INNOVARE JOURNAL OF MEDICAL SCIENCES

Vol 10, Issue 5, 2022

ISSN - 2321-4406

Research Article

A CRUCIAL: CLUSTER AND FACTOR ANALYSIS APPROACH TO CLASSIFY ZIKA VIRUS DATASETS IN USA AND INDIA

VENU PARITALA*, HARSHA THUMMALA

Department of BioTechnology, Vignan's Foundation for Science, Technology and Research, Guntur, Andhra Pradesh, India.
Email: vvenuparitala@gmail.com

Received: 11 June 2022, Revised and Accepted: 30 July 2022

ABSTRACT

Objective: The aim of this research is to analyze the relationships between the counts of cases and the deaths due to Zika Virus in USA, India, countries that are severely affected from this pandemic disease.

Methods: Cluster correlation is used to determine the relationships among these countries. Then, factor analysis is applied to categorize these countries based on their relationships. R (reproducible research with R and R studio, second edition, 2018) is server which is hosted in a cloud platform. For the development of analytical pipeline, various R packages were used.

Results: The novel analysis which results in factor analysis it reports the count of positive cases in hugely Kerala and Trivandrum in India and American Samoa, Puerto Rico, and us Virgin Islands territories in the USA.

Conclusion: However, as fast as viruses spread, the detection of pandemics, and taking early, these analyses help to suggest virus affected states in disparate countries, then government take care of them.

Keywords: Cluster analysis, Factor analysis, Zika virus, WHO, CDC, Coefficient correlation, Cases analysis.

Mathematics Subject Classification: 62-07, 62H30, 62H25, 62-04, 62P10, 62P35

 DOI: http://dx.doi.org/10.22159/ijms.2022v10i5.45425. Journal homepage: https://innovareacademics.in/journals/index.php/ijms

INTRODUCTION

The zika virus is a flavivirus spread by mosquitos that were discovered in monkey in Uganda in 1947. It was later discovered in humans in Uganda and the united republic of Tanzania in 1952. Zika virus outbreaks have been reported in Africa, America, Asia, and the Pacific. From 1980's, a few occasional cases of human infections were identified in Africa and Asia, usually accompanied by minor diseases. It affects fetal development and results in serious neurodevelopmental problems. The virus has a significant influence on medical, economic, and sociologic circumstances in many nations. [1,2], it does not survive to present the highest mortality rate. The fundamental goal in most of this research is to track and anticipate how the virus will move across the country. Researchers have worked hard to estimate how long the outbreak would last. As a result, numerous research groups have used various modeling strategies for prediction and have come up with many intriguing conclusions [3].

The primary goal of this research is to improve monitoring of infected locations, which will be critical in identifying the severity of the new Zika virus's spread . They will then help to reduce the number of sick and deceased people. A cluster method is employed in this research to find correlation which is used to determine the relationships among these countries [4,5]. Then, the factor analysis is applied to categorize these countries based on their relationships. The analysis also proves results out the mean, median of counts of cases and the deaths due to Zika virus in USA and India [6].

Data description

The initial statistics used here are acquired from the particular site generated by who (https://www.who.int/emergencies/disease-outbreak-news/item/zika-virus-disease-india) and cdc (https://www.cdc.gov/zika/reporting/2017-case-counts.html). Due to the generally low number of cases in many countries at the start of the pandemic, the 1st day used here is the July 08, 2021, while the last day used is the August 04, 20201 in India, and the 1st day used here is the of July 2015, while the last day used is the December 4, 2021 in the United States. All time-series data were collected and integrated using excel 2019. An algorithm to provide consistent clustering of various countries concerning positive cases, death rate, and active cases per different territories in India and the USA was developed.

MATERIALS AND METHODS

Cluster analysis

The k-means clustering analysis was carried out, which is a non-linear unsupervised method for clustering data based on similarities or groups. It makes an effort to categorize the data into a preset number of instances [7,8]. This clustering analysis is performed by r, r is a programming language used by data miners and statisticians to construct statistical applications and data analysis [9,10]. Clustering is a set of methods for locating subsets of observations in a data set. The response variable, this is an unsupervised technique. Its aim is to discover data correlations without being taught by a response variable. Clustering assists us in determining which observations are related and labeling them accordingly. K-means clustering is the simplest and most commonly used clustering method for splitting a dataset into a collection of k groups (Fig. 1). The hierarchical and k-means algorithms are two of the most often used methods. We chose both of these clustering algorithms in our work due to their simply interpretable visualization and sensible interpretation.

Data preparation

1. Observations (individuals) are represented by rows, while variables are represented by columns.
2. Any missing data value must be deleted or approximated.

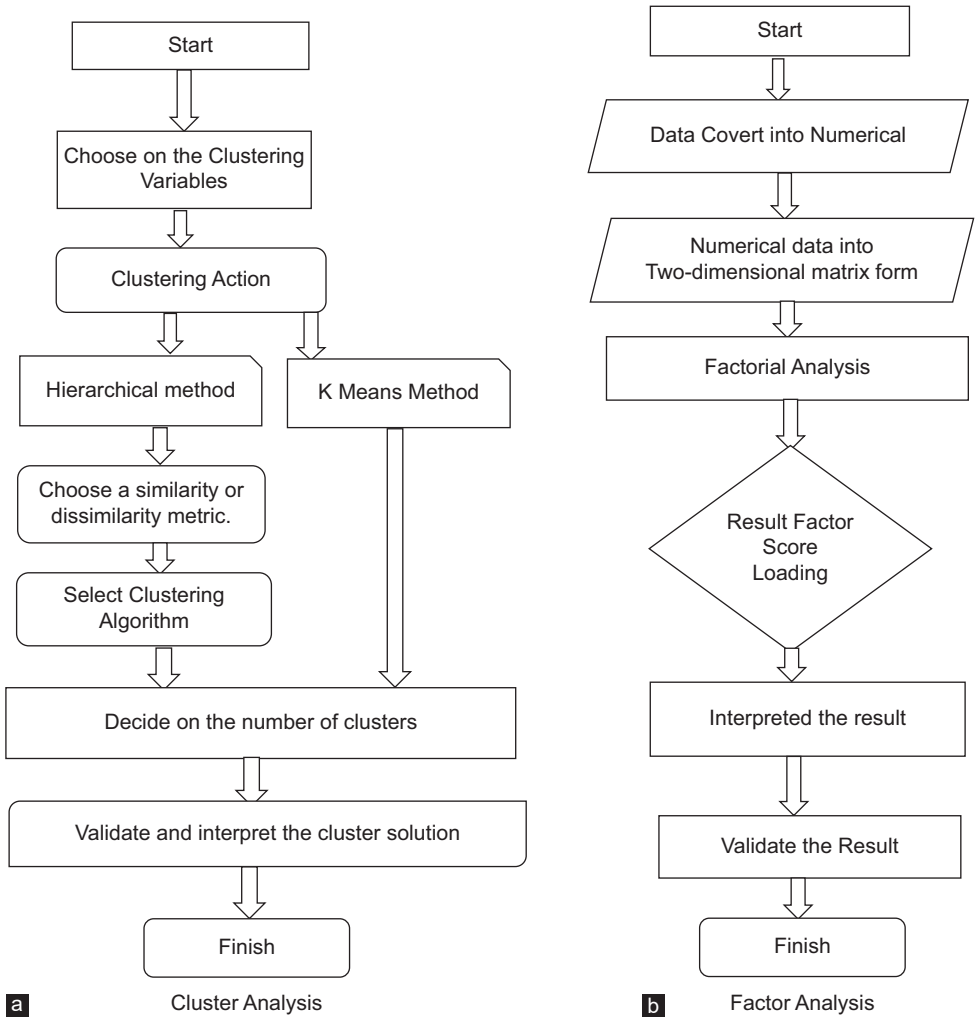

Fig. 1: (a and b) Flow charts of cluster and factorial algorithm

Table 1: Parameters of Zika virus in India

S. No	Parameters
1	Year
2	Countries
3	State
4	Cases
5	Test
6	Laboratory

Table 2: Parameters of Zika virus in the USA

S. No	Parameters
1	Year
2	Country
3	Territories
4	No. cases
5	Deaths
6	Presumptive viremic blood
7	Presumptive viremic blood (%)

Table 3: Standard deviation of the USA Zika virus dataset

S. No	Cases	Year	Presumptive. viremic. blood	Presumptive. viremic. blood (%)
Min	1	2015	0	0
1st Quality	1	2016	0	0
Median	5	2017	0	0
Mean	5.78	2017	22.07	13.33
3rd Quality	11.9	2018	0	0
Max	12	2021	325	100

Table 4: Standard deviation of India

S. No	Cases
Min	1
1st Quality	1
Median	2
Mean	5.78
3rd Quality	11.9
Max	13

3. To make variables comparable, the data must be normalized (i.e., scaled). Remember that standardization entails altering the attributes likewise they possess a mean of zero and a standard deviation of one.

Grouping of scrutiny necessitates the use of various methods for calculating the distance or (dis) parallelism between every individual set of scrutiny. A disparallelism or distance matrix is an outcome of this algorithm. There are several techniques for calculating these distance particulars; the selection of distance measurements is an important stage in clustering. It specifies how parallelism of paired components (a, b) is determined, and it has an impact on the form of the clusters. Choice of distance calculated is a censorious pace in clustering.

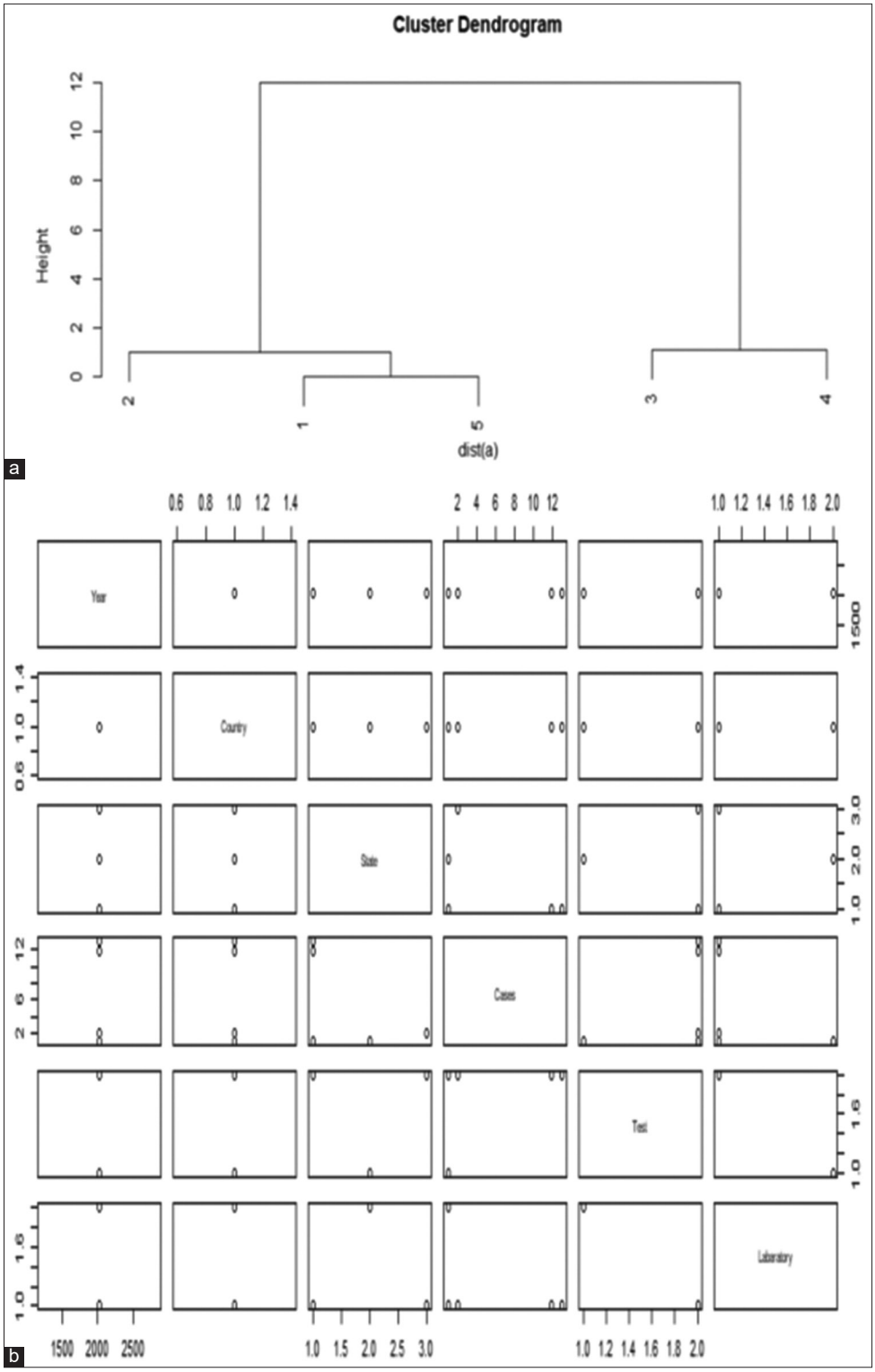

Fig. 2: (a and b) Hierarchical of cluster analysis in India

It defines in what way the parallelism of paired components (a, b) is deliberate and will influence the structure of the clusters. The classical techniques for distance measures are Euclidean and Manhattan distances, which are defined as follow:

Euclidean distance

$$d_{distance}(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (1)$$

Manhattan distance

$$d_{man}(a,b) = \sum_{i}^{n} |(a_i - b_i)| \quad (2)$$

The goal of the k-means clustering is to define clusters in likewise a way that the total intra-cluster variance (also known as a total within-cluster variation) is minimized. Several k-means algorithms are available. The Hartigan-Wong method (1979) is the standard approach, and it

defines entire within-cluster variance as the sum of squared euclidean distances between items and the associated centroid:

$$C_k = \sum_{c_k} (a_i - \mu_k)^2 \quad (3)$$

Factorial analysis

Factor analysis is an approach in favor of decreasing an enormous number of attributes to a small number of elements. The strategy takes greatest ordinary difference of all attributes to combining into a single score. For future research, we may utilize this score as an index of all factors. Exploratory factor analysis (EFA) is a technique for finding the factor structure of a measure and evaluating its internal reliability. When researchers have no assumptions about the nature of their measure's underlying factor structure, EFA is typically recommended. The main purpose of this study is to identify the key variables that influence the Zika virus data set. To investigate such a sophisticated algorithm, an objective strategy encompassing many techniques (statistics, r, gis, descriptive analysis, and graphical visualization) was adopted [11]. Factor analysis (FA) is concerned with variable correlations so that variables within a factor are highly associated with one another, but variables within other factors are significantly uncorrelated. Factor analysis generates straight amalgamations of components to extract the attributes underlying commonality. To the expanse that the attributes possess an underlying commonality, fewer components constitute the bulk of the difference in the data set. This allows us to explain an underlying concept in a model by aggregating a large number of observable variables, making the data more understandable. The flexibility of our data, aa, is supplied by, and its approximate is composed of the flexibility indicated by variables explained by a linear combination of factors and the flexibility that cannot be described by a linear combination of factors.

Table 5: K-means of cluster analysis India

Year	Cases	Cluster
2015	0.8620622	3
2018	−0.9407894	2
2021	−0.7833351	1

Table 6: Coefficient correlation of USA

No cases (%)	Presumptive viremic blood	Sumptive viremic blood (%)
0.8218183	−0.2632681	−0.3789324
−1.2259016	−0.2632681	−0.3789324
0.8218183	−0.2632681	−0.3789324
0.8422955	−0.2632681	−0.3789324
−1.16447	3.6141701	2.4630604
0.7603867	−0.2632681	−0.3789324
0.8218183	−0.2632681	−0.3789324
−1.0825613	−0.1916846	2.4630604
0.6784779	−0.2632681	−0.3789324
0.8218183	−0.2632681	−0.3789324
−1.2054244	−0.2632681	−0.3789324
0.8013411	−0.2632681	−0.3789324
−1.2259016	−0.2632681	−0.3789324
−1.1849472	−0.2632681	−0.3789324
0.7194323	−0.2632681	−0.3789324

Table 7: K-Means cluster analysis of USA

S. No	Deaths	Presumptive viremic blood	Sumptive viremic blood (%)
1	−1.1644700	6141701	2.4630604
2	0.7876896	−0.2632681	−0.3789324
3	−1.1849472	−0.2489514	0.1894662

$$\Sigma = AA^T + \Psi \quad (4)$$

To, calculates an initial estimate of Ψ^Ψ^ and factors S−Ψ^S−Ψ^, or R−Ψ^R−Ψ^ for the equivalence array. Regrouping the estimated covariance and equivalence array with the estimated AA array yields:

S−Ψ=X
A=S−Ψ (5)
R−Ψ=X

of S−Ψ^S−Ψ^ or R−Ψ^R−Ψ^. Ψ^Ψ^ is a diagonal array is the principal component technique, the h^2ih^i2, is equal to sii−ψ^isii−ψ^I for S−Ψ^S−Ψ^ and 1−ψ^i1−ψ^I for A−Ψ^R−Ψ^. The correlation of KK or AA is replaced by their respective commonalities in ψ^iψ^i which gives us the following forms:

$$S - \Psi^{\wedge} = \begin{bmatrix} k1\ a1\ s1 \\ s2\ k2\ a2 \\ a3\ s3\ kn \end{bmatrix} \quad (6)$$

$$R - \Psi^{\wedge} = \begin{bmatrix} k1\ p1\ p1 \\ p2\ k2\ p2 \\ p3\cdots, kn \end{bmatrix} \quad (7)$$

The squared multiple correlations between the observation vector YIYI and the other p1p1 variables are used to make an initial estimate of the commonalities. In the instance of PP, the squared multiple correlations are comparable to the following:

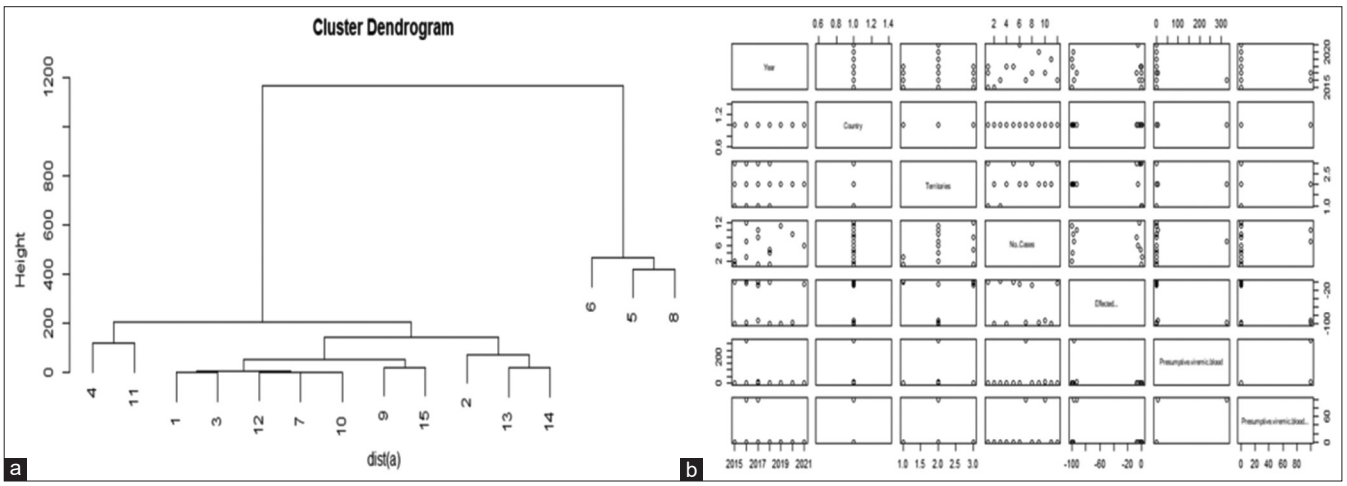

Fig. 3: (a and d) Hierarchical cluster in USA

Table 8: Factorial analysis of Zika virus in India

S. No	Vars.	n	mean	SD	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	Se
Year	1	5	2021	0	2021	2021	0	2021	2021	0	NaN	NaN	0
Country*	2	5	1	0	1	1	0	1	1	0	NaN	NaN	0
Test*	5	5	1.8	0.4472136	2	1.8	0	1	2	1	−1.0733126	−0.92	0.2
Laboratory*	6	5	1.2	0.4472136	1	1.2	0	1	2	1	1.0733126	−0.92	0.2
State*	3	5	1.6	0.8944272	1	1.6	0	1	3	2	0.6037384	−1.67	0.4
Cases	4	5	5.78	6.1148998	2	5.78	1.4826	1	13	12	0.2914057	−2.232069	2.734666

Table 9: Factorial analysis of USA

S. No	Vars	n	Mean	SD	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	Se
Year	1	15	2017.2	1.820518	2017	2017.076923	1.4826	2015	2021	6	0.519746	−0.825977	0.4700557
Country*	2	15	1	0	1	1	0	1	1	0	0	0	0
Territories*	3	15	2	0.7559289	2	2	1.4826	1	3	2	0	−1.366667	0.19518
No. Cases*	4	15	5.4	3.9242834	5	5.2307692	5.9304	1	12	11	0.2562797	−1.532938	1.0132456
Effected.*	5	15	6.733333	3.7122705	7	6.8461538	5.9304	1	11	10	−0.155554	−1.661757	0.9585041
Presumptive. viremic. blood	6	15	22.066667	83.8182274	0	0.4615385	0	0	325	325	3.13099	8.379036	21.6417732

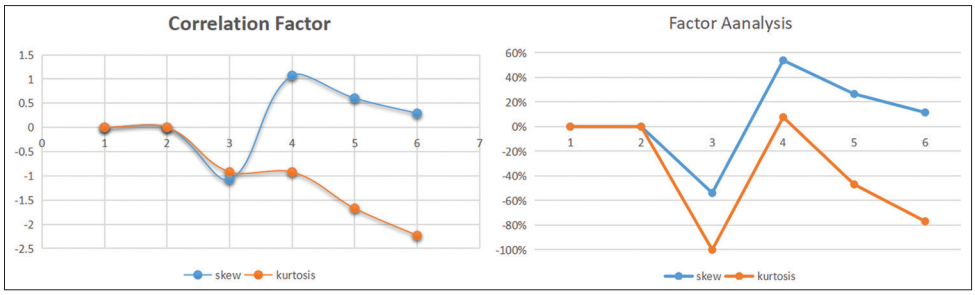

Fig. 4: Factorial analysis in India

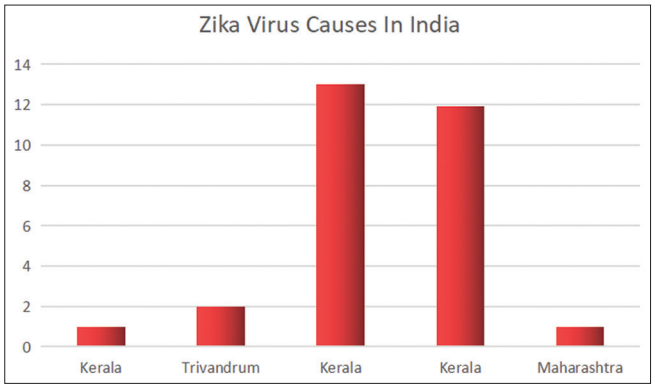

Fig. 5: Count number of positive cases in India

$$k = 1 - \frac{1}{p} \quad (8)$$

Choice of parameters in India

The data set was retrieved from who (who.gov.in), the dataset contains stating month of January 2021, essentially due to the increase of cases in different states in India. We acquire the best fit to the noticed cases and effects in states of Zika virus by utilizing the following set of parameters that are constituents in Table 1.

Choice of a parameter in USA

The data set was retrieved from the cdc (https://www.cdc.gov/zika/reporting/2017-case-counts.html), the dataset contains stating 2015 in the USA and ending 2021, essentially due to the increase of cases in different territories in the USA. We acquire the best fit to the observed cases, death rate, and effects in territories of Zika virus by utilizing the following set of parameters, which are constituents in Table 2.

RESULTS

Standard deviation of India and USA dataset

Tables 3 and 4 show the rates of cases and deaths caused by the virus in various regions that have been severely affected. The cases of correlation factor observer in the year of 2016 were reported 1% and 13.33% of arithmetic means and median of cases and deaths due to Zika virus in 2017. It also observes minimum and maximum rate of effect due to the virus. The order of analysis cases reported in USA minimum is 1% and maximum is 12%, meanwhile in India minimum 1% and maximum 13%.

Cluster analysis in India

The hierarchical clustering consistently generated viable results it observes in Fig. 2. The inclusion of random instance in the construction of data set resulted in a normal distribution of inter-cluster distances, where inter-cluster distance is a random variable allowing the simulation to more accurately represent in this research for interpretable analysis, we generated data points by taking mean of each data points of 1000 repetitions and regressing them on dummy codes for methods, number of clusters, number of observations, and conditional probability. The results of a linear model of the mean are a cluster analytic techniques. The number of clusters, number of observations (sample size), conditional probability, and interactions between the method and each of these other variables are presented in Table 5. The total difference in accuracy across clustering algorithms was not statistically significant. Higher conditional probabilities linking the indicators to the clusters were related with increased accuracy

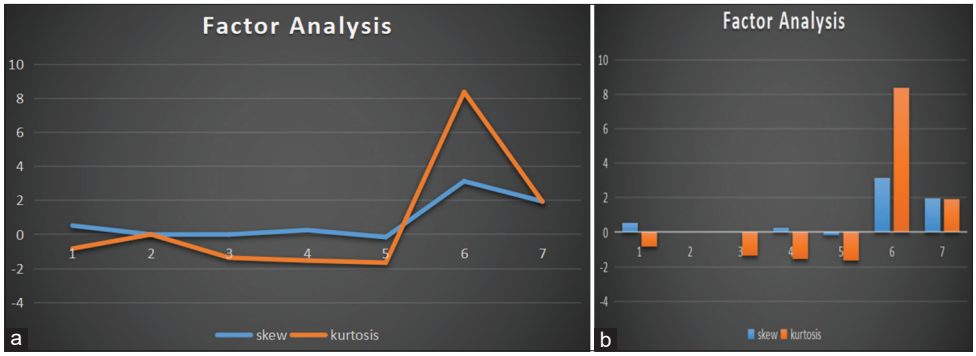

Fig. 6: (a and b) Factorial analysis of Zika virus in the USA

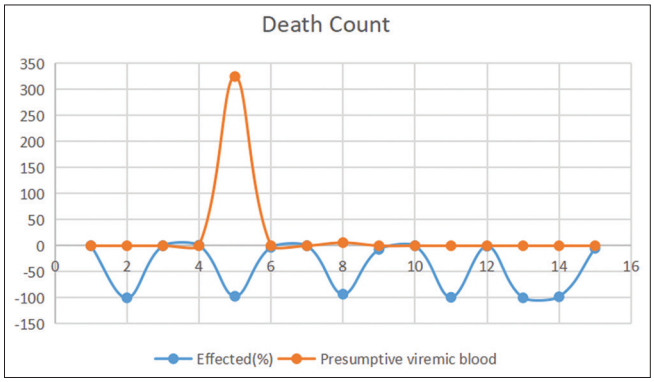

Fig. 7: Count positive cases in USA

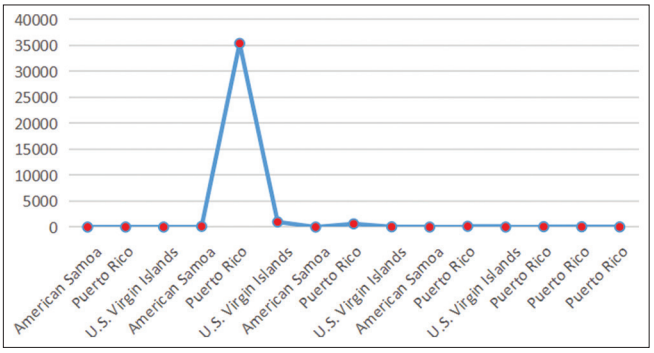

Fig. 9: Count of Cases and deaths affecting due to Zika virus in different terrorizes in USA

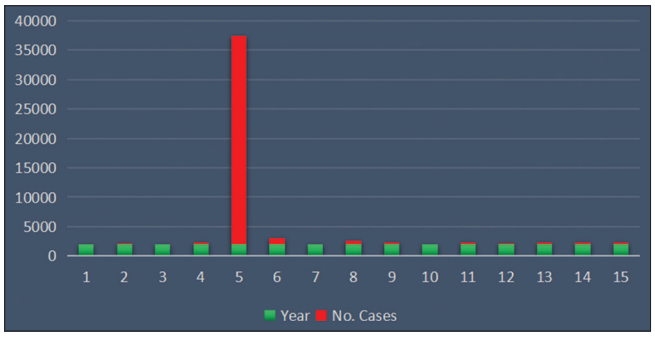

Fig. 8: Count of death cases in USA

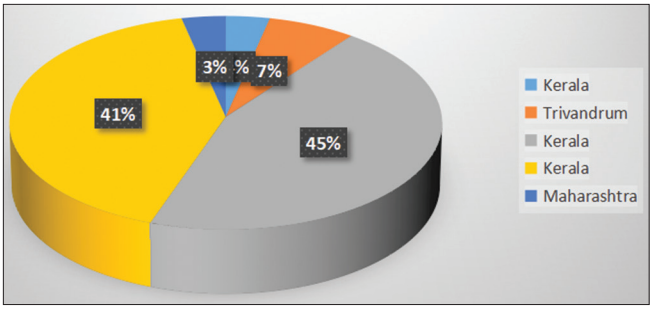

Fig. 10: Count of cases and deaths affecting of Zika virus in different states in India

(p<0.87); however, this was significantly less for hierarchical clustering (b=–0.9).

Cluster analysis in USA

Hierarchical clustering consistently produced feasible findings using data from the Zika virus, as shown in Fig. 3. The use of arbitrary instance in the data set building ensured in a usual apportionment of inter-cluster distances permits the simulation to more precisely depict research. Table 6 and 7 shows the findings of a linear structure of the mean of cluster analytic approaches. Greater conditional probabilities connecting the indicators to the clusters were related with increased accuracy (p<0.11); however, this was significantly less for hierarchical clustering (b=–0.8).

Factorial analysis in India

This section summarizes the findings of a factorial analysis method for dealing with a large number of instances, cases annually, and states in India based on exploratory factors (Table 8). It should be noted number of primary factors in FA that was calculated by multiplying the number of the Eigen value of the correlation analysis with greater affinity; furthermore, bartlett's test using Zika virus India dataset to confirm the value is p<0.05; then, the data set is suitable for normally distributed. It observes Fig. 4.

Count number of positive cases due to Zika virus

Fig. 5 shows to categorize the increasing positive cases rate at mostly Kerala and Trivandrum in India based on the counts of cases. The outputs show the statistical differences in the positive cases report on different' states in India.

Factorial analysis in the USA

Factorial analysis is dealing with a high number of occurrences and fatalities per year, in different territories in USA. The number of major components in FA was obtained by multiplying the number of eigenvalues of correlation analysis with a larger affinity which is observed in Table 9. Furthermore, Bartlett's test using the Zika virus USA dataset to validate value is p<0.045; therefore, the data set is adequate for normally distributed. It observes in Fig. 6.

Count positive cases in USA

Fig. 7 depicts FA technique to categorize the virus's expanding spread across the United States based on cumulative number of case. The

findings reveal statistical variations in the associations between the cases count.

Count death cases in USA

Fig. 8 depicts FA technique to categorize the virus's expanding spread across the United States based on cumulative number of death count. The findings reveal statistical variations in the associations between the deaths count.

CONCLUSION

The virus has a significant influence on medical, economic, and sociologic circumstances in many nations. it does not survive to present the highest mortality rate. Many diverse disciplines are attempting to identify solutions and drive strategies for a wide range of highly crucial problems. This study presents a novel analysis which results in factor analysis it reports the count of positive cases in hugely Kerala and Trivandrum in India (Figs. 9 and 10) and American Samoa, Puerto Rico, and US Virgin Islands territories in the USA (Fig. 9). Claustral analysis results observe to interrelationship of the count of cases and deaths of territories in the USA and different states in India.

ACKNOWLEDGMENT

I sincerely thank Vignan University for its constant support.

AUTHOR'S ROLE

Venu paritala conceived of the presented idea and developed the theory and performed the computations. Harsha thummala verified the analytical methods, investigate and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

CONFLICTS OF INTEREST

The author declares no conflicts of interest.

REFERENCES

1. Chatterjee K, Chatterjee K, Kumar A, Shankar S. Healthcare impact of the Covid-19 epidemic in India: A stochastic mathematical model. Med J Armed Forces India 2020;76:147-55.
2. Sengupta P, Ganguli B, SenRoy S, Chatterjee A. An analysis of Covid-19 clusters in India: Two case studies on Nizamuddin and Dharavi. BMC Public Health 2021;21:631.
3. Shumway RH, Stoffer DS. Time Series Analysis and its Applications. New York: Springer Verlag; 2000.
4. Zarikas V, Poulopoulos SG, Gareioud Z, Zervas E. Clustering analysis of countries using the Covid-19 cases dataset. Data Brief 2020;31:10578.
5. Brereton RG. Multivariate Pattern Recognition in Chemometrics: Illustrated by Case Studies. New York: Elsevier; 1992.
6. Henry D, Dymnicki AB, Mohatt N, Allen J, Kelly JG. Clustering methods with qualitative data: A mixed methods approach for prevention research with Small Samples. Prev Sci 2015;16:1007-16.
7. Majcherek D, Weresa MA, Ciecierski C. A cluster analysis of risk factors for cancer across EU countries: Health policy recommendations for prevention. Int J Environ Res Public Health 2021;18:8142.
8. Noorbakhsh F, Abdolmohammadi K, Fatahi Y, Dalili H, Rasoolinejad M, Rezaei F, *et al*. Zika virus infection, basic and clinical aspects: A review article. Iran J Public Health 2019;48:20-31.
9. Paritala V, Reddy SR, Sukesh K, Pasupuleti B. IMMUNE DB: A large internet framework for collecting and analysing biological immune results. J Appl Bioinform Comput Biol 2021;10:6.
10. Paritala V, Reddy SR, Kalva S. Neglected disdb: A broad internet framework for gathering and analysing data from neglected diseases. J Appl Bioinform Comput Biol 2021;5:2.
11. Omuya OS, Tayo AO. Analytical framework to minimize the latency in tele-herbal healthcare service. J Eng Res Sci 2022;1:39-50.