Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Physica D 432 (2022) 133158

Contents lists available at ScienceDirect

Physica D

journal homepage: www.elsevier.com/locate/physd

Comparing the dynamics of COVID-19 infection and mortality in the United States, India, and Brazil

Nick James [a,*], Max Menzies [b], Howard Bondell [a]

[a] School of Mathematics and Statistics, University of Melbourne, Victoria, Australia
[b] Beijing Institute of Mathematical Sciences and Applications, Tsinghua University, Beijing, China

ARTICLE INFO

Article history:
Received 18 August 2021
Received in revised form 6 December 2021
Accepted 8 January 2022
Available online 19 January 2022
Communicated by Víctor M. Pérez-García

Keywords:
COVID-19
Time series analysis
Population dynamics
Nonlinear dynamics
Federal states

ABSTRACT

This paper compares and contrasts the spread and impact of COVID-19 in the three countries most heavily impacted by the pandemic: the United States (US), India and Brazil. All three of these countries have a federal structure, in which the individual states have largely determined the response to the pandemic. Thus, we perform an extensive analysis of the individual states of these three countries to determine patterns of similarity within each. First, we analyse structural similarity and anomalies in the trajectories of cases and deaths as multivariate time series. Next, we study the lengths of the different waves of the virus outbreaks across the three countries and their states. Finally, we investigate suitable time offsets between cases and deaths as a function of the distinct outbreak waves. In all these analyses, we consistently reveal more characteristically distinct behaviour between US and Indian states, while Brazilian states exhibit less structure in their wave behaviour and changing progression between cases and deaths.

1. Introduction

The United States (US), India and Brazil have each been severely impacted by COVID-19 and lead the world in both case and death counts. While the three countries have quite different cultures and levels of economic and technological development, they each have a similar federation structure, with governing responsibilities divided between federal and state governments. In all three countries, government responses have consistently differed between constituent states and over time [1–3], yielding different levels of virus transmission and impact on communities. Thus, a careful analysis of the most and least successful states is of great relevance to a response to the ongoing threat of COVID-19. Moreover, it is worthwhile to compare and contrast the state-by-state behaviours of the pandemic between the three countries as a whole.

In the US, India and Brazil, as well as throughout the world, the scientific response to COVID-19 has been as multifaceted and as significant as the government response. Medical researchers have uncovered numerous means of treating infections [4–7], culminating in the production of vaccines [8,9]. Outside the medical field, analytical approaches to model and study the virus and its impact have been broad. First, many models based on existing mathematical models, such as the Susceptible–Infected–Recovered (SIR) model and the reproductive ratio R_0, have been proposed and systematically collated by researchers [10,11]. These have been utilised for various purposes, including diagnosis and prognosis of COVID-19 patients, studies of the efficacy of medications, and vaccine development. Next, nonlinear dynamics researchers have proposed several sophisticated extensions to the classical predictive SIR model, including analytic techniques to find explicit solutions [12,13], modifications to the SIR model with additional variables [14–19], incorporation of Hamiltonian dynamics [20] or network models [21], and a closer analysis of uncertainty in the SIR equations [22]. Other mathematical approaches to prediction and analysis include power-law models [23–25], forecasting models [26], fractal approaches [27–29], neural networks [30], Bayesian methods [31], distance analysis [32], network models [33–36], analyses of the dynamics of transmission and contact [37,38], clustering [39,40] and many others [41–45]. Finally, numerous articles have been devoted to understanding the spatial components of the virus' spread, in numerous countries [46–49].

We have a different motivation and approach relative to the aforementioned work. Numerous works have studied trends in COVID-19 prevalence on a country-by-country basis [50] or state-by-state basis, frequently within the US or Brazil [2,51]. However, we are unaware of any work to consider more than one federation of states at once. We were motivated to compare the US, India and Brazil for several reasons. First, these are the three countries most impacted by COVID-19, both in case and death counts. Secondly, the level of human development varies drastically from country to country, but less so within each federation

* Corresponding author.
E-mail address: nick.james@unimelb.edu.au (N. James).

https://doi.org/10.1016/j.physd.2022.133158
0167-2789/

of states. Third, during the COVID-19 pandemic, international movement drastically decreased, leaving such large federal states almost as self-contained regions in which COVID-19 spread independently from what was occurring in other countries. Thus, tracking the heterogeneity of COVID-19 prevalence and behaviour within and between federations could be used to distinguish the effects of policies at the state and federal level. For example, countries whose federal government had less of a policy role could see more heterogeneity of behaviours with states, if states implemented drastically differing policies.

This work could assist various researchers in different fields. Analysing and predicting the spread of COVID-19 is consistently challenging due to the inability to establish true control groups; indeed, it is practically impossible to split entire countries into different regions where certain mitigation measures are or are not implemented. By comparing states within and between different federations, policy researchers can approximate the existence of control groups, and investigate which socioeconomic features and interventions were associated with better and worse outcomes. For policymakers, a comparison of different states within each federation can provide opportunities for state governments to learn from each other's triumphs and setbacks. Across the three countries, this analysis could reveal relationships between COVID-19 spread and the intervention of the national government or the underlying level of economic development.

This paper is structured in such a way as to thoroughly investigate numerous aspects of the spread and human cost of COVID-19 in the three federations. First, Section 2 investigates the structural similarity and anomalies in the trajectories of cases, deaths and rolling mortality rate on a state-by-state basis in the three countries. We explore commonalities in virus behaviour within the three countries as well as the extent of heterogeneity across each country as a whole. Next, Section 3 performs a closer analysis of a highly significant aspect of COVID-19 epidemiology: differing waves of the outbreak. Using a newly introduced turning point algorithm and distance between finite sets, we perform clustering on all the individual states of the US, India and Brazil to identify characteristic wave behaviours across the entire collection. Finally, Section 4 draws upon the previous two sections to address a highly pertinent metric — the average progression between cases and deaths. This paper introduces a variety of novel optimisation methods to estimate this, and takes a new approach, separating this feature according to the mathematically determined waves of the pandemic. We employ five different optimisation methods, each of which uses state-by-state data [52–54], to estimate an appropriate offset between case and death time series for the US, India and Brazil as a whole. This allows us to track the changing nature of COVID-19 mortality among the different waves of the pandemic. We summarise all our findings and insights in Section 5.

In addition to the above motivation and specific questions we study, the methodologies used in this paper have applicability well beyond the COVID-19 pandemic, and could be used in any setting of multivariate time series. In particular, Section 2 presents a new approach to carefully quantify the extent of heterogeneity in a multivariate time series (or in other spaces more generally) that handle the existence of outlier elements well, while Section 4 could be used to study various other time series where lagging is to be expected. Given the fourth wave of COVID-19 that Europe is currently facing, scientists should seek to learn from the countries most severely impacted by COVID-19, and their prior waves of COVID-19 cases. This manuscript provides computational tools and findings that would be of great relevance to this audience.

2. Trajectory analysis, structural similarity and anomaly detection

In this section, we explore the similarity and structure between case, death and rolling mortality time series for the US, India and Brazil. Our data spans 26 Feb 2020 to 23 May 2021, a period of $T = 452$ days. For each country, let the multivariate time series of new COVID-19 cases and deaths be $x_i(t)$ and $y_i(t)$, where $t = 1, \ldots, T$ indexes the days and $i = 1, \ldots, N$ indexes states under consideration. Throughout this manuscript, we will examine either one country at a time, with $N = 51$ states (including the District of Columbia) for the US, $N = 36$ states (including union territories) for India, $N = 27$ states (including the Federal District) for Brazil, or the entire collection of individual states, with $N = 114$.

In addition, we define a 30-day rolling mortality rate for each state as follows:

$$r_i(t) = \frac{\sum_{j=t-29}^{t} y_i(j)}{\sum_{j=t-29}^{t} x_i(j)}, \quad t = 30, \ldots, T. \tag{1}$$

We wish to examine the three aforementioned multivariate time series to determine the structure and degree of heterogeneity within each country's states and collectively, between all countries' underlying states. To a case time series $x_i(t), t = 1, \ldots, T$ we associate the following probability distribution:

$$f_i = \frac{1}{\sum_{s=1}^{T} x_i(s)} \sum_{t=1}^{T} x_i(t)\delta_t, \tag{2}$$

where δ_t is the Dirac delta distribution at t. That is, f_i is a distribution that apportions to day t the weight of the new cases observed on that day as a proportion of the total cases across the whole period. Then, we define

$$M_{ij}^{C} = W_1(f_i, f_j), \tag{3}$$

where W_1 is the L^1-Wasserstein metric [55] between distributions on $\mathbb{R}$. Analogously, we associate distributions g_i and h_i to death and mortality time series y_i and r_i, respectively. We define *trajectory distance matrices* between state trajectories for deaths and mortality analogously as follows:

$$M_{ij}^{D} = W_1(g_i, g_j); \tag{4}$$

$$M_{ij}^{R} = W_1(h_i, h_j). \tag{5}$$

This distance has several advantageous properties over previously used discrepancy measures between normalised trajectories. Previous work [51] has used the L^1 norm and metric between normalised trajectories, defined as follows:

$$\|x_i\|_1 = \sum_{t=1}^{T} x_i(t) \tag{6}$$

$$\mathbf{v}_i = \frac{x_i}{\|x_i\|_1} \tag{7}$$

$$d_{ij} = \|\mathbf{v}_i - \mathbf{v}_j\|_1 \tag{8}$$

This treats each time series $x_i(t)$ as a vector in $\mathbb{R}^T$, normalises by its L^1 norm, and compares these normalised vectors with the L^1 metric [56]. This distance is suitable in most instances but has some undesirable properties when quantifying discrepancy between noisy time series. Specifically, this L^1 distance d_{ij} has maximal possible value equal to 2 when $x_i(t)$ and $x_j(t)$ have disjoint support. Practically, this would mean that two states' trajectories would receive a large L^1 discrepancy measure if the cases were simply reported to fall on different days. For example, if state i and state j had broadly similar trends in cases, but in

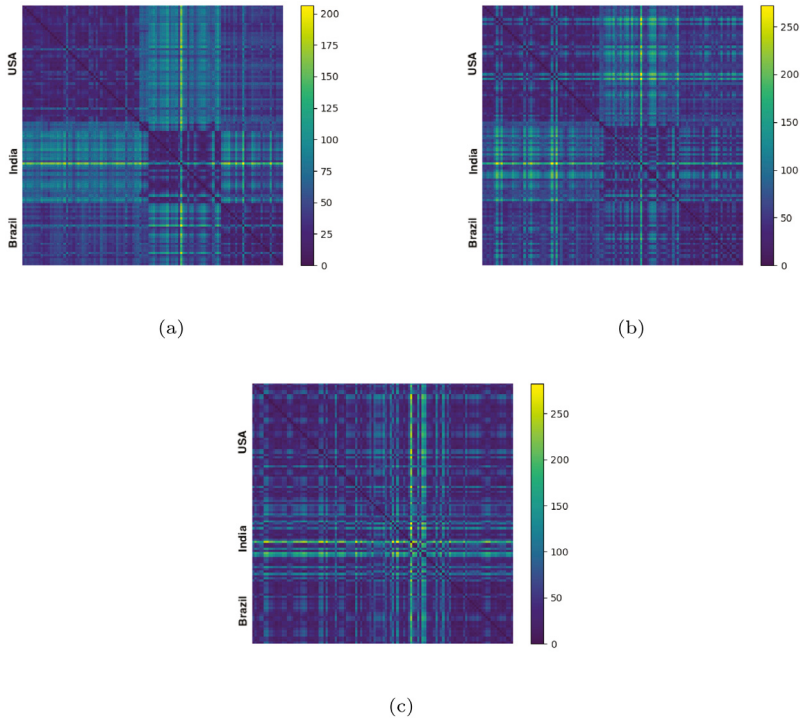


Fig. 1. Trajectory distance matrices, as defined in Section 2, with respect to (a) cases, (b) deaths and (c) mortality rate time series. Each matrix is computed using the entire collection of 114 states and ordered with US states first, then Indian states, then Brazil. Darker values indicate smaller entries of the matrix, signifying greater similarity between states. India exhibits particular heterogeneity between mortality rates.

state i cases were reported more on Mondays and Wednesdays while state j reported more on Tuesdays and Thursdays, then the L^1 distance measure would be larger than their similarity. Smoothing and 7-day averaging can resolve some of these issues, but the Wasserstein metric ameliorates this issue even more, as it is robust to small translations of distributions. That is, if f is a distribution and $f_\delta(x) = f(x+\delta)$, then $W_1(f, f_\delta) = |\delta|$, as shown in [57]. This means the Wasserstein metric assigns a low value in the case that states i and j have similar trajectories where cases just fall on nearby but distinct days.

We will examine the matrices defined above (M^C, M^D and M^R) for each individual country (with $N = 51$ for the US, 36 for India, 27 for Brazil) as well as the entire collection of states, with $N = 114$. In Fig. 1, we display the matrices M^C, M^D, and M^R each for the totality of the collection. In Table 1, we record the L^1-norms $\|M^C\|$, $\|M^D\|$, and $\|M^R\|$ each restricted to one of the three federations. For example, for the US, M^C, M^D and M^R are 51×51 matrices, whereas they are 36×36 matrices for India. For an $N \times N$ matrix A, we define its norm by

$$\|A\| = \frac{1}{N(N-1)} \sum_{i,j=1}^{N} |A_{ij}|. \tag{9}$$

This calculates a total magnitude of the matrix, appropriately normalised for the number of non-zero elements. For our distance matrices M^C, M^D, and M^R, these norms reflect the heterogeneity among trajectories within each country. As the Wasserstein distance is taken between appropriately normalised distributions, it is possible to compare between case, death and mortality time series. Due to the normalisation coefficient, it is possible to compare this between different countries.

Table 1 reveals that India exhibits the highest heterogeneity between states regarding all three behaviours, with norms of 40.09, 55.08 and 76.45 for cases, deaths and mortality, respectively. For case trajectories, the US and Brazil have similar levels of total homogeneity. For deaths and mortality trajectories, however, Brazil's norms of 35.33 and 29.67 are rather less than the US' scores of 43.70 and 43.03. This highlights the relative homogeneity in death and mortality trajectories among Brazilian states.

Next, we wish to further examine the heterogeneity between states of each country, as well as identify the presence of any outlier states that may be influencing the total norms recorded in Table 1. Given each country's trajectory matrix M (with respect to cases, deaths or mortality rates), we perform the following procedure to sequentially identify the most anomalous state, remove it, and compute the resulting norm of the reduced collection. This is described in Algorithm 1.

Algorithm 1 Anomalous trajectory identification

1: Input: a distance matrix M between all states for a candidate country.
2: Initialise empty lists for norm scores and state anomaly ranking.
3: Set matrix M to current iteration, M^c.
4: **for** $r = 0$ to $N-1$ **do**
5: Compute number of rows, $m^c = M^c - r$, in current distance matrix, M^c.
6: Generate current norm of matrix M^c, $\nu^c = \frac{1}{m^c(m^c-1)} \sum_{i,j=1}^{m^c} |M^c_{ij}|$.
7: Append ν^c to norm scores list.
8: **for** $l = 1$ to m^c **do**
9: Compute $a_l = \sum_{i=1}^{m^c} |M^c_{il}|$.
10: Let $a_k = \max_l\{a_l\}$
11: Append name of state k to state anomaly ranking list.
12: Update matrix M^c by removing kth column and row corresponding to state with highest anomaly score, a_k.
13: Generate norm scores trajectory and state anomaly ranking.

In Fig. 2, we display the sequence of norm scores ν^c for each matrix M^C, M^D, and M^R for the US, India and Brazil. By removing the greatest a_k in each step of the algorithm, this

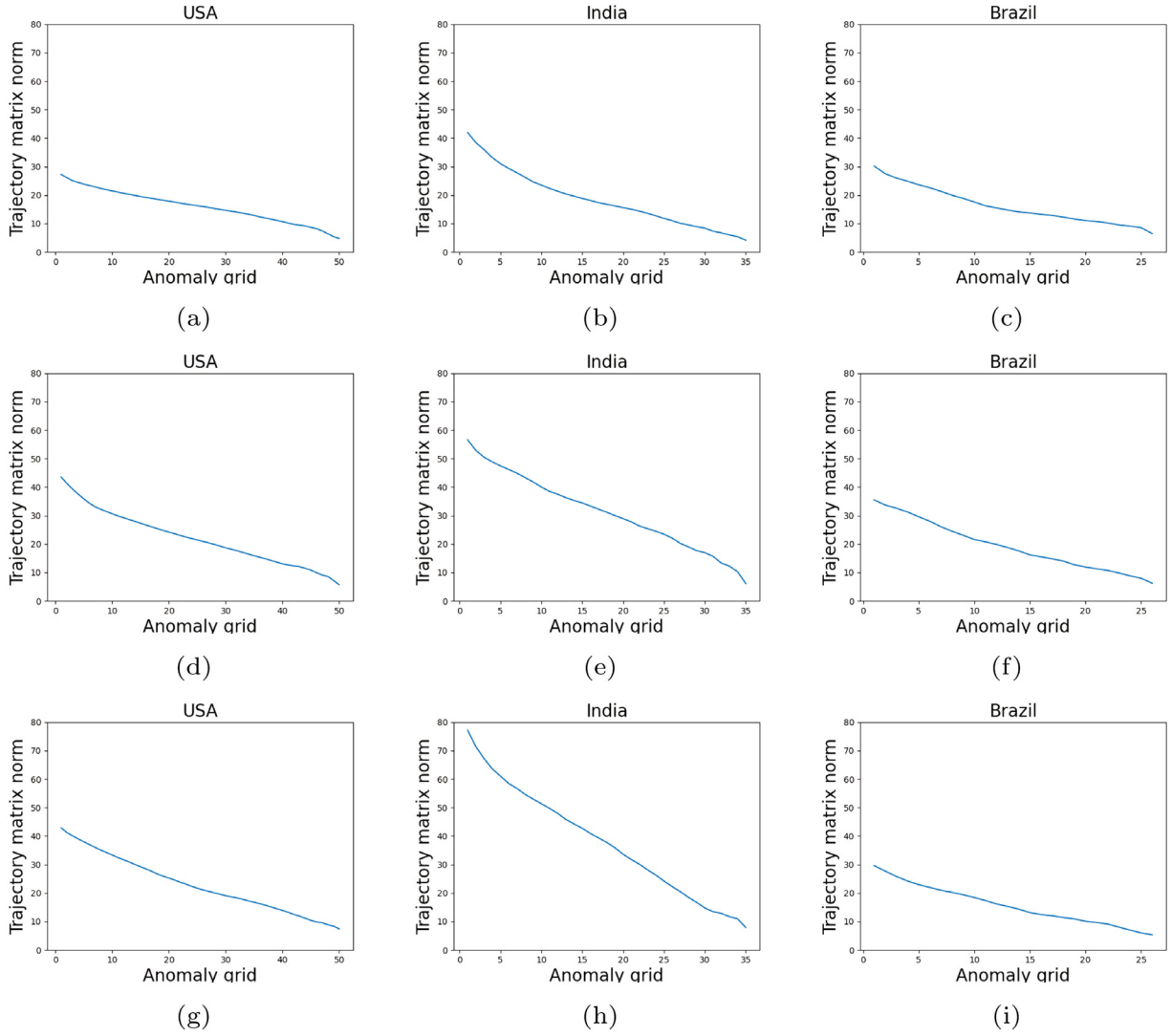


Fig. 2. Sequences of decreasing matrix norms ν^c, as determined in Algorithm 1, for (a) US cases (b) Indian cases (c) Brazilian cases (d) US deaths (e) Indian deaths (f) Brazilian deaths (g) US mortality rates (h) Indian mortality (i) Brazilian mortality. These norms are obtained sequentially after removing the most anomalous country at each step. India exhibits sharper drops at the start, indicating a small collection of highly anomalous states, particularly for mortality rates.

Table 1
Normalised matrix norms $\|M^C\|$, $\|M^D\|$, and $\|M^R\|$, as defined in Section 2, for each of the three countries with respect to case, death and mortality time series. The higher values for India indicate greater heterogeneity between its states.

Trajectory distance matrix norms			
Country	Cases	Deaths	Mortality rate
US	27.39	43.70	43.03
India	40.09	55.08	76.45
Brazil	30.30	35.33	29.67

sequence of norm scores is necessarily decreasing. As all norms are appropriately normalised, we may compare these decreasing sequences between all our different countries and time series. Several insights can be gained from these figures. First, India consistently produces the largest anomaly score for all three attributes. This can be seen by the magnitude of the decreasing trend for India throughout the plots. This is consistent with the analysis in Table 1, but ensures that it is not due simply to the presence of a small number of outlier states. Second, relative to cases and deaths, mortality rate trajectories are significantly more dissimilar in the case of India. For the US and Brazil, there is greater uniformity in anomaly trajectories among each of the three attributes. When examining the nine sequential norm trajectories, it is pertinent to look for sharp drops, which would indicate that a particular state accounts for a disproportionate amount of heterogeneity. This effect is seen in the Indian mortality rate norms (Fig. 2(e)) and to a lesser extent in the cases and deaths norms, (Figs. 2(b) and Fig. 2(e), respectively).

Table 2 records the five most anomalous states in each country with respect to cases, deaths and mortality rates, as determined by Algorithm 1, and also reveals several insights. In the US, there is a pronounced geographic trend in all three attributes' anomaly trajectories. Northeastern states New York, New Jersey, Connecticut and Vermont are identified as anomalous in at least two attributes' trajectories each. Several other Northeastern states appear, such as New Hampshire, Maine, Massachusetts and DC. In addition, there is substantial consistency in the states exhibiting anomalous behaviours in cases, deaths and mortality. In India, the state Lakshadweep is the most anomalous in cases, deaths and mortality, but otherwise relatively less repetition is observed among the most anomalous states. Lakshadweep's status as an anomaly can also explain the sharp drops observed for India in Fig. 2, but not for the US or Brazil. Brazil exhibits even greater variability in the most anomalous states than the US or India, with

Table 2
The five most anomalous states in each country with respect to case, death and mortality rate time series, as determined by Algorithm 1.

Country	Cases	Deaths	Mortality
US_1	Vermont	New York	Oklahoma
US_2	Maine	New Jersey	Vermont
US_3	New Hampshire	Connecticut	New Jersey
US_4	New York	DC	Connecticut
US_5	Michigan	Massachusetts	New York
$India_1$	Lakshadweep	Lakshadweep	Lakshadweep
$India_2$	Andaman & Nicobar Islands	Tripura	Mizoram
$India_3$	Tripura	Andhra Pradesh	Nagaland
$India_4$	Arunachal Pradesh	Odisha	Himachal Pradesh
$India_5$	Assam	Dadra and Nagar Haveli	Gujarat
$Brazil_1$	Maranhão	Pernambuco	Pernambuco
$Brazil_2$	Roraima	Paraná	Piauí
$Brazil_3$	Amapá	Minas Gerais	Ceará
$Brazil_4$	Distrito Federal	Rio Grande do Sul	Distrito Federal
$Brazil_5$	Minas Gerais	Santa Catarina	Paraíba

little consistency in the states exhibiting anomalous behaviours among cases, deaths and mortality.

3. Wave behaviour analysis

In this section, we investigate one of the most significant aspects of the spread of COVID-19, the tendency for the virus to exhibit multiple distinct waves of prevalence. As in the last section, we analyse either each country on a state-by-state basis (with $N = 51$, 36, and 27 states) or the entire collection of states across the three countries together ($N = 114$ states). To each state, we apply a newly introduced turning point algorithm [51] to identify non-trivial local maxima (peaks) and minima (troughs) in the new case time series.

We first apply a *Savitzky–Golay filter* to each new case time series $x_i(t)$ to generate a smoothed collection of time series $\hat{x}_i(t)$, $t = 1, \ldots, T$ and $i = 1, \ldots, N$. We then apply a two-stage turning point algorithm, detailed in the Appendix, to generate non-empty sets P_i and T_i of non-trivial local maxima (peaks) and local minima (troughs), respectively. These turning points alternate between a trough and peak, beginning with a trough at $t = 1$, when there are no cases.

Next, we use an appropriate distance measure to quantify the similarity between two sets of turning points. We apply the semi-metric first introduced in [57]. Given two non-empty finite sets A, B, this is defined as

$$D(A, B) = \frac{1}{2}\left(\frac{\sum_{b\in B} d(b, A)}{|B|} + \frac{\sum_{a\in A} d(a, B)}{|A|}\right), \tag{10}$$

where $d(b, A)$ is the minimal distance from $b \in B$ to the set A. The distance measure $D(A, B)$ is symmetric, non-negative, and zero if and only if $A = B$. We then define $N \times N$ turning point distance matrices M^{TP} by

$$M^{TP}_{ij} = D(P_i, P_j) + D(T_i, T_j). \tag{11}$$

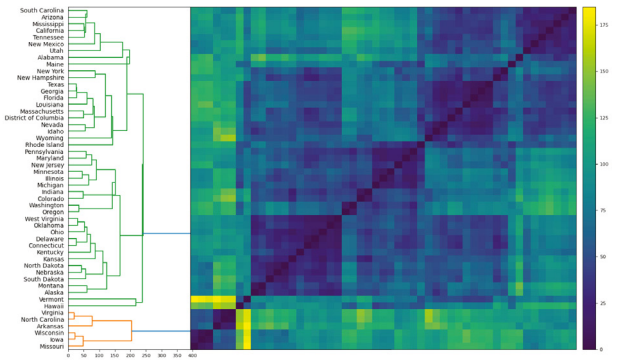
As before, this may be computed for the entire collection ($N = 114$) or one specific country. In Figs. 3(a)–3(c), respectively, we display hierarchical clustering on the three obtained turning point matrices M^{TP} restricted to the states of the US, India and Brazil separately.

Examining these three dendrograms reveals a similar cluster structure between the US and India. Both countries display a dense majority cluster and a small collection of outlier states. Brazil, by contrast, exhibits quite a different structure, with two similarly sized clusters that contain the majority of elements, and then some outliers. We can further examine the cluster-split behaviour of Brazil by examining the results of clustering all $N = 114$ states in our collection in Fig. 4. This total dendrogram contains a majority cluster containing $\sim$ 90% of all states, and two small outlier clusters of five and four states (clusters B and C respectively). The majority cluster contains two subclusters (A1 and A2), featuring a break between US and Indian states, with almost no intersection between the two countries. However, Brazil's states are far more widely distributed. Not only do the outlier clusters B and C consist only of Brazilian states, but Brazil's states are spread throughout both A1 and A2, interleaving between US and Indian states. This finding suggests that US and Indian states exhibit higher intra-collection homogeneity and inter-collection heterogeneity in their wave behaviours when compared to Brazilian states.

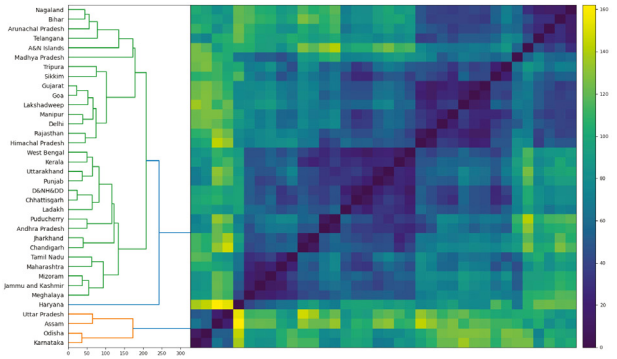
Table 3
Median and standard deviation of the length of the first wave T_1, defined in Section 3 and measured by the first non-trivial trough. The US has the shortest first wave, India has the longest, while Brazil exhibits the greatest variability.

Country	Median T_1	Standard deviation T_1
US	92	76.9
India	231	63.6
Brazil	143	109

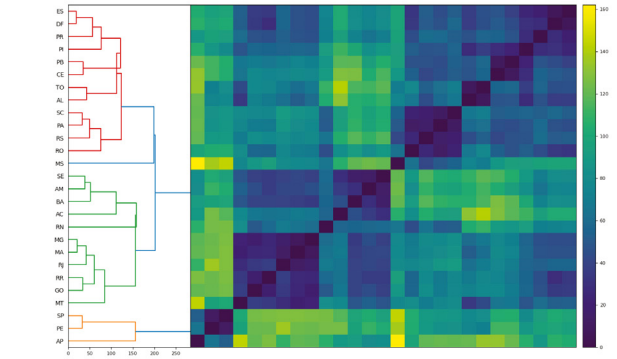
To elucidate the reasons behind these state clustering patterns, we study the distribution of the location of the first non-trivial trough, T_1. This trough indicates the end of the first wave; thus, the value T_1 gives the total length of the first wave in each state. Table 3 documents the median and standard deviation of T_1 among each country's states, while Fig. 5 displays kernel density estimates of the full distribution of values. There is significant variability between the states' first wave lengths between the three countries. The US has a median value of 92 and a standard deviation of 76.9, indicating that most states experienced a short first wave. By contrast, Indian states mostly experienced a long first wave, with a median value of 231 and a standard deviation of 63.6. This suggests that the first wave of COVID-19 cases in Indian states was on average 2.5 times longer than US states, with limited variance between states. As in Fig. 3, Brazil does not exhibit as strong a characteristic behaviour, with a median T_1 score of 143 and a significantly higher standard deviation among Brazilian states of 109. Notably, the median T_1 value of Brazilian states is located between the US and Indian median values. Also of note is the highly skewed distribution for the Brazilian states, with a substantial number of high values despite the relatively lower peak. When viewed in conjunction with Fig. 4, one can see how the heterogeneous turning point behaviours of Brazilian states are classified into predominantly US or Indian subclusters (A1 and A2, respectively). Fig. 5 shows in more detail that the lengths of the first wave among Brazilian states are broadly positioned between those of US and Indian states.



(a)



(b)



(c)

Fig. 3. Hierarchical clustering on the matrix M^{TP}, defined in Section 3, between the individual sets of (a) US states (b) Indian states (c) Brazilian states. A broadly similar cluster structure is observed between the US and India, while Brazil's structure is quite different.

4. Offsets between cases and deaths

In this section, we combine the motivating questions from the previous two sections: the different wave behaviour of the virus, and the time-varying properties of cases, deaths and mortality rates by states. Here, we investigate various methods to quantify and analyse the changing offset between cases to deaths in the different waves of the pandemic in the three countries under consideration. To standardise our comparison of offsets between constituent states, we consider a uniform partition into waves for each entire country. That is, let $x(t)$ be the new daily case time series for an entire country (total counts for the US, India, or Brazil). As in the previous section, we use the methodology of [51], detailed in the Appendix, to divide each aggregated country's case time series into a first, second and possibly third wave. Let $T_0 = 1$, T_1 be the first non-trivial trough, and T_2 be the second non-trivial trough, if it exists. For India and Brazil, this does not exist, so we set $T_2 = T$. Then the interval $T_0 \le t \le T_1$ represents the first wave, $T_1 \le t \le T_2$ the second wave, and in the case of the US only, $T_2 \le t \le T$ represents the third wave. For notational convenience, we set $T_3 = T$ for the US. Thus, the kth wave can be described by the interval $[T_{k-1}, T_k]$, where $k = 1, 2$ for India and Brazil and $k = 1, 2, 3$ for the US. These turning points for the three country's aggregated cases are displayed in Fig. 6.

We apply five different methods to estimate suitable values of the offset τ_k between case and death time series for each wave in each country. Each method determines an appropriate offset using case and death data only between $a = T_{k-1}$ and $b = T_k$. Let $T^* = b - a$ be the length of this interval. We describe the five methods below.

1. **Affinity matrices**: For a given wave and country, let the offset α_k be chosen as follows: on each day t, let $D_X(t), D_Y(t)$ be the matrices of differences between cases and deaths, respectively. That is, $D_X(t)$ is an $N \times N$ matrix defined by $D_X(t)_{ij} = |x_i(t) - x_j(t)|$, where indices i, j range over the states of one country, and similarly for $D_Y(t)$. To any distance matrix D, we can assign a corresponding affinity matrix A defined by
$$A_{ij} = 1 - \frac{D_{ij}}{\max D}. \tag{12}$$
Let $\mathrm{Aff}_X(t)$ and $\mathrm{Aff}_Y(t)$ be the affinity matrices corresponding to $D_X(t), D_Y(t)$, respectively. Given an offset α, with $0 < \alpha < T^*$, let the *normalised total affinity difference* be defined as
$$\frac{1}{T^* - \alpha} \sum_{a \le t \le b-\alpha} \|\mathrm{Aff}_X(t) - \mathrm{Aff}_Y(t + \alpha)\|. \tag{13}$$
The matrix norm is the same as defined in (9). Then, the *affinity offset* of a wave is defined as the value α_k that minimises this total difference.
2. **Probability density function (PDF)**: For a given wave and country, let the offset β_k be chosen as follows: on each day, let $p_X(t), p_Y(t)$ be the probability vector for new cases and deaths on day t. That is, $p_X(t)$ is a length N vector defined by $p_X(t)_i = \frac{x_i(t)}{\sum_{j=1}^N x_j(t)}$ where i ranges over the states of one country. Given an offset β, let the *normalised total pdf difference* be defined as
$$\frac{1}{T^* - \beta} \sum_{a \le t \le b-\beta} \|p_X(t) - p_Y(t + \beta)\|_1, \tag{14}$$
where $\|\|_1$ is the L^1 norm between vectors. Then the *pdf offset* of a wave is defined as the value β_k that minimises this total difference.
3. **Wasserstein distance**: Again, we assume a given country and wave $a \le t \le b$ is under consideration. For each constituent state i, let λ_i be the offset that minimises the Wasserstein distance,
$$W_1(f_i(a : b - \lambda_i), g_i(a + \lambda_i : b)), \tag{15}$$
where $f_i(a : b - \lambda_i)$ is the distribution associated over the interval $a \le t \le b - \lambda_i$, as in (2), and similarly for $g_i(a + \lambda_i : b)$. Then, let $\lambda_k = [\frac{1}{N} \sum_{i=1}^N \lambda_i]$ be the nearest integer to the mean of the estimated offsets for each state i.

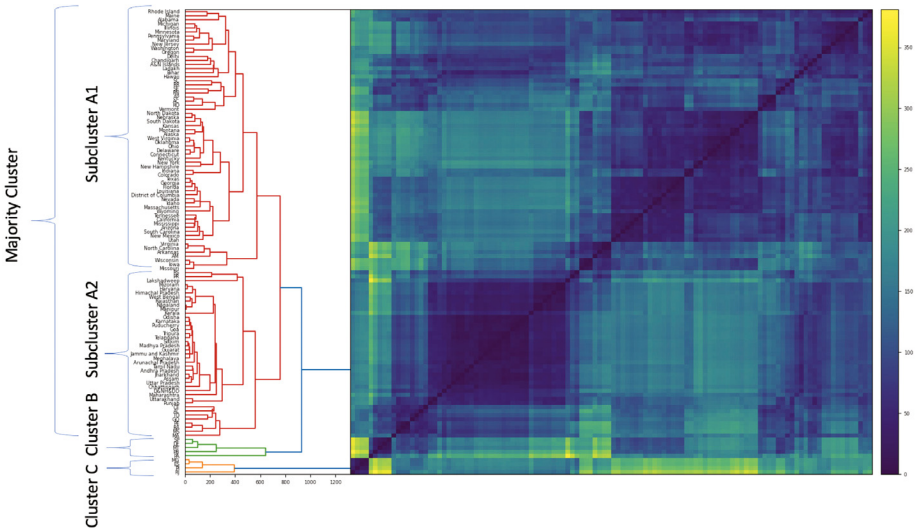


Fig. 4. Hierarchical clustering on the matrix M^{TP}, defined in Section 3, for all $N = 114$ states in our collection. The majority cluster contains two subclusters A1 and A2, broadly consisting of US states and Indian states, respectively. Brazilian states, labelled with two letters for visibility, are interleaved among A1 and A2 and also the two outlier clusters B and C.

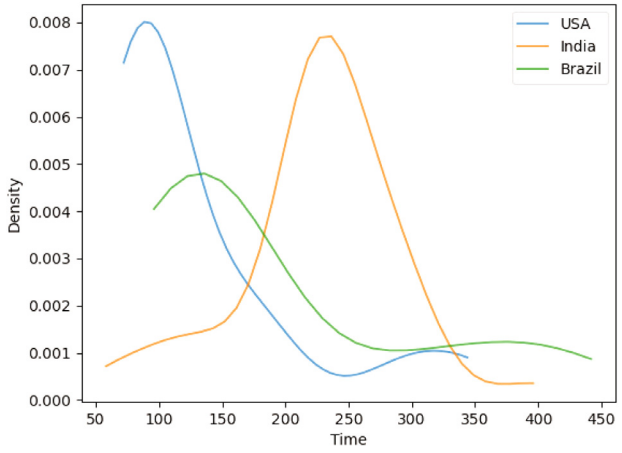


Fig. 5. Kernel density estimates of distributions of the first wave length T_1, defined in Section 3, over each country. The US exhibits the smallest first wave length, India the greatest, while Brazil has the greatest variability.

4. **Energy distance**: Using similar notation as the above method, for each constituent state i, let μ_i be the offset that minimises the *energy distance* [58],

$$D^2(f_i(a : b - \mu_i), g_i(a + \mu_i : b)), \tag{16}$$

where $f_i(a : b - \mu_i)$ and $g_i(a + \mu_i : b)$ are distributions defined above and D^2 is the L^2 integral norm between the associated cumulative distribution functions [58]. Then, let $\mu_k = [\frac{1}{N}\sum_{i=1}^{N} \mu_i]$ analogously as before.

5. **Normalised inner product**: Using similar notation as the above method, for each constituent state i, let ν_i be the offset that minimises the normalised inner product $\langle ., . \rangle_n$, defined as

$$\langle x_i(a : b - \nu_i), y_i(a + \nu_i : b)\rangle_n \tag{17}$$

$$= \frac{x_i(a)y_i(a + \nu_i) + \cdots + x_i(b - \nu_i)y_i(b)}{(x_i(a)^2 + \cdots + x_i(b - \nu_i)^2)^{\frac{1}{2}}(y_i(a + \nu_i)^2 + \cdots + y_i(b)^2)^{\frac{1}{2}}}. \tag{18}$$

Then, let $\nu_k = [\frac{1}{N}\sum_{i=1}^{N} \nu_i]$ analogously as before.

Thus we have offsets $\tau_k \in \{\alpha_k, \beta_k, \lambda_k, \mu_k, \nu_k\}$, for each country and wave k. Each of these methods considers case and death data on a state-by-state basis, taking into account the federal structure of each country. We remark that the affinity matrix and PDF methods share common features of analysing relationships between different states' proportional sizes of case and death counts. Also, the Wasserstein and energy methods share common features of truncating time series and computing distances between distributions.

Before we present the results of this methodology, we present a proposition that demonstrates our methods work well in the case of simulated data.

Proposition 4.1. *Let the multivariate time series of cases and deaths for a federation be $x_i(t)$ and $y_i(t)$. Suppose they have the property that there exists a consistent and proportionate progression from cases to deaths after a time lag of τ_0. That is,*

$$y_i(t) = \begin{cases} \gamma x_i(t - \tau_0), t = \tau_0 + 1, \ldots, T \\ 0, t \leq \tau_0, \end{cases} \tag{19}$$

where $\gamma < 1$ and $\tau_0 \in \mathbb{Z}_{>0}$ are constants. Then, for any wave $[T_{k-1}, T_k]$ of length at least τ_0, all five methods above return $\tau_k = \tau_0$. That is, all five methods identify the correct offset for the following simulated example.

Proof. Let $[T_{k-1}, T_k]$ be a fixed interval of length T^*. Then the normalised total affinity difference (13), evaluated for $\alpha = \tau_0$, produces the value

$$\frac{1}{T^* - \tau_0}\sum_{t=T_{k-1}}^{T_k-\tau_0} \|\text{Aff}_X(t) - \text{Aff}_Y(t + \tau_0)\|. \tag{20}$$

By (19), $y_i(t + \tau_0) = \gamma x_i(t)$ for all t in the interval $[T_{k-1}, T_k - \tau_0]$. Thus, $D_Y(t + \tau_0) = \gamma D_X(t)$. Due to the normalisation process of computing the affinity matrix, this implies $\text{Aff}_X(t) = \text{Aff}_Y(t + \tau_0)$ for all t. Thus, the normalised total affinity difference for the value $\alpha = \tau_0$ produces the minimal possible value of zero, so the method selects $\alpha_k = \tau_0$.

Next, for the PDF method, the normalised total pdf difference evaluated for $\beta = \tau_0$ produces

$$\frac{1}{T^* - \tau_0}\sum_{t=T_{k-1}}^{T_k-\tau_0} \|p_X(t) - p_Y(t + \tau_0)\|_1. \tag{21}$$

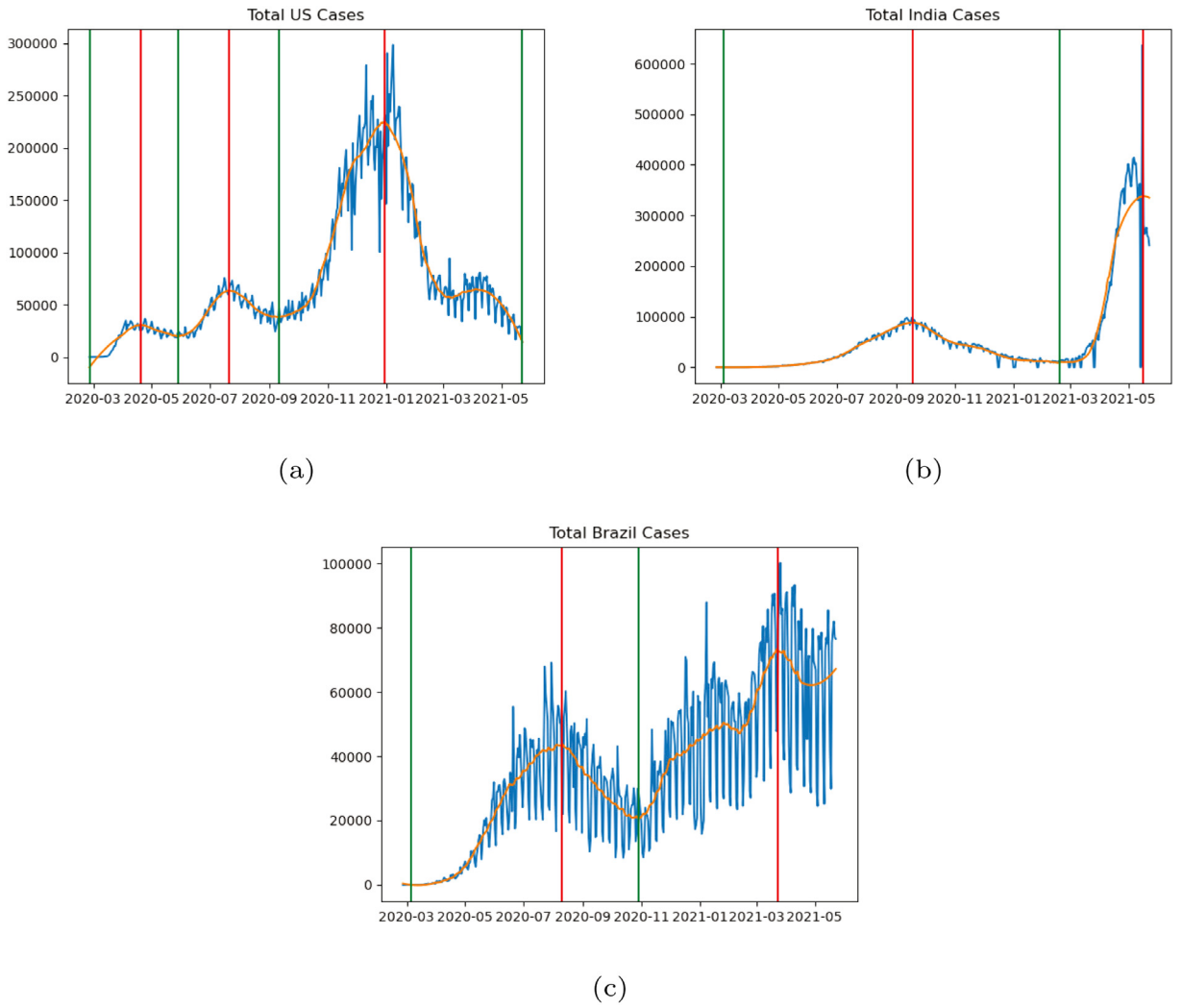


Fig. 6. New daily case time series and determined turning points, defined in Section 3, for (a) the US (b) India (c) Brazil.

Again by (19), we have $y_i(t+\tau_0) = \gamma x_i(t)$ for all t in the interval $[T_{k-1}, T_k - \tau_0]$, so $p_X(t) = p_Y(t+\tau_0)$ for all $t \in [T_{k-1}, T_k - \tau_0]$. Thus, the normalised total pdf difference for the value $\beta = \tau_0$ produces the minimal possible value of zero, so the method selects $\beta_k = \tau_0$.

Next, we turn to the Wasserstein and Energy distance methods. Here, we can again show that for the selected offset $\lambda_i = \tau_0$, the corresponding Wasserstein distance

$$W_1(f_i(T_{k-1} : T_k - \tau_0), g_i(T_{k-1} + \tau_0 : T_k)), \tag{22}$$

is equal to zero. Indeed, $x_i(t)_{T_{k-1} \le t \le T_k - \tau_0}$ is a scalar multiple of $y_i(t)_{T_{k-1}+\tau_0 \le t \le T_k}$, so when both are normalised to distributions $f_i(T_{k-1} : T_k - \tau_0)$ and $g_i(T_{k-1} + \tau_0 : T_k)$ respectively, they coincide. Thus, $\lambda_i = \tau_0$ produces the minimal possible value of zero for the Wasserstein distance and so the method selects $\lambda_i = \tau_0$ for each state i, hence $\lambda_k = \tau_0$. The same argument holds *mutatis mutandis* for the Energy distance.

Finally, for the normalised inner product method, the same reasoning shows that the normalised inner product achieves its maximal value of 1 when $\nu_i = \tau_0$, so the method selects $\nu_i = \tau_0$ for each state i. Hence, ν_k is analogously chosen to be equal to τ_0.

We remark that the procedure of truncating the interval $[T_{k-1}, T_k]$ to $[T_{k-1}, T_k - \tau_k]$ for the case time series and $[T_{k-1+\tau_k}, T_k]$ for the death time series is essential for the proof to work as above. Indeed, in this simulated example, the death time series $y_i(t)$ has exactly τ days of leading zeros before it coincides with a shifted constant times $x_i(t)$, and the truncation is necessary for the methods to select the correct offset. □

Table 4
Offsets between cases and deaths by country and wave of the pandemic, computed with five different methods, as described in Section 4. Only the US is determined to have a third wave within our period of analysis.

Methodology	Wave 1	Wave 2	Wave 3
Affinity (US)	6	37	16
PDF (US)	5	23	16
Wasserstein (US)	11	19	41
Energy (US)	9	17	38
Inner product (US)	10	20	29
Affinity (India)	11	8	n/a
PDF (India)	8	7	n/a
Wasserstein (India)	32	5	n/a
Energy (India)	32	5	n/a
Inner product (India)	13	8	n/a
Affinity (Brazil)	9	9	n/a
PDF (Brazil)	9	9	n/a
Wasserstein (Brazil)	18	13	n/a
Energy (Brazil)	15	11	n/a
Inner product (Brazil)	12	21	n/a

Table 4 documents the wave-specific offsets for all three countries among our five methods. We observe broad similarity across all countries and waves between the results obtained by pairs

of related methods (affinity and PDF, Wasserstein and energy). Each country presents a unique pattern in the length of their progression from cases to deaths for each wave of the pandemic. First, the US is the only country determined to experience three waves of COVID-19 cases within our analysis window. For all five methods, the first wave produces a significantly lower offset than the second and third waves of COVID-19. The timing of the first wave corresponds to the first half of 2020, when many US states (especially those located in the Northeast) were overwhelmed by early case numbers. As a result, many cases went undetected, and hospitals were unable to administer optimal care to patients. Furthermore, early in the pandemic, there was greater uncertainty within the medical community on suitable treatments for COVID-19 patients.

India, which exhibits two waves of COVID-19 in our analysis window, features almost the opposite observation. As shown in Table 3, the length of the first wave in India was $\sim$2.5 times that of the US, and it exhibited a more gradual progression (and subsequent decline) in daily cases until states reached their first peak and trough, respectively. Although much shorter, the second wave was more severe among Indian states — with universally rapid growth in cases and deaths. All five optimisation methods determined the offset of the second wave to be shorter than that of the first wave. This mirrors our finding in the case of the US: when states are overwhelmed with COVID-19, hospitals become overwhelmed with cases, and many patients go undetected — this leads to a decrease in the length of the offset between cases and deaths. This can most likely be explained by latent COVID-19, the inability to access critical equipment (such as ventilators), and inferior treatment within hospitals.

Brazil has quite a different finding again, with little consistency in the offset trend between its first and second waves. Several reasons may explain the variability in our estimates. First, the Brazilian data is quite noisy, with more missing data and reporting issues than the US and India. Second, the variability in the distribution of states' T_1 values may suggest limited collective consistency in offset trends among the Brazilian states. Accordingly, we see no clear trend in offset behaviours as we progress from the first to the second wave of the outbreak.

5. Discussion

In this paper, we perform a detailed analysis of the three countries most impacted by COVID-19, the US, India and Brazil. Given COVID-19's severe yet varied impact on countries worldwide, our motivation is to understand the differences in the dynamics of the virus' propagation among the world's three worst affected countries. We seek to study both internal structural similarity between states within each country and differences between the countries with respect to several attributes around COVID-19. Comparing the structural dynamics of separate countries' COVID-19 outbreaks may provide insights into the influence different governments, cultures and healthcare systems have had in the evolution of the pandemic. In addition to this explicit contrast, we wanted to explore variability within each country, namely similarity between countries' constituent states.

First, we study the similarity between case, death and mortality rate trajectories produced by each of our three countries' constituent states. In Section 2, we offer methodological contributions as well as non-trivial findings regarding heterogeneity between states in each federation. Our procedure in Algorithm 1 not only identifies a sequence of the most anomalous elements (in this case states) of a collection, it also produces an easily interpretable decreasing curve quantifying the collective heterogeneity. This procedure is robust to the existence of one or even several outlier elements. By the scale of the curves displayed in Fig. 2, one can immediately see that India exhibits the greatest heterogeneity between states with respect to the three trajectories analysed, particularly rolling mortality rates. This is a robust finding that consistently holds even when we remove anomalous states, and highly non-trivial given the findings of Section 3 discussed below. The specific identification of the most anomalous states is also non-obvious, revealing different patterns in each federation. In the US, we find that the most anomalous behaviour is consistently located in the Northeast. In India, the state Lakshadweep is consistently identified as most anomalous in cases, deaths and mortality. In Brazil, there is less consistency in the type of anomalies identified among our three attributes.

The insights generated above concern broad structure in the data on a state-by-state basis. We have combined existing statistical learning methodologies (such as clustering), a new distance between trajectories as well as a new algorithmic approach to identify specific states and quantify overall heterogeneity, with robustness to outliers. The insights presented in this manuscript would not be possible without a combination of existing (rather sophisticated) and new (rather bespoke) procedures, all carefully considered for the application. More broadly, most COVID-19 data consumed by the general public is reported at the national level; most variation within states is ignored, especially a detailed quantification of heterogeneity. Our methods combine non-trivial mathematical investigation with data sets that are typically not examined in detail at the state level.

In Section 3, we apply our turning point algorithm to study wave behaviours among the three countries. In the US, where three waves of COVID-19 cases are observed, a median first wave length of 92 days is found among the distribution of US states. By contrast, Indian states produced a median first wave length of 231 days, with a lower variance than the US, and just two waves of COVID-19 cases overall. In Brazil, where two waves of the cases were also identified, the median length of states' first wave was 143, with high variance. Our analysis suggests that US and Indian states exhibit stronger characteristic behaviours than those exhibited by Brazil. Indeed, clustering reveals that the US and India are quite dissimilar in wave behaviour, almost entirely clustering among themselves, while Brazil is quite heterogeneous, with some states similar to US states, some similar to Indian states, and some outlier states.

These findings are highly non-trivial without undertaking judicious mathematical analysis as we have done. Numerous papers on COVID-19 simply estimate the duration of waves by inspection or other unreliable methods, while we use a careful algorithm to do so. Unlike most work, we do so on a state-by-state basis, and thus must deal with data issues such as anomalous counts and missing values. Our findings contrast notably with Section 2 and are highly non-trivial to guess. While it is predictable that US and Indian states exhibit relatively strong characteristic wave behaviours among themselves, it is certainly non-trivial that Brazilian states interleave between US and Indian states with respect to wave behaviour, and that the distribution of first wave length among Brazilian states (Fig. 5) is so broad. Further, it is striking that Section 2 reveals the greatest heterogeneity between Indian states in terms of trajectories, but Section 3 demonstrates the least variance in first wave length (Table 3). This is not necessarily contradictory but is highly non-obvious: case and death curves exhibit substantial differences but the overall wave pattern is more uniform across India.

Finally, Section 4 introduces new optimisation methodologies to study the progression of COVID-19 cases to deaths in each of our three countries' waves of the pandemic. We believe this is the first work to explicitly acknowledge that the progression from cases to deaths may vary between different waves of the pandemic and aim to study this. In the US, we highlight a significantly longer period between diagnosis and death in the second

and third waves of COVID-19 cases. This finding is consistent among all five optimisation methods. In India, all five methods demonstrate a sharp reduction in the length of this offset as we progress from the first to the second wave. In Brazil, we find limited consistency among our methods, with no clear takeaway regarding the change in the length of the COVID-19 case life cycle, in the first and second waves. In aggregate, our analysis suggests that when countries become overwhelmed with COVID-19 cases, the length of the case-to-death progression decreases. This may be due to overwhelmed hospital systems, sub-optimal medical treatment, limited access to medical resources such as ventilators and an increase in undetected cases. We also include theoretical validation of our methodology, which is non-trivial due to the truncation of time series inherent in the case and death data (that is, death data lag behind cases and non-zero counts begin later).

There are several reasons why these determinations of offsets between cases and deaths are not particularly obvious. First, they are computed in a high dimensional manner with several methods that use the federal structure of the three countries. Second, the changes between waves of these offsets are different for all three federations, which we believe shows the impossibility of a straightforward prediction of their behaviour. Algorithmic techniques must be used to identify time series turning points (corresponding to waves of the pandemic), and the relationship between cases and deaths is fluid — varying over time, across countries and between countries' constituent states and territories. Although the offset in the progression from COVID-19 cases to deaths is only one facet of a hugely complex global pandemic, it is of great importance to understand for the future treatment and management of COVID-19 cases. COVID-19 data follows a causal structure: any COVID-19 case will ultimately progress into either the recovered or death category. This causal structure is typically modelled via SIRD models and their variants described in Section 1. These have their utility, but are not ideal to study the multi-wave dynamics of COVID-19 brought about by regularly shifting government restrictions and community behaviour. We choose to exclusively address the transition from cases to deaths without the strong parametric assumptions in SIRD models; we believe this progression to be of direct importance in treating COVID-19 patients currently burdening many countries' healthcare systems.

5.1. Future work

There are many avenues for potential future work, in both methodological and applied contexts. First, one could investigate the reasons for more or less heterogeneity among constituent states for various countries. For example, one could explore why Brazil's states experienced rather different outcomes relative to wave behaviours and progression from cases to deaths. In this paper, we highlight that these differences are far more significant than the USA and India. Indeed, Brazil's human development index (HDI) of 0.765 is between that of the US (0.926) and India (0.645), and it is conceivable that development among Brazilian states differs more than that among the US or India. This, along with other predictors, may help construct supervised and unsupervised learning algorithms where relationships can be learned and associations can be formed, respectively.

Next, the methods that are introduced in this paper could be extended. Although the offsets in this paper have been implemented in discrete time partitions, these methods could conceivably be implemented in a rolling manner, where a continuous (time-varying) offset may be estimated. Furthermore, the theoretical aspects of these estimators could be further investigated, and tested on data generated from a variety of data generating processes. This may include noise generated from a wide variety of distributions, adversarial data such as extreme points and outliers, and so on. In addition, future work could further explore the aforementioned causal structure in the data, including offsets between time series of COVID-19 cases, counts of recovered patients (including those who experience "long Covid" [59]) and COVID-19 deaths. One could compare the offsets between COVID-19 cases and deaths, and COVID-19 cases and recovered patients separately — and then study whether there is a latent relationship between these two offsets, and more specifically, study how they evolve with time. Our descriptive and nonparametric analysis could conceivably be incorporated with judiciously chosen SIRD models on a wave by wave basis.

At the time of writing this paper, many parts of the world are currently experiencing a fourth wave of COVID-19 cases. Many European countries such as Austria and Germany are attracting a substantial amount of publicity, regarding their growth in new daily COVID-19 cases. It would be of great interest to compare the heterogeneity of COVID-19 epidemiology within differing states or regions of these countries, and estimate the offset in the progression from cases to deaths during the fourth wave of the pandemic. In particular, with the appropriate data, one could distinguish between the vaccinated and unvaccinated populations.

6. Conclusion

Overall, we have identified numerous features that characterise the nature of the pandemic within the US, India and Brazil. India exhibits the greatest heterogeneity in its trajectories, and yet simultaneously the most homogeneity in its wave behaviours due to a very long first wave and a rapid second wave in almost every state. The US and India cluster quite separately in trajectory and wave behaviours, while Brazilian states are interleaved between them, characterised by the greatest variance in wave lengths. A similar distinction is observed in offsets, where the US case-to-death progressions drastically lengthen between first and subsequent waves, the reverse holds for India, while Brazil is again a mixture of the two.

Throughout this work, we have identified specific states within the three federations as the most anomalous and determined various non-trivial features in the federations' COVID-19 behaviour, including heterogeneity of trajectories, wave behaviour, and the progression from cases to deaths. New methodologies have been presented for this purpose, including the ability to more robustly determine distances between trajectories and determine patterns in overall heterogeneity without too much vulnerability to outliers. We have identified numerous avenues for future work to apply these methods in new contexts, such as Europe's fourth wave, or to undertake closer analysis with researchers from other disciplines to investigate some of the policy measures or regional features that could be contributing to these patterns.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding sources

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

Data availability

Daily COVID-19 case and death counts for the US, India and Brazil can be found at the New York Times [52], PRS Legislative Research [53] and the Brazilian Ministry of Health [54], respectively.

Appendix. Turning point methodology

In this section, we provide more details for identifying turning points of a new case time series $x(t)$. First, some smoothing of the counts is necessary due to data irregularities and discrepancies between different data sources. There are consistently lower counts on the weekends and some negative counts due to retroactive adjustments. A Savitzky–Golay filter ameliorates these issues by combining polynomial smoothing with a moving average computation — this moving average eliminates all but a few small negative counts; we then replace these negative counts with zero. This yields a smoothed time series $\hat{x}(t) \in \mathbb{R}_{\geq 0}$. Subsequently, we perform a two-step process to select and then refine a non-empty set P of local maxima (peaks) and T of local minima (troughs).

Following [51], we apply a two-step algorithm to the smoothed time series $\hat{x}(t)$. The first step produces an alternating sequence of troughs and peaks, beginning with a trough at $t = 1$, when there are zero cases. The second step refines this sequence according to chosen conditions and parameters. The primary conditions to identify a peak or trough, respectively, in the first step, are the following:

$$\hat{x}(t_0) = \max\{\hat{x}(t) : \max(1, t_0 - l) \leq t \leq \min(t_0 + l, T)\}, \tag{A.1}$$

$$\hat{x}(t_0) = \min\{\hat{x}(t) : \max(1, t_0 - l) \leq t \leq \min(t_0 + l, T)\}, \tag{A.2}$$

where l is a parameter to be chosen. Following [51], we select $l = 17$, which accounts for the 14-day incubation period of the virus [60] and less testing on weekends. Defining peaks and troughs according to this definition alone has several flaws, including the potential for two consecutive peaks.

Instead, we implement an inductive procedure to select an alternating sequence of peaks and troughs. Suppose t_0 is the last determined peak. We search in the period $t > t_0$ for the first of two cases: if we find a time $t_1 > t_0$ that satisfies (A.2) and a non-triviality condition $\hat{x}(t_1) < \hat{x}(t_0)$, we add t_1 to the set of troughs and proceed from there. If we find a time $t_1 > t_0$ that satisfies (A.1) and $\hat{x}(t_0) \geq \hat{x}(t_1)$, we ignore this lower peak as redundant; if we find a time $t_1 > t_0$ that satisfies (A.1) and $\hat{x}(t_1) > \hat{x}(t_0)$, we remove the peak t_0, replace it with t_1 and proceed from t_1. A similar process applies from a trough at t_0.

At this point, a time series is assigned an alternating sequence of troughs and peaks. However, some turning points are immaterial and should be excluded. The second step is a flexible approach introduced in [51] for this purpose. In this paper, we introduce new conditions within this framework. First, let t_m be the global maximum of $\hat{x}(t)$. If this is not unique, we declare t_m to be the first global maximum. This point t_m is always declared a peak during the first step detailed above. Given any other peak t_1, we compute the peak ratio $\frac{\hat{x}(t_1)}{\hat{x}(t_m)}$. We select a parameter δ, and if $\frac{\hat{x}(t_1)}{\hat{x}(t_m)} < \delta$, we remove the peak t_1. If two consecutive troughs t_0, t_2 remain, we remove t_0 if $\hat{x}(t_0) > \hat{x}(t_2)$, and remove t_2 if $\hat{x}(t_0) \leq \hat{x}(t_2)$. That is, we ensure the sequence of peaks and troughs remains alternating. In our implementation, we choose $\delta = 0.01$. Unlike [51], we remove earlier peaks, not just subsequent peaks, according to this condition.

Finally, we use the same *log-gradient* function between times $t_1 < t_2$, defined as

$$\text{log-grad}(t_1, t_2) = \frac{\log \hat{x}(t_2) - \log \hat{x}(t_1)}{t_2 - t_1}. \tag{A.3}$$

The numerator equals $\log(\frac{\hat{x}(t_2)}{\hat{x}(t_1)})$, a "logarithmic rate of change". Unlike a standard rate of change given by $\frac{\hat{x}(t_2)}{\hat{x}(t_1)} - 1$, the logarithmic change is symmetrically between $(-\infty, \infty)$. Let t_1, t_2 be adjacent turning points (one a trough, one a peak). We choose a parameter $\epsilon = 0.01$; if

$$|\text{log-grad}(t_1, t_2)| < \epsilon, \tag{A.4}$$

that is, the average logarithmic change is less than 1%, we remove t_2 from our sets of peaks and troughs. If t_2 is not the final turning point, we also remove t_1.

References

[1] R.L. Haffajee, M.M. Mello, Thinking globally, acting locally - the U.S. response to Covid-19, N. Engl. J. Med. 382 (22) (2020) e75, http://dx.doi.org/10.1056/nejmp2006740.

[2] R.M. da Silva, C.F.O. Mendes, C. Manchein, Scrutinizing the heterogeneous spreading of COVID-19 outbreak in large territorial countries, Phys. Biol. 18 (2) (2021) 025002, http://dx.doi.org/10.1088/1478-3975/abd0dc.

[3] I. Bharali, et al., India's policy response to COVID-19, 2020, The Center for Policy Impact in Global Health, June, 2020.

[4] M. Wang, et al., Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro, Cell Res. 30 (3) (2020) 269–271, http://dx.doi.org/10.1038/s41422-020-0282-0.

[5] E.M. Bloch, Convalescent plasma to treat COVID-19, Blood 136 (6) (2020) 654–655, http://dx.doi.org/10.1182/blood.2020007714.

[6] X. Xu, et al., Effective treatment of severe COVID-19 patients with tocilizumab, Proc. Natl. Acad. Sci. 117 (20) (2020) 10970–10975, http://dx.doi.org/10.1073/pnas.2005615117.

[7] B. Cao, et al., A trial of lopinavir-ritonavir in adults hospitalized with severe Covid-19, N. Engl. J. Med. 382 (19) (2020) 1787–1799, http://dx.doi.org/10.1056/nejmoa2001282.

[8] F.P. Polack, et al., Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine, N. Engl. J. Med. 383 (27) (2020) 2603–2615, http://dx.doi.org/10.1056/nejmoa2034577.

[9] E.E. Walsh, et al., Safety and immunogenicity of two RNA-based Covid-19 vaccine candidates, N. Engl. J. Med. 383 (25) (2020) 2439–2450, http://dx.doi.org/10.1056/nejmoa2027906.

[10] L. Wynants, et al., Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal, BMJ (2020) m1328, http://dx.doi.org/10.1136/bmj.m1328.

[11] E. Estrada, COVID-19 and SARS-CoV-2. Modeling the present, looking at the future, Phys. Rep. 869 (2020) 1–51, http://dx.doi.org/10.1016/j.physrep.2020.07.005.

[12] N.S. Barlow, S.J. Weinstein, Accurate closed-form solution of the SIR epidemic model, Physica D 408 (2020) 132540, http://dx.doi.org/10.1016/j.physd.2020.132540.

[13] S.J. Weinstein, M.S. Holland, K.E. Rogers, N.S. Barlow, Analytic solution of the SEIR epidemic model via asymptotic approximant, Physica D 411 (2020) 132633, http://dx.doi.org/10.1016/j.physd.2020.132633.

[14] K.Y. Ng, M.M. Gui, COVID-19: Development of a robust mathematical model and simulation package with consideration for ageing population and time delay for control action and resusceptibility, Physica D 411 (2020) 132599, http://dx.doi.org/10.1016/j.physd.2020.132599.

[15] C. Vyasarayani, A. Chatterjee, New approximations, and policy implications, from a delayed dynamic model of a fast pandemic, Physica D 414 (2020) 132701, http://dx.doi.org/10.1016/j.physd.2020.132701.

[16] M. Cadoni, G. Gaeta, Size and timescale of epidemics in the SIR framework, Physica D 411 (2020) 132626, http://dx.doi.org/10.1016/j.physd.2020.132626.

[17] A.G. Neves, G. Guerrero, Predicting the evolution of the COVID-19 epidemic with the A-SIR model: Lombardy, Italy and São Paulo state, Brazil, Physica D 413 (2020) 132693, http://dx.doi.org/10.1016/j.physd.2020.132693.

[18] A. Comunian, R. Gaburro, M. Giudici, Inversion of a SIR-based model: A critical analysis about the application to COVID-19 epidemic, Physica D 413 (2020) 132674, http://dx.doi.org/10.1016/j.physd.2020.132674.

[19] T. Sun, Y. Wang, Modeling COVID-19 epidemic in Heilongjiang province, China, Chaos Solitons Fractals 138 (2020) 109949, http://dx.doi.org/10.1016/j.chaos.2020.109949.

[20] A. Ballesteros, A. Blasco, I. Gutierrez-Sagredo, Hamiltonian structure of compartmental epidemiological models, Physica D 413 (2020) 132656, http://dx.doi.org/10.1016/j.physd.2020.132656.

[21] S. Liu, M.Y. Li, Epidemic models with discrete state structures, Physica D 422 (2021) 132903, http://dx.doi.org/10.1016/j.physd.2021.132903.

[22] N.M. Gatto, H. Schellhorn, Optimal control of the SIR model in the presence of transmission and treatment uncertainty, Math. Biosci. 333 (2021) 108539, http://dx.doi.org/10.1016/j.mbs.2021.108539.

[23] C. Manchein, E.L. Brugnago, R.M. da Silva, C.F.O. Mendes, M.W. Beims, Strong correlations between power-law growth of COVID-19 in four continents and the inefficiency of soft quarantine strategies, Chaos: Interdisciplinary J. Nonlinear Sci. 30 (4) (2020) 041102, http://dx.doi.org/10.1063/5.0009454.

[24] B. Blasius, Power-law distribution in the number of confirmed COVID-19 cases, Chaos: Interdisciplinary J. Nonlinear Sci. 30 (9) (2020) 093123, http://dx.doi.org/10.1063/5.0013031.
[25] B.K. Beare, A.A. Toda, On the emergence of a power law in the distribution of COVID-19 cases, Physica D 412 (2020) 132649, http://dx.doi.org/10.1016/j.physd.2020.132649.
[26] M. Perc, N.G. Miksić, M. Slavinec, A. Stožer, Forecasting COVID-19, Front. Phys. 8 (2020) 127, http://dx.doi.org/10.3389/fphy.2020.00127.
[27] S. Boccaletti, W. Ditto, G. Mindlin, A. Atangana, Modeling and forecasting of epidemic spreading: The case of Covid-19 and beyond, Chaos Solitons Fractals 135 (2020) 109794, http://dx.doi.org/10.1016/j.chaos.2020.109794.
[28] O. Castillo, P. Melin, Forecasting of COVID-19 time series for countries in the world based on a hybrid approach combining the fractal dimension and fuzzy logic, Chaos Solitons Fractals 140 (2020) 110242, http://dx.doi.org/10.1016/j.chaos.2020.110242.
[29] O. Castillo, P. Melin, A novel method for a COVID-19 classification of countries based on an intelligent fuzzy fractal approach, Healthcare 9 (2) (2021) 196, http://dx.doi.org/10.3390/healthcare9020196.
[30] P. Melin, J.C. Monica, D. Sanchez, O. Castillo, Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico, Healthcare 8 (2) (2020) 181, http://dx.doi.org/10.3390/healthcare8020181.
[31] D. Manevski, N.R. Gorenjec, N. Kejžar, R. Blagus, Modeling COVID-19 pandemic using Bayesian analysis with application to slovene data, Math. Biosci. 329 (2020) 108466, http://dx.doi.org/10.1016/j.mbs.2020.108466.
[32] N. James, M. Menzies, Trends in COVID-19 prevalence and mortality: A year in review, Physica D 425 (2021) 132968, http://dx.doi.org/10.1016/j.physd.2021.132968.
[33] K. Shang, B. Yang, J.M. Moore, Q. Ji, M. Small, Growing networks with communities: A distributive link model, Chaos: Interdisciplinary J. Nonlinear Sci. 30 (4) (2020) 041101, http://dx.doi.org/10.1063/5.0007422.
[34] A. Karaivanov, A social network model of COVID-19, PLoS One 15 (10) (2020) e0240878, http://dx.doi.org/10.1371/journal.pone.0240878.
[35] J. Ge, D. He, Z. Lin, H. Zhu, Z. Zhuang, Four-tier response system and spatial propagation of COVID-19 in China by a network model, Math. Biosci. 330 (2020) 108484, http://dx.doi.org/10.1016/j.mbs.2020.108484.
[36] L. Xue, S. Jing, J.C. Miller, W. Sun, H. Li, J.G. Estrada-Franco, J.M. Hyman, H. Zhu, A data-driven network model for the emerging COVID-19 epidemics in Wuhan, Toronto and Italy, Math. Biosci. 326 (2020) 108391, http://dx.doi.org/10.1016/j.mbs.2020.108391.
[37] F. Saldaña, H. Flores-Arguedas, J.A. Camacho-Gutiérrez, I. Barradas, Modeling the transmission dynamics and the impact of the control interventions for the COVID-19 epidemic outbreak, Math. Biosci. Eng. 17 (4) (2020) 4165–4183, http://dx.doi.org/10.3934/mbe.2020231.
[38] A. Danchin, G. Turinici, Immunity after COVID-19: Protection or sensitization? Math. Biosci. 331 (2021) 108499, http://dx.doi.org/10.1016/j.mbs.2020.108499.
[39] J.A.T. Machado, A.M. Lopes, Rare and extreme events: the case of COVID-19 pandemic, Nonlinear Dynam. (2020) http://dx.doi.org/10.1007/s11071-020-05680-w.
[40] N. James, M. Menzies, P. Radchenko, COVID-19 second wave mortality in Europe and the United States, Chaos: Interdisciplinary J. Nonlinear Sci. 31 (2021) 031105, http://dx.doi.org/10.1063/5.0041569.
[41] C.N. Ngonghala, E.A. Iboi, A.B. Gumel, Could masks curtail the post-lockdown resurgence of COVID-19 in the US? Math. Biosci. 329 (2020) 108452, http://dx.doi.org/10.1016/j.mbs.2020.108452.
[42] J. Cavataio, S. Schnell, Interpreting SARS-CoV-2 seroprevalence, deaths, and fatality rate — making a case for standardized reporting to improve communication, Math. Biosci. 333 (2021) 108545, http://dx.doi.org/10.1016/j.mbs.2021.108545.
[43] N. James, M. Menzies, Efficiency of communities and financial markets during the 2020 pandemic, Chaos: Interdisciplinary J. Nonlinear Sci. 31 (8) (2021) 083116, http://dx.doi.org/10.1063/5.0054493.
[44] L.O. Náraigh, A. Byrne, Piecewise-constant optimal control strategies for controlling the outbreak of COVID-19 in the irish population, Math. Biosci. 330 (2020) 108496, http://dx.doi.org/10.1016/j.mbs.2020.108496.
[45] D.H. Glass, European and US lockdowns and second waves during the COVID-19 pandemic, Math. Biosci. 330 (2020) 108472, http://dx.doi.org/10.1016/j.mbs.2020.108472.
[46] Y. Zhou, et al., A spatiotemporal epidemiological prediction model to inform county-level COVID-19 risk in the United States, Harv. Data Sci. Rev. (2020) http://dx.doi.org/10.1162/99608f92.79e1f45e.
[47] P. Melin, J.C. Monica, D. Sanchez, O. Castillo, Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps, Chaos Solitons Fractals 138 (2020) 109917, http://dx.doi.org/10.1016/j.chaos.2020.109917.
[48] Y. Wang, Y. Liu, J. Struthers, M. Lian, Spatiotemporal characteristics of the COVID-19 epidemic in the United States, Clin. Infect. Dis. 72 (4) (2020) 643–651, http://dx.doi.org/10.1093/cid/ciaa934.
[49] N. James, M. Menzies, H. Bondell, Understanding spatial propagation using metric geometry with application to the spread of COVID-19 in the United States, EPL (Europhys. Lett.) 135 (4) (2021) 48004, http://dx.doi.org/10.1209/0295-5075/ac2752.
[50] N. James, M. Menzies, Association between COVID-19 cases and international equity indices, Physica D 417 (2021) 132809, http://dx.doi.org/10.1016/j.physd.2020.132809.
[51] N. James, M. Menzies, COVID-19 in the United States: Trajectories and second surge behavior, Chaos: Interdisciplinary J. Nonlinear Sci. 30 (2020) 091102, http://dx.doi.org/10.1063/5.0024204.
[52] Coronavirus (Covid-19) data in the United States, 2021, The New York Times, https://github.com/nytimes/covid-19-data. (Accessed 24 July 2021).
[53] Details on cases, 2021, PRS Legislative Research, https://prsindia.org/covid-19/cases. (Accessed 24 July 2021).
[54] Painel coronavírus, 2021, Ministério da Saúde, https://covid.saude.gov.br. (Accessed 24 July 2021).
[55] E. del Barrio, E. Giné, C. Matrán, Central limit theorems for the Wasserstein distance between the empirical and the true distributions, Ann. Probab. 27 (2) (1999) 1009–1071, http://dx.doi.org/10.1214/aop/1022677394.
[56] H. Minkowski, Geometrie Der Zahlen, Chelsea, 1953.
[57] N. James, M. Menzies, L. Azizi, J. Chan, Novel semi-metrics for multivariate change point analysis and anomaly detection, Physica D 412 (2020) 132636, http://dx.doi.org/10.1016/j.physd.2020.132636.
[58] G.J. Székely, M.L. Rizzo, Energy statistics: A class of statistics based on distances, J. Stat. Plan. Inference 143 (8) (2013) 1249–1272, http://dx.doi.org/10.1016/j.jspi.2013.03.018.
[59] E. Mahase, Covid-19: What do we know about "long covid"? BMJ (2020) m2815, http://dx.doi.org/10.1136/bmj.m2815.
[60] S.A. Lauer, et al., The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application, Ann. Intern. Med. 172 (9) (2020) 577–582, http://dx.doi.org/10.7326/m20-0504.