

Optimized monitoring systems in COVID-19 infected States in India using data mining techniques

R. Jebitha *

V. Deneshkumar †

Department of Statistics

Manonmaniam Sundaranar University

Tirunelveli 627012

Tamil Nadu

India

M. Manikandan §

Department of Community Medicine

Pondicherry Institute of Medical Sciences (PIMS)

Puducherry 605014

India

K. Senthamarai Kannan †

Department of Statistics

Manonmaniam Sundaranar University

Tirunelveli 627012

Tamil Nadu

India

Abstract

The COVID-19 pandemic, caused by the Corona Virus Disease 2019, had a global impact with over 71 million infections and more than 1 million deaths by December 31, 2020[1]. To categorize Indian States/Union Territories (UTs) according to their COVID-19

* E-mail: jebitharajan1998@gmail.com (Corresponding Author)

ORCID ID: 0000-0002-6525-7347

† E-mail: vdenesh77@gmail.com

ORCID ID: 0000-0001-8047-043X

§ E-mail: manikandanmsu@gmail.com

ORCID ID: 0000-0001-8425-016X

† E-mail: senkannan2002@gmail.com

ORCID ID: 0000-0003-1984-242X

statistics, we utilized hierarchical clustering analysis. This method groups the data based on their similarities, considering confirmed, recovered cases and COVID-19 related deaths. The results were presented in the form of informative tables and graphs, which can be valuable in devising strategies to prevent the further spread of the virus.

Subject Classification: 62P10.

Keywords: COVID-19, Data mining, Cluster analysis & Infected states.

1. Introduction

COVID-19 (Coronavirus disease 2019) was first identified in Wuhan, China, in December 2019. Since its initial detection, the virus has rapidly spread worldwide, prompting the World Health Organization (WHO) to officially declare it a pandemic outbreak on March 11th, 2020. COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) whose symptomatic cases include fever, cough, fatigue, and shortness of breath [2, 3]. The severity of the disease has led many countries to implement a range of preventive measures. These measures include strict lockdowns, quarantine protocols, screening tests, practicing social distancing, wearing masks, and promoting hand sanitization, among others. As of May 5, 2021, the worldwide tally of confirmed COVID-19 cases and deaths stood at 20,282,833 and 222,408, respectively. The pandemic has had a profound impact on people's lives and the global economy [3], causing significant disruptions. In India the first person reported from the state of Kerala in January 2020 and according to his travel history, he returned from Wuhan, China. Subsequently, there has been a notable surge in the number of cases and deaths across various states in India.

The cumulative Covid-19 infection tally has now reached 22,992,517 and the death toll to 249,992. Currently, India has 3,715,221 active COVID-19 cases, which make up 16.53% of the total cases. A significant number of 19,027,304 patients have been successfully discharged after receiving treatment for Covid-19. According to official data, the national recovery rate stands at 82.39%, while the mortality rate is 1.09%. Notably, approximately half of the total cases and deaths in the country have been reported in Maharashtra, Karnataka, Kerala, Uttar Pradesh, Tamil Nadu, Delhi, and West Bengal. India's COVID-19 situation has significant regional and global implications. The nation accounts for over 90% of both cases and deaths in the South Asian region, and within the past month, it represents 46% of global cases and 25% of global deaths [3]. Data mining

aims to discover non-trivial and concealed patterns, extracting valuable knowledge from various types of data. Cluster analysis is the most important method of the Data Mining to identifying the groups of observation that is organized and separated from other groups. Cluster analysis has been increased recently due to the emergence of several new areas of applications [4]. Many research has studied the COVID-19 pandemic at the country, state, city, or district level [5, 6, 7, 8, 9] using the Susceptible–Exposed–Infected–Removed (SEIR) model [10], or time series and stochastic models [11, 12], which could capture the dynamics of an infectious disease such as COVID-19. Diverse statistical approaches have been employed to identify the occurrence of Covid-19 cases and assess its associated risk factors, but very few studies adopted the data mining techniques for classifying the cases according to severity of pandemic with different states. Data mining techniques was used in classification and clustering of gene expression dataset and supervise and unsupervised methods that used recently to classify gene expression for the purpose of diagnosis and prognosis of terrifying diseases [13]. In this study, we conducted a cluster analysis on the states/Union Territories of India, utilizing variables such as the number of confirmed cases, the number of cured/recovered individuals, and the number of COVID-19 related deaths. Through the application of the Hierarchical Clustering method, we endeavored to categorize various states in India based on the severity of COVID-19 cases. This classification can prove beneficial in enhancing government policies, decision-making processes, medical facilities, and treatment protocols. Additionally, it aids in comprehending the gravity of the novel coronavirus (COVID-19) spread. The insights gained from this study can contribute to mitigating the number of cases and deaths among individuals affected by Covid-19. The rest of the paper organized as follows, Methodology, results and discussion and finally conclusion of the results.

2. Methodology

Data pertaining to COVID-19, spanning from January 30, 2020, to May 5, 2021, was collected from the website “www.covid19india.org”. The dataset encompassed information from 34 states affected by COVID-19, namely Tamil Nadu, Puducherry, Telangana, Karnataka, Kerala, Tripura, Uttar Pradesh, and West Bengal. Andhra Pradesh, Arunachal Pradesh, Assam, Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli, Andaman and Nicobar Islands, Goa, Daman/Diu, Delhi, Gujarat,

Haryana, Lakshadweep, Madhya Pradesh, Maharashtra, Manipur, Meghalaya, Mizoram, Nagaland, Odisha, Punjab, Rajasthan, Sikkim, Himachal Pradesh, Jammu and Kashmir, Jharkhand, Ladakh. The dataset comprised the total number of confirmed cases, the total number of cured / discharged cases, the total number of deaths, and the total number of active cases for each state.

2.1 Hierarchical Clustering analysis

The process of hierarchical clustering analysis involves initially forming an initial partition of m clusters and then gradually reducing the number of clusters one by one until all m observations are ultimately consolidated into a single cluster. Three distinct similarity measures, namely single linkage, complete linkage, and average linkage, are employed to identify similar groups. For this paper, the average linkage method was utilized to identify and form the clusters.

Hierarchical clustering is a significant clustering method that involves recursively agglomerating or splitting the data, resulting in a nested class hierarchy or class pedigree. This approach is instrumental in organizing and structuring the data into meaningful clusters. Hierarchical clustering methods work by grouping data objects into a tree of clusters. It could be further classified into “agglomerative” grouping small clusters from large, and “divisive” splitting big clusters into small ones, depending on whether the hierarchical decomposition is formed in a bottom up or top down fashion. Both agglomerative and divisive methods yield results that can be visualized using a dendrogram. These methods offer the advantage of detecting clusters of various shapes and obtaining multi-level structures of clustering with different sizes. Agglomerative hierarchical clustering algorithms, in particular, exhibit high clustering quality as they align well with the inherent characteristics of the data.

In this paper, Covid-19 datasets are classified using hierarchical clustering methods. The average linkage method is employed to identify clusters among the states/union territories. This method utilizes the Euclidean distance as the primary measure of similarity.

1. Begin with N clusters, each containing a single entity, and create an $N \times N$ symmetric matrix of distances (or similarities).
2. Explore the distance matrix to find the nearest (most similar) pair of clusters, denoted as U and V , with a distance of “ d ”.

3. Merge clusters U and V to form a new cluster (UV). Update the distance matrix by (a) removing the rows and columns corresponding to clusters U and V, and (b) adding a row and column that represent the distance between cluster (UV) and the remaining clusters.
4. Repeat steps 2 and 3 a total of N-1 times until all objects are in a single cluster. Keep track of the merged clusters and the levels (distances or similarities) at which the merges occur.

The dendrograms of cluster analysis calculated based on all the variables separately for the confirmed cases, recovered cases, deaths and tests are given in the Figure 5-9.

Average Linkage

Average linkage clustering involves calculating the distance between two clusters as the average distance between objects from the first cluster and objects from the second cluster. The algorithm follows the general agglomerative approach, where we initially identify the most similar objects, such as U and V, from the distance matrix $D = \{d_{ij}\}$. These objects are then merged to create the cluster (UV). During step 3 of the general agglomerative algorithm, the distances between cluster (UV) and any other cluster, such as W, are determined by calculating their respective average distances.

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}$$

The distance between object “i” in cluster (UV) and object “k” in cluster W is denoted by d_{ik} , where $N_{(UV)}$ and N_W are represent the number of items in cluster (UV) and cluster W, respectively.

Euclidean distance measure

Euclidean distance is most widely used measure of similarity when trying to groups among multivariate patterns (Johnson and Wichin (1982)). The Euclidean distance d of x_1 and x_2 is given below

$$d(x_1, x_2) = \sqrt{(x_{11} - x_{12})^2 + (x_{12} - x_{22})^2 + \dots + (x_{1p} - x_{2p})^2}$$

The popularity of this method stems from its intuitive appeal as a similarity measure. The similar observations are expected to be grouped

together in relatively small clusters, while dissimilar observations are separated by relatively larger clusters.

4. Results and Discussions

Data pertaining to COVID-19 in India was collected from January 30, 2020, to May 5, 2021, from Kaggle.com’s open-source database. The dataset covered all states and Union territories of India and included information on the total number of confirmed cases, total number of cured/recovered cases, and total number of deaths. As of May 5, 2020, total number of confirmed was 15855124, total number of recovered was 12840577, total number of deaths was 154515 and the total number of tests was 287105662. The summary statistics of confirmed cases, recovered cases, active cases and deaths are given in the Table 1. Figure 1, 2 and 3 clearly shows there are four groups of states such as Maharashtra, Karnataka, Kerala, Andhra Pradesh, Tamil Nadu, Delhi, Uttar Pradesh and West Bengal were highly affected. Whereas, the remaining states the affected were less than the above-mentioned states. During the data analysis, we did not remove outlier values in the data because these could give the worst affected states in the country. Also, this could be useful for the health department and decision makers for controlling spread of the disease. Figure 1,2 3, and 4 shows the top 15 highly affected states in India, these shows Maharashtra is severely affected, followed by Andra Pradesh, Karnataka, Tamil Nadu, Kerala, Delhi, West Bengal, Odisha, Telangana, Bihar, Assam, Rajasthan, Chhattisgarh and Gujarat.

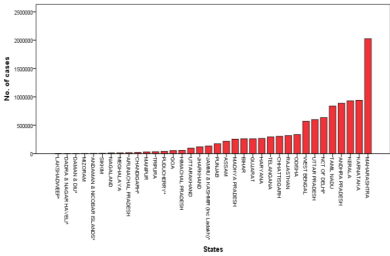


Figure 1
No. of confirmed cases

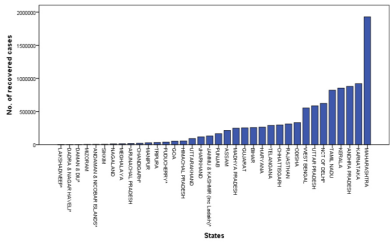


Figure 2
No. of recovered cases

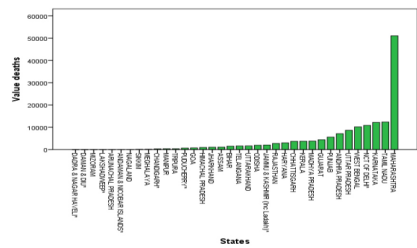


Figure 3
No. of deaths

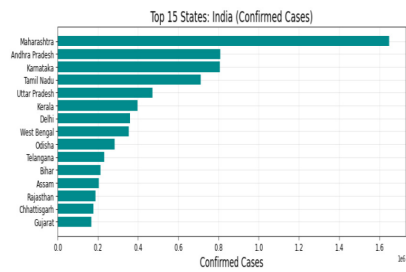


Figure 4
Top 15 states in India (confirmed cases)

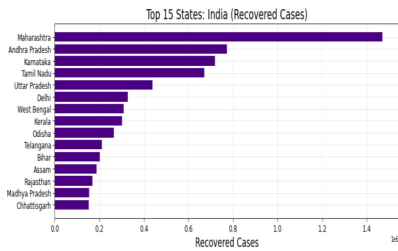


Figure 5
Top 15 states in India (recovered cases)

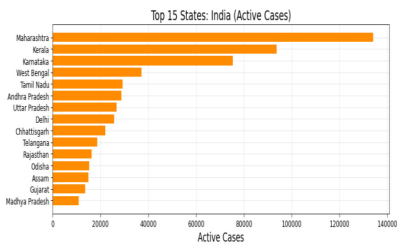


Figure 6
Top 15 states in India (Active cases)

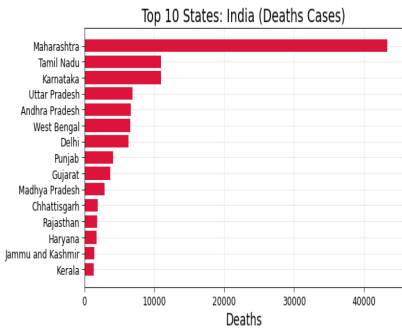


Figure 7
Top 15 states in India (deaths)

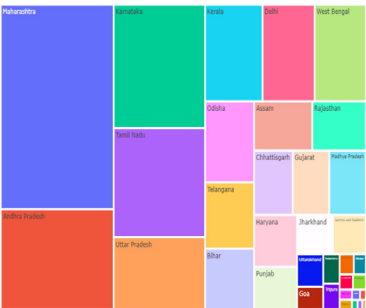


Figure 8
Tree Plot

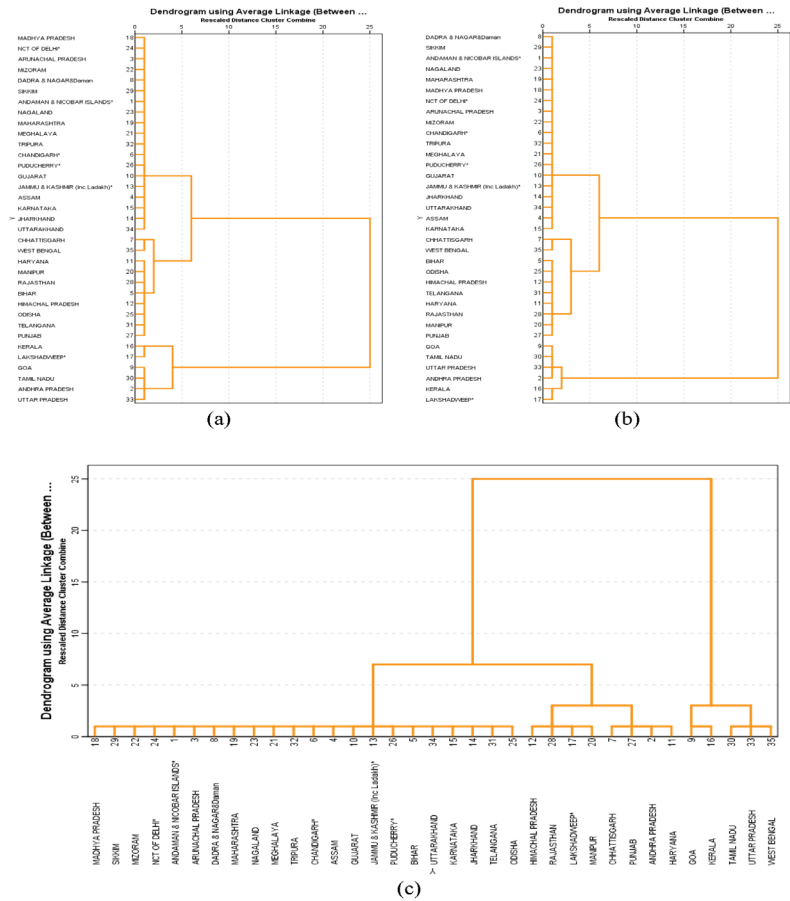


Figure 9
Dendrogram –(a) number of confirmed cases, (b) cured, (c)deaths

Table 1
Clusters

Initial Stage	Combined Cluster		Coefficients	First Appears - Cluster Stage		Next Stage
	1 st Cluster	2 nd Cluster		1 st Cluster	2 nd Cluster	
1	18	29	0	0	0	6
2	8	19	4	0	0	4
3	1	3	121	0	0	8

Contd...

4	8	23	145	2	0	8
5	21	32	729	0	0	13
6	18	22	1156	1	0	7
7	18	24	2497.333	6	0	14
8	1	8	3143.167	3	4	14
9	4	10	3364	0	0	17
10	14	31	4761	0	0	18
11	12	28	7569	0	0	25
12	5	34	7921	0	0	16
13	6	21	11738.5	0	5	19
14	1	18	15616.2	8	7	19
15	7	27	25600	0	0	28
16	5	15	56970.5	12	0	23
17	4	13	61357	9	0	22
18	14	25	113750.5	10	0	23
19	1	6	138028.9	14	13	26
20	17	20	245025	0	0	25
21	2	11	260100	0	0	28
22	4	26	374008.7	17	0	26
23	5	14	485571.1	16	18	29
24	30	33	662596	0	0	30
25	12	17	933637.5	11	20	31
26	1	4	1447969	19	22	29
27	9	16	1473796	0	0	32
28	2	7	2415386	21	15	31
29	1	5	5487679	26	23	33
30	30	35	6222170	24	0	32
31	2	12	13194510	28	25	33
32	9	30	15964882	27	30	34
33	1	2	40364809	29	31	34
34	1	9	16300000	33	32	0

Conclusion

This study utilized cluster analysis to categorize Indian states/Union Territories according to the severity of COVID-19 infections and mortality.

The hierarchical cluster analysis grouped states and UTs into two clusters based on Covid-19 confirmed, recovered, and death cases. Cluster I demonstrated a need for increased monitoring, including measures like screening, closures, curfews, lockdowns, evacuations, and legal actions. These findings are useful for the government, public health department, policymakers, and other stakeholders involved in combating the Covid-19 pandemic.

Acknowledgements

The first author expresses gratitude to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli - 627012, for their generous financial support through the prestigious Bharat Ratna Dr. M.G.R Centenary Research Fellowship, which enabled the successful completion of this work.

References

- [1] Gorbalenya, A. E. et al. The species severe acute respiratory syndrome-related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* 5, 536 (2020). <https://doi.org/10.1038/s41564-020-0695-z>.
- [2] Lai, C.-C., Shih, T.-P., Ko, W.-C., Tang, H.-J. & Hsueh, P.-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges. *Int. J. Antimicrob. Agents* 55, 105924 (2020). <https://doi.org/10.1016/j.ijantimicag.2020.105924>.
- [3] World Health Organization. COVID-19 Weekly Epidemiological Update 22. World Health Organ [Internet]. 2021; (December):1–3. Available from: https://www.who.int/docs/default-source/coronavirus/situation-reports/weekly-epidemiological_update_22.pdf
- [4] Mining WI. Data mining: Concepts and techniques. Morgan Kaufmann (2006).
- [5] M. Chinazzi, J.T. Davis, M. Ajelli, C. Gioannini, M. Litvinova, S. Merler, A. Pastore y Piontti, K. Mu, L. Rossi, K. Sun, C. Viboud, X. Xiong, H. Yu, M.E. Halloran, I.M. Longini, A. Vespignani, The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak, *Science* (2020), eaba9757.

- [6] K. Prem, Y. Liu, T.W. Russell, A.J. Kucharski, R.M. Eggo, N. Davies, M. Jit, P. Klepac, S. Flasche, S. Clifford, C.A.B. Pearson, J.D. Munday, S. Abbott, H. Gibbs, A. Rosello, B.J. Quilty, T. Jombart, F. Sun, C. Diamond, A. Gimma, K. van Zandvoort, S. Funk, C.I. Jarvis, W.J. Edmunds, N.I. Bosse, J. Hellewell, The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study, *Lancet Publ. Health* (2020).
- [7] A.J. Kucharski, T.W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R.M. Eggo, F. Sun, M. Jit, J.D. Munday, N. Davies, A. Gimma, K. van Zandvoort, H. Gibbs, J. Hellewell, C.I. Jarvis, S. Clifford, B.J. Quilty, N.I. Bosse, S. Abbott, P. Klepac, S. Flasche, Early dynamics of transmission and control of COVID-19: A mathematical modelling study, *Lancet Infect. Dis.* (2020).
- [8] Keeling MJ, Rohani P. Stochastic dynamics. In Modeling infectious diseases in humans and animals 2011. Princeton University Press.
- [9] J.A. Firth, J. Hellewell, P. Klepac, S. Kissler, A.J. Kucharski, L.G. Spurgin, Using a real-world network to model localized COVID-19 control strategies, *Nat. Med.* 26, 1616–1622 (2020).
- [10] J. Stehle, N. Voirin, A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Regis, J.-F. Pinton, N. Khanafer, W. Van den Broeck, P. Vanhems, Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees, *BMC Med.* 9, 87 (2011).
- [11] Agrebi S, Larbi A. Use of artificial intelligence in infectious diseases. Artificial Intelligence in Precision Health. 2020:415–38. doi: 10.1016/B978-0-12-817133-2.00018-5. Epub 2020 Mar 13. PMID: PMC7153335.
- [12] Rizvi SA, Umair M, Cheema MA. Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators. *Chaos Solitons Fractals*. 2021 Oct;151:111240. doi: 10.1016/j.chaos.2021.111240. Epub 2021 Jul 8. PMID: 34253943; PMID: PMC8264526.
- [13] Audu Musa Mabu, Rajesh Prasad & Raghav Yadav. Mining gene expression data using data mining techniques: A critical review, *Journal of Information and Optimization Sciences*, 41:3, 723-742 (2020), DOI: 10.1080/02522667.2018.1555311.