Cluster-based dual evolution for multivariate time series: Analyzing COVID-19

Cite as: Chaos **30**, 061108 (2020); doi: 10.1063/5.0013156
Submitted: 8 May 2020 · Accepted: 11 June 2020 ·
Published Online: 30 June 2020

View Online Export Citation CrossMark

Nick James[1] and Max Menzies[2,a)]

AFFILIATIONS

[1]School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia
[2]Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China

[a)]**Author to whom correspondence should be addressed:** max.menzies@alumni.harvard.edu

ABSTRACT

This paper proposes a cluster-based method to analyze the evolution of multivariate time series and applies this to the COVID-19 pandemic. On each day, we partition countries into clusters according to both their cases and death counts. The total number of clusters and individual countries' cluster memberships are algorithmically determined. We study the change in both quantities over time, demonstrating a close similarity in the evolution of cases and deaths. The changing number of clusters of the case counts precedes that of the death counts by 32 days. On the other hand, there is an optimal offset of 16 days with respect to the greatest consistency between cluster groupings, determined by a new method of comparing affinity matrices. With this offset in mind, we identify anomalous countries in the progression from COVID-19 cases to deaths. This analysis can aid in highlighting the most and least significant public policies in minimizing a country's COVID-19 mortality rate.

COVID-19 has resulted in a global pandemic with severe human, social, and economic costs. In order to manage the economic ramifications of prioritizing citizen safety, policymakers have sought a multi-level approach involving social distancing, business closures, and movement restrictions. For this purpose, a careful identification of the most and least successful countries at responding to the spread of COVID-19 is of great relevance. This paper meets such a demand by developing a new method to analyze *multivariate time series*, in which the variables are the cumulative cases and death counts of each country on each day. We have three goals: first, we analyze the cases and death counts on a country by country basis; second, we analyze the two multivariate time series in conjunction to elucidate their similarity further; and third, we determine anomalous countries relative to cases and deaths.

I. INTRODUCTION

Understanding the trajectories of COVID-19 cases and death counts assists governments in anticipating and responding to the impact of the pandemic. As the disease spreads, the timely identification of anomalous countries, both successful and unsuccessful, provides opportunities to determine effective response strategies. This analysis can be difficult as death counts naturally lag behind case counts.

This paper builds on the extensive literature of *multivariate time series analysis*, developing a new mathematical method and a more extensive analysis of COVID-19 dynamics than previously performed. Existing methods of time series analysis include parametric models,[1] such as exponential[2] or power-law models,[3] and nonparametric methods, such as distance analysis,[4] distance correlation,[5–7] and network models.[8] Both parametric and nonparametric methods have been used to model COVID-19.[9,10]

Cluster analysis is another common statistical method with successful applications to COVID-19 and more broadly, epidemiology. Designed to group data points according to similarity, cluster analysis has been used to study non-communicable diseases,[11,12] infectious diseases,[13,14] and epidemic outbreaks such as Ebola,[15] SARS,[16] and COVID-19.[10] Clustering algorithms are highly varied—common examples are K-means[17] and spectral clustering,[18] which partition elements into discrete sets, and hierarchical clustering,[19,20] which does not specify a precise number of clusters. In this paper, we will use hierarchical clustering,[19,20] K-means,[17] and its optimal one-dimensional variant Ckmeans.1d.dp.[21] K-means and Ckmeans.1d.dp require an initial choice of the number of clusters k. We draw upon several methods to address the subtle question of how to select this k. The goal of this paper is to use a dynamic and smoothed implementation of cluster analysis to study the worldwide spread of COVID-19, track the relationships between different countries' cases and death counts, and make inferences regarding the most successful strategies in managing the progression from cases to deaths.

This paper is structured as follows: in each of the following three sections, we introduce portions of our methodology and present our results. Section II investigates the multivariate time series of cases and deaths individually. Section III analyzes the two time series in conjunction, determining suitable offsets for the number of clusters and the cluster memberships. Section IV determines anomalous countries with respect to cases and deaths. Section V summarizes the results and the new findings regarding COVID-19.

II. INDIVIDUAL ANALYSIS OF COVID-19 CASES AND DEATHS

A. Time-varying cluster analysis methodology

The most general setup of our methodology is as follows: let $x_i^{(t)}$ be a multivariate time series over an interval of length T, for $i = 1, \dots, n$ and $t = 1, \dots, T$, with each $x_i^{(t)}$ belonging to a common normed space $\mathfrak{X}$. Slightly different procedures apply if $\mathfrak{X}$ is one-dimensional, namely, $\mathbb{R}$, or higher-dimensional.

In this paper, the two multivariate time series we present are the cumulative daily counts of cases and deaths on a country by country basis. We order the countries by alphabetical order and denote these counts by $x_i^{(t)}, y_i^{(t)} \in \mathbb{R}$, respectively. We choose cumulative counts to best analyze the evolution of the disease over time. Our data spans 12/31/2019 to 04/30/2020, a period of $T = 122$ days across $n = 208$ countries.

Given the exponential nature of the data, we choose a logarithmic difference as our metric. First, we do the following data preprocessing: any entry in the data that is empty or 0—before any cases are detected—we replace with a 1, so that the log of that number is defined. Then, we define a distance on case and death counts by $d(x, y) = |\log(x) - \log(y)|$. Effectively, this pulls back the standard metric on $\mathbb{R}$ under the homeomorphism $\log : \mathbb{R}^+ \to \mathbb{R}$ and makes the positive real numbers a one-dimensional normed space.

The goal is to partition the counts $x_1^{(t)}, \dots, x_n^{(t)}$ into a certain number of clusters at each time t. We wish to carefully choose the number of clusters in such a way that provides us meaningful inference on how the data change. A wildly varying number of clusters would obscure inference on individual countries' cluster memberships changing with time. Thus, we combine several methods of choosing this number to reduce the bias in our estimator and perform additional exponential smoothing to yield a suitably changing number with time. In our experiments, we use six methods outlined in Appendix A. These have been chosen after experimentation and consultation with the literature, but our method is flexible and could use any combination of methods. Given cluster numbers $k_1^{(t)}, \dots, k_6^{(t)}$ offered by these methods, we compute the average $k_{av}^{(t)} = \frac{1}{6}\sum_{j=1}^{6} k_j^{(t)}$. This is not necessarily an integer; we do not compute clusters directly with this value.

In our implementation, this average value $k_{av}^{(t)}$ exhibits itself as approximately locally stationary. Thus, we apply exponential smoothing to $k_{av}^{(t)}$ to produce a smoothed integer value $\hat{k}^{(t)}$. We use this value $\hat{k}^{(t)}$ at each t to obtain a clustering at that time. As the daily case and death data are one-dimensional, the most appropriate clustering method is the optimal implementation of K-means specific to one-dimensional data, Ckmeans.1d.dp.[21] We implement this algorithm to group daily counts into $\hat{k}^{(t)}$ clusters and sort the clusters according to the ordering on $\mathbb{R}$.

Similar experiments can also be performed for higher-dimensional data. Analyzing three-day rolling counts of cases and deaths $\tilde{\mathbf{x}}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)} \in \mathbb{R}^3$ requires the use of standard K-means clustering. These yield similar results to the daily analysis and can be seen in Appendix B.

B. Matrix analysis of multivariate time series

We record the results of this analysis in several sequences of matrices. Having performed the data preprocessing described above, first let $D^{(t)}$ be the $n \times n$ matrix of (logarithmic) distances between counts $x_i^{(t)}$ at time t, that is, $D_{ij}^{(t)} = |\log(x_i^{(t)}) - \log(x_j^{(t)})|$. Next, let $\text{Aff}^{(t)}$ and $G^{(t)}$ be two different $n \times n$ *affinity matrices* defined as follows:

$$\text{Aff}_{ij}^{(t)} = 1 - \frac{D_{ij}^{(t)}}{\max D^{(t)}}, \tag{1}$$

$$G_{ij}^{(t)} = \exp\left(\frac{-m^2\left(D_{ij}^{(t)}\right)^2}{2(\max D^{(t)})^2}\right). \tag{2}$$

We term $\text{Aff}^{(t)}$ and $G^{(t)}$ *standard* and *Gaussian affinity matrices*, respectively. These definitions are motivated by standard constructions, but we appropriately normalize G for subsequent analysis. We vary $m = 1, 2, 3$ in experiments so that the matrix entries mimic Gaussian spreads over 1, 2, 3 standard deviations, respectively. Then, let $\text{Adj}^{(t)}$ be an $n \times n$ *adjacency matrix* defined as follows:

$$\text{Adj}_{ij}^{(t)} = \begin{cases} 1, & x_i^{(t)} \text{ and } x_j^{(t)} \text{ are in the same cluster,} \\ 0, & \text{else.} \end{cases}$$

Finally, we define a distance on the set of *dates* $t = 1, \dots, T$. Let the Frobenius norm of an $n \times n$ matrix A be defined as $\|A\| = \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}}$. Given $s, t \in [1, \dots, T]$, let $d(s, t) = \|\text{Adj}^{(t)} - \text{Adj}^{(s)}\|$. Performing hierarchical clustering on these distances $d(s, t)$ produces a dendrogram on the set of dates that we term the *cluster evolution dendrogram*. This groups moments in time according to similarity in the evolving cluster structures. In Appendix C, we include an algorithmic presentation of the steps taken in Secs. II A and II B. In Appendix D, we include a list of mathematical objects and their respective definitions used in this paper.

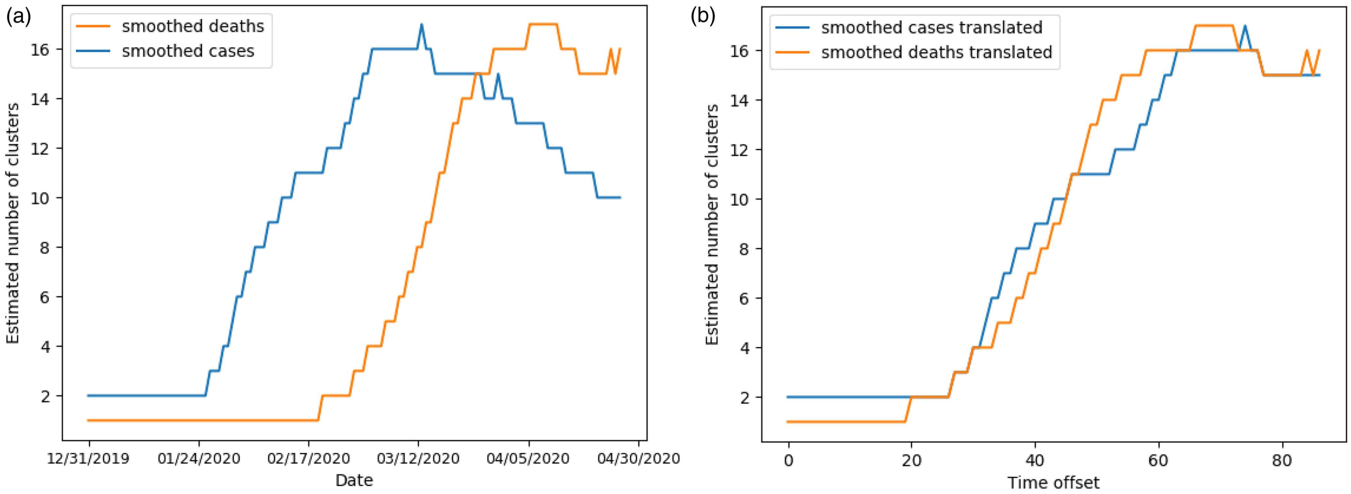

FIG. 1. Smoothed number of clusters $\hat{k}^{(t)}$ as a function of time, defined in Sec. II A. In (a), the blue and orange curves track the number of clusters for cases and deaths, respectively, from 12/31/2019 to 04/30/2020. In (b), the curves are shown after translation by the optimal *series evolution offset*, defined in Sec. III, computed to be $\delta = 32$. There is a strong similarity between the two curves up to this offset: both peak at 17 clusters before declining, suggesting reduced spread in the data.

C. Results for time series of cases

In this section, we implement Ckmeans.1d.dp[21] on daily counts of cases. Experiments using standard K-means on three-day rolling counts of cases produce similar results included in Appendix B. Our analysis supports several aspects of the empirically observed natural history regarding the spread of COVID-19 cases. The smoothed number of clusters $\hat{k}^{(t)}$, depicted in Fig. 1(a), ranges between $\{2, \ldots, 17\}$. Until the end of January, there were only two clusters, with China being the only country severely impacted by the virus. However, as the virus has spread around the world, reported counts have changed day by day, with the number of clusters increasing rapidly toward a peak in early March. As depicted in Fig. 2(a), Italy

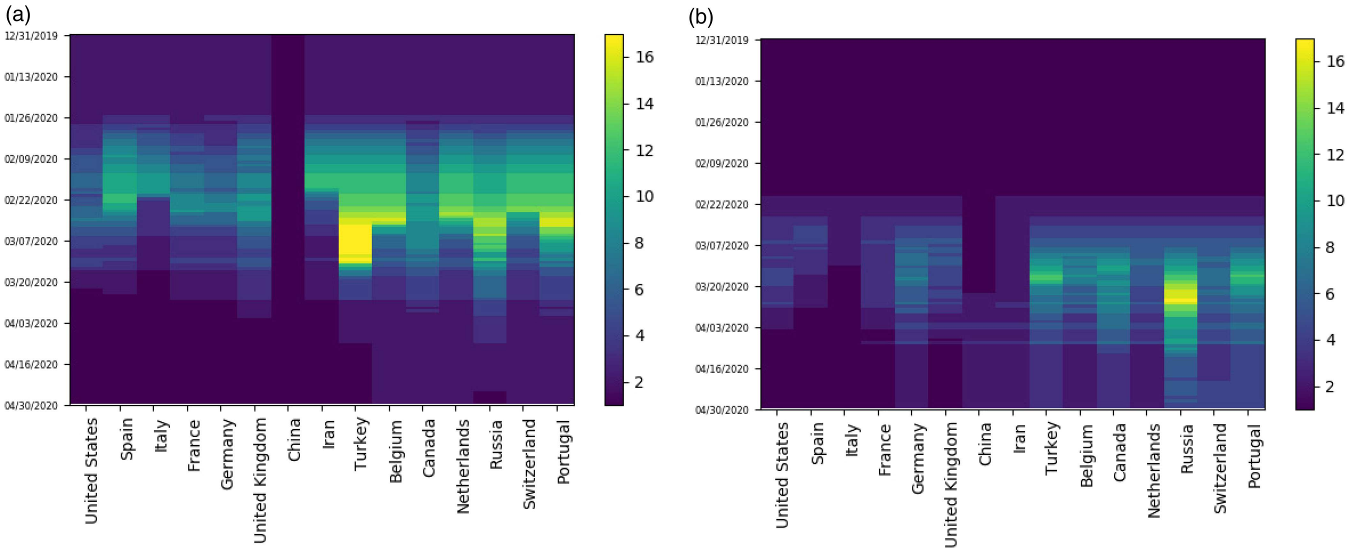

FIG. 2. Heat maps track the changing cluster membership of the 15 most severely impacted countries with respect to their counts of (a) cases and (b) deaths, respectively. Cluster membership, determined by Ckmeans.1d.dp, depicts COVID-19 severity relative to the rest of the world. Clusters are ordered with 1 being the worst impacted at any time. Darker and lighter colors correspond to smaller and greater numbered cluster labels and represent worse and less affected clusters, respectively.

was the first country to join the most severely impacted cluster, with the United States (US), Spain, France, Germany, Iran, and the United Kingdom (UK) all joining by late March. Subsequently, cluster numbers slowly declined until the end of our analysis window and appear to have stabilized. Indeed, the ranking of worst affected countries has largely stabilized in April, producing more consistent clustering results.

In Fig. 3(a), we depict the *cluster evolution dendrogram* for the daily cases, defined in Sec. II B, to study the evolution of the cluster structure. This uses hierarchical clustering to determine similarity between adjacency matrices at different times, which encode the cluster structure on each day. We exclude the first 50 days, in which the cluster structure and associated adjacency matrices are all identical, with only China in its own cluster. The dendrogram identifies two distinct clusters, the larger of which contains two meaningful sub-clusters. All three (sub-)clusters identified are contiguous intervals of dates, 02/19–03/01, 03/02–03/14, and 03/15–04/30. This reveals a marked transition in cluster behavior on 03/02 for the case counts, with a smaller transition on 03/15.

D. Results for time series of deaths

In this section, we implement Ckmeans.1d.dp[21] on daily counts of deaths. The smoothed number of clusters $\hat{k}^{(t)}$, depicted in Fig. 1(a), ranges between $\{1, \ldots, 17\}$. The trajectory for number of death clusters follows a similar pattern to that of cases, with a lag of approximately one month. As with the case counts, our analysis highlights the key takeaways in severely impacted countries. Although we have highlighted a one-month offset in the general evolution of COVID-19 cases and deaths, there are dissimilarities regarding the membership of the worst affected cluster. In mid-March, China moved out of the worst cluster into the second death cluster, demonstrating its relative success in responding to the pandemic. On the other hand, the US, Spain, Italy, France, and the UK have recently moved into the worst cluster, as depicted in Fig. 2(b). Examining cluster constituencies of cases and deaths over time confirms that China has managed potential COVID-19 deaths relatively effectively, while Italy, Spain, the UK, and the US have been ineffective.

In Fig. 3(b), we depict the *cluster evolution dendrogram*, defined in Sec. II B, for the daily deaths. We exclude the first 66 days, in which the cluster structure and associated adjacency matrices are all identical. Figures 3(a) and 3(b) show near-identical hierarchical clustering results for cases and deaths, respectively. Again, two distinct clusters are identified, with two meaningful sub-clusters within the larger cluster. All three (sub-)clusters are again contiguous intervals of dates, 03/06–03/18, 03/19–03/30, and 03/31–04/30. This reveals there is a marked transition in cluster behavior on 03/19 for the death counts, with a smaller transition on 03/31. These are 17 and 16 days later than the corresponding breaks for the case counts.

III. SERIES OFFSET ANALYSIS

In this section, we describe further analysis on two related multivariate time series $x_i^{(t)}$ and $y_i^{(t)}$ valued in a common normed space $\mathfrak{X}$. With the application to COVID-19 in mind, we develop a new method that can determine if there is an appropriate time offset between the two time series. We perform several analyses for this purpose; in Sec. IV, we can subsequently study anomalous individual countries. We adopt our notation from Sec. II, using subscripts X or Y to refer to mathematical objects pertaining to the cases or deaths counts.

First, we have already observed a clear offset in the evolution of $\hat{k}^{(t)}$ for the time series of cases and deaths and wish to determine it precisely. We define the *series evolution offset* with respect to the changing number of clusters as follows: let $f(t) = \hat{k}_X^{(t)}$ and $g(t) = \hat{k}_Y^{(t)}$ be the smoothed number of clusters for each time series. Given an offset δ, let f_δ be the *translated function* defined by $f_\delta(t) = f(t + \delta)$. Let the series evolution offset be the integer δ that minimizes the L^1 distance between functions,

$$\|f_\delta - g\|_{L^1} = \int |f_\delta(t) - g(t)| dt.$$

For our application, this offset is $\delta = 32$, confirming the one-month offset observation in Fig. 1(a).

Next, we determine the offset that minimizes the discrepancy between affinity matrices Aff_X and Aff_Y of the two time series. Given an offset τ, let the *normalized total offset difference* between affinity matrices be defined as follows:

$$\frac{1}{T - |\tau|} \sum_{1 \leq s,t \leq T, t-s=\tau} \|\mathrm{Aff}_X^{(s)} - \mathrm{Aff}_Y^{(t)}\|. \tag{3}$$

We normalize by the number of terms in this sum, which varies with τ, for an appropriate comparison. When $\tau > 0$ we can rewrite this as follows:

$$\frac{1}{T - \tau} \sum_{t=1}^{T-\tau} \|\mathrm{Aff}_X^{(t)} - \mathrm{Aff}_Y^{(t+\tau)}\|.$$

Let the *cluster consistency offset* be the integer τ that minimizes the normalized total offset difference. We can also do the same for the offset with respect to the Gaussian affinity or adjacency matrices G and Adj, respectively. All these matrices are normalized, so a comparison of their values is appropriate. We choose the normalization parameter of the Gaussian affinity matrix in Eq. (2) for this purpose. We standardize notation such that δ always refers to the series evolution offset, while τ refers to the cluster consistency offset.

Results are displayed in Table I, with the optimal affinity matrix offset determined in Fig. 4. To illustrate the flexibility of the method, we choose different start dates for our offset analysis. The first 30 days carry some triviality in the cluster structure, with few cases observed outside China, so it may be desirable to exclude them from the analysis. Fortunately, the optimal offset differs only slightly with different start dates.

The optimal cluster consistency offset is overwhelmingly around 16. This confirms known medical findings[22] indicating time from diagnosis to death has generally been around 17 days. Moreover, this is consistent with the results of Fig. 3, where two breaks in the cluster behavior occurred 17 and 16 days later in the death counts relative to the case counts. This is quite different from the series evolution offset of 32 days. While the cluster consistency offset seeks to align the similarity of case and death counts among individual countries, the series evolution offset seeks to quantify the overall spread of the data as a function of time.

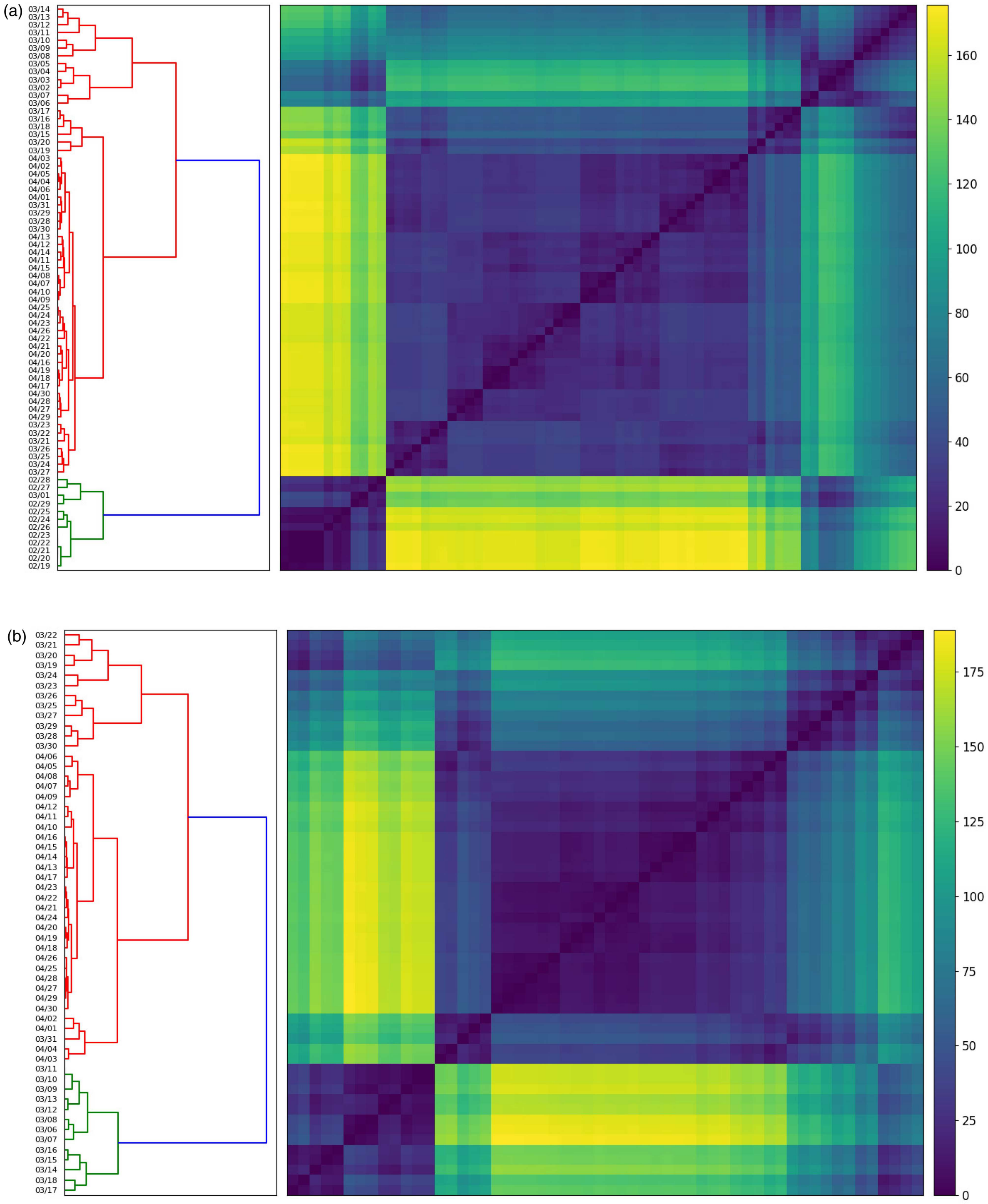

FIG. 3. *Cluster evolution dendrograms*, defined in Section II B for (a) cases and (b) deaths. These apply hierarchical clustering to the distance $d(s,t)$ between adjacency matrices $\mathrm{Adj}^{(t)}$ at varying times t, thereby grouping different *dates* according to the cluster structures at these times. The y-axis excludes the first 50 days for cases and 66 days for deaths, as the cluster structure of counts is trivial before these periods, respectively. Each cluster is an unbroken interval of dates. There is a clear break in the cluster structure between 03/01 and 03/02 for cases, and 03/18 and 03/19 for deaths, with a 17-day difference.

TABLE I. Cluster consistency offset for various adjacency and affinity matrices at different starting dates. These are determined by minimizing the normalized total offset difference in Eq. (3), as well as its analog for Gaussian and adjacency matrices. The parameter *m* is defined in Eq. (2).

	Optimal cases vs deaths offset				
	Gaussian	Gaussian	Gaussian		
Start date	$m = 1$	$m = 2$	$m = 3$	Adj	Aff
12/31/2019	16	16	16	20	16
01/13/2020	12	13	14	20	15
01/21/2020	12	13	14	19	15
01/31/2020	12	13	14	19	15

IV. ANOMALY ANALYSIS

Having identified a suitable offset between two multivariate time series, one can then investigate the existence of any anomalies. In this case, we use $\tau = 16$ as the cluster consistency offset relative to affinity matrices, as depicted in Table I and then perform a closer analysis of the affinity matrices to identify anomalous countries. Let $\mathrm{Inc}^{(t)}$ be the $n \times n$ *inconsistency matrix* defined entry-wise by $\mathrm{Inc}^{(t)}_{ij} = |\mathrm{Aff}^{(t)}_{X,ij} - \mathrm{Aff}^{(t+\tau)}_{Y,ij}|$, where the absolute value of each entry is taken. Smaller entries indicate greater consistency between cases and deaths, while greater entries indicate anomalous (inconsistent) countries. Let the *anomaly score* of any individual country be defined as $a^{(t)}_j = \sum_{j=1}^{n} \mathrm{Inc}^{(t)}_{ij}$. Larger values indicate more anomalous countries and the sequence of anomaly scores can reveal the emergence and disappearance of anomalies over time. Let the *lag-adjusted death rate* for each country be defined as follows:

$$LDR^{(t)}_j = \frac{y^{(t)}_j}{x^{(t-\tau)}_j},\ j = 1, \dots, n;\ t = \tau + 1, \dots, T.$$

These ratios may be orders of magnitude higher than standard reported death rates and are no longer bound between 0 and 1. This measure provides insight into the rate of spread and a country's success in minimizing the number of deaths, conditional on a given number of cases τ days prior.

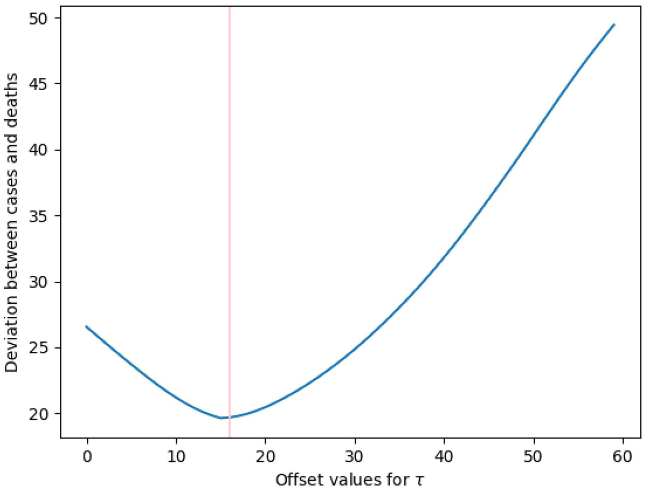

FIG. 4. The normalized total offset difference as a function of the offset τ, defined in Eq. (3). The convex nature of this plot indicates that $\tau = 16$ is a globally optimal value.

In Table II, we depict the results of ordering the ten most anomalous countries, by anomaly score, from 01/28/2020 to 04/27/2020. In Fig. 5, we display the affinity matrices for cases and deaths and the inconsistency matrix for 04/27/2020, with an offset of $\tau = 16$ from Table I. We only include countries that had at least 5000 cases as of 04/30/2020. Anomalies may signify either disproportionately high or low number of deaths relative to the number of cases.

This analysis supports several aspects of the empirically observed spread of COVID-19, identifying the most and least successful countries in the progression of cases to deaths. Early in

TABLE II. The ten most anomalous countries in progression from cases to deaths as defined by their anomaly score from Sec. IV and a lag of $\tau = 16$. AE: United Arab Emirates, AT: Austria, AU: Australia, BD: Bangladesh, BE: Belgium, BY: Belarus, CA: Canada, CN: China, DE: Germany, DO: Dominican Republic, ES: Spain, FR: France, ID: Indonesia, IE: Ireland, IL: Israel, IN: India, IR: Iran, IT: Italy, JP: Japan, KR: South Korea, ME: Mexico, MY: Malaysia, NL: Netherlands, NO: Norway, QA: Qatar, SG: Singapore, SW: Sweden, TR: Turkey, UA: Ukraine, UK: United Kingdom, US: United States, ZA: South Africa.

	Ten most anomalous countries: inconsistency matrix analysis									
Date	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
01/28/2020	US	UK	IT	IL	IE	IR	ID	IN	DE	FR
02/07/2020	US	DO	IT	IL	IE	IR	ID	IN	DE	FR
02/17/2020	SG	JP	KR	AU	MY	US	DE	FR	AE	CA
02/27/2020	IR	SG	MY	IT	AU	US	DE	UK	AE	CA
03/08/2020	IT	IR	SG	MY	DE	AE	CA	JP	ES	US
03/18/2020	ES	SG	IT	IR	AE	UK	NL	FR	US	KR
03/28/2020	QA	ES	TR	UK	SG	KR	AE	BY	US	IT
04/07/2020	QA	SG	KR	UK	CN	UA	NO	ZA	AU	TR
04/17/2020	BD	QA	SG	UK	AU	KR	BE	ZA	AT	FR
04/27/2020	QA	SG	BD	ME	AU	UK	SW	BE	DE	IL

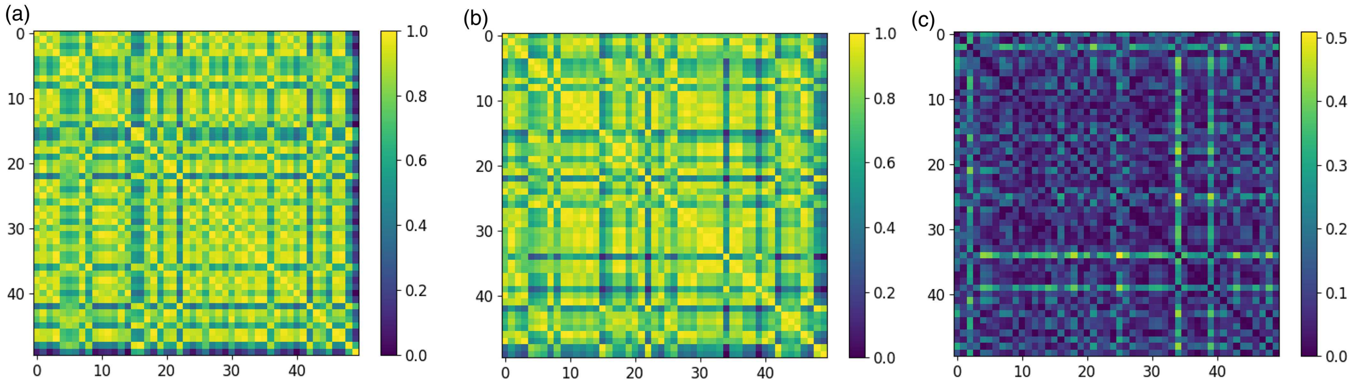

FIG. 5. (a) depicts the affinity matrix for case counts at 04/27/2020, (b) depicts the deaths affinity matrix for 04/11/2020, and (c) depicts the inconsistency matrix with an offset of $\tau = 16$ from Table I. Only countries with greater than 5000 cases at 04/30 are included and ordered alphabetically along the axes. The more prominent the respective row and column in the inconsistency matrix, the more anomalous the country. The three most prominent anomalies in (c) are Qatar, Singapore, and Bangladesh.

the global spread of COVID-19, Iran and Italy were internationally known as countries that were struggling to contain the number of deaths.[23] Table II identifies both as anomalous on 02/27/2020 and 03/08/2020, reflecting their sharp rise in deaths even before other severely impacted countries. On the other hand, Singapore is identified as anomalous during this period due to its relatively small number of deaths. As at 03/07/2020, Singapore had 130 COVID-19 cases and 0 deaths.

A similar trend continued until late March, during which Spain and Italy are identified as the most consistently anomalous countries due to their high death rates. The lag-adjusted death rates for Spain and Italy are 227% and 73.3%, respectively. Indeed, the number of deaths in Spain on 03/28/2020 was more than two times greater than the number of cases 16 days earlier. This confirms the severity of the COVID-19 pandemic: Spain and Italy suffered a large number of deaths within a short window. As of late March, Singapore was still identified as anomalous due to the relatively small number of deaths. Toward the end of our analysis window, Qatar and Australia are also identified as anomalous with low death rates, while the UK is identified as anomalous due to a high death rate. The lag-adjusted death rates for Qatar and Australia as of 04/27/2020 are 0.398% and 1.33%, respectively. The lag-adjusted death rate for the UK is 34.2%.

V. CONCLUSION

In this paper, we introduce a new method of analyzing a multivariate time series via cluster analysis. Unlike typical applications of time series analysis to epidemiology, it is nonparametric; and unlike existing applications of cluster analysis to time series, we produce a dynamically smoothed number of clusters that changes over time. The analysis of case and death counts over time produces two multivariate time series, which we partition into clusters on each day. While previous studies examine fewer countries over shorter time windows,[9,10] we study 208 countries over 4 months. Individual countries' cluster membership tracks their severity of counts relative to the rest of the world, while the number of clusters reflects the overall spread of the data.

The high degree of similarity between the two time series facilitates the identification of anomalous countries in the progression of cases to deaths. We introduce another method herein, using inconsistency matrices and lag-adjusted death rates to highlight the sequential emergence and disappearance of such anomalies over time. These may be used to evaluate a country's effectiveness at handling the pandemic, taking into account an appropriate time offset in mortality due to the disease. Our inconsistency matrices provide a multivariate method with greater generality than the included application. For this reason, they do not identify high or low mortality rates, which are only applicable in a one-dimensional context. The lag-adjusted death rate meets this purpose in our application and any other one-dimensional setting. Last, this methodology is flexible: different metrics between data, clustering methods, and means of learning offset could all be used to study related multivariate time series and identify changing similarity and anomalies.

Our analysis also provides new insights into the spread of COVID-19 across countries and over time. We show a strong similarity between the evolution of case and death counts, identifying a suitable time offset of 16 days for cluster membership between the two time series. This confirms known medical findings,[22] indicating time from diagnosis to death as approximately 17 days. The cluster evolution dendrograms provide further support of a distinct lag between cases and deaths. These dendrograms are highly similar, also up to an offset of 16 days, and demonstrate sharp transition points at 03/02/2020 and 03/19/2020 for cases and deaths, respectively, again with a 17-day difference. These transitions reflect the natural history of the spread of COVID-19 cases and deaths, respectively. On 03/02/2020, numerous countries began to report their first instances of COVID-19 cases, predominantly imported from Iran and Italy. On 03/19/2020, Italy's death toll surpassed that of China.[24] Less pronounced transitions exist on 03/15/2020 and 03/31/2020 for cases and deaths, respectively. Again, a 16-day offset is observed.

On the other hand, the time offset between the evolution of the number of clusters is 32 days. One explanation for the series evolution offset being longer is that there is an additional delay between cluster membership changes with respect to cases and deaths that can be attributed to stresses on a country's healthcare resources. First, the number of cases may increase significantly, placing a country into a different cluster relative to cases and overwhelming its healthcare resources, thereby leading to a greater number of death counts. That is, the progression from elevation in case clusters to death clusters is not necessarily due to a natural progression from infection to death, but involves mediating factors like stresses on hospital capacity. Perhaps the initial wave of patients can be treated with ventilators, but these may quickly run out, causing more deaths from later instances of cases. Regardless, it is an interesting observation that the offset of 32 days in the number of clusters does not minimize the offset in affinity or adjacency matrix norm differences.

This analysis may assist in identifying the characteristics of the most and least successful government strategies for managing COVID-19. In particular, Singapore, Qatar, Australia, and South Korea are four countries whose policies have been most successful in minimizing COVID-19 mortality. Each of these countries provided a substantial amount of easily accessible testing in the early stages of COVID-19 development.[25] Singapore and Australia also closed their borders to travel before a critical mass in total case counts was established and were early to implement strict lockdown procedures.[26]

By contrast, Italy, Spain and the UK are three countries whose policies managed the progression from COVID-19 cases to deaths least effectively. Many argue that lockdown procedures in Italy and Spain, although severe once in place, were implemented too late.[27] Similarly, the UK initially elected not to shut down large gatherings or introduce social distancing measures in an attempt to build herd immunity among the community. Ultimately, however, the UK did implement strict lockdown policies as mortality rates rose.[28]

These findings suggest that the timeliness of various lockdown procedures is perhaps more important than their severity. Countries with easy access to early testing also appear to manage the progression from cases to deaths more effectively. Conversely, countries that struggled to minimize their COVID-19 mortality rate also exhibit some general similarities. First, these countries were slow to implement measures that would restrict people's movements. Second, many of these countries carried an early high case burden, suggesting that mediating factors such as undue stress from finite healthcare resources may contribute to the mortality rate.

Overall, this paper introduces a new method for analyzing multivariate time series individually and in conjunction, thereby providing new insights into the caseload and mortality rate affecting different countries. As the pandemic evolves, it is the objective of emerging research to facilitate timely and appropriate means of producing effective government strategies for minimizing the extensive human, social, and cultural costs of COVID-19.

ACKNOWLEDGMENTS

The authors thank Kerry Chen and Alex Judge for helpful comments and edits.

APPENDIX A: EXISTING CLUSTER THEORY

In this section, we provide an overview of the three clustering algorithms used in the body of the paper: hierarchical clustering, K-means, and its optimal one-dimensional variant Ckmeans.1d.dp. In our most general setup, $x_1, \ldots, x_n$ are elements of a normed space $\mathfrak{X}$.

Hierarchical clustering[19,20] is an iterative clustering technique that does not specify discrete groupings of elements. Rather, it seeks to build a hierarchy of similarity between elements. Hierarchical clustering is either agglomerative, where each element x_i begins in its own cluster and branches between them are successively built, or divisive, where all elements begin in one cluster and are successively split. The results of hierarchical clustering are commonly displayed in *dendrograms*. Hierarchical clustering does not require the choice of a number of clusters k. In this paper, hierarchical clustering is exclusively used to produce the dendrograms of Fig. 3. There, we implement agglomerative clustering.

K-means clustering seeks to minimize an appropriate sum of square distances. With k chosen *a priori*, we investigate all possible partitions (disjoint unions) $C_1 \cup C_2 \cup \cdots \cup C_k$ of $\{x_1, \ldots, x_n\}$. Let z_j be the *centroid* (average) of the subset C_j. One seeks to minimize the sum of square distances within each cluster to its centroid,

$$\sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2.$$

For a normed space with dimension at least 2, it is NP-hard to find the global optimum of this problem. The K-means algorithm[17] is an iterative algorithm that converges quickly and suitably to a locally optimal solution. It is usually sufficient for applications. In this paper, multivariate K-means is exclusively used in Fig. 6.

On the other hand, the K-means optimization problem is efficiently solvable in the one-dimensional case—when x_i are real numbers, they are equipped with an ordering, which considerably simplifies the problem. To cluster n elements of $\mathfrak{X} = \mathbb{R}$ into k clusters requires one to order the elements and then determine $k - 1$ breaks in the ordering. This is far less computationally intensive than the higher-dimensional analog. Ckmeans.1d.dp[21] is a dynamic programing algorithm that guarantees optimal clustering in one dimension, choosing k *a priori*.

How to best choose the number of clusters k for the K-means algorithm is a difficult problem. Different methods for estimating k may produce considerably differing results. In this paper, we draw upon six methods to determine the appropriate number of clusters before implementing K-means, in both the one and higher-dimensional cases. These methods are well-known: Ptbiserial index,[30] silhouette score,[31] KL index,[32] C index,[33] McClain–Rao index,[34] and Dunn index.[35] We have chosen these methods based upon consultation with the literature and our own experiments. However, our methodology is flexible, and any combination of existing methods may be used. For one-dimensional data, it is often regarded as unsuitable to use multivariate clustering methods, as optimal alternatives exist. Since we study one-dimensional data in this paper, it is necessary to use these methods to choose the number k before implementation of Ckmeans.1d.dp.

In the body of the paper, we choose the smoothed number of clusters $\hat{k}^{(t)}$, depicted in Fig. 1, by applying exponential smoothing to the average of the six choices of cluster number listed above. We then apply Ckmeans.1d.dp to divide daily counts of data into $\hat{k}^{(t)}$

clusters. This determines our results in Fig. 2. In Fig. 6, we display analogous results for three-day rolling counts, clustering the corresponding elements of $\mathbb{R}^3$ using standard K-means. The results are highly similar.

APPENDIX B: THREE-DAY ROLLING COUNTS

In this section, we briefly show the applicability of our method to higher-dimensional data. We present two multivariate time series of the cumulative three-day rolling counts of cases and deaths on a country by country basis. We order the countries by alphabetical order and denote these three-day rolling counts by $\tilde{\mathbf{x}}_i^{(t)}, \tilde{\mathbf{y}}_i^{(t)} \in \mathbb{R}^3, i = 1, \dots, 208$. We proceed exactly as in Sec. II, applying standard K-means instead of Ckmeans.1d.dp. In Fig. 6, we depict the same countries' changing cluster membership as were depicted in Fig. 2. The similarity shows the robustness and generality of our method.

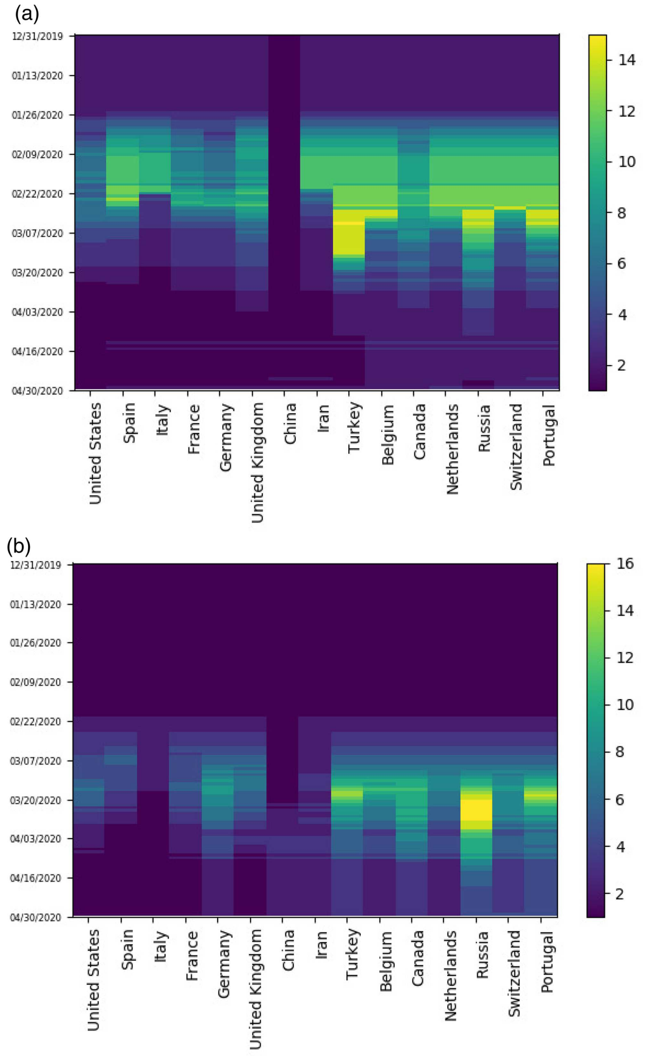

FIG. 6. Heat maps track the changing cluster membership of the 15 most severely impacted countries with respect to their three-day rolling counts of (a) cases and (b) deaths, respectively. Cluster membership, determined by K-means, depicts COVID-19 severity relative to the rest of the world. There is a strong similarity relative to Fig. 2.

APPENDIX C: ALGORITHMIC DESCRIPTION OF METHODOLOGY

In this section, we provide an algorithmic presentation of the computational steps taken for the analysis of an individual multivariate time series, described in Secs. II A and II B.

Algorithm Cluster-based evolution analysis

```
Given: a multivariate time series $x_i^{(t)} \in \mathbb{R}_{\geq 0}$
Data preprocessing:
**if** $x_i^{(t)} = 0$ or NaN **then**
    $x_i^{(t)} = 1$
Data transformation:
$x_i^{(t)} = \log x_i^{(t)}$
**for** $t = 1$ to $T$ **do**
    Compute $k_1^{(t)}, \dots, k_6^{(t)}$
    $k_1^t = \text{Ptbiserial}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_2^{(t)} = \text{Silhouette score}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_3^{(t)} = \text{KL index}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_4^{(t)} = \text{C index}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_5^{(t)} = \text{McClain-Rao index}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_6^{(t)} = \text{Dunn index}\left( (x_{i:1,N}^{(t)})_{1\leq i\leq N} \right)$
    $k_{av}^{(t)} = \frac{1}{6} \sum k_i^{(t)}$
**End for**
$\hat{k}^{(t)}$ = simple exponential smoothing($k_{av}^{(t)}$)
**for** $t = 1$ to $T$ **do**
    Ckmeans.1d.dp sort($x_i^{(t)}$) into $\hat{k}^{(t)}$ clusters
    Record and sort cluster labels.
    Let $\text{Adj}_{ij}^{(t)}$ be adjacency matrix.
    **if** $x_i^{(t)}$ and $x_j^{(t)}$ are in same cluster **then**
        $\text{Adj}_{ij}^{(t)} = 1$
    **else**
        $\text{Adj}_{ij}^{(t)} = 0$
**End for**
Compute affinity matrix, $\text{Aff}_{ij}^{(t)} = 1 - \frac{D_{ij}^{(t)}}{\max D^{(t)}}$
Compute Gaussian matrix $G_{ij}^{(t)} = \exp\left( \frac{-m^2 \left( D_{ij}^{(t)} \right)^2}{2(\max D^{(t)})^2} \right)$
Compute $d(s,t) = \|\text{Adj}^{(t)} - \text{Adj}^{(s)}\|$.
**do** hierarchical clustering on $d(s,t), 1 \leq s,t \leq T$.
```

TABLE III. Mathematical objects and definitions.

Mathematical objects glossary	
Object	Description
$D^{(t)}$	Distance matrix between log counts
$\text{Aff}^{(t)}$	Standard affinity matrix
$G^{(t)}$	Gaussian affinity matrix
$k_{av}^{(t)}$	Unsmoothed number of clusters obtained as average of six methods
$\hat{k}^{(t)}$	Smoothed number of clusters
$\text{Adj}^{(t)}$	Adjacency matrix coding cluster outputs for $\hat{k}^{(t)}$ clusters
$d(s, t)$	Frobenius distance between adjacency matrix of various dates
δ	Series evolution offset with respect to number of clusters
τ	Cluster consistency offset with respect to cluster membership
$\text{Inc}^{(t)}$	Lag-adjusted inconsistency matrix
$a_j^{(t)}$	Anomaly score of country j
$LDR_j^{(t)}$	Lag-adjusted death rate of country j
$\|f_\delta - g\|_{L^1}$	L^1 norm between functions

APPENDIX D: GLOSSARY OF MATHEMATICAL OBJECTS

In this brief section, we include a glossary of mathematical objects presented in the paper and their respective definitions in Table III.

DATA AVAILABILITY

The data that support the findings of this study are openly available at Our World in Data, Ref. 29.

REFERENCES

[1] H. W. Hethcote, "The mathematics of infectious diseases," SIAM Rev. **42**, 599–653 (2000).

[2] G. Chowell, L. Sattenspiel, S. Bansal, and C. Viboud, "Mathematical models to characterize early epidemic growth: A review," Phys. Life Rev. **18**, 66–97 (2016).

[3] A. Vazquez, "Polynomial growth in branching processes with diverging reproductive number," Phys. Rev. Lett. **96**, 038702 (2006).

[4] R. Moeckel and B. Murray, "Measuring the distance between time series," Physica D **102**, 187–194 (1997).

[5] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," Ann. Stat. **35**, 2769–2794 (2007).

[6] C. F. Mendes and M. W. Beims, "Distance correlation detecting Lyapunov instabilities, noise-induced escape times and mixing," Phys. A Stat. Mech. Appl. **512**, 721–730 (2018).

[7] C. F. O. Mendes, R. M. da Silva, and M. W. Beims, "Decay of the distance autocorrelation and Lyapunov exponents," Phys. Rev. E **99**, 062206 (2019).

[8] K. Shang, B. Yang, J. M. Moore, Q. Ji, and M. Small, "Growing networks with communities: A distributive link model," Chaos **30**, 041101 (2020).

[9] C. Manchein, E. L. Brugnago, R. M. da Silva, C. F. O. Mendes, and M. W. Beims, "Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategies," Chaos **30**, 041102 (2020).

[10] J. A. T. Machado and A. M. Lopes, "Rare and extreme events: The case of COVID-19 pandemic," Nonlinear Dyn. **100**, 2953–2972 (2020).

[11] T. Cassetti, F. L. Rosa, L. Rossi, D. D'Alò, and F. Stracci, "Cancer incidence in men: A cluster analysis of spatial patterns," BMC Cancer **8**, 344 (2008).

[12] H. Alashwal, M. E. Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The application of unsupervised clustering methods to Alzheimer's disease," Front. Comput. Neurosci. **13**, 31 (2019).

[13] X. Xiao, A. J. van Hoek, M. G. Kenward, A. Melegaro, and M. Jit, "Clustering of contacts relevant to the spread of infectious disease," Epidemics **17**, 1–9 (2016).

[14] R. M. McCloskey and A. F. Y. Poon, "A model-based clustering method to detect infectious disease transmission outbreaks from sequence variation," PLoS Comput. Biol. **13**, e1005868 (2017).

[15] H. Muradi, A. Bustamam, and D. Lestari, "Application of hierarchical clustering ordered partitioning and collapsing hybrid in Ebola virus phylogenetic analysis," in *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (IEEE, 2015).

[16] R. Rizzi, P. Mahata, L. Mathieson, and P. Moscato, "Hierarchical clustering using the arithmetic-harmonic cut: Complexity and experiments," PLoS One **5**, e14067 (2010).

[17] S. Lloyd, "Least squares quantization in PCM," IEEE Trans. Inform. Theory **28**, 129–137 (1982).

[18] U. von Luxburg, "A tutorial on spectral clustering," Stat. Comput. **17**, 395–416 (2007).

[19] J. H. Ward, "Hierarchical grouping to optimize an objective function," J. Am. Stat. Assoc. **58**, 236–244 (1963).

[20] G. J. Szekely and M. L. Rizzo, "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method," J. Classif. **22**, 151–183 (2005).

[21] H. Wang and M. Song, "Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming," R J. **3**, 29–33 (2011).

[22] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study," Lancet **395**, 1054–1062 (2020).

[23] K. Mayberry, F. Najjar, and V. Pietromarchi, "Italy coronavirus death toll to 107, 3089 cases: Live updates," see https://www.aljazeera.com/news/2020/03/italy-death-toll-jumps-global-o utbreak-deepens-live-updates-200303233420584.html, Al Jazeera (last accessed March 5, 2020).

[24] C. Kantis, S. Kiernan, and J. Bardi, "Updated: Timeline of the coronavirus," see https://www.thinkglobalhealth.org/article/updated-timeline-coronavirus, Think Global Health (last accessed April 25, 2020).

[25] J. McCurry, "Test, trace, contain: how South Korea flattened its coronavirus curve," see https://www.theguardian.com/world/2020/apr/23/test-trace-contain-how-s outh-korea-flattened-its-coronavirus-curve (2020), The Guardian (last accessed April 23, 2020).

[26] S. McDonell, "Coronavirus: US and Australia close borders to Chinese arrivals," see https://www.bbc.com/news/world-51338899 (2020), BBC (last accessed February 2, 2020).

[27] A. McCann, N. Popovich, and J. Wu, "Italy's virus shutdown came too late. What happens now?," see https://www.nytimes.com/interactive/2020/04/05/world/europe/italy-coro navirus-lockdown-reopen.html, The New York Times (last accessed April 5, 2020).

[28] O. Matthews, "Britain drops its go-it-alone approach to coronavirus," https://foreignpolicy.com/2020/03/17/britain-uk-coronavirus-response-j ohnson-drops-go-it-alone/ (2020), Foreign Policy (last accessed March 19, 2020).

[29] "Our world in data," see https://ourworldindata.org/coronavirus-source-data (last accessed April 30, 2020).

[30] G. W. Milligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," Psychometrika **45**, 325–342 (1980).

[31] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math. **20**, 53–65 (1987).

[32] W. J. Krzanowski and Y. T. Lai, "A criterion for determining the number of groups in a dataset using sum-of-squares clustering," Biometrics **44**, 23–34 (1988).

[33] L. J. Hubert and J. R. Levin, "A general statistical framework for assessing categorical clustering in free recall," Psychol. Bull. **83**, 1072–1080 (1976).

[34] J. O. McClain and V. R. Rao, "CLUSTISZ: A program to test for the quality of clustering of a set of objects," J. Mark. Res. **12**, 456–460 (1975).

[35] J. C. Dunn, "Well-separated clusters and optimal fuzzy partitions," J. Cyber. **4**, 95–104 (1974).