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

Clinical Epidemiology and Global Health 10 (2021) 100690

Contents lists available at ScienceDirect

Clinical Epidemiology and Global Health

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

Original article

Spatial mapping of COVID-19 for Indian states using Principal Component Analysis

Vasna Joshua [a,*], J. Sylvia Grace [b], J. Godwin Emmanuel [c], S. Satish [a], A. Elangovan [a]

[a] ICMR-National Institute of Epidemiology, Chennai, 600077, India
[b] KCG College of Technology, Chennai, 600097, India
[c] Vellore Institute of Technology, Vellore, Tamil Nadu 632014, India

ARTICLE INFO

Keywords:
COVID-19
Principal component analysis
Spatial mapping
Indian states

1. Background

The first case of the pandemic outbreak of Coronavirus disease 'COVID-19' was reported in Wuhan, China, in November 2019. The pandemic outbreak has spread very quickly to 210 countries, including territories across the globe.[1]

In India, the first case of the COVID-19 was reported on January 30, 2020, originating from China. As of Oct 12, 2020, the Ministry of Health and Family Welfare has confirmed 7,011,388 cases, 6,149,535 recoveries, and 109,150 deaths in the country.[2] The infection rate of COVID-19 in India has slowed down, and the growth of the infections has become more or less linear and not exponential.[3]

The outbreak has been declared an epidemic in more than a dozen states and union territories, where provisions of the Epidemic Diseases Act, 1897 have been invoked, and educational institutions, tourist's places, Shopping malls, recreational centres, foreign consulates, and many commercial establishments have been shut down.[4] According to Centres for Disease Control and Prevention (CDC), persons at higher risk for the severe illness of COVID-19 are older adults and persons of any age who have serious ailments and under medication like Asthma, HIV, etc., pregnant people, experiencing homeless dwellers, and persons with disabilities.[5] The present study's objective was to identify the regions at greater risk of developing the disease for Indian states using COVID-19 data and its risk-related factors using the Principal Component Analysis technique.

2. Materials and methods

Study population: We retrieved the latest data available on the official website of the Ministry of Health and Family Welfare (MoHFW),[2] India; Census of India[6]; National Institution for Transforming India Aayog[7]; National AIDS Control Organization[8]; National Health Mission[9]; National Health Profile 2018[10]; National Family Health Survey 4[11]; Handbook of Social Welfare Statistics, Ministry of Social Justice and Empowerment[12]; Source State of forest report 2019[13] and published articles[14,15,16].

The information on COVID-19 active cases, deaths, and confirmed cases were collected on Oct 12, 2020.[2] The selection of the risk related factors of COVID-19 was based on a review of the literature and essentially with the available data. They were retrieved for 37 Indian States, including Union territories. The risk related factors extracted were population, percentage of geographical region, population density, number of households, the proportion of males, average family size, persons per room, percentage of illiterates, percentage of the elderly population (60 or more years), percentage of the homeless population, percentage of slum population, net migration rate, persons below poverty line (BPL), disability rate, the prevalence of diabetes, common cancers and hypertension among attending NCD clinics and adult HIV prevalence.

* Corresponding author. ICMR_National Institute of Epidemiology, Second Main Road Tamil Nadu Housing Board Ayapakkam, Near Ambattur, Chennai, 600 077, India.
E-mail addresses: vasnajoshua@nie.gov.in (V. Joshua), ssylviagrace@gmail.com (J. Sylvia Grace), jgodwinemmanuel@gmail.com (J. Godwin Emmanuel), yessatish@gmail.com (S. Satish), elangopunitha@gmail.com (A. Elangovan).

https://doi.org/10.1016/j.cegh.2020.100690
Received 20 October 2020; Received in revised form 27 November 2020; Accepted 27 December 2020
Available online 5 January 2021

3. Statistical analysis

Factor analysis was used to reduce the large data set into a smaller subset without losing much information. Principal Component Analysis (PCA) technique was used to achieve it. The objective of the PCA is to take a larger number of variables, say N variables X_1, X_2, …, X_N and find combinations of these to produce principal components Z_1, Z_2, …, Z_N that are uncorrelated in order of their importance, and to describe the variation in the data. The ith principal component is a linear combination given by

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{iN}X_N$$

N of these components and the coefficients a_{ij}'s are given by the eigenvector $\mathbf{a}_i$ corresponding to the ith largest eigenvalue λ_i of the correlation matrix of the X variables. When doing so, there is always a possibility that most of the principal components' variances may turn to be negligible. In that case, most of the full data set variation can be adequately described by the few Z components with variances that are not negligible. The best results are obtained when the original variables are highly correlated, either positively or negatively. The original set of 20 or more variables can be adequately represented by few (three or four) principal components. The first principal component has the highest variance, whereas the other components all have variances that are much less than the highest, which means that the first principal component is the most important, followed by (two/three) other components for representing the variation in the measurements of the (20 or more) variables. For better interpretability, the factors are improved using varimax rotation, which is widely used, maximizing the sum of the variances of all factors used.

For further analysis, it is usual to use only the first few principal components, providing that the sum of their variances is a high percentage (e.g., 80% or more) of the sum of the variances for all N components. A factor score can be obtained as a linear combination of standardized factors. The factor coefficient of the factors is called the factor score coefficient. Using the variance percentages as weights on the factor scores, the initial score is computed[17,18,19,20,21,22].

Our ultimate aim was to make the original data set into relatively fewer independent factors and estimate the factor scores. The original data set contained an array of dimension 37 states x 21 factors. These factors were examined using the correlation matrix and for a meaningful representation. The risk related factors were of different units of measurement; hence they were standardized. The PCA reduced the 19 risk-related factors (after omission of net migration and prevalence of common cancers) into ten highly correlated factors. Hence the final data set used in the analysis was the size (37 states x10 factors) (Table 1) further, a smaller subset of four factors extracted using the eigenvalue greater than one. Varimax rotation was used to improve the factors, and finally, the factor scores were obtained. Percentage of variation was used as weights, and the initial score for each state was obtained. For the sake of comparison, the initial scores were standardized and listed. The above analysis was done using the SPSS software.[23]

4. Spatial mapping using Inverse Distance Weighting (IDW) interpolation technique

A simple spatial interpolation method, namely the Inverse Distance Weighting method (IDW),[24] was applied to predict unmeasured locations using the available information from the measured locations. Here we have information in the form of derived scores for 37 locations (states), and the weights were assigned as the inverse of the distance between known and unknown locations. An IDW power coefficient of 2 with 12 nearest neighbourhood was used for the analysis.

The locations (longitude, latitude) of each state and the derived score were integrated into the ArcGIS version 10 software (ESRI, Redlands, CA, USA)[25] to predict values in the unmeasured locations.

5. Results

The PCA identified four factors, which together explained about 83% of the total variation. All the factors selected for the analysis were examined. It was found to be highly correlated as required for the factor analysis. The four-factor loadings that are larger ($\geq$0.64) are listed in

Table 2
Principal Component analysis - Varimax rotation factor matrix.

	Factor				
	I	II	III	IV	Communalities
Homeless population			.884		.829
Illiteracy		.787			.671
Elderly citizens			.638		.768
Disability rate				.898	.816
Population			.660		.877
Mean persons per room		.807			.726
Slum population				.834	.784
Active cases	.966				.944
deaths	.950				.922
Confirmed cases	.939				.946
Eigenvalue (>1)	3.080	1.880	1.672	1.652	
Percent of variation explained	30.800	18.801	16.721	16.523	
Total variation explained	82.845				

Table 1
State-level summary statistics of the factors studied for the Indian States.

S. No	Factor understudy	Definition	Data from Reference No	Min	Max	Mean	Median	Mode
1	Population	de facto population 2018	14	71218	228959599	36092294	18345784	71218
2	Illiterates	percentage of illiterates As per Census 2011	6	6.00	38.20	22.73	23.74	32.84
3	Elderly population	percentage of the elderly population (60 or more years) Census 2011	12	4.04	12.55	7.86	7.84	7.36
4	Homeless population	percentage of the homeless population, Census 2011	6	0.00	8.96	0.77	0.15	0.02
5	Slum population	percentage of slum population, Census 2011	6	0.00	45.00	18.98	18.98	0
6	Persons per room	The average number of people per room in an occupied housing unit, Census 2011	6	1.80	3.40	2.63	2.70	2.70
7	Disability rate	Census 2011	12	0.90	5.40	2.19	2.21	1.75
8	**COVID-19** active cases as of 12th Oct 2020	Persons currently with the disease	2	0	221637	23293	9275	51
9	**COVID-19** deaths as on 12th Oct 2020	Persons died due to the disease	2	0	40349	2950	816	0
10	**COVID-19** confirmed cases as of 12th Oct 2020	Persons with laboratory confirmation of COVID-19 infection, irrespective of clinical signs and symptoms.	2	0	1487877	189497	91738	0

Table 2.

The first factor consists of the disease COVID-19 highly correlated statistics, namely active cases, number of deaths, and confirmed cases. The second factor consists of the illiterate population and the mean number of persons used per room. The third factor consists of the residential population, homeless population, and elderly population aged 60 or more years, and the fourth factor consists of disability rate and slum population.

The initial score for various states are listed in Table 3, wherein the last column represents the corresponding standardized score in descending order.

States Maharashtra, Uttar Pradesh, Andhra Pradesh, Karnataka, and Tamil Nadu stood above the average. It had a standardized score of 50 or above, indicating greater interventional care needed to bring down the COVID-19 transmission in India.

States NCT of Delhi, West Bengal, Bihar, Telangana, Madhya Pradesh, Odisha, Rajasthan, Chhattisgarh, Uttarakhand, Punjab, Gujarat, Jammu Kashmir, and Haryana, which had a score between 50 and 25 needs the next priority care and the last nineteen states which had a score of less than 25 needs less care as on the date of the investigation.

The map obtained (Fig. 1) showed an optimal unbiased representation of multiple risk-related factors of the disease COVID-19 transmission with the Inverse Distance weighted estimates. The figure shows the regional variation and the disease high risk concentrated regions (hot spots) and regions at the greater risk of developing the infection. The estimates showed the high-risk concentrated regions as the central part of India with hot spots in Maharashtra, Uttar Pradesh, Andhra Pradesh, Karnataka, and Tamil Nadu. The transmission appeared to be lower in the North-Eastern part of India, Himachal Pradesh, and Dadra & Nagar Haveli.

Table 3
The initial scores and standardized scores for the Indian states, 2020.

S. No	State name	Initial Score	Standardized score	Rank
1	Maharashtra	154.8	100.0	1
2	Uttar Pradesh	69.0	59.4	2
3	Andhra Pradesh	62.9	56.5	3
4	Karnataka	48.5	49.6	4
5	Tamil Nadu	48.1	49.5	5
6	NCT of Delhi	38.8	45.0	6
7	West Bengal	32.2	41.9	7
8	Bihar	28.9	40.4	8
9	Telangana	27.3	39.6	9
10	Madhya Pradesh	27.3	39.6	10
11	Odisha	25.0	38.5	11
12	Rajasthan	17.4	34.9	12
13	Chhattisgarh	16.3	34.4	13
14	Uttarakhand	11.7	32.2	14
15	Punjab	10.1	31.5	15
16	Gujarat	2.5	27.8	16
17	Jammu Kashmir	−1.2	26.1	17
18	Haryana	−2.8	25.4	18
19	Ladakh	−9.3	22.3	19
20	Jharkhand	−9.5	22.2	20
21	Kerala	−13.6	20.2	21
22	Assam	−21.6	16.4	22
23	Puducherry	−23.9	15.3	23
24	Goa	−27.2	13.8	24
25	Mizoram	−28.2	13.3	25
26	Himachal Pradesh	−28.4	13.2	26
27	Tripura	−32.5	11.3	27
28	Sikkim	−33.0	11.0	28
29	Arunachal Pradesh	−34.4	10.4	29
30	Meghalaya	−38.1	8.6	30
31	Manipur	−40.2	7.6	31
32	Chandigarh	−40.2	7.6	32
33	Dadara and Nagar Havelli	−40.3	7.6	33
34	Nagaland	−43.8	5.9	34
35	Andaman and Nicobar Island	−44.4	5.6	35
36	Lakshadweep	−52.0	2.0	36
37	Daman and Diu	−56.3	0.0	37

Minimum initial score (MIN_INS); Maximum initial score (MAX_INS).
Standardized Score = [(INITIAL SCORE of the state – MIN_ INS)/(MAX_INS - MIN_INS)]*100.

6. Discussion

The states have been classified with zones/districts as red if there are a sizeable number of covid-19 cases or with hotspots, the green zone is areas with zero confirmed cases in the last 21 days, and left-outs are orange zone with a limited number of cases, and thereby people's movements are restricted.[26] Looking at the raw data and the magnitude of confirmed cases, Maharashtra stood first, followed sequentially by Andhra Pradesh, Karnataka, Tamil Nadu, Uttar Pradesh, Delhi, West Bengal, Kerala, and Odisha. The above exercise brings out a red alert state in sequential order as Maharashtra, Uttar Pradesh, Andhra Pradesh, Karnataka, and Tamil Nadu COVID-10 risk-related factors in a multivariate set up. The map shows the hot spot regions, mainly in the central part of India. It also showed a few cold spots in 'Seven Sister' states.[27] Apart from the data COVID-19, the study also brings out the proxy determinants as illiteracy and mean number of persons using per room; followed by residential population, homeless population and elderly population 60 years or more; disability rate and slum population. Even though eighteen variables (including chronic disease rates) were included in the study, only the above seven variables (other than COVID-19 cases) showed a higher correlation value of more than 0.5 with the infection cases. Directly or indirectly, all the seven variables are a function of the variable 'social distancing'. Public health officials emphasize social distancing as they are considered an important measure for mitigating the pandemic COVID-19. In a country like India, 'Social distancing poses unique challenges.[28]

The study had used the state as a unit of study. If finer grid points like districts or taluks are considered, it would have been more precise to pinpoint the country's pockets for the remedial measure. The COVID-19 risk-related data have been used from multiple sources with different years, which could be one of the study's limitations.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cegh.2020.100690.

Appendix 1

1. The original data set (37*21) was extracted from various sources (shown in supplementary material).
2. The selection of 10 factors was based on the following.
   (i) Net migration rate state-wise was readily available only for the year 1991–2001 hence not included and
   (ii) Prevalence of common cancers from 01.01.2017 to 31.12.2017 attending NCD clinics was missing for 5 states hence was also not included for the further analysis.

Hence the original data set was reduced to (37*19), and it was further examined.

   (iii) The basic assumption of factor analysis is to identify highly correlated factors. It also brings out the number of factors required to represent the major portion provided by all the observed factors. It is done by expressing each factor as the best linear combination of a small number of unknown common unobserved factors. The success of any factor analysis depends on obtaining really meaningful factors.
3. Based on the above assumption, the original extracted data set was reduced (37*10 factors). Hence the final data set (37*10 factors) was used in the analysis.

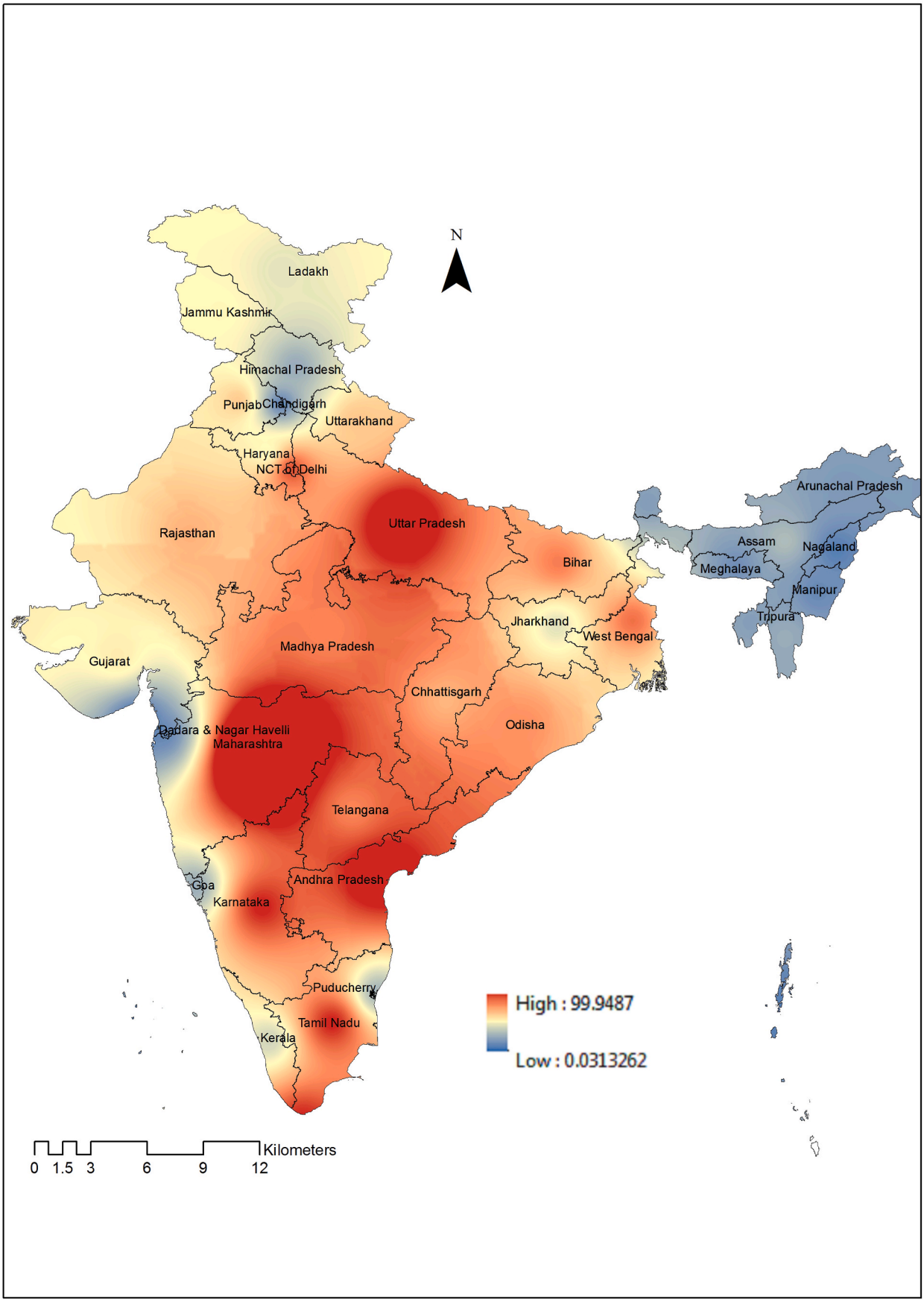

Fig. 1. Inverse Distance weighted estimates based on several high risk related factors of COVID-19, India, 2020.

Steps involved:

(i) As a first step, the factors were standardized to eliminate the effect of different scales of measurement. The standardized dataset (shown in the supplementary material) was given as an input to the factor analysis program in SPSS software.
(ii) The next step was to examine the correlation matrix between the factors. The values ranged from 0.1 to 0.9 in absolute value. The majority of the correlates was $\geq$0.3
(iii) The communality values for the factors $\geq$0.7 are shown in the last column of Table 2.
(iv) Further suitability of the data set or analysis was assessed using Bartlett's test of sphericity, Kaiser-Meyer-Olkin measure of sampling adequacy, and inspection of residuals and rotated factor loadings.
(v) Bartlett's test of sphericity, which tests that the correlation matrix is the identity on the assumption of multivariate normality, was found to be highly significant of P($<$0.001)
(vi) Kaiser-Meyer-Olkin measure of sampling adequacy was 0.61, which represents an acceptable value for factor analysis.
    (i) Further, a smaller subset of four factors was extracted using the eigenvalue greater than one. Varimax rotation was used to improve the factors and readily identifiable, and finally, the factor scores were obtained (shown in supplementary material).
    (ii) Percentage of variation (is shown in the last column of Table 2) was used as weights, and the initial score for each state was obtained (is shown in the last column of Table 3).
    (iii) Let Minimum initial score denoted by (MIN_INS) and Maximum initial by (MAX_INS)
    Then Standardized Score = [(INITIAL SCORE of the state – MIN_ INS)/(MAX_INS –MIN_INS)]*100. The standardized score is shown in the last column of Table 3.

References

1 https://edition.cnn.com/interactive/2020/health/coronavirus-maps-and-cases/ accessed on Oct 12, 2020.
2 https://www.mohfw.gov.in/accessed on Oct 12, 2020.
3 https://www.livemint.com/news/india/covid-19-infection-rate-in-india-has-slowed-down-to-flatten-the-curve-govt-11587667324493.html accessed on April 15, 2020.
4 https://www.ndtv.com/india-news/coronavirus-impact-visas-to-india-suspended-till-april-15-2193382 accessed on April 15, 2020.
5 https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html accessed on April 15, 2020.
6 https://www.census2011.co.in/accessed on April 15, 2020.
7 https://niti.gov.in/state-statistics accessed on April 15, 2020.
8 http://naco.gov.in/sites/default/files/HIV%20Estimations%202017%20Report_1.pdf accessed on April 15, 2020.
9 https://nhm.gov.in/index4.php?lang=1&level=0&linkid=457&lid=686 accessed on April 15, 2020.
10 http://www.cbhidghs.nic.in/Ebook/National%20Health%20Profile-2018%20(e-Book)/files/assets/common/downloads/files/NHP%202018.pdf accessed on April 15, 2020.
11 http://rchiips.org/nfhs/nfhs4.shtml accessed on April 15, 2020.
12 http://socialjustice.nic.in/writereaddata/UploadFile/HANDBOOKSocialWelfareStatistice2018.pdf accessed on April 15, 2020.
13 http://164.100.117.97/WriteReadData/userfiles/ISFR2019%20Vol-I.pdf accessed on April 15, 2020.
14 http://statisticstimes.com/demographics/population-of-indian-states.php accessed on April 15, 2020.
15 https://thieme-connect.com/products/ejournals/pdf/10.1055/s-0038-1676242.pdf accessed on April 15, 2020.
16 https://link.springer.com/article/10.1007/s10389-019-01072-6/tables/1 accessed on April 15, 2020.
17 Manly BFJ. *Multivariate Statistical Methods - A Primer*. third ed. New York: Chapman & Hall; 1986:76–89.
18 Cattell RB. *Factor Analysis an Introduction and Manual for the Psychologist and Social Scientist*. Harper & Brothers; 1952.
19 Harman HH. *Modem Factor Analysis*. third ed. Chicago: The University of Chicago Press; 1976.
20 Sekhar CC, Indrayan A, Gupta SM. Development of an index of need for health resources for Indian States using factor analysis. *Int J Epidemiol*. 1991;20:246–250.
21 Bhagavandas M, Joshua V. Mapping co-variates of mortality up to age of five years for Indian states. *Indian J Publ Health*. 2003;47:22–26.
22 Vasna J, Gupte MD, Adhikary R, et al. Index based mapping of high-risk behaviors for HIV among female sex workers in India. *Indian J Med Res*. 2012;136:14–22.
23 *Statistical Package for the Social Sciences (SPSS) for Windows (Version 18)*. Chicago, Illinois, USA: SPSS Inc; 2009.
24 Teegavarapu RSV, Chandramouli V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *J Hydrol*. 2005;312:191–206.
25 Environmental Systems Research Institute. *Inc. ArcGIS Desktop: Release 10*. Redlands, CA, USA: ESRI, Inc.; 2010.
26 https://www.timesnownews.com/india/article/coronavirus-zones-and-their-meanings-covid-19-containment-plan-what-are-red-orange-green-zones/580094 accessed on Oct 10, 2020.
27 https://www.jagranjosh.com/general-knowledge/north-eastern-states-at-a-glance-seven-sisters-of-india-1450951506-1 accessed on April 15, 2020.
28 https://science.thewire.in/health/coronavirus-social-distancing/accessed on April 15, 2020.