https://doi.org/10.1007/s42979-022-01083-3

ORIGINAL RESEARCH

Clusters of COVID-19 Indicators in India: Characterization, Correspondence and Change Analysis

Aniket Raj[1] · Pramit Bhattacharyya[1] · Gagan Raj Gupta[1]

Received: 13 June 2021 / Accepted: 4 March 2022 / Published online: 5 April 2022
© The Author(s), under exclusive licence to Springer Nature Singapore Pte Ltd 2022

Abstract

We conduct a long-term epidemiology study of COVID-19 in India from Mar 2020 to May 2021 using a number of indicators such as active cases, daily new cases, and deaths, on a micro (district level, per capita) and macro level (state level). Our automated shape-based cluster discovery of the per capita daily new cases (*case rate*) during the *first wave* in India (between Mar 2020 and Jan 2021) revealed four distinct shape patterns: sharp-rise and decline, steady-rise and decline, plateau and multiple relatively high peaks. These clusters exhibit a strong geographical correlation. To determine the correspondence between clusters obtained by different indicators, we design a novel metric for determining edge-weights in their *intersection graph*. This is used for comparative analysis and to develop informative *hierarchical* cartographic visualizations. We then perform dynamic cluster analysis for different time windows to answer some pertinent questions. Is the *second wave* similar to or different from the *first wave*? How has the relative ranking (on micro- and macro-level indicators) of the states varied over the last one year? How much medical resources have been stressed during the peak? We demonstrate that using multiple indicators, we can assess the impact of the epidemic holistically in a particular geography. Our analysis techniques and insights obtained can help the local and state governments in monitoring and managing COVID-19 situation and fine-tuning the ongoing vaccination drive in India.

Keywords Covid-19 · Agglomerative-clustering · Cluster-correspondence · Change-analysis

Introduction

Since early 2020 until the time of writing this paper, the world is facing deep human, social and economic crises due to the spread of COVID-19 pandemic. To manage the pandemic, governments have sought a multi-level approach involving various social restrictions. The first COVID case was registered in India on 30 January 2020 and the number of active cases grew to over a million active cases in September 2020. The number of active cases started declining in most parts of the country around January 2021. However, around Feb–Mar 2021, a devastating second wave of COVID-19 cases emerged in India, overwhelming the healthcare infrastructure and resulting in unprecedented number of deaths. This shows the unpredictable nature of the pandemic which still remains a global threat and needs to be very carefully analyzed, contained and managed. For improving the management strategy, a careful understanding of various factors related to the pandemic such as its intensity, growth, spread, change, medical resource consumption etc. is needed.

This is challenging because of a wide variety in density of population, awareness among people, health care resources: skilled medical staff, testing equipment, ICU beds, etc. Thus, simply analyzing the absolute numbers at the state level, hides the situation in states with low population, even though the per-capita impact (which assesses the bottom line health quality) may be much higher. Similarly, the healthcare resources in the country are very unevenly distributed with

"This article is part of the topical collection "Social Data Science: Research Challenges and Future Directions" guest edited by Sarbani Roy, Chandreyee Chowdhury and Samiran Chattopadhyay".

✉ Aniket Raj
aniket.raj1947@gmail.com

Pramit Bhattacharyya
callpramit@gmail.com

Gagan Raj Gupta
gagan@iitbhilai.ac.in

[1] IIT Bhilai, Raipur, Chhattisgarh, India

respect to the population of each state or district. There are also some variations in the new active cases per day due to inconsistency in the testing. For example, the number of tests conducted/reported on Sundays is always smaller than the weekly average as shown in Fig. 2, causing a sharp decline in the plot, giving it a seasonal behavior. Similarly, the reporting of deaths can be delayed or inaccurate and such delays may be different in different states and districts.

To help in this process, we develop new techniques to automate cluster discovery, assign severity ranking to clusters, establish cluster correspondences and visualize cluster dynamics for hierarchical time-series data. These techniques help us in performing epidemiology study of COVID-19 in India using a number of indicators (Table 1) such as active cases, daily new cases, deaths etc. on a micro (district level, per capita) and macro level (state level).

Using the above techniques, we generate the severity ranks for each state for smaller time periods (e.g. every 30 days) for each indicator. We visualize these indicators using heatmaps which reveal the relative evolution patterns of COVID19 in each state over the last 14 months. We find that some states (e.g. Delhi) experienced more than two distinct waves, the time periods (from on-set to decline) of the COVID-19 waves were different in different parts of the country. Not only was the second wave much higher in terms of the absolute numbers, some states like Chhattisgarh and Punjab were relatively more severely impacted by the second wave. Thus, in many ways, second-wave was different than the first-wave. Finally, we assess the medical resources (using the number of beds as an indicator) and identify regions that have relatively poor health infrastructure to deal with the pandemic, especially if the third wave happens and is larger than the second one. We now summarize the key contributions of the paper as follows:

1. Time-series clustering analysis of normalized COVID-19 data across all the states and districts in India. To our knowledge, this is the first study that uses *case rate* for each district in India for a period of more than 400 days. Previous studies, e.g., [1–6] were relatively short term and did not use normalized metrics and hence their results have a bias towards more populous regions.
2. Design of a novel cluster similarity (edge-weight) metric for hierarchical dataset and a novel application of cluster correspondence using maximum edge-weighted matching to find the best mapping between clusters of states and districts. This helps us to present them together in a informative *hierarchical* cartographic visualization and eventually mine states and districts with relatively high *case rate* intensities than their neighbors or national and state averages.
3. Clustering of COVID-19 *mortality rate* data across all the states and its correspondence with clusters of *case rate* to estimate the relative mortality risks in a state. Our analysis techniques and insights obtained can help the local and state governments in monitoring and managing COVID-19 situation and fine-tuning the ongoing vaccination drive in India.
4. Dynamic cluster analysis based on the COVID-19 indicators and answering pertinent questions related to the spread of the epidemic and its changing nature.
5. Analysis of variegated (highly unequal resource distribution) medical infrastructure and its impact.

Related Work

Time series data occurs in a variety of contexts and there has been a considerable amount of research in this field [7–9]. Clustering of time-series data is frequently used to discover interesting patterns in the data. Clusters are formed by grouping objects with maximum *similarity* within the same cluster and ensuring that they have minimum similarity with objects in other clusters. Such clustering may also be useful in discovering anomalies or unexpected patterns in the dataset.

Several researchers [1–3] have studied important problems related to COVID19 in different contexts. However, they study much shorter periods than what we have analyzed in this paper and since they do not normalize the data, there is a bias with the size of population. For example, in [3], the authors analyze data in the U.S. context and reveal that the long-standing inequity issue in the U.S. stands in the way of the effective implementation of social distancing measures. In [1], the authors study multivariate time series of COVID-19 cases and deaths across all countries in the world during 12/31/2019 to 04/30/2020 and study the similarity in the evolution of cases and deaths. While they clustered

Table 1 Definitions of different indicators studied in the paper along with section reference

Indicator	Definition	Population group
Absolute new cases	Daily new cases in absolute numbers	State, District (Sect. 6.1)
Absolute new deaths	Daily new deaths in absolute numbers	State, District (Sect. 6.1)
Case rate	Daily new cases normalized by the population	State, District (Sect. 5.1)
Mortality rate	Daily new deaths normalized by the population	State (Sect. 5.3)
New cases by beds	Daily new cases normalized by total beds	State, District (Sect. 6.2)

the countries based on the daily statistics and performed change analysis, we cluster the states and districts of India according to the pattern/shape of the time series in a given period and use cluster correspondence to understand the changes between clusters of different metrics. Our methods are flexible and can be applied to shorter or longer periods as desired.

There have been several studies [4–6] in the Indian context as well. None of the studies thus far, have jointly analyzed COVID-19 trends in all the districts of India over a long period of time. [4], provides a good historical account of COVID-19 in India until May 1, 2020. In this paper, several statistical models were developed to categorize the states as severe, moderate, or controlled. However, the methodology for categorization was not systematic.

Several research papers [10, 11] have analyzed the stability of a given clustering algorithm while varying its parameters, and to compare clusters yielded by different algorithms, using comparison schemes based on matchings, information theory, and use of various indices (Rand, Jaccard). This was generalized to accommodate many-to-many matchings between clusters, via the *D-family-matching* on the *intersection graph*, with D as the upper bound on the diameter of the graph induced by the clusters of any meta-cluster by [12]. While this problem is NP-complete and hard to approximate, a polynomial time, spanning tree based heuristic was presented. In Sect. 4, we present an example illustrating that the optimal *D-family matching* is not guaranteed to cover all the nodes (clusters) in the *intersection graph* even when a max-weighted matching can do so. To the best of our knowledge, most of the cluster correspondence studies have focused on the clusters produced on the same dataset, while we generalize the application of maximum weighted matching to clusters produced on two different datasets (e.g. district clusters and state clusters) by defining the edge weights in an intuitive and novel manner.

Data Collection and Preparation

The primary source of data used in this paper are [13–15]. The data for daily new cases and deaths have been collected from March 2020 to May 2021 for each district and state in India. The states are referred by their respective codes as given in [16] and national average by ‘TT’. The collected data is in the form of a time series.

Since we are dealing with normalized data at a granular level, the time series of districts having a small population are too erratic and unstable for proper analysis of the clusters. Based on the data of population of districts, the threshold for the chunk of districts with small population comes out to be 75,000 as shown in Fig. 1. Therefore, districts with population lying below the threshold have been omitted in the analysis and clustering.

The time series for the daily new cases in districts have been up-sampled to weekly new cases to counter instability in reporting of cases and ‘*Sunday anomalies*’ (a consistent drop in daily new cases reported on Mondays compared to other weekdays. This is due to lower number of tests performed on Sundays) observed due to administrative reasons as shown in Fig. 2. The time-series for the states have been clustered on a daily basis.

Many of the studies and reporting of COVID-19 data uses the states as a dimension of reporting. However, each state varies widely in its area and population and, therefore, causes bias towards larger states. In general, states with large population will usually have larger number of COVID-19 patients as evident in Fig. 3. Thus, to remove this bias, we decided to normalize the data by dividing the number of COVID-19 patients with the total population of each state. This allows us to assess the relative situation of regions with a small population like north-eastern states and union territories of India.

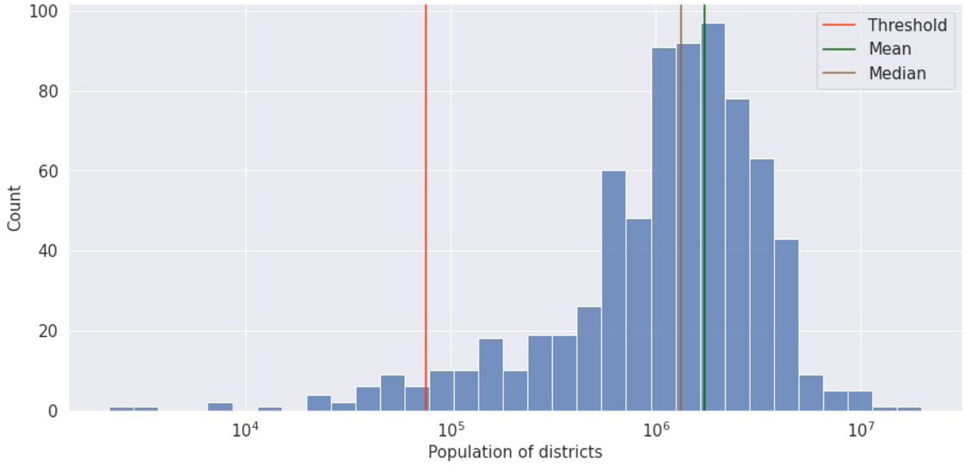


Fig. 1 A log scaled histogram of population of districts in India along with highlighted mean, median and threshold used for filtering districts

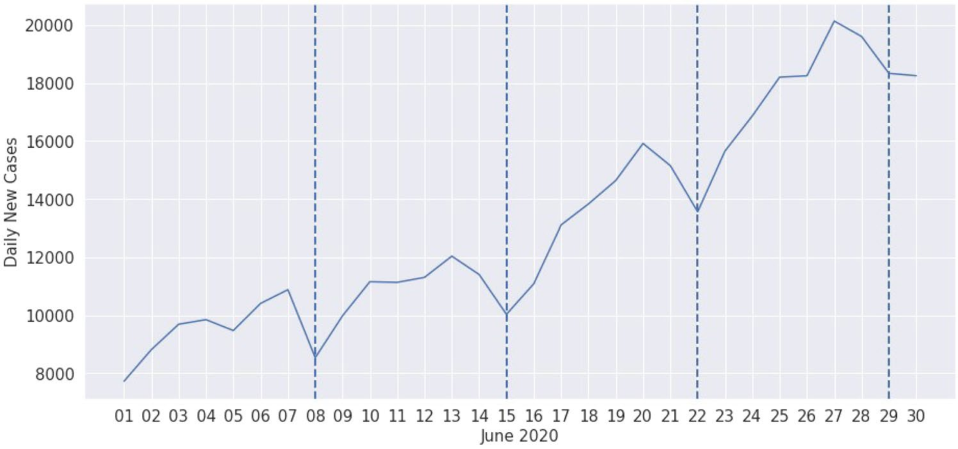


Fig. 2 The above plot shows the absolute number of daily new cases in India in the month of June 2020. The dotted lines denote the Mondays where a dip is observed due to lower number of tests performed on Sundays

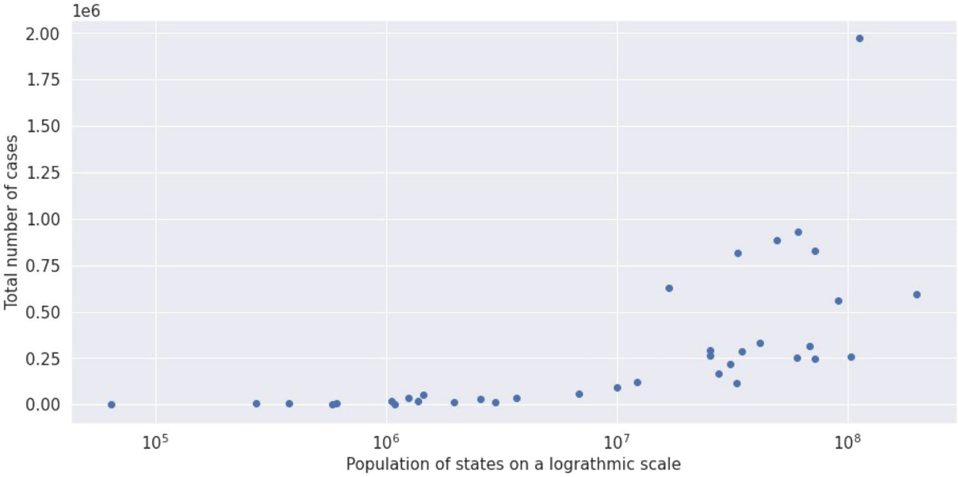


Fig. 3 The above scatter plot shows the total number of cases in states up to January 2021 against their population

Model and Techniques

We now describe the major techniques used in the paper.

Agglomerative Clustering

Hierarchical clustering is a method of cluster analysis used in data mining and statistics, to build a hierarchy of clusters. Agglomerative Clustering is a strategy for hierarchical clustering in which each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. We use agglomerative clustering [8] on the normalized time series.

We define *case rate* for region *r* on day *t* as

$$c_t^r = \frac{\text{Total number of new cases in region } r \text{ on day } t}{\text{Population of region } r} \tag{1}$$

We define *mortality rate* for region *r* on day *t* as

$$m_t^r = \frac{\text{Total number of deaths in region } r \text{ on day } t}{\text{Population of region } r} \tag{2}$$

Distance Matrix

Our primary objective is to compare the patterns of rise and decline of COVID19 across various states and districts in India; therefore, we have used a shape-based approach for time series clustering. Some of the districts and states also experienced multiple waves of COVID-19 infections. The DTW metric as defined by [17] is suitable for this analysis. As noted in [18, 19], this metric has better accuracy than the traditional Euclidean distance metric for time-series clustering applications. Dynamic Time Warping (DTW) [17] score helps us to measure the similarity between two time series based on their shapes. The lower the DTW score $DTW(x,y)$ of a pair of time series x and y, the similar they are, with $DTW(x,x)=0$. DTW is a symmetric function, with $DTW(x,y)=DTW(y,x)$. The distance matrix for the above clustering was prepared based on this score.

Dendrogram

The results of a hierarchical clustering is commonly represented by a dendrogram, a tree-like diagram which describes the series of steps taken by the clustering technique from n distinct singleton clusters to a single cluster containing all n individuals. The algorithm to return k clusters for a given dendrogram is given in Algorithm 1. This function will return k clusters if at least k leaf nodes are present in the tree otherwise it returns the number of clusters equal to the number of leaf nodes. Here, we always divide the largest subtree available to us at each point of time, in case it can be divided further.

Once the optimal number of clusters, K has been determined, we use the algorithm discussed in Sect. 4.3 to automatically generate the K clusters and their members. This is especially important for studying cluster dynamics in Sect. 6 where we generate clusters automatically over multiple time windows. It is also important for our dashboard which will update automatically.

Algorithm1 FindKClusters

```
Input:
    root : the root of the full binary tree representing the dendrogram.
    k: the number of clusters that is needed to be created.
Output:
    clusters : list of k clusters, where each cluster is the list of nodes present in the cluster.
procedure Segregate Clusters
    if root is NULL then return                              ▷ Clusters cannot be formed.
    if root is a leaf node then return [root]                ▷ There exists only one cluster.
        ▷ rootNodes: list of roots of the candidate subtrees which can be further divided.
    rootNodes ← [root]                                       ▷ It is initialised with root.
        ▷ The following loop continues till all the candidate subtrees are exhausted or k clusters
    are already obtained.
    while length of rootNodes > 0 and k > 1 do
        maxRootNode ← node in rootNodes with maximum size of the subtree rooted at it.
        if maxRootNode is a leaf node then break ▷ Other's are definitely leaf nodes.
            ▷ maxRootNode is removed from rootNodes, divided into two subtrees and root(if exists)
        of each subtrees(left and right) is inserted into rootNodes and k is decreased by 1.
        delete maxRootNode from rootNodes
        if maxRootNode->left is not NULL then insert maxRootNode->left to rootNodes
        if maxRootNode->right is not NULL then insert maxRootNode->right to rootNodes
        k ← k − 1
        ▷ When the above while loop breaks, all the k clusters (if possible) are obtained, with
    roots of each of them present in the list rootNodes.
    clusters ← [][]                    ▷ clusters: list of lists of leafnodes present in each cluster.
        ▷ For each of the nodes present in rootNodes, leaf nodes of the subtree of rooted at that
    node is obtained using any tree traversal method(e.g, DFS).
    i ← 0
    while i < length of rootNodes do
        clusters[i] ← leaf nodes in subtree rooted at rootNodes[i]
        i ← i + 1
    return clusters
```

Elbow Method

Elbow method is a heuristic method used in *time series K-means clustering* to get the optimal number of clusters based on a cost function. We define the following cost function with *Residual Sum of Squares (RSS)* with an added term, which penalizes high number of clusters (k).

$$\mathrm{Cost} = RSS + \alpha \times \log\ k \tag{3}$$

Cluster Correspondence

For the clusters obtained through the above method for states and districts in *case rate* analysis and states in *mortality rate* analysis, we use a novel cluster correspondence method based on maximum-weighted matching. As discussed in Sect. 2, *D-family-matching* is the generalized way to draw cluster correspondences. Here we present an example to show that the optimal *D-family matching* is not guaranteed to cover all the clusters in the *intersection graph*. As shown in

Fig. 4, the clusters *c* and *d* remain unmatched in the optimal *D-family matching* with $D = 2$ even when some weak relationships are present between them. The underlying assumption of the similarity metric in *D-family matching* (taken as the intersection of the two clusters) is that the clusters are produced from the same dataset.

In this paper, we are dealing with clusters obtained through two different datasets which are therefore, non-uniform in size and could represent different metrics. We present a novel way of using a specialized edge-weighing formulation in intersection graph and to find the correspondences through maximum-weighted bipartite matching of the

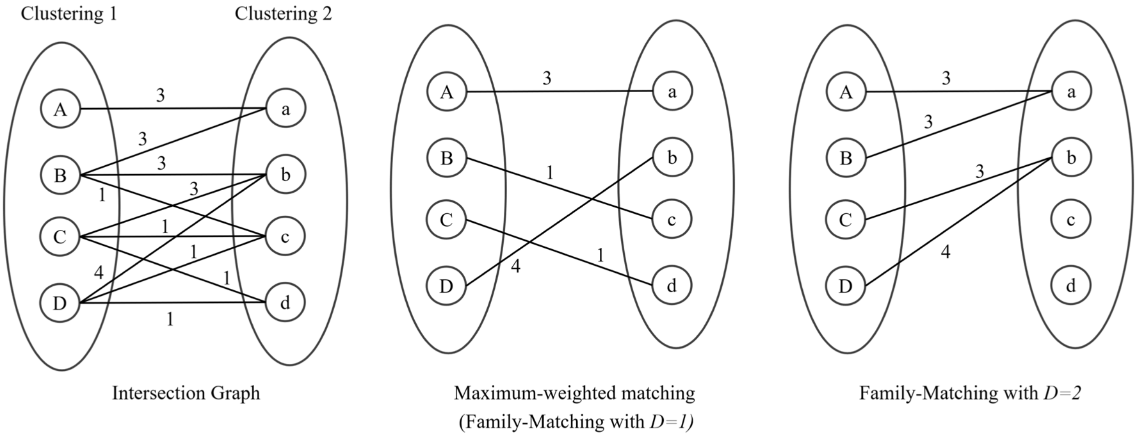


Fig. 4 Example to show that the optimal *D-family matching* is not guaranteed to cover all the clusters for all values of D

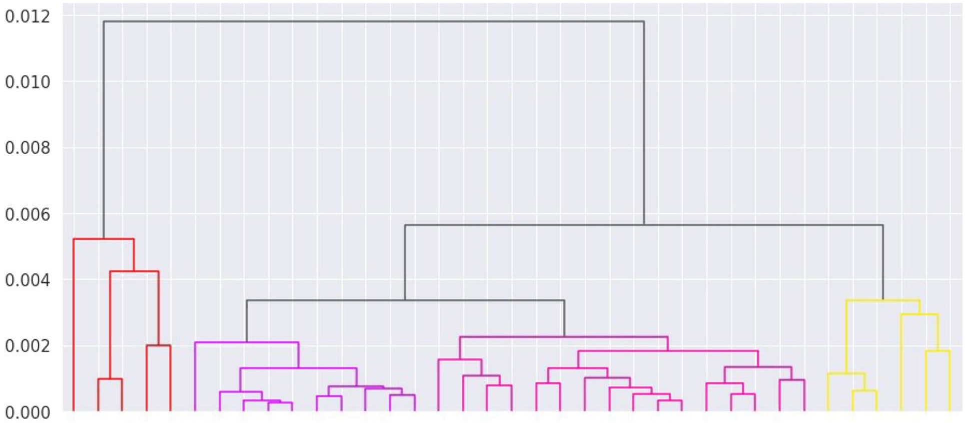


Fig. 5 Dendrogram for states using Complete Linkage in *Case Rate* analysis, showing 4 distinct clusters (red, purple, pink and yellow)

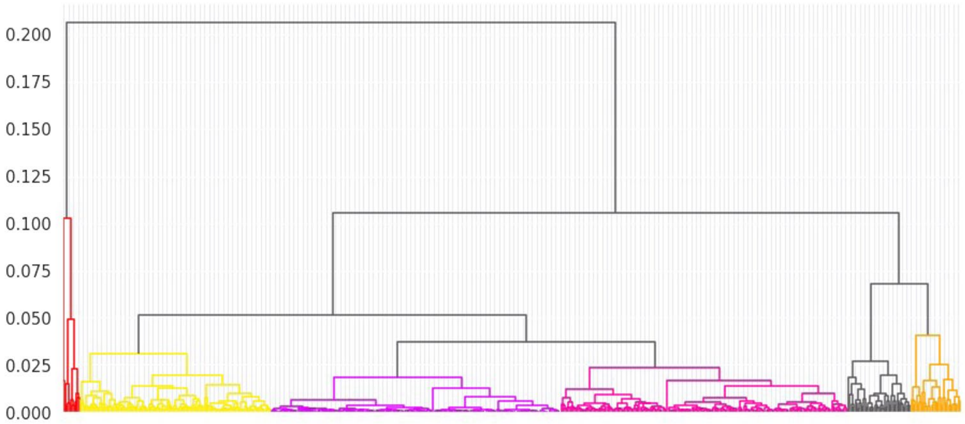


Fig. 6 Dendrogram for districts using Complete Linkage in *Case Rate* analysis, showing 6 distinct clusters (red, yellow, purple, pink, grey and orange)

graph in section [5.2, 5.3], figure [12, 15]. It is noteworthy that our edge-weight metric can be generalized to compute similarity between two sets containing objects at different levels of hierarchies in a hierarchical dataset.

Clustering Analysis

In this section, we apply the agglomerative time series clustering using DTW scores for the *case rate* of states and *case rate* of districts. The DTW score matrix for the *case rate* clustering can be visualized through a heatmap in Fig. 7. After obtaining the clusterings of states and districts, we follow the cluster correspondence between the two as specified in Sect. 4. A similar correspondence is also performed with the *mortality rate* clusters against the *case rate* clusters of the states.

Clusters of States

In DTW function, we can use a parameter known as Sakoe-Chiba radius *r*, which indicates the off-diagonal elements to be considered, also called warping window size. The advantage of using this is to get temporal grouping and two time series would belong to the same cluster iff they show similar shapes within a time range of *r* days. For the purpose of this paper, we choose *r* = 7 while computing DTW score for every pair of states. This makes the clustering robust to Sunday anomalies and helps us in coping up with any delay in reporting of cases. However, our data was not sensitive to

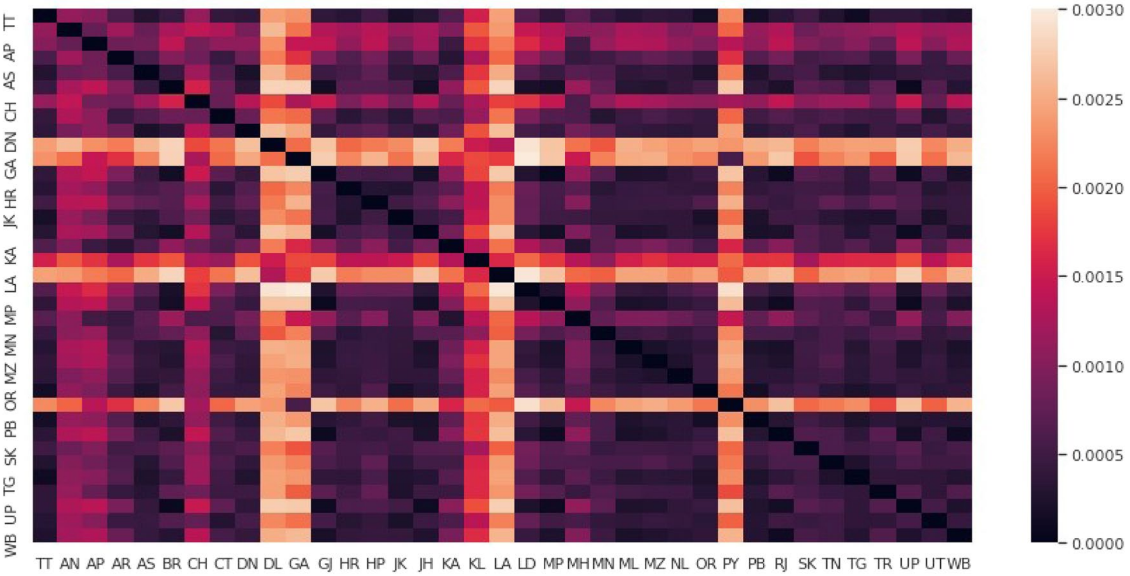
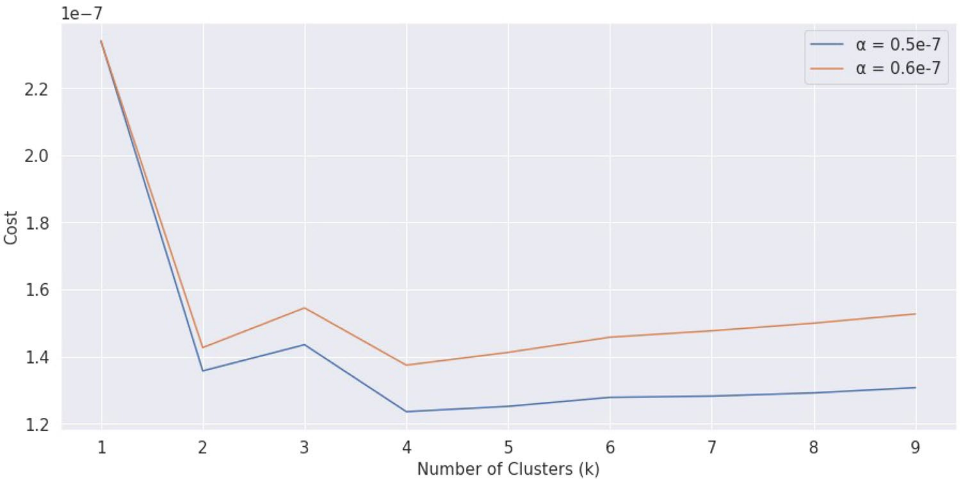


Fig. 7 Heatmap of DTW scores for the states computed on the *case rate* time series during 14 March 2020 to 11 Jan 2021, with Sakoe-Chiba radius = 7

Fig. 8 Deducing the number of clusters using elbow method in Time Series K-means clustering. The cost function used here is defined in section [4]



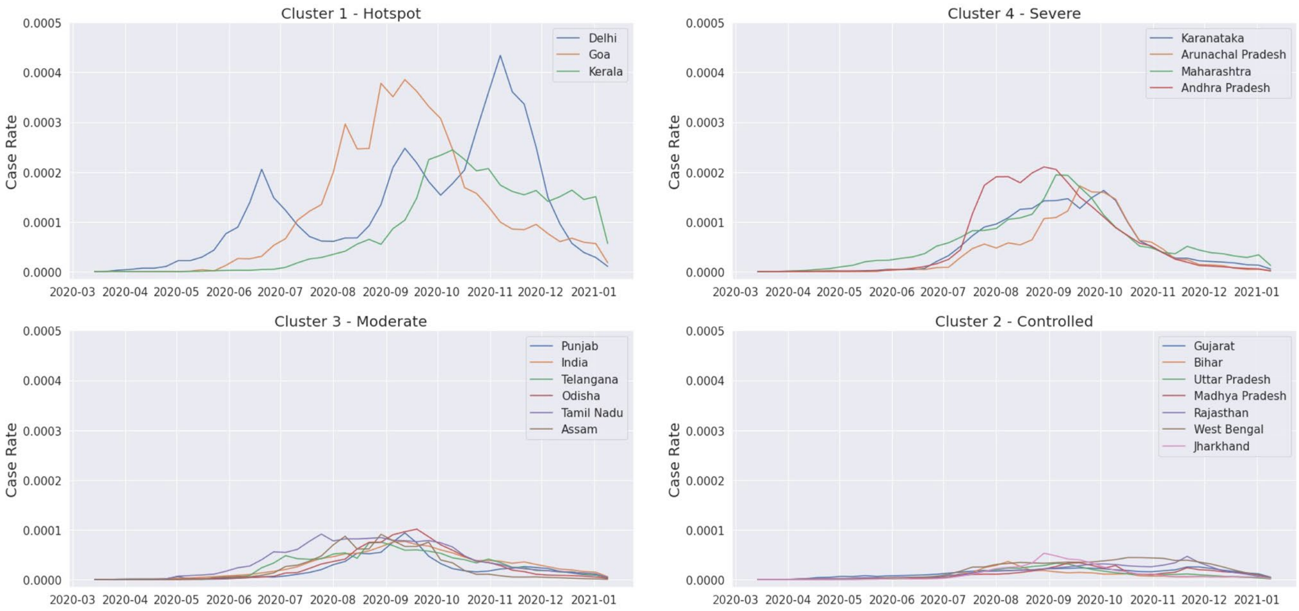
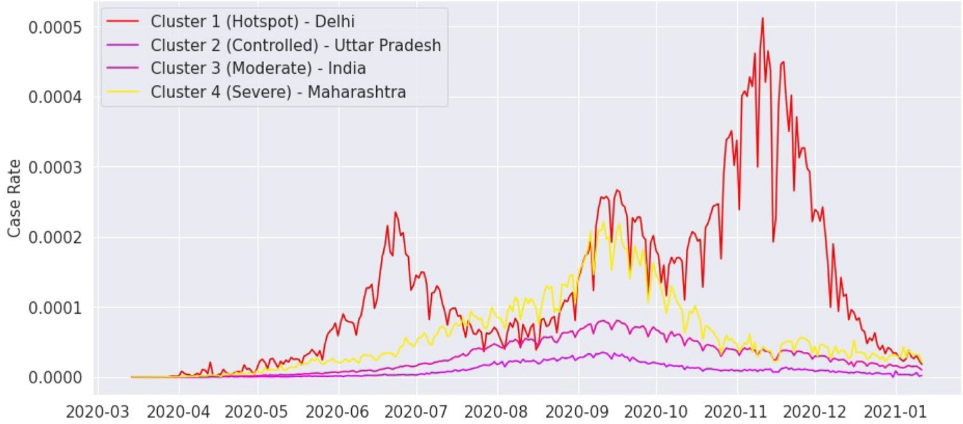


Fig. 9 *Case Rate* of the four state clusters are shown in the above plots with the same limits on the *y*-axis

Fig. 10 The inter-cluster plot for the states in *Case Rate* analysis to visually highlight the difference among the cluster representatives



"*r*" as we got negligible changes in final clustering, which indicates that our clustering is also temporally coherent.

After getting the above DTW score matrix for districts and states, the dendrogram is obtained using complete linkage [20] in the agglomerative clustering as shown in the Figs. 5 and 6. Complete linkage is preferred in this case since it is less susceptible to noise, outliers and tends to break large clusters. The elbow method as discussed in Sect. 4.4 is deployed in the above clustering using K-means (Fig. 7). It indicates the optimal number of clusters to be 4 as shown in Fig. 8. Therefore, if we consider top-down traversal of the dendrogram shown in Fig. 5 as a binary tree, we get the left subtree (red) and the right subtree (purple, pink and yellow). Since, the right subtree is large in size, we split it up. We move forward with the left sub-subtree formed in the second split and divide the same into two smaller subtrees—purple and pink. Hence, we obtain four clusters, as shown with different colors in Fig. 5.

We now provide a brief explanation of the time series clusters obtained and characterize the same.

Cluster 1: This cluster has the highest peak of *case rate*. The states of this cluster may well be labeled as the "hotspot states" since they experienced the highest *case rate* with multiple peaks, shown in Fig. 9.

Cluster 2: The plateau like nature with no sharp peaks depicted by the curve of the states belonging to this cluster, indicates that these states performed very well in comparison to other states (Fig. 10). Thus, these states have a controlled and subdued outbreak of the pandemic. Interestingly, these states represent a compact geographical region in the northern plains of India as shown in Fig. 11.

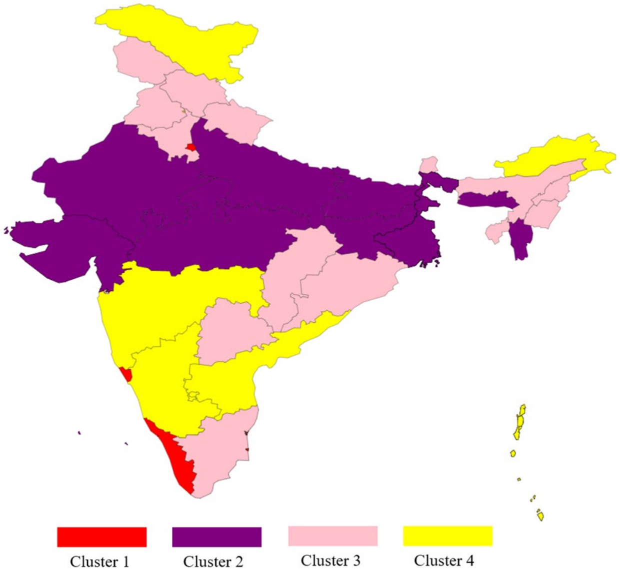


Fig. 11 State Clusters on national map for *Case Rate* analysis showing states distributed in four distinct clusters: “Hotspot” (Red), “Controlled” (Purple), “Moderate”(Pink) and “Severe” (Yellow). It highlights geographical coherence of the clusters

Cluster 3: This cluster indicates steady rise and decline of the *case rate*. The national average, being an entity, also belongs to this cluster. Hence we can safely conclude that this cluster represents a moderate outbreak close to the national average.

Cluster 4: This cluster may be regarded as the set of states which have experienced significantly high *case rate* than the national average but have still performed better than the hotspot states. As evident in Fig. 9, there is a consistent rise and decline in the states with the peak being just relatively lower than the hotspot states. Geographically, most of these states represent the south-central region of India.

We can now try to summarise the entire picture using the below inter-cluster plot of states shown in Fig. 10. For this figure, a representative state has been picked up from each of the above four clusters. Note that for the third cluster, we have picked the country’s average as the representative since it was also an entity in the above clustering which is in the third cluster.

To compare the current model for clustering, we implement the standard Eucledian distance instead of DTW score between two time-series as discussed in Sect. 4.2. On comparing the clusters produced by DTW model with that of the clusters produced by Eucledian model, we observe the following differences:

1. Kerala shifts from hotspot cluster to severe cluster.
2. Nagaland shifts from moderate cluster to controlled cluster.

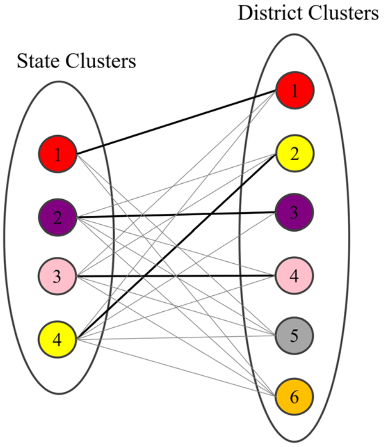


Fig. 12 The maximum matching is shown in the above bipartite graph of cluster correspondence between State Clusters and District Clusters. It shows all the edges with positive weights. The highlighted edges are the matching edges

This is mainly due to sakoe chiba radius (explained is Sect. 5.1), which is used in calculation of the DTW score, being absent in this case. This makes the Eucledian distance between Kerala and other time series of *Hotspot Cluster* larger than that of Kerala and other time series of *Severe Cluster*. So, Kerala shifts to Cluster 4 from Cluster 1 in this case. The same argument remains valid for Nagaland.

Clusters of Districts

After a top-down traversal of the *dendrogram*, we find out six different clusters as shown in Fig. 6. However, on applying *elbow method* for the time series of *case rate* in districts, we find out the optimal number of clusters is 4. We apply Cluster Correspondence technique, specified in Sect. 4, to find correspondence between the state clusters and the district clusters and to get the four prominent district clusters out of the six. Here the state clusters belong to one group of nodes in a bipartite graph and the district clusters occupies the other group. Each node of a group is connected with every node of the other. For any state cluster C_i, C_i^d is defined as the set of all districts in all the states belonging to cluster C_i. The weight of the edge between state cluster C_i and district cluster D_j, denoted by w_{ij}, is defined as the ratio of the number of unique districts of cluster D_j that belongs to any of the states of cluster C_i to the total number of distinct districts of cluster D_j along with all the districts of all the states of cluster C_i. It is given by,

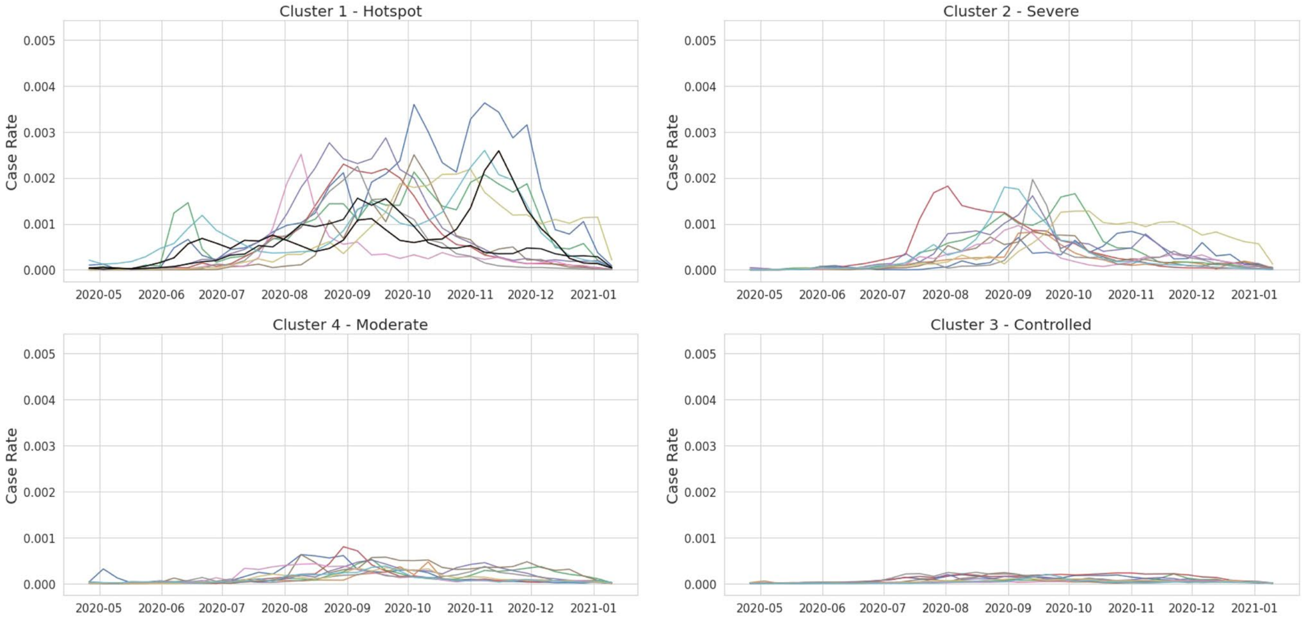
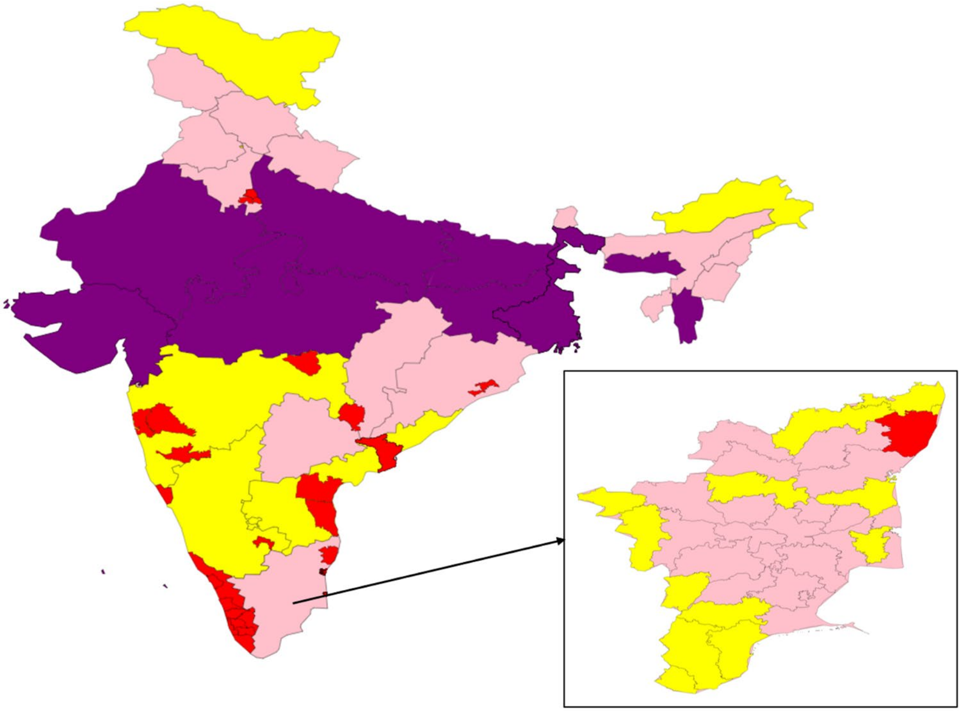


Fig. 13 *Case Rate* of the four district clusters are shown in the above plots with same limits on *y*-axis by taking 10 random districts in each cluster

Fig. 14 States Cluster on national map overlayed with hotspot districts (Red) for *Case Rate* analysis which shows big cities are more vulnerable than its rural counterparts. States are distributed in four distinct clusters: "Hotspot" (Red), "Controlled" (Purple), "Moderate" (Pink) and "Severe" (Yellow). A magnified visualisation of Tamil Nadu with district clusters is shown



$$w_{ij} = \frac{\left|C_i^d \cap D_j\right|}{\left|C_i^d \cup D_j\right|}$$

where $C_i^d = \bigcup_{s_p \in C_i} \bigcup_{d_q \in s_p} d_q$

The above formulation of the edge weights is related to the *Jaccard* similarity metric and can be extended to any hierarchical dataset.

Based on the maximum weighted matching, we find a one-to-one correspondence of four out of six of the district clusters with the four clusters obtained in case of states as shown in Fig. 12. Furthermore, we observe that the time series in district Clusters 5 (grey) and 6 (orange) have a relatively small distance to State Clusters 1 and 2. As *elbow method* had indicated four clusters as being optimal, there is not much benefit to keep them as separate clusters. Therefore, we pick up the entities of the remaining two clusters (grey and orange), and find the average distance of each of these entities from the other four clusters. We assign each entity to one of the four clusters whose average distance is minimum from it. This integration is also evident from Fig. 13 where two districts, one from Cluster 5 and the other from Cluster 6 which got merged with Cluster 1 are plotted in *black* colour. Both of these curves gel well with the other Cluster 1 districts.

Some of the salient observations in district clusters:

- In general, the districts of major cities of the country are present in District Cluster 1: *Delhi, Pune, Nagpur, Gurugram* and *Bengaluru Urban* among other major city districts. It is interesting to note that there does not exist even a single district from *State Cluster 2—Controlled* which captures a significant geographical area in Central India as shown in Fig. 14.
- In Cluster 4, we get districts from 23 states of India which is the highest among all the clusters. This diversification serves as a strong indicator of the fact that this cluster represents the country's average. Moreover, the national average *case rate* time series ends up in this cluster, in the case of districts as well.

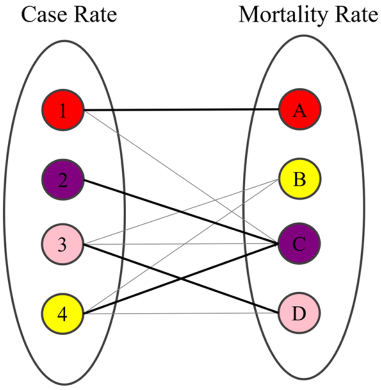


Fig. 15 The maximum matching is shown in the above bipartite graph of cluster correspondence between *Case Rate* and *Mortality Rate* analysis. It shows all the edges with positive weights and the highlighted edges are the matching edges

Below is the map of India with the state clusters as discussed above. On top of this, the hotspot districts are also overlayed on the map in Fig. 14. Note that we observe no or very few red patches in "Controlled Cluster" or "Moderate Cluster".

Also, we take a closer look at one of the states of India, *Tamil Nadu* in Fig. 14 where districts clusters are shown. Tamil Nadu shares its borders with Kerala in the west, which belongs to *Hotspot Cluster (Red)*. In north, Tamil Nadu shares its border with Andhra Pradesh, which belongs to *Severe Cluster (Yellow)*. The districts on the west border and north border of this state mostly belongs to the *Severe Cluster (Yellow)*. This district clusters distribution can be explained through its proximity with severely affected regions as mentioned above (Fig. 15).

Mortality Rate Analysis

We apply the above methods for clustering and analysis of *mortality rate* timeseries at the state level. We obtain four distinct clusters of the states. The *mortality rate* time series for a representative state from each of the clusters is plotted in Fig. 16.

To draw the cluster correspondence between the *case rate* and the *mortality rate*, we construct a similar bipartite graph. The *mortality rate* clusters are labelled *A, B, C* and *D*. The weight of the edge between *case rate* cluster c_i and *mortality*

Fig. 16 The inter cluster plot for the states in *Mortality Rate* analysis to visually highlight the difference among the cluster representatives

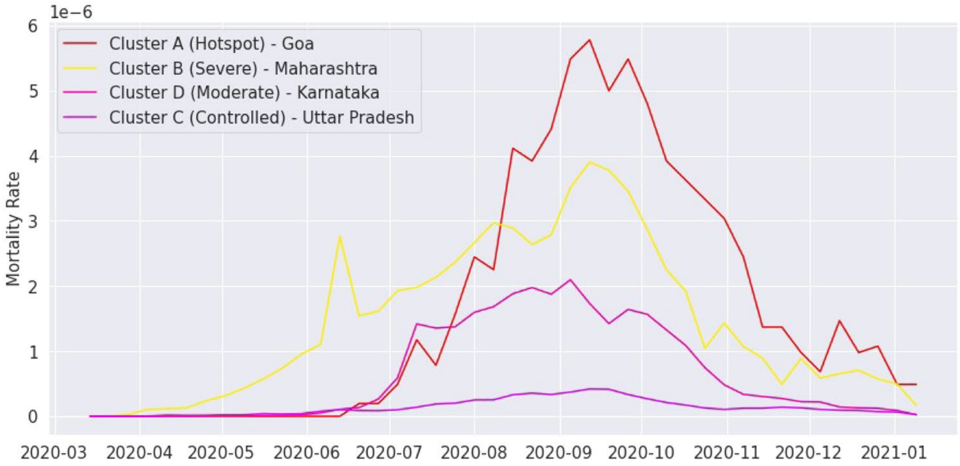


Table 2 Intersection of clusters from *Case Rate Clusters(1,2,3,4)* and *Mortality Rate Clusters(A,B,C,D)*

	Cluster A: Hotspot	Cluster B: Severe	Cluster C: Controlled	Cluster D: Moderate
Cluster 1: Hotspot	*DL, PY, GA, LA*		KL	
Cluster 2: Controlled			*RJ, MP, JH, GJ, ML, LD, WB, BR, MZ, UP*	
Cluster 3: Moderate		SK	AS, TT, HR, NL, OR, TG, DN	*CT, MN, TN, TR, HP, UT, PB, JK*
Cluster 4: Severe		*CH, MH, AN*	AR	AP, KA

rate cluster m_j, is again given by Jaccard's similarity denoted by w_{ij}. It is defined as the ratio of the number of states of cluster m_j that belongs to cluster c_i to the total number of distinct states of cluster c_i and cluster m_j. It is given by,

$$w_{ij} = \frac{\left|c_i \cap m_j\right|}{\left|c_i \cup m_j\right|}$$

Based on the maximum weighted matching obtained, we get a one-to-one correspondence of *case rate* and *mortality rate* clusters as shown in Fig. 15. However, some states, on an individual level, do not follow the correspondence. This implies that states with a high *case rate* might have low *mortality rate*. Such inconsistencies might be interpreted in terms of states having distinct healthcare facilities, median age-group or other factors determining mortality risks of COVID-19 [1]. In Table 2, each cell (*i,j*) represents the intersection of cluster of *i*th row and *j*th column. The italicized cells denote one-to-one correspondence between the related clusters. The ordering of the clusters according to the mortality rate, from highest to the lowest is, $A > B > D > C$. Suppose a cluster of *i*th row has a one-to-one correspondence with cluster of *j*th column, then any state in cell (*i,p*) has performed relatively bad if cluster of *p*th column is higher in the ordering than cluster of *j*th column and vice versa. For example, Kerala(KL) in cell $(1, C)$ has performed better in terms of *mortality rate* as Cluster 1 has a one-to-one correspondence with cluster A, but Kerala ends up in Cluster *C*. On the other hand, Sikkim(SK) in cell $(3, B)$ has performed worse in terms of *mortality rate* as Cluster 3 has a one-to-one correspondence with cluster *D*, but Sikkim ends up in cluster *B*.

A major takeaway of the clustering analysis is that the clusters tends to form a compact geographical region. This suggests that distance contiguity is a major factor in the spread of COVID-19. As expected, we get the districts of the big cities of the country in the hotspot cluster like Delhi, Pune and Bengaluru. This is another indicator of the fact that regions with good connectivity and heavy economic activities are the most susceptible regions. Moreover, we see that the number of cases in a region is not a true indicator of its capability to handle the pandemic. As discussed in the analysis of mortality rate, certain regions are more vulnerable in terms of fatalities. We now look at more questions such as the following:

1. Can we distinguish the waves of COVID-19? Did each state experience the COVID-19 waves at the same time?
2. Were some states badly hit by second wave in comparison to the first? Did some states manage the second wave better than the others?
3. How much medical resources were utilized during the peak of second wave?
4. How prepared is a particular state/district for another wave which might affect more people?

Cluster Dynamism

In the analysis so far, the clusters across all the indicators were generated for the entire time period of the first wave until January 2021. In this section, we perform clustering for smaller time windows for several indicators. For every time window, generate a dendrogram using hierarchical clustering. We use the same process to determine optimal number of clusters as discussed in Sect. 5. For every cluster, we assign a rank according to its intensity for the given indicator (e.g. case rate). We visualize these ranks for all the time windows in a heatmap to get insights about the granular transitions of the states and their relative ranking for a particular indicator. We use the Algorithm 1 in Sect. 4 to return *k* clusters from the agglomerative clustering. We perform this for daily new cases normalised by population, absolute daily new cases, daily new deaths normalised by population, absolute daily new deaths.

After obtaining the *k* clusters using Algorithm 1, we rank them according to the ranking method given in the following equation. In this method, the ranking is done based on the sum of the top *n* values of each time-series, $data_j$ in a cluster,

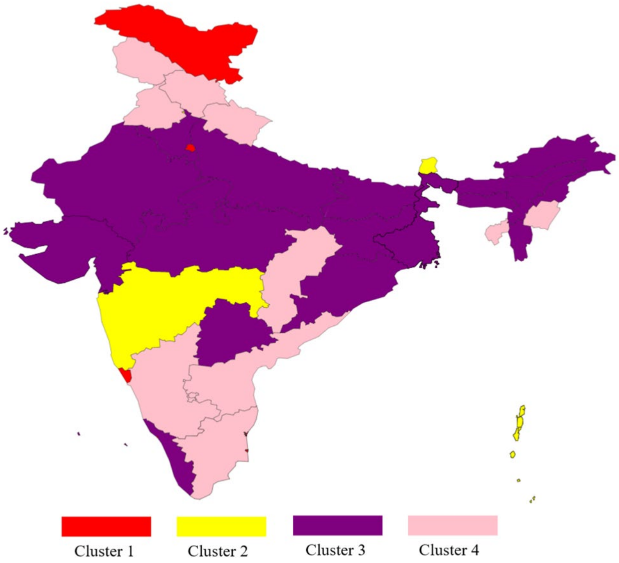


Fig. 17 State Clusters of *Mortality Rate* showing states distributed in four distinct clusters

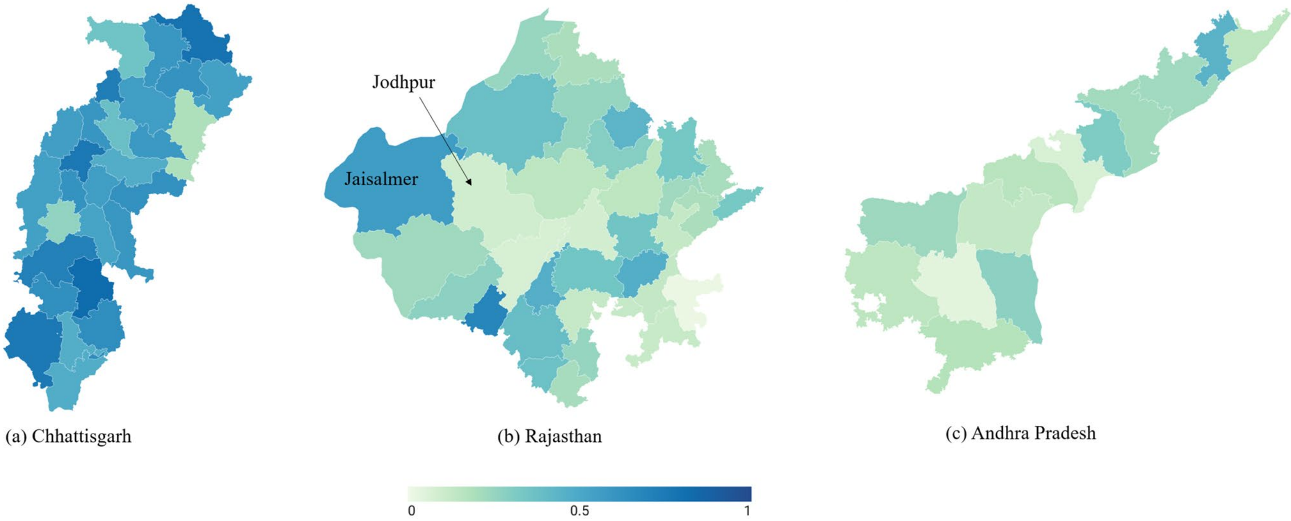
normalised by the size of the corresponding cluster. Choosing top n values (taken as $n=5$ in this case) helps us focus on the highest intensity of the indicator under investigation and at the same time reduces the effect of outliers in the data. Therefore, the ranking of the ith cluster is done based on its score S_i as shown in the equation. The cluster with highest intensity is assigned the smallest rank.

$$Score\ (S_i) = \frac{\sum_{j=1}^{\text{size of cluster } i} \text{Sum of top } n\ (\text{data}_j)}{size\ of\ cluster\ i}$$

Analysis of Dynamic Clusters Across All the Indicators

We visualize cluster dynamics with the help of heatmaps in figures [20, 21, 24]. The rows correspond to states and columns to contiguous non-overlapping time windows (Fig. 17). Furthermore, the rows are sorted (ascending) by the average rank in the whole period. This helps us quickly identify interesting patterns such as: worst affected states, periods of relative high intensity in each state, improvement or degradation relative to other regions. We have highlighted some of the interesting patterns in the figures and provided some salient observations below. The reader is encouraged to study the chart and make further observations for the periods, states and/or indicators of their interest.

We first study cluster dynamics for the absolute number of daily deaths clustered per 30 days from Mar 15, 2020 to May 6, 2021 in Fig. 22. While most deaths occurred in highly populous states like Maharashtra and Uttar Pradesh, there are notable exceptions like Delhi and Punjab. Delhi is nineteenth in terms of population in India, but is on second rank in terms of daily deaths over the whole period. Unlike most of the other states, Delhi shows more than two distinct waves. In other words, the situation seems to improve and degrade again. Gujarat, the ninth most populous state in India, is also on ninth rank in this heatmap. While in the first few months (Mar–Jun, 2020), it was one of the worst affected states, its situation improved subsequently. On the other hand, Haryana and Chhattisgarh performed poorly in second wave compared to first wave. Bihar is another exception because in-spite of its high population, the number of daily deaths is relatively low. Under-reporting of daily new deaths may cause such anomalous results. This justifies the



(a) Chhattisgarh

(b) Rajasthan

(c) Andhra Pradesh

Fig. 18 Ratio of availability of beds

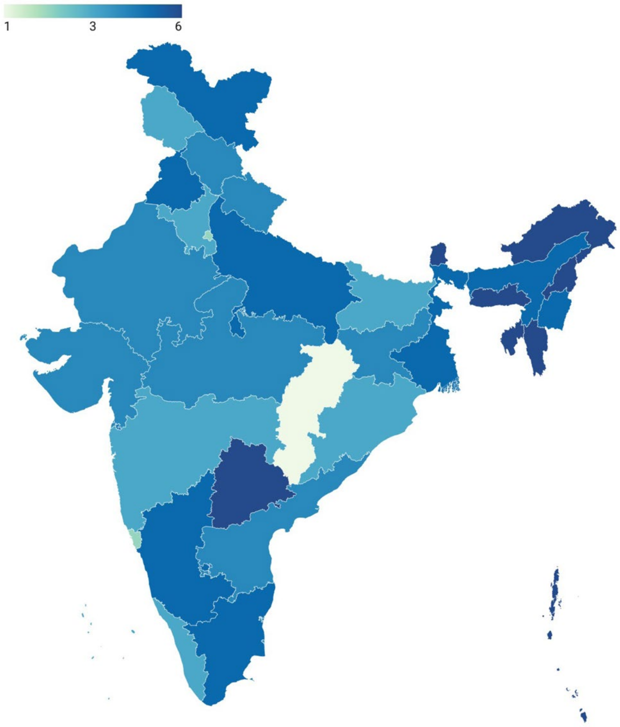


Fig. 19 Map showing the relative ranking (1–6) of the states for cases normalized by total beds in April 2021

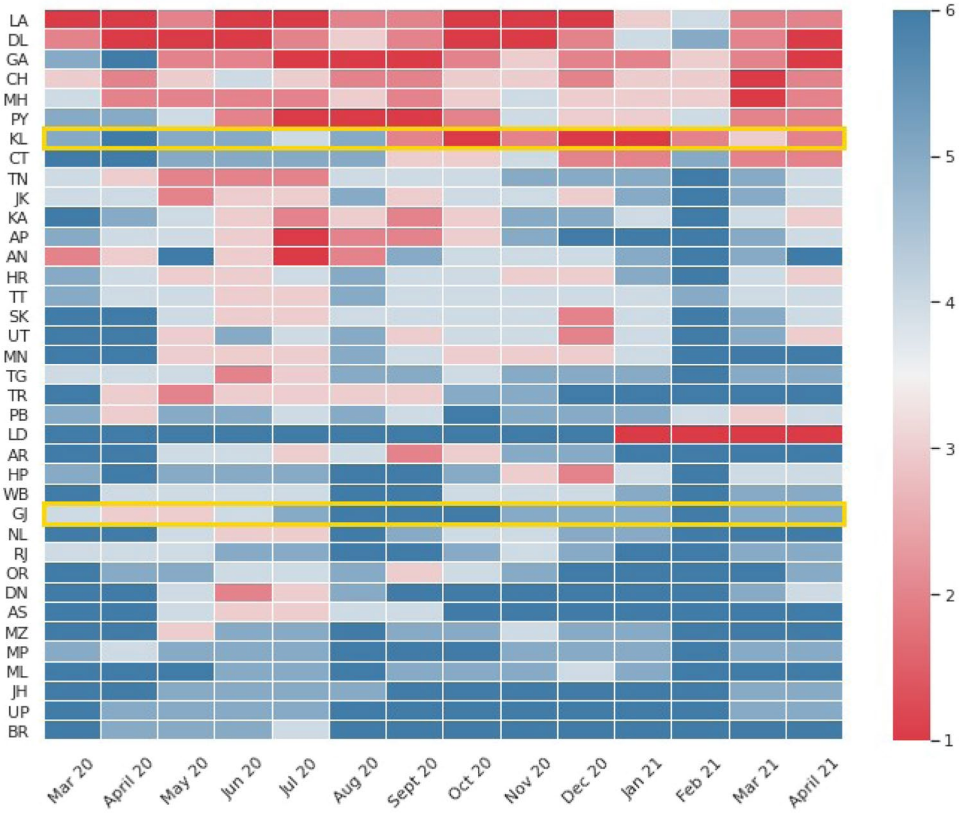
need for analysis of data with the help of normalized metrics such as *death rate* (Fig. 23).

By observing cluster dynamics of death rate in Fig. 24, we can immediately notice states with relatively smaller populations such as Chandigarh, Goa, Puducherry and Ladakh have been significantly impacted with COVID-19. The situation has been worse specially from Aug 2020 to May 2021. This figure also shows the time periods when the *mortality rate* is relatively high in a particular state. We note that each state is unique in this respect. We again note that Gujarat has managed to keep the *mortality rate* relatively low since Jun 2020. This is in part explained by the fact that Gujarat had low *case rate* since Jun 2020 as seen in Fig. 20. However, Kerala which had manged *mortality rate* during April 2020 to Nov 2020 had an opposite trend during Dec 2020 to April 2021 when the *case rate* were much higher than the rest of the country (Fig. 25).

Analysis of the Availability of Medical Resources

As discussed above, both the absolute and normalised figures do not capture the actual medical stress in a region. Normally one would expect the number of doctors, hospitals and beds in a state or district to be roughly in proportion with the population. This assumption does not hold in most parts of the country as the distribution appears to be highly skewed. To assess the situation, we take the total number

Fig. 20 Heatmap for ranking of states based on *Case Rate*



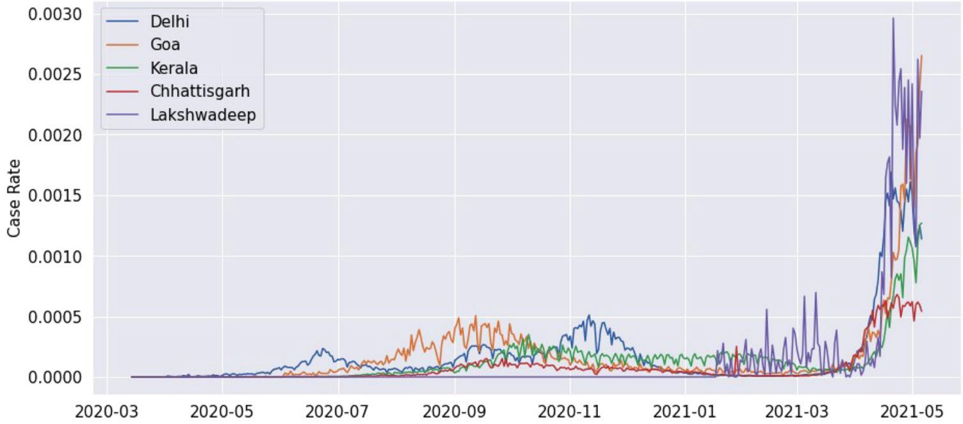


Fig. 21 Case Rate of a few states

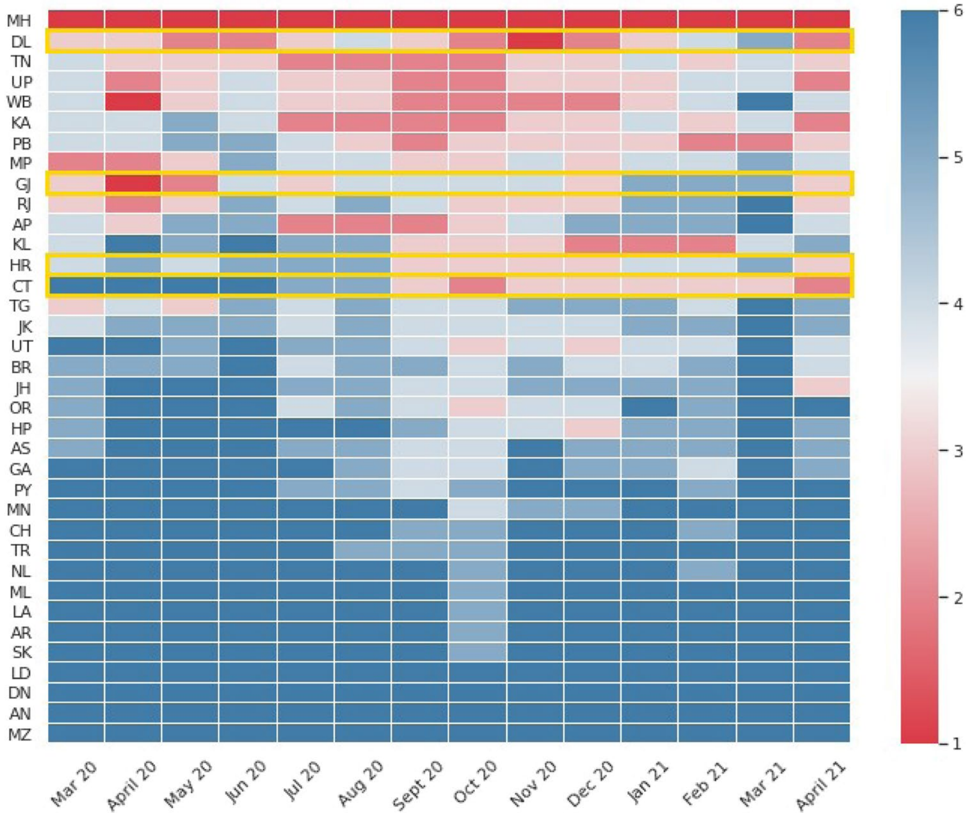


Fig. 22 Heatmap for ranking of states based on absolute daily new deaths

of beds in a state to normalise the absolute figures instead of the population while assuming that the ratio of infected people who require hospitalization remains almost same in every state. The heatmap for the daily new cases normalized by the total number of beds in the respective states is shown in Fig. 26. The data for the total number of beds in the Indian states is taken from [21].

The map for this indicator is shown in Fig. 19 for April 2021. One interesting observation from this analysis is the abrupt rise of ranking of states in the northern plains of India—Bihar, Jharkhand and Madhya Pradesh. While these states are placed around the bottom of the ranking in the daily new cases heatmap normalized by population (Fig. 20), they are placed significantly higher in the average ranking in the heatmap normalized by total number of beds (Fig. 26). This suggests that these states do not have the required number of beds in proportion with its population.

Another important metric while trying to assess the medical crunches is the availability ratio of the total beds in a region. This volatile number changes for every district every day depending on the case load and the recovery rate. We have collected the snapshot of this data on 11 May 2021 for

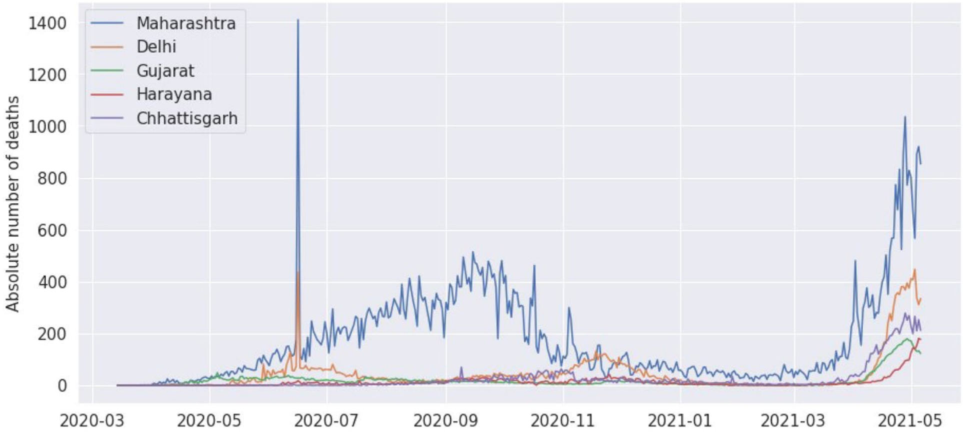


Fig. 23 Absolute number of daily new deaths for a few states

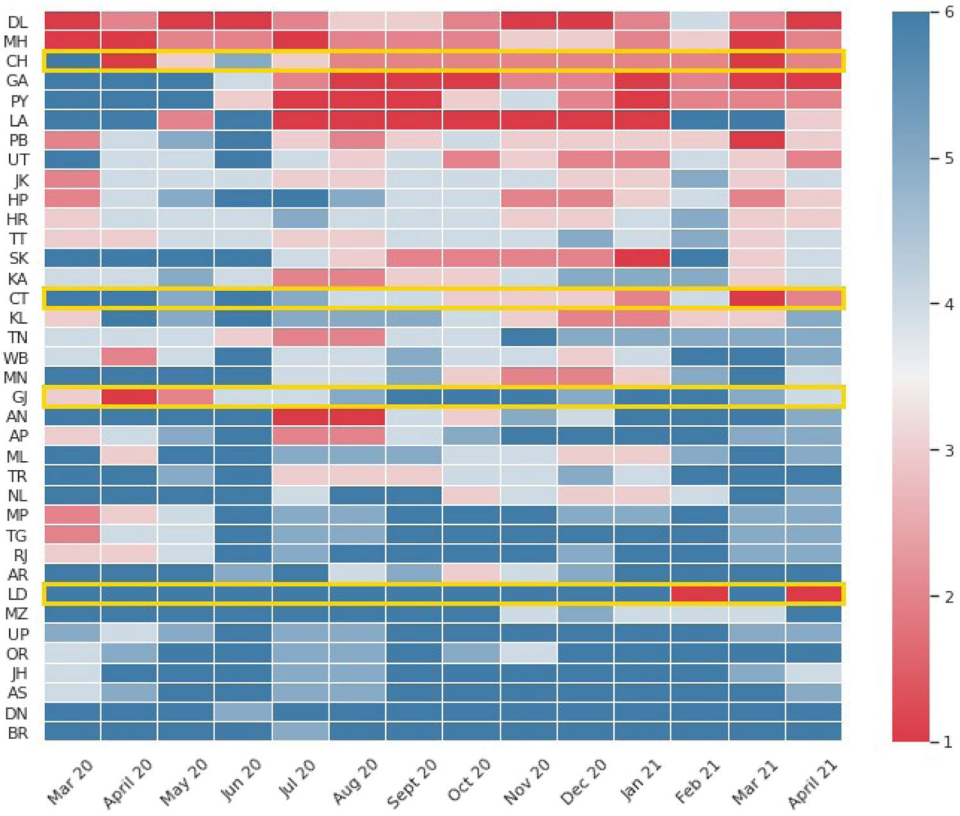


Fig. 24 Heatmap for ranking of states based on *Mortality Rate*

three states—Chhattisgarh, Rajasthan and Andhra Pradesh from their respective state government's COVID19 portal [22–24]. As evident from the district maps for these states shown in Fig. 18, there are extensive differences within a state where there are many districts with almost no beds available while a handful number of districts in the same state have this ratio close to 1. This indicates poor distribution of resources since the neighbouring regions of an overloaded district are not necessarily overloaded as one would expect. For example, Jaisalmer district in West Rajasthan has the bed availability ratio equal to 0.6 while its neighbour district, Jodhpur has this ratio equal to 0.09. Since these maps are drawn based on the snapshot of the data during the peak of the second wave, these could serve as a general distribution of availability of beds for any potential waves in the future. The state governments may plan expansion of their medical infrastructure and installment of new beds based on the same.

Fig. 25 Mortality Rate of a few states

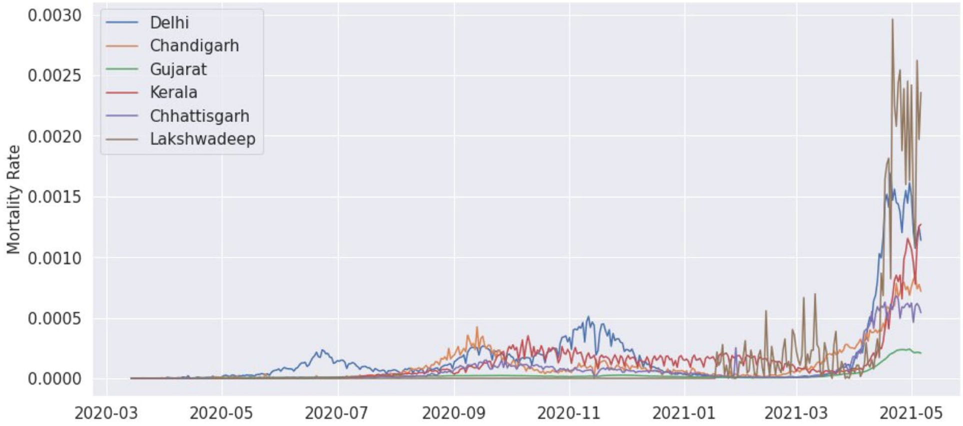
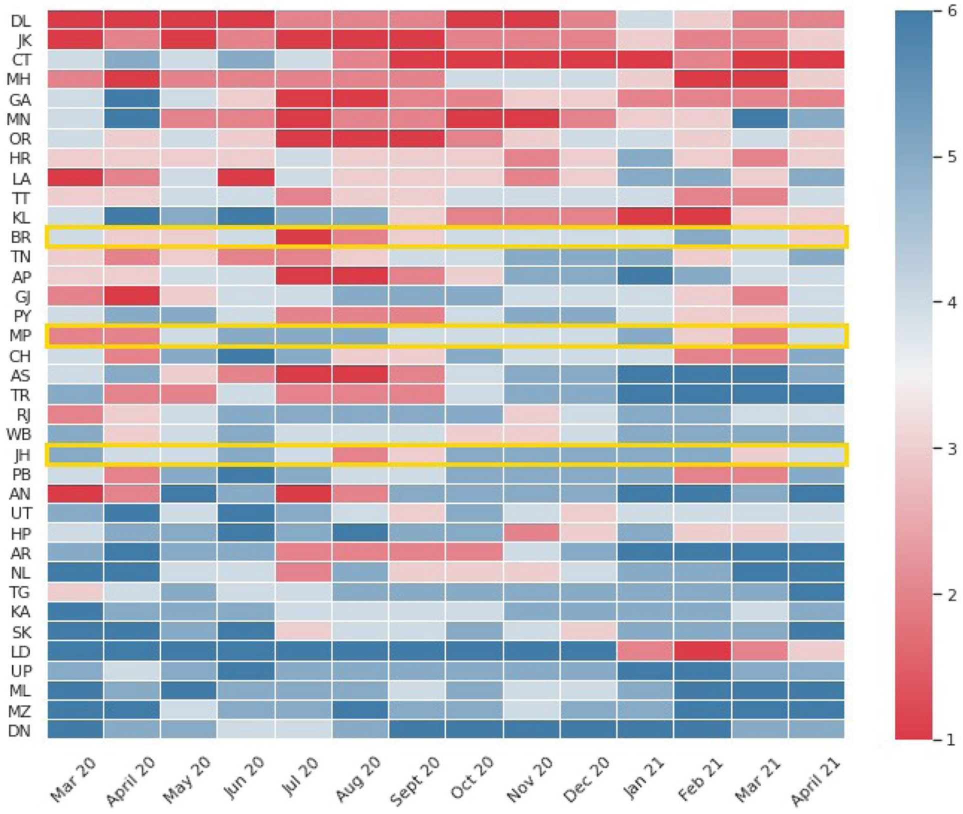


Fig. 26 Heatmap for ranking of states based on daily new cases normalized by total beds



Comparison of the First and the Second Wave

Through the cluster dynamism, we also attempt to compare the first and the second wave up to their respective peak values. Here we find the difference between the ranks of a particular state in the first wave and the second wave. Formally, we have the ranks for n time windows of a state in the first wave and the corresponding n time windows of the same state in the second wave. Since the first wave is far stretched than the second wave, the window size of the first wave is set of two weeks whereas the window size of the second wave is set to a single week for clustering.

Therefore we get 9 time windows of the first wave from 22nd May, 2020 to 17th September 2020 (18 weeks). For the second wave, the time window ranges from 4th March, 2021 to 6th May, 2021 (9 weeks).

From Fig. 27, we can observe that Chhattisgarh, a state in central India, and Kerala in southern India have a high positive value for each of the time windows which suggests that they have performed very well during the initial days of the first wave compared to the second wave.

In Table 3, we present the summary through the rankings of a few states in April 2021 for all the discussed indicators.

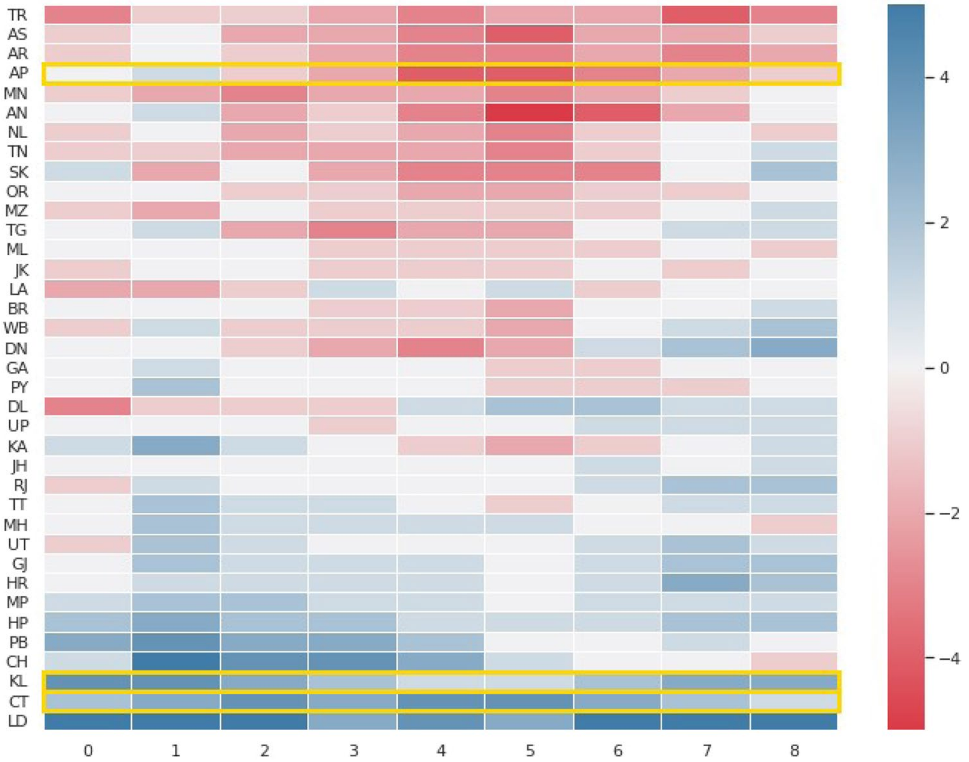


Fig. 27 Heatmap showing the comparison between the first and the second wave

Table 3 Rankings of a few states in April 2021 for all the discussed indicators

	Case rate	Mortality rate	Absolute new cases	Absolute new deaths	New cases by beds
Andhra Pradesh	4	5	3	4	4
Bihar	6	6	3	4	3
Chhattisgarh	2	2	3	2	1
Delhi	1	1	2	2	2
Maharashtra	2	2	1	1	3
Rajasthan	5	5	3	3	4

Case Rate: Daily new cases normalized by the population (Sect. 5.1), *Mortality Rate*: Daily new deaths normalized by the population (Sect. 5.3), *Absolute New Cases*: Daily new cases in absolute numbers (Sect. 6.1), *Absolute New Deaths*: Daily new deaths in absolute numbers (Sect. 6.1), *New Cases by beds*: Daily new cases normalized by total beds in the state (Sect. 6.2)

Conclusion and Future Work

We now summarize the results and key findings of the paper as follows:

1. We applied suitable techniques to cluster over 700 time-series of *case rate* across districts and states of India in a coherent manner and identified four prominent shape patterns that emerged during a period of more than 300 days. Thus, we gained additional insights for states such as Uttar Pradesh as compared to the previous work [25] where the study was on a shorter duration and data was not normalized.
2. Our novel cluster similarity (edge-weight) metric and the application of cluster correspondence using maximum edge-weighted matching was very effective in mapping between similar clusters of *case rates* of states and districts. We were able to produce informative hierarchical cartographic visualization and mine states and districts with relatively high case rate intensities.
3. We also established a useful correspondence between clusters of *case rate* and *mortality rate* at the state level and identified that some of the states like Kerala did well in reducing the mortality risks, whereas Sikkim is the only state which had relatively higher mortality risks than expected.

Most of the countries including India have opted to roll out the vaccination program in a phased manner based on decreasing order of the age-group of its citizens. This approach is based on studies [26] that suggest age and co-morbidities are the major factors of mortality risks associated with COVID-19. On top of this, a strategic vaccination drive can be planned based on the results of this paper. For this purpose, the *mortality rate* clustering can be taken as the reference since the primary objective of the vaccination drive is to mitigate the number of fatalities. In particular, the hotspot regions might be allocated relatively higher proportions of available vaccines at disposal. This strategy can be made adaptive as and when the clusters transform with time. Moreover, our techniques could be applied to other metrics like positivity rate (ratio of positive cases over total number of tests conducted), R_0-factor(number of people that one infected person is likely to spread the infection to) and percentage of critical cases. As a part of this ongoing effort, we plan to create a dashboard to monitor the various clusters, perform correspondence and dynamic change analysis.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest in this paper.

References

1. James N, Menzies M. Cluster-based dual evolution for multivariate time series: analyzing covid-19. Chaos Interdisciplinary J Nonlinear Sci. 2020;30(6):061108. https://doi.org/10.1063/5.0013156.
2. Rojas F, Valenzuela O, Rojas I. Estimation of covid-19 dynamics in the different states of the United States using time-series clustering, medRxiv 2020. https://doi.org/10.1101/2020.06.29.20142364.
3. Huang X, Li Z, Lu J, Wang S, Wei H, Chen B. Time-series clustering for home dwell time during covid-19: what can we learn from it?, ISPRS Int J GeoInf. 2020. https://www.mdpi.com/2220-9964/9/11/675
4. Ghosh P, Ghosh R, Chakraborty B. COVID-19 in India: statewise analysis and prediction. JMIR Public Health Surveill. 2020;6(3):e20341.
5. Shastri S, Singh K, Kumar S, Kour P, Mansotra V. Time series forecasting of covid19 using deep learning models: India-USA comparative case study. Chaos Solit Fract. https://doi.org/10.1016/j.chaos.2020.110227.
6. Tiwari A. Modelling and analysis of covid-19 epidemic in India. J Saf Sci Resilience. 2020;1(2):135–40.
7. Maharaj E, Caiado J, D'Urso P. Chapman and Hall/CRC. 2019. https://doi.org/10.1201/9780429058264.
8. Aghabozorgi S, Seyed Shirkhorshidi A, Ying Wah T. Time-series clustering—a decade review. Inf Syst. 2015;53:16–38.
9. Xu R, Wunsch D. Survey of clustering algorithms. Trans Neur Netw. 2005;16(3):645–78. https://doi.org/10.1109/TNN.2005.845141.
10. U. Von Luxburg, Clustering stability: an overview (2010).
11. Meila M. Comparing clusterings—an information based distance. J Multivar Anal. 2007;98(5):873–95. https://doi.org/10.1016/j.jmva.2006.11.013.
12. Cazals F, Mazauric D, Tetley R, Watrigant R. Comparing two clusterings using matchings between clusters of clusters. ACM J Exp Algorithm. 2019. https://doi.org/10.1145/3345951.
13. Covid-19 India API, Accessed on March 30, 2021. https://api.covid19india.org/
14. Covid-19 India API Sources. Accessed March 30, 2021. https://telegra.ph/Covid-19-Sources-03-19
15. India Census 2011 Dataset, Kaggle. Accessed March 30, 2021. https://www.kaggle.com/danofer/india-census
16. List of State Abbreviation, Directorate of Economics and Statistics, India. Accessed March 30, 2021. https://ddvat.gov.in/docs/List%20of%20State%20Code.pdf
17. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans Acoust Speech Signal Process. 1978;26(1):43–9. https://doi.org/10.1109/TASSP.1978.1163055.
18. Chu S, Keogh E, Hart D, Pazzani M. Iterative deepening dynamic time warping for time series, pp. 195–212. https://doi.org/10.1137/1.9781611972726.12
19. Ratanamahatana CA, Keogh E. Three myths about dynamic time warping data mining, pp. 506–510. arXiv: https://doi.org/10.1137/1.9781611972757.50
20. Tan P-N, Steinbach M, Karpatne A, Kumar V. Introduction to data mining. 2nd ed. London: Pearson; 2018.
21. Kapoor G, Hauck S, Sriram A, Joshi J, Schueller E, Frost I, Balasubramanian R, Laxminarayan R, Nandi A. State-wise estimates of current hospital beds, intensive care unit (ICU) beds and ventilators in India: are we prepared for a surge in covid-19 hospitalizations? medRxiv (2020). https://www.medrxiv.org/content/early/2020/06/18/2020.06.16.20132787.full.pdf. https://doi.org/10.1101/2020.06.16.20132787. https://www.medrxiv.org/content/early/2020/06/18/2020.06.16.2013278723.
22. Covid-19 hospital beds monitoring portal, Chhattisgarh. Accessed May 11, 2021. https://cg.nic.in/health/covid19/RTPBedAvailable.aspx
23. Covid-19 hospital beds monitoring portal, Rajasthan. Accessed May 11, 2021. https://covidinfo.rajasthan.gov.in/COVID19HOSPITALBEDSSTATUSSTATE.aspx
24. Covid-19 hospital beds monitoring portal, Andhra Pradesh. Accessed May 11, 2021. http://dashboard.covid19.ap.gov.in/ims/hospbedreports/
25. Pooja Sengupta SSAC, Ganguli B, et al. An analysis of covid-19 clusters in India. Nature. 2021. https://doi.org/10.1186/s12889-021-10491-8.
26. Williamson EJ, et al. Factors associated with covid-19-related death using open safely. Nature. 2020;584(7821):430–6. https://doi.org/10.1038/s41586-020-2521-4.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.