Asian Journal of Advances in Medical Science

3(1): 32-38, 2021

USE OF CLUSTER ANALYSIS TO MONITOR NOVEL CORONA VIRUS (COVID-19) INFECTIONS IN INDIA

SHABIR A. ZARGAR[1], TAJAMUL ISLAM[1*], ISHFAQ UL REHMAN[1] AND DIGVIJAY PANDEY[2]

[1]Department of Botany, University of Kashmir, Srinagar-190006, Jammu and Kashmir, India.
[2]Department of Technical Education, IET, Lucknow, India.

AUTHORS' CONTRIBUTIONS

This work was carried out in collaboration among all authors. Author SAZ wrote the first draft of manuscript. Author TI designed the study and managed the analysis of the study. Author IUR edited the text. Author DP carried out corrections. All the authors read and approved the final manuscript.

Received: 01 January 2021
Accepted: 29 January 2021
Published: 22 February 2021

Original Research Article

ABSTRACT

Objectives: In December 2019, in Wuhan, China, a novel coronavirus disease (COVID-19), a highly infectious disease, was first described. The disease has spread to 210 countries and territories across the world and more than two million people have been infected (confirmed). In India, the disease was first detected on 30 January 2020 in Kerala in a student who returned from Wuhan. The disease has been continuously spreading all the state of India. The main objective of this study was to identify and classify affected districts into real clusters on the basis of observations of similarities within a cluster and dissimilarities among different clusters so that government policies, decisions, medical facilities (ventilators, testing kits, masks, treatment etc.), etc. could be improved for reducing the number of infected and deceased persons and hence cured cased could be increased.
Materials and Methods: We concentrated on the COVID-19 affected states and UTs of India in the report. To fulfill the task, we applied cluster analysis, one of the data mining techniques. The study of variations among various clusters for each of the variables was performed using box plots. We used PAST software for getting for getting a scatter plot for each of the variables.
Results: Results obtained from the clustering analysis and box plot methods for each of the variables. For confirmed cases, cluster I corresponded to the states AP, AR, AS, BR, CG, GA, GJ, HR, HP, JH, KA, KL, MP, MH, MN, ML, MZ, NL, OR, PB, RJ, SK, TN, TG, TR, UP, UK, WB, AN, CH, DNDD, DL, JK, LA, LD, PY. For cured cases, cluster II and for death cases, cluster III corresponded to all the states and UTs of India.
Conclusions: The study showed that the state MH, AP, AR, DL and KL under cluster I have a high number of confirmed cases. The box plots and histogram shows variations among different clusters of the three cases. The trend in box plots and histograms showed a good percentage of cured cases in some of the states and UTs. It was observed that the states (MH, UP, KR, TN, DL and WB) under clusters III had severe conditions which need optimization of monitoring techniques which could help the government in making improvement government policies, actions, etc. to reduce the number of infected persons.

Keywords: Coronavirus disease-19; India; cluster analysis; box plot; data mining.

**Corresponding author: Email: islamtajamul66@gmail.com;*

1. INTRODUCTION

The coronavirus disease (COVID-19) originated from Seafood general Market in Wuhan City, Hubei Province of China in late 2019 [1,2]. Latter World Health Organization reported that a quite distinctive coronavirus (2019 nCoV), which was named as Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV2) by the International Committee on Taxonomy of Viruses (ICTV) on 11 February 2020, because of the causative virus by Chinese authorities on 7 January [3]. Coronavirus is zoonotic in nature and bat is said to be the origin of coronavirus [4], and transmission from humans to humans occurs from those people who come in contact with a COVID-19 infected person. The Municipal Health Commission, Wuhan, China, on December 31, 2019, reported a cluster of transmittable pneumonia cases and it was identified as novel coronavirus disease (COVID-19). Further, other provinces of China have got spread of it and till today almost all the countries around the world have been affected due to the spread of the COVID-19.

In India, a country of 1.3 billion people, reported the first case of the COVID-19 from the state of Kerala on January 30, 2020. The Indian government declared this outbreak an epidemic in all the states and union territories (UTs). All educational institutions and commercial offices were shutdown. On March 22, 2020, India announced a 14 h public curfew. Further, the Indian government on March 24, 2020, ordered a nationwide lockdown for 21 days (till April 14, 2020) and after the completion of the period of this lockdown, the central government extended the lockdown up to May 3, 2020 and after that several phases of lockdown imposed. Several other types of actions were taken by the state and UT governments to control the spread of the virus COVID-19 [5]. Apart from lockdown, people have certain conjectures about possible reasons behind India's relative success, e.g., measures like the travel ban relatively early, use of BCG vaccination to combat tuberculosis in the population that may have secondary effects against COVID-19 [6,7].

India seems to have managed the COVID-19 pandemic better compared to many other countries. In India, COVID-19 has spread in all states across the country, with 10.2 million cases and 148,000 deaths reported till December 26, 2020 [8]. Frontline healthcare workers (HCWs), including doctors, nurses, technicians, ward, and sanitation workers, are experiencing acute challenges in effectively managing patients while also protecting themselves and their families from COVID-19. Globally, tens of thousands of HCWs have been infected with the coronavirus (SARS-CoV-2), and hundreds have died in the line of duty [9.] In India, according to the Indian Council of Medicine Research, 5.2% of the COVID-19 infected patients are healthcare workers [10]. The virus can infect anyone, including age, except for the elderly and other people existing with problems such as diabetes, disorder, immunosuppressive state etc. The people are vulnerable to virus when people are in touch with one another [11]. The daily infection-rate (DIR) for a given day is defined as:

$$\text{DIR} = \frac{\text{Total active cases in a day} - \text{total active cases in the previous day}}{\text{Total active cases in the previous day}}$$

Note that India may have seen fewer COVID-19 cases till now, but the war is not over yet. There are many states like Maharashtra (MH), Karnataka (KR), Andhra Pradesh (AP), Tamil Nadu (TN), Delhi (DL), West Bengal (WB) and Uttar Pradesh (UP) who are still at high risk. These states may see a huge jump in confirmed COVID-19 cases in the coming days if preventive measures are not implemented properly [12]. On the positive side, many states and UTs have shown how to effectively "flatten" or even "crush the curve" of COVID-19 cases. We hope India can be free of COVID- 19 with a strong determination as already shown by the central and respective state Governments. There are a few works that are based explicitly on Indian COVID-19 data. Das [13] has used the epidemiological model to estimate the basic reproduction number at national and some state levels. Ray et al. [14] used a predictive model for case-counts in India. There are some preventive measures to control the spread of coronavirus pandemics [15].

2. MATERIALS AND METHODS

2.1 Study Area

Study area India which is located in the Southern portion of Asia. India is the second most populous country in the world, after China. As per consensus of India 2011, the population of India is 1,210,193,422. The country is situated north of the equator between 8°4' north to 37°6' north latitude and 68°7' east to 97°25' east longitude [16]. It is the seventh-largest country in the world, with a total area of 3,287,263 km^2 [17,18,19]. It has 28 states and 9 Union territories: Andhra Pradesh (AP), Arunachal Pradesh (AR), Assam (AS), Bihar (BR), Chhattisgarh (CG), Goa (GA), Gujarat (GJ), Haryana (HR), Himachal Pradesh (HP), Jharkhand (JH), Karnataka (KA), Kerala (KL), Madhya Pradesh (MP), Maharashtra (MH), Manipur (MN), Meghalaya (ML), Mizoram (MZ), Nagaland (NL), Odisha (OR), Punjab (PB), Rajasthan (RJ), Sikkim (SK), Tamil Nadu (TN),

Telangana (TG), Tripura (TR), Uttar Pradesh (UP), Uttarakhand (UK), West Bengal (WB), Andaman and Nicobar Islands (AN), Chandigarh (CH), Dadra and Nagar Haveli (DN) and Daman and Diu (DD), Delhi (DL), Jammu and Kashmir (JK), Ladakh (LA), Lakshadweep (LD), Puducherry (PY).

Uttar Pradesh is the most populous state having population around 199,812,341 at the density of 690/ km^2 and having an area about 240928 km^2. Whereas Lakshadweep is the least populous state with population around 64,473 at the density of 2149/ km^2 and possessing an area about 32.62 km^2.

2.2 Methodology

We distribute the whole study into three parts. Part I consists of a collection of data and its exploratory analysis; part II consists of a performance of statistical analysis of COVID-19 data set using cluster analysis (Euclidean); and part III consists of deviations within clusters for each of the cases using a box plots/histograms.

Part I: Data collection and exploratory analysis

We collected data related to COVID-19 from January 30, 2020 to December 26, 2020 in India from the website of "Johns Hopkins Coronavirus Resource Center" https://coronavirus.jhu.edu/map.html [8]. Some related information is also supported by https:// en.wikipedia.org [20]. We included 28 different states and 9 Union territories (COVID-19 affected): AP, AR, AS, BR, CG, GA, GJ, HR, HP, JH, KA, KL, MP, MH, MN, ML, MZ, NL, OR, PB, RJ, SK, TN, TG, TR, UP, UK, WB, AN, CH, DN, DD, DL, JK, LA, LD, PY.

The data consist of three variables: The total number of confirmed cases, the total number of cured/discharged cases, and the total number of death cases. The total number of confirmed, cured, and deaths cases during the period mentioned above are 1,020,0000, 97,60,000 and 1,48,000 respectively. However, the Lakshadweep (LD) union territory has no confirmed case. An exploratory analysis of all the three variables is given in Table 1, which summarizes basic statistics for the variables mentioned above. We also represent the characteristics of the three variables of the all states and UTs of India using histograms/box plots in Fig. 5 and further by histograms for each of the variables in Fig. 6.

Part II: Cluster analysis (CS)

Cluster Analysis is one of the data analyzing techniques which clusters the sample observations into classes depending on the essential similarities within a class and dissimilarities among different classes found in the data set [21,22]. Ward [23] suggested agglomerative hierarchical cluster analysis which is based on a squared Euclidean distance. The ward method is the simplest and the most commonly used method which requires no prior assumption and uses the analysis of variance to calculate distances among clusters [24]. In this study, we used the PAST software (version 3) to perform the cluster analysis. We scaled the data set before carrying out the cluster analysis. Principle Component Analysis (PCA) using PAST software for getting a scatter plot for each of the variables given in Fig. 4.

Part III: Analysis using box plot and histograms

To measure the deviation within clusters for each of the variables, we analyzed it statistically using PAST software and for the purpose, we used box plots and histograms for representing the deviation in each of the cases. The observations related to the variables are skewed which were shown in Figs. 5 & 6, so the median is more appropriate to use [25]. It is well known that the box plot is the most powerful tool for showing median, range, as well as the shape of the underlying distribution of the data.

3. RESULTS

From the Fig. 6, it was seen that there was a great difference between minimum and maximum number of observations for all of the variables. Further, from Fig. 5, it was observed that the data related to each of the variables was skewed. Extreme observations were also present in the data set. The dendrograms of cluster analysis calculated on the basis of all the variables separately for the COVID-19 data set are given in Figs. 4-6, respectively for confirmed cases, cured cases, and death cases for the visual representation. For confirmed cases, cluster I corresponded to the states AP, AR, AS, BR, CG, GA, GJ, HR, HP, JH, KA, KL, MP, MH, MN, ML, MZ, NL, OR, PB, RJ, SK, TN, TG, TR, UP, UK, WB, AN, CH, DNDD, DL, JK, LA, LD, PY. For cured cases, cluster II and for death cases, cluster III corresponded to all the states and UTs of India.

4. DISCUSSION

All the states and UTs except Lakshadweep under cluster I, II and the states and UTs excluding Lakshadweep (LD), Arunachal Pradesh (AN), Dadra and Nagar Haveli and Daman and Diu (DNDD), Mizoram (MZ), Nagaland (NL) under cluster III

were in the severe zone. Maharashtra (MH), Karnataka (KR), Utter Pradesh (UP), Tamil Nadu (TN), West Bengal (WB) and Delhi (DL) under cluster III had high severity of death cases around 1.52 %. Scatter plot depicts the trend of confirmed, cured and death cases towards the affected states and UTs of India.

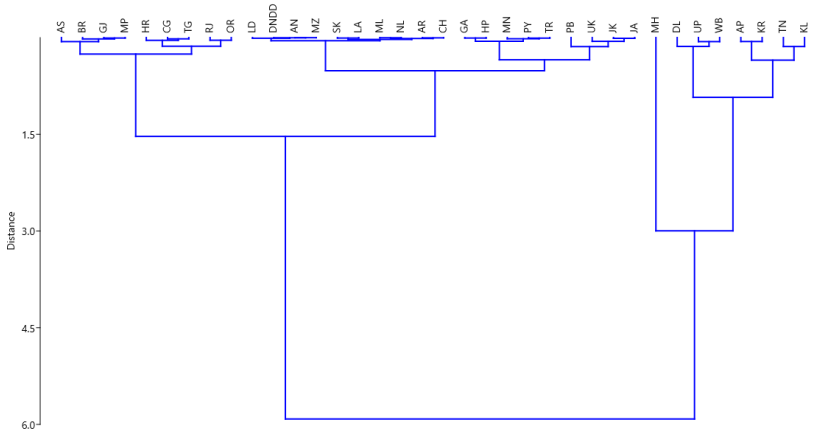

Fig. 1. A dendrogram showing clustering of districts for confirmed cases from coronavirus disease-19

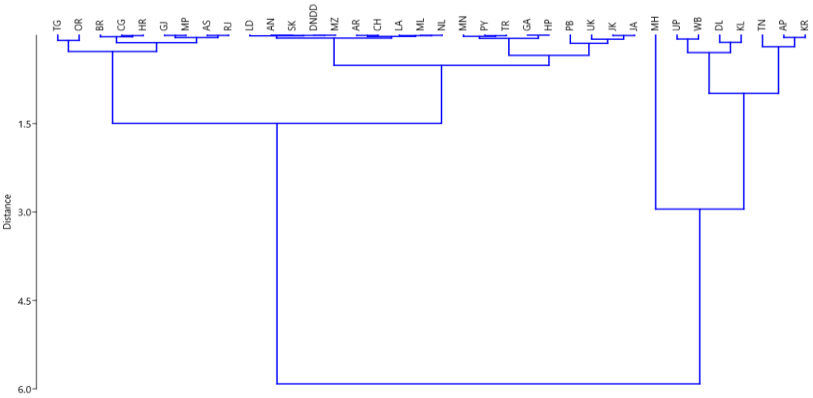

Fig. 2. A dendrogram showing clustering of districts for recovered cases from coronavirus disease-19

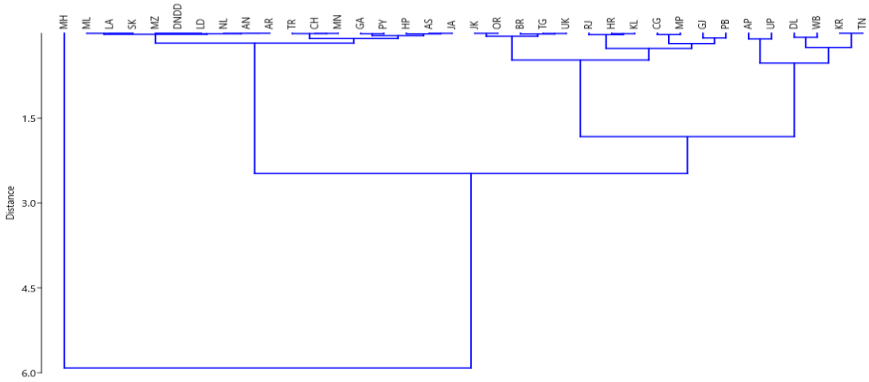

Fig. 3. A dendrogram showing clustering of districts for death cases from coronavirus disease-19

Table 1. A summary of COVID-19 status of 27 states and 9 UTs in India

S. No.	States/UTs	Cases	Recovered	Deaths
1	Maharashtra	1920000	1810000	49189
2	Karnataka	915000	890000	12051
3	Andhra Pradesh	881000	870000	7092
4	Tamil Nadu	813000	792000	12059
5	Kerala	736000	669000	2951
6	Delhi	622000	605000	10437
7	Uttar Pradesh	581000	557000	8293
8	West Bengal	546000	522000	9569
9	Odisha	328000	324000	1857
10	Rajasthan	305000	209000	2664
11	Telangana	285000	277000	1531
12	Chhattisgarh	274000	257000	3275
13	Haryana	261000	253000	2865
14	Bihar	249000	243000	1379
15	Gujarat	241000	226000	4275
16	Madhya Pradesh	237000	224000	3545
17	Assam	216000	211000	1035
18	Punjab	165000	155000	5281
19	Jammu and Kashmir	120000	115000	1867
20	Jharkhand	114000	112000	1018
21	Uttarakhand	89218	82298	1476
22	Himachal Pradesh	54280	49076	913
23	Goa	50595	48913	731
24	Puducherry	37947	36962	630
25	Tripura	33237	32695	385
26	Manipur	27976	26331	344
27	Chandigarh	19423	18754	315
28	Arunachal Pradesh	16687	16477	56
29	Meghalaya	13371	12985	138
30	Nagaland	11897	11545	78
31	Ladakh	9394	9051	126
32	Sikkim	5799	5157	125
33	Andaman and Nicobar Islands	4912	4789	62
34	Mizoram	4182	4046	8
35	Dadra and Nagar Haveli and Daman and Diu	3372	3362	2
36	Lakshadweep	0	0	0

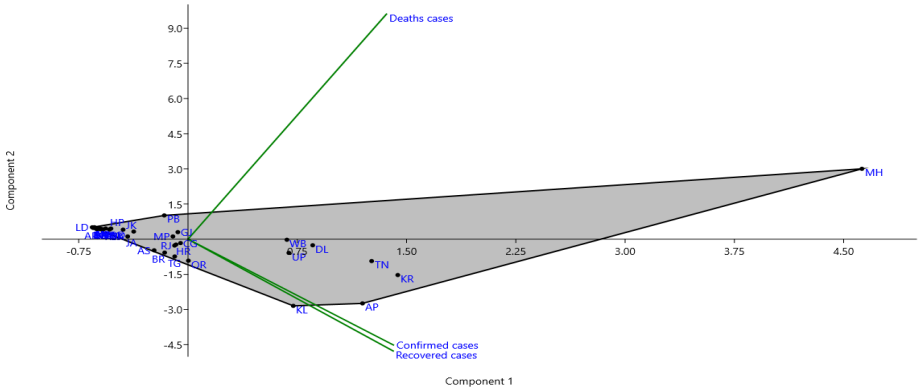

Fig. 4. Principle Component Analysis (PCA) for the three cases: confirmed, cured and death cases of coronavirus disease-19

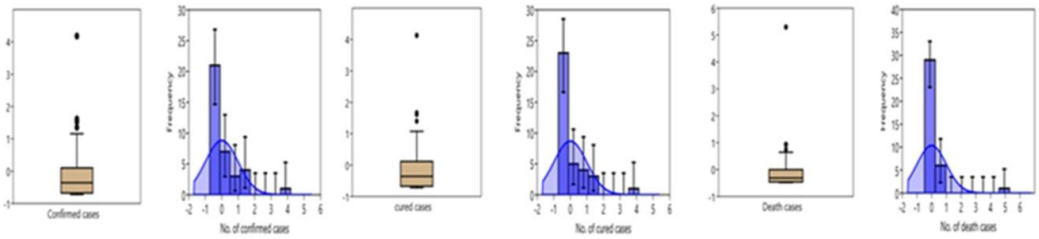

Fig. 5. Box plots and histograms for number of confirmed, cured and death cases of coronavirus disease-19

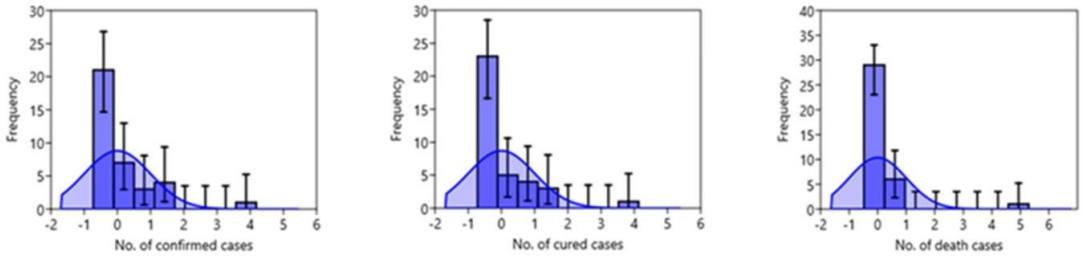

Fig. 6. Histograms for number of confirmed, cured and death cases of coronavirus disease-19

5. CONCLUSIONS

In this study, we performed a Euclidean hierarchical cluster analysis to classify states and UTs of India on the basis of the various status of COVID-19. The technique grouped 36 different affected states and UTs into three clusters (I-III) for each of the cases. All states and UTs except Lakshadweep (LD) under cluster I, II and III were affected with COVID-19, where the state MH, AP, AR, DL and KL under cluster I have a high number of confirmed cases. The box plots and histogram shows variations among different clusters of the three cases. The trend in box plots and histograms showed a good percentage of cured cases in some of the states and UTs. It was observed that the states (MH, UP, KR, TN, DL and WB) under clusters III had severe conditions which need optimization of monitoring techniques which could help the government in making improvement government policies, actions, etc. to reduce the number of infected persons.

CONSENT

It is not applicable.

ETHICAL APPROVAL

It is not applicable.

ACKNOWLEDGEMENT

The authors are grateful to Rayees Ahmad Rather for providing the valuable comments about the manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y. Early transmission dynamics in Wuhan, China, of novel coronavirus-infected pneumonia. N Engl J Med; 2020.
   Available:https://doi.org/10.1056/NEJMoa2001316, 2020.
2. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. https://doi.org/; 2020.
   DOI:10.1016/S0140-6736(20)30183-5,
3. World Health Organization. Coronavirus. World Health Organization; 2020.
   Available:https://www.who.int/health-topics/coronavirus.
4. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature;2020.
   Available:https://doi.org/
   DOI:10.1038/s41586-020-2012-7,.
5. Gulia A, Panda PK, Parikh P. India and COVID-19 pandemic- standing at crossroad! Indian J Med Sci. 2020;72:1-2.

6. Hegarty PK, Service NH, Kamat AM, Dinardo A. BCG vaccination may be protective against Covid-19; 2020. DOI:10.13140/RG.2.2.35948.10880.
7. Bacille Calmette-Guérin (BCG) vaccination and COVID-19; 2020. Available:https://www.who.int/news-room/commentaries/detail/bacille-calmette-guérin-(bcg)-vaccination-and-covid-19. (accessed on 21 December 2020).
8. Johns Hopkins Coronavirus Resource Center; 2020. Available:https://coronavirus.jhu.edu/map.html
9. Gun AM, Gupta MK, Dasgupta B. Fundamentals of Statistics. Kolkata: World Press Private. 2008;1.
10. Abraham P, Aggarwal N, Babu GR, Barani S, Bhargava B, Bhatnagar T, et al. Laboratory surveillance for SARS-CoV-2 in India: Performance of testing and descriptive epidemiology of detected COVID-19. Indian J Med Res 2020;151:23640. Available:http://www.ijmr.org.in/ preprintarticle. asp?id=285361. (accessed on 03 December 2020).
11. Abere OJ. Survival Analysis of Novel Corona Virus (2019-Ncov) Using Nelson Aalen Survival Estimate. International Journal 0f Business Education And Management Studies. 2020;3(1):30-40.
12. Kumar S. Use of cluster analysis to monitor novel coronavirus-19 infections in Maharashtra, India. Indian Journal of Medical Sciences. 2020;72(2):44.
13. Das S. Prediction of COVID-19 Disease Progression in India: Under the Effect of National Lockdown; 2020. http://arxiv.org/abs/2004.03147. (accessed on 21 December 2020).
14. Ray D, Salvatore M, Bhattacharyya R, et al. Predictions, role of interventions and effects of a historic national lockdown in India's response to the COVID-19 pandemic: data science call to arms. medRxiv; 2020; DOI:10.1101/2020.04.15.20067256.
15. MoHFW | Home; 2020. Available:https://www.mohfw.gov.in/. (accessed on 21 December 2020).
16. India Yearbook. Publications Division, Ministry of Information & Broadcasting, Govt. Of India; 2007. ISBN 978-81-230-1423-4.
17. India. Encyclopædia Britannica. Retrieved 17 July 2012. Total area excludes disputed territories not under Indian control; 2020.
18. India at a Glance: Area. Ministry of Home Affairs: Government of India; 2001. (accessed 9 December 2020).
19. Jammu and Kashmir - CIA (PDF). Central Intelligence Agency; 2002. (accessed on 9 December 2020.
20. Dilts D, Khamalah J, Plotkin A. Using cluster analysis for medical resource decision making. Med Decis Mak. 1995;15:333-47.
21. McLachlan GJ. Cluster analysis and related techniques in medical research. Stat Methods Med Res. 1992;1:27-48.
22. Romesburg HC. Cluster Analysis for Researchers. Belmont: Lifetime Learning Publications; 1984.
23. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236-46.
24. Ministry of Health and Family Welfare. Government of India; 2020. Available from: https://www.mohfw.gov.in. (accessed on 03 December 2020].
25. Reuters. Over 90,000 Health Workers Infected with Covid-19 Worldwide: Nurses Group; 2020. Available:https://www.thehindubusinessline.com/news/over-90000-healthworkers-infected-with-covid-19-worldwide-nurses-group/article31519587.ece. (accessed on 03 December 2020).