based approach to mapping COVID-19 cases in the United States', *Geo-spatial Information Science* **24**(3), 333–339.
URL: *https://doi.org/10.1080/10095020.2020.1871306*

Kennedy, A. P. & Yam, S. C. P. (2020), 'On the authenticity of COVID-19 case figures', *PloS One* **15**(12), e0243123.

Kiaghadi, A., Rifai, H. S. & Liaw, W. (2020), 'Assessing COVID-19 risk, vulnerability and infection prevalence in communities', *PLOS ONE* **15**(10), 1–21.
URL: *https://doi.org/10.1371/journal.pone.0241166*

Kumar, S. (2020), 'Monitoring novel corona virus (COVID-19) infections in India by cluster analysis', *Annals of Data Science* **7**, 417–425.

Lee, D., Robertson, C. & Marques, D. (2021), 'Quantifying the small-area spatio-temporal dynamics of the Covid-19 pandemic in Scotland during a period with limited testing capacity', *Spatial Statistics* . doi: 10.1016/j.spasta.2021.100508.
URL: *https://www.sciencedirect.com/science/article/pii/S221167532100018X*

Li, Z., Wang, L., Huang, L.-s., Zhang, M., Cai, X., Xu, F., Wu, F., Li, H., Huang, W., Zhou, Q. et al. (2021), 'Efficient management strategy of COVID-19 patients based on cluster analysis and clinical decision tree classification', *Scientific Reports* **11**(1), 1–13.

Machado, J. A. T. & Lopes, A. M. (2020), 'Rare and extreme events: the case of COVID-19 pandemic', *Nonlinear dynamics* **100**, 2953–2972.

Mandelbrot, B. (1953), 'An informational theory of the statistical structure of language', *Communication Theory* **84**, 486–502.

Mandelbrot, B. (1961), 'On the theory of word frequencies and on related Markovian models of discourse', *Structure of Language and Its mathematical Aspects* **12**, 190–219.

McDonell, S. (2020), 'Coronavirus: Sharp increase in deaths and cases in Hubei', BBC.
URL: *https://www.bbc.co.uk/news/world-asia-china-51482994*

Middelburg, R. A. & Rosendaal, F. R. (2020), 'COVID-19: How to make between-country comparisons', *International Journal of Infectious Diseases* **96**, 477–481.
URL: *https://www.sciencedirect.com/science/article/pii/S1201971220303738*

Moein, S., Nickaeen, N., Roointan, A., Borhani, N., Heidary, Z., Javanmard, S. H., Ghaisari, J. & Gheisari, Y. (2021), 'Inefficiency of SIR models in forecasting COVID-19 epidemic: a case study of Isfahan', *Scientific Reports* **11**(1), 1–9.

Nabi, K. N. (2020), 'Forecasting COVID-19 pandemic: A data-driven analysis', *Chaos, Solitons & Fractals* **139**, 110046.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011), 'Scikit-learn: Machine Learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Prasanth, S., Singh, U., Kumar, A., Tikkiwal, V. A. & Chong, P. H. (2021), 'Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach', *Chaos, Solitons & Fractals* **142**, 110336.
URL: *https://www.sciencedirect.com/science/article/pii/S0960077920307311*

Rios, R. A., Nogueira, T., Coimbra, D. B., Lopes, T. J. S., Abraham, A. & de Mello, R. F. (2021), 'Country transition index based on hierarchical clustering to predict next COVID-19 waves', *Scientific Reports* **11**(1), 1–13.

Rizvi, S. A., Umair, M. & Cheema, M. A. (2021), 'Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators', *Chaos, Solitons & Fractals* **151**, 111240.
URL: *https://www.sciencedirect.com/science/article/pii/S0960077921005944*

Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. (2020), 'Coronavirus pandemic (COVID-19)', *Our World in Data,* . https://ourworldindata.org/coronavirus.

Rousseeuw, P. J. (1987), 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics* **20**, 53–65.
URL: *https://www.sciencedirect.com/science/article/pii/0377042787901257*

Schneble, M., De Nicola, G., Kauermann, G. & Berger, U. (2021), 'Nowcasting fatal COVID-19 infections on a regional level in Germany', *Biometrical Journal* **63**(3), 471–489.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202000143*

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H., Hussain, F., Khatoon, K. & Ahmad, S. (2020), 'Correlation between temperature and COVID-19 (suspected, confirmed and death) cases based on machine learning analysis', *Journal of Pure Applied Microbiology* **14**(suppl 1), 1017–24.

Small, C. & Sousa, D. (2021), 'Spatiotemporal evolution of COVID-19 infection and detection within night light networks: comparative analysis of USA and China', *Applied Network Science* **6**(1), 1–20.

Tang, F., Feng, Y., Chiheb, H. & Fan, J. (2021), 'The Interplay of Demographic Variables and Social Distancing Scores in Deep Prediction of U.S. COVID-19 Cases', *Journal of the American Statistical Association* . doi: 10.1080/01621459.2021.1901717.

Tian, T., Tan, J., Luo, W., Jiang, Y., Chen, M., Yang, S., Wen, C., Pan, W. & Wang, X. (2021), 'The Effects of Stringent and Mild Interventions for Coronavirus Pandemic', *Journal of the American Statistical Association* . doi: 10.1080/01621459.2021.1897015.

Tuli, S., Tuli, S., Tuli, R. & Gill, S. S. (2020), 'Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing', *Internet of Things* **11**, 100222.

Vadyala, S. R., Betgeri, S. N., Sherer, E. A. & Amritphale, A. (2021), 'Prediction of the number of covid-19 confirmed cases based on K-means-LSTM', *Array* **11**, 100085.
URL: *https://www.sciencedirect.com/science/article/pii/S2590005621000333*

Vasconcelos, G. L., Cordeiro, L. P., Duarte-Filho, G. C. & Brum, A. A. (2021), 'Modeling the Epidemic Growth of Preprints on COVID-19 and SARS-CoV-2', *Frontiers in Physics* **9**, 125.

Zarikas, V., Poulopoulos, S. G., Gareiou, Z. & Zervas, E. (2020), ‘Clustering analysis of countries using the COVID-19 cases dataset’, *Data in Brief* **31**, 105787.
URL: *https://www.sciencedirect.com/science/article/pii/S2352340920306818*

Zhang, T. & Lin, G. (2021), ‘Generalized k-means in GLMs with applications to the outbreak of COVID-19 in the united states’, *Computational Statistics & Data Analysis* **159**, 107217.
URL: *https://www.sciencedirect.com/science/article/pii/S0167947321000517*

Zhao, F., Zhang, P., Zhang, Y. & Ma, Z. (2020), ‘Time to lead the prevention and control of public health emergencies by informatics technologies in an information era’, *Journal of Biosafety and Biosecurity* . doi: 10.1016/j.jobb.2020.06.001.

Zubair, M., Iqbal, M. A., Shil, A., Haque, E., Hoque, M. M. & Sarker, I. H. (2020), An Efficient K-means Clustering Algorithm for Analysing COVID-19, *in* ‘International Conference on Hybrid Intelligent Systems’, Springer, pp. 422–432.