

Identifying Rainfall Diversity through Clustering

Vikas Singh^{*1} and Mukesh Kumar²

Department of Statistics,
Central University of South Bihar,
Gaya-824236, Bihar.

^{*1}Email : singhvikasbhu@gmail.com

²Email : Mukseh.kumar@cub.ac.in

Abstract

The uneven distribution of rainfall at different regions is an important issue now days. Particularly the places of low rainfall are of serious concerns. These places are lacking in drinking water and lot of other related problems like unsatisfactory irrigation, sanitation, etc. The rainfall diversity in India through clusters (similar regions) has been observed under this study. The rainfall distribution in India is varied over seasons, and high-rainfall observed under monsoon season. In this paper, we have analyzed rainfall data in the context of India through divisive clustering technique K-means and model based probability clustering. The techniques are used over rainfall data given by the Indian Meteorology Department (IMD), and a comparative finding has been reported. The annual and seasonal rainfalls data for the year 2016-17 has been taken for this study. Optimal numbers of clusters are obtained by using the Elbow method in case of K-means clustering and Bayesian Information Criterion (BIC) is used for model based clustering. Moreover, the results obtained through K-means and model based cluster analysis have been reported in terms of spatial distribution. The main advantage of splitting the states of rainfall into homogenous groups based on similar features is that it eliminates the necessity to carry out a detailed drought characteristic analysis for any location of interest. Therefore detection of rainfall pattern in a spatial domain is very important & useful for policy makers, climatologists, government, hydrologists and local and regional planners & managers.

Keywords: Clustering, K-means, Model based clustering, EM-algorithm, BIC, rainfall.

Mathematics Subject Classification (MSC): 62H30, 68T10, 62Hxx, 62H12

1. INTRODUCTION

The rainfall and other water resources play a vital role in survival. These are sometimes called as lifeline of survival. As far as Indian rainfall is concerned it is unevenly distributed across the country. Therefore, a few parts of the country suffer from huge floods while others are drought regions. Also, rainfall and weather in India are not uniform throughout the year. The climate of India covers a complete assortment of climatic conditions across a tremendous topographical scale. Generally climate in south of the nation is hotter than north of India. The countries most prominent part doesn't encounter temperatures beneath 10°C (50°F) in winter and the temperature surpasses 40 °C (104 °F) during summer the nation over. The India Meteorological Department (IMD, Ministry of Earth Sciences, Govt. of India) follows the worldwide climate of 4-seasons with a few neighbourhood alterations. It depicts the four climatic seasons as (1) winter (December to February). December and January are the coolest months, when

mean temperatures lie between 10–15°C for example (50–59°F) in the north-west, temperatures increment as one proceeds towards the equator, coming to in the middle of 20–25°C (68–77°F) in southeast India. (2) Summer or pre-storm season (March to May). April and May are the regularly hot month for north locales of India. The expected temperatures stay in the reach 32–40°C (90–104 °F) in the fringe limit. (3) Monsoon or blustery season (June to September). This season is overwhelmed by the dampness which gradually clears the nation around late May or early June. Storm downpours start to subside from North India toward the start of October. South India regularly gets more precipitation when contrasted with northern pieces of India. (4) Post-storm or harvest time season (October to November). In the north-west of India, October and November usually are cloudless.

The rural economy of India primarily based on agricultural production which quietly based on rainy season. Rainfall is one of the important factor and sources of water supply for agricultural based activity in India. Therefore, finding the places of high and low rainfall in the country might be important & useful from various viewpoints. Particularly, as all the states observe different quantity of rainfalls, a grouping into similar states are likely to decrease the problems involved in managing the rainfall among the various states of India.

2. LITERATURE REVIEW

The studies are regularly carried out for, taking the consistent decision for management of flood, drought, irrigation problems etc. Finding the homogenous groups of places based on rainfall may minimize the efforts in management as dealing with a small no. of homogeneous groups is always better. In literature, studies are found for describing sub-stations of a spatial province by rainfall and useful for managing water resources and planning of land use [Raziei et al. (2008), Shirin and Thomas (2016)]. Meddi et al. (2013) indicated that separation into smaller group by rainfall is important to recognize the pattern of weather and Indian climate.

As far as researches which focuses for use of cluster analysis are enormous. The method varies from simple heuristic to some advance methods like model based clustering approach. The popularity of these methods is increasing day by day in various studies and particularly in finding clusters of rainfall data Some contributions are due to Modarres (2006), Sarhadi and Heydarizadeh (2014)), Goyal and Gupta (2014), etc.

3. DATA SOURCES

The data comprises of records of annual and seasonal rainfall of 35 states of India collected for the year 2016-2017. This study uses data on seasonal rainfall of various states as recorded by Indian Meteorological Department (IMD). The average rainfall of India is around 24993.17 mm. For each of the states, annual rainfall shall be computed by adding the month wise rainfall over the all the months in a year. Similarly for each of the states, the rainfall in different season like the winter, summer (pre-monsoon), rainy session and autumn sessions are calculated by adding the monthly rainfall over the months January to February, March to May, June to September and October to December, respectively.

4. STATISTICAL TECHNIQUES

Techniques are available which might be used for the analysis of rainfall data. However for finding the high rainfall and low rainfall regions are primarily important from resource management point of view. Clustering techniques are one of the most important and widely used techniques for finding homogeneous groups called clusters. One commonly used clustering technique known as K-means clustering. This clustering technique is based on simple distance measures. Another more advanced technique is based on statistical models called Model based clustering. In clustering finding the optimal no. of cluster is a challenging task. This study uses Elbow method (Kaufman and Rousseeuw, 1990) and BIC (Schwartz, 1978) criterion to decide the optimal numbers of clusters for K-mean (MacQueen, 1967) and model based clustering (Banfield & Raftery, (1993), McLachlan & Basford, 1988) respectively. A brief discussion on these methods has been presented below.

5. CLUSTER ANALYSIS

Cluster analysis is an unsupervised learning techniques (which does not include any response variable) that splits a set of objects or data into homogeneous groups. It is also well known as optimal grouping of objects or items based on some characteristic of interest. The objective is to put more similar objects in a group and dissimilar objects in different groups (Johnson and Wicharn 2015).

In the method of clustering, a matrix $X(n \times p)$ of multivariate observations comprises of n -observations with p - variable or characteristics. A proximity matrix ($n \times n$) that gives the amount of dissimilarity or similarity between all possible pairs of data matrix X . In case of all metric data some of the commonly used distances are City block, Euclidean distance, etc. but Euclidean distance is most widely used in distance matrix calculations. Let there be two vectors $X = (x_1, x_2, \dots, x_p)'$ and $Y = (y_1, y_2, \dots, y_p)'$, then the Euclidean distance is given as

$$d(X, Y) = \sqrt{(X - Y)'(X - Y)} = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}$$

For adjusting the variances among p - variables, sometimes we use the statistical distance which is defined as

$$d(X, Y) = \sqrt{(X - Y)'S^{-1}(X - Y)}$$

Where S is the variance-covariance matrix of the sample data points. In general, the well-known (heuristic) methods of cluster analysis is based on the similarity/dissimilarity measures like distance matrix. Due to simplicity and wide applicability, cluster analysis is frequently used in so many areas. Among the existing methods, the agglomerative hierarchical and K-means are commonly used for finding the clusters of rainfall observations (Modarres (2006)). On the other hand, the probability distribution based method mainly used by statistician known as model based clustering. These are quite advanced and new emerging methods of clustering. Here we have tried K means (heuristic) method and the model based approach to find rainfall clusters. The K-mean clustering technique is governed by a dissimilarity matrix called distance matrix. Suppose we are interested in K clusters or groups, say $K = k$. The process begins by allocating the feature (variable) vectors in k groups in random. Therefore, for every possible cluster or group, the initial mean (centroid) of the groups is obtained by

taking average of the variable or features (Johnson and Wicharn (2015)). On the other hand, the method of model-based probability clustering is based on some statistical probability distribution and its associated inference. The data points are assumed to coming from a mixture of finite (normal) distribution and no. of cluster is obtained through BIC. The advantage of these methods is that they automatically choose the best model and hence able to find most appropriate no. of clusters.

5.1. K-MEAN CLUSTERING

In K-means cluster analysis (MacQueen, 1967), each of group or cluster is denoted by their means of sample observations belonging to the cluster. This technique is used to split the dataset into k groups/clusters, where k is the no. of cluster/group specified by the statistician/researcher/data analyst in advance. There are many procedures to apply k-means method however; one standard procedure is Hartigan-Wong technique (1979), which describes the total within sum of squares (WSS) i.e. Euclidean distances between the objects and their respective cluster mean. It split the datasets in a several groups in such a manner that the items within the same groups are more similar that is the highest intra-class similarity, whereas data from different groups are dissimilar with lowest inter-class similarity. The working methodology of K means method is presented below.

K- Mean Clustering Algorithm for Partitioning

The following steps involved in K-mean algorithm

Step 1. Choose the no. of clusters (K) in advance.

Step 2. Choose k items randomly from the given n data points as the initial cluster centroid.

Step 3. Assign rest of the data sets to the nearby cluster based on some distance measure between point and cluster point/center.

Step 4. For each k cluster update the cluster mean by computing the new mean values of all observations in the group. The mean of K^{th} cluster is a vector of length p (number of variables) containing the averages of all variables in the k^{th} cluster.

Step 5. Repeatedly minimize the total WSS, i.e. repeat the step 3 and 4 till the assignments of cluster stop varying.

5.2. MODEL BASED CLUSTER ANALYSIS METHOD

Model-based clustering methods are based on mixture of finite probability density function. In the theory of model-based probability clustering, it is assumed that the observations are coming from a mixture of finite density function. Generally, multivariate normal distributions give exciting results when used as a component distribution for mixture model. Many researchers including McLachlan & Basford, 1988, Melnykov & Maitra, 2010, etc. have contributed in this field and obtained interesting results. A brief theory and methodology have been presented below.

Let there be a set of n , p -dimensional vectors $y_1, y_2, \dots, y_n$ of observations coming from a mixture of k -component density functions and each component having some unknown mixing proportions or weights $\pi_1, \pi_2, \dots, \pi_k$ (McLachlan & Peel, 2004). Assuming that there is finite no. of clusters in the data and the data belonging to a particular component distribution are forming a separate cluster. The probability

density function (PDF) of finite mixture density (see Raftery & Dean, 2006 and Celeux & Govaert, 1995) is given by

$$f(y_j; \psi) = \sum_{i=1}^k \pi_i f_i(y_j; \theta_i)$$

Where $f_i(y_j; \theta_i)$ is the PDF of component distributions in the finite mixture model. Here $0 \leq \pi_i \leq 1$ and $\sum_{i=1}^k \pi_i = 1$ for all $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, n$. The vector $\psi = (\pi, \theta)$ of parameter includes all the parameters of finite mixture density. The parameters can be estimated using the ML (maximum likelihood) method through EM-algorithm (Demster et al. (1977)).

Banfield & Raftery, (1993) recommended that when the component density taken as a multivariate normal in model-based method of clustering, the connection between observations can be obtained by variance-covariance matrix and is decomposed in terms of eigenvalue as

$$\Sigma_i = \lambda_i O_i D_i O_i^T$$

Where $\lambda_i = |\Sigma_i|^{-\frac{1}{p}}$ is the scaler which is the highest among all eigenvalue of Σ_i . O_i is the orthogonal matrix of eigenvectors and it gives the orientation of cluster, see (Banfield & Raftery, 1993) D_i is diagonal matrix in such a way that $|D_i| = 1$, with the normalize eigenvalue of Σ_i on the diagonal in descending order and it determines the shape of cluster. λ_i measure the volume of corresponding i^{th} cluster. Mclust (Fraley and Raftery (1999)) programme is used for conducting the analysis using R-software (R Core Team, 2018).

6. FINDING THE NUMBER OF OPTIMAL CLUSTER

There is no specific method in the literature, which is said to be complete or best for finding optimal number of clusters. The optimum no. of cluster is someway personal and depends on the technique used for assessing similarities through distance matrix. Finding the appropriate and optimum no. of clusters is one of the important problems under cluster analysis. In methods like k-mean clustering or any other techniques the researcher need to specify the no. of clusters k in advanced. The useful method for finding the optimal no. of cluster in K-means (heuristic) clustering is given as

6.1. ELBOW METHOD

In methods like k-mean, there is need to express cluster in such a way that the total within-cluster sum of square (WSS) or total intra-cluster variability is minimized. The total WSS measures the compactness (variability) of cluster. The Elbow technique (Kaufman & Rousseeuw, 1990) contains the total WSS as a function of the no. of cluster. We generally finalize the no of cluster so where the addition of more no. of cluster do not improve the total variability. The steps behind the optimal no. of cluster suggested by elbow is as follows

1. Obtained the cluster through k means method for some k ($k=1,2,3,\dots$).
2. For each of k value, compute the total WSS
3. Draw the curve between WSS and the no. of cluster k .
4. The bend (knee) of the plot is usually taken as a sign of the optimal no. of cluster.

6.2. BAYESIAN INFORMATION CRITERION

This criterion is mostly used under the model based cluster analysis. The no. of clusters and most appropriate mixture density can be decided by the Bayesian Information Criterion (BIC) see (Schwartz, 1978). The BIC is the most popular choice in Gaussian Mixture Modeling see (Fraley & Raftery, 1998) and given by

$$BIC = -2\log L(\hat{\psi}) + d\log n$$

Where, $L(\hat{\psi})$ is the maximum value of log-likelihood with K component, n is the no. of observations in the matrix and d is the no. of estimated parameters from the given density. The least value of BIC has taken as criterion for optimal number of cluster in the study.

7. DATA ANALYSIS AND RESULT

In this study we have analyzed the rainfall data of 35 states of India for the year 2016–2017. For the purpose of analysis, we have used R-statistical software. In particular R libraries factoextra, mclust, are used for performing cluster analysis. The annual and seasonal trend has been obtained through MS-Excel. We have also shown the trend of annual and seasonal rainfall in India which as follows. Figure (1) gives the annual rainfall (in mm) in various Indian states while the figure (2) shows total rainfall distributed under various seasons. Figure (3) shows season wise rainfall in various states.

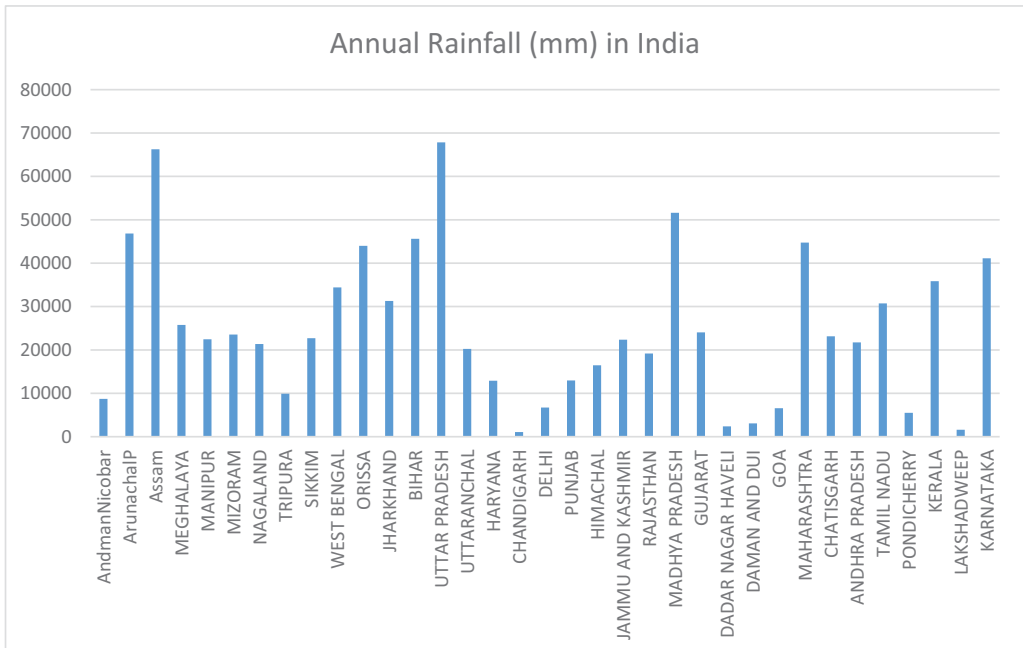


Figure 1. Rainfall (in mm) in India states wise

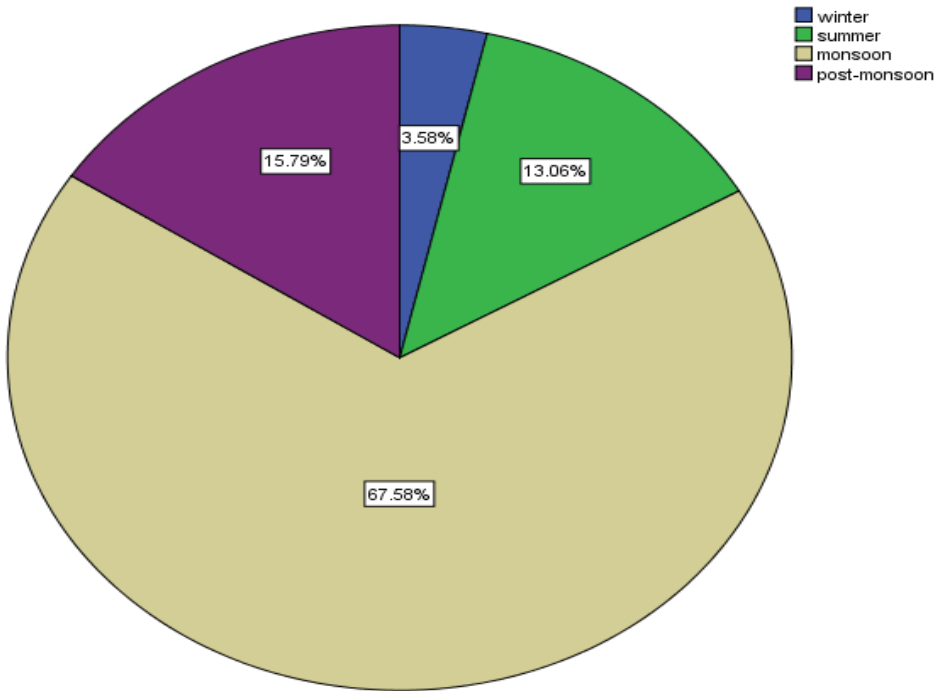


Figure 2. Percentage share of total Rainfall (in mm) in different seasons of India

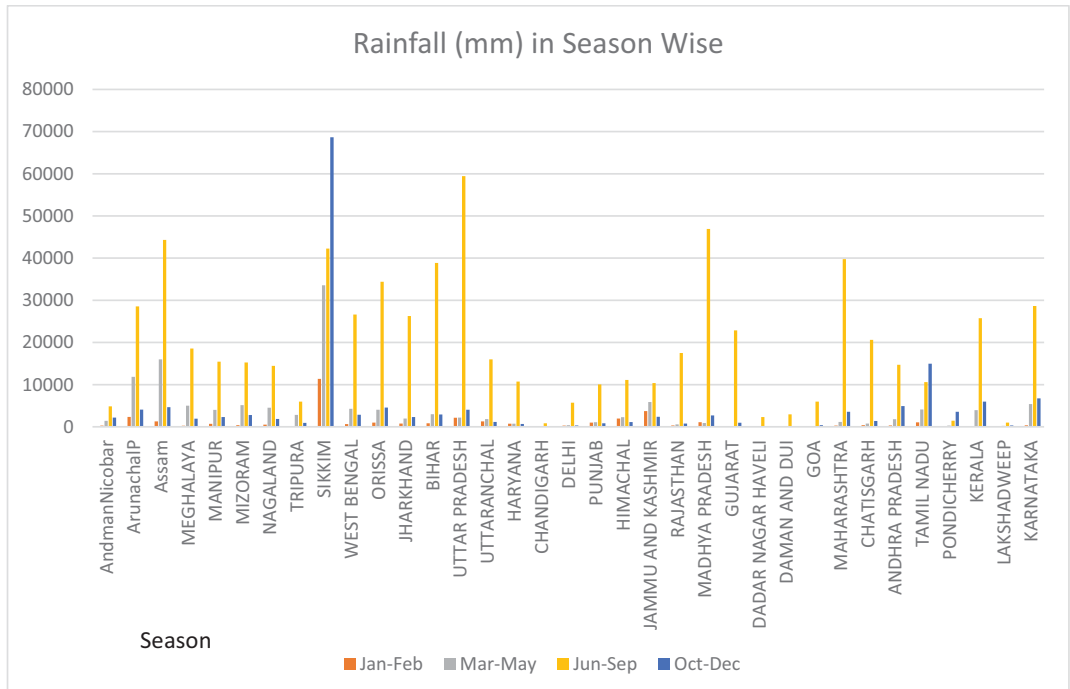


Figure 3. Season wise rainfall (mm) in India states wise

For finding the optimal numbers of cluster using K-mean clustering and model based clustering, we have used Elbow method and Bayesian Information Criterion (BIC) respectively as discussed above. We are getting 4 rainfall cluster in respect of India using Elbow method in case of k-mean clustering and is given in figure no. (4). The number of cluster using K-means technique is presented in figure (5). We can see from the figure that all the 35 states of the country is divided in 4 group or cluster of sizes 11, 16, 7, 1 and it gives the complete idea of homogeneity of rainfall station. The spatial spreading of rainfall in India is presented in figure (6) which shows only 4 cluster of rainfall data.

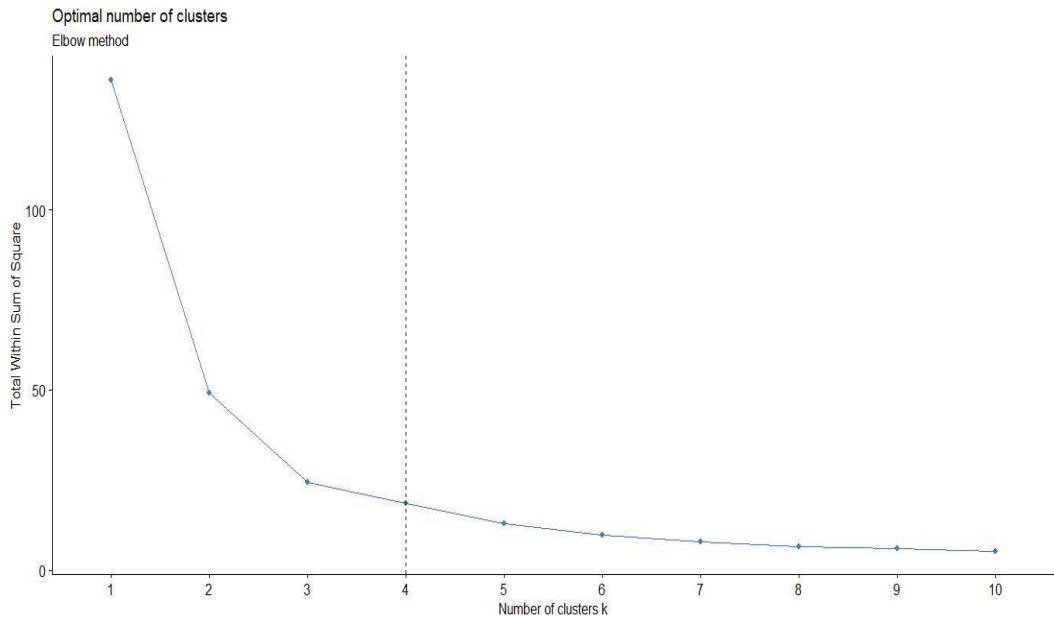


Figure 4. Optimal no. of cluster using Elbow method for K-mean

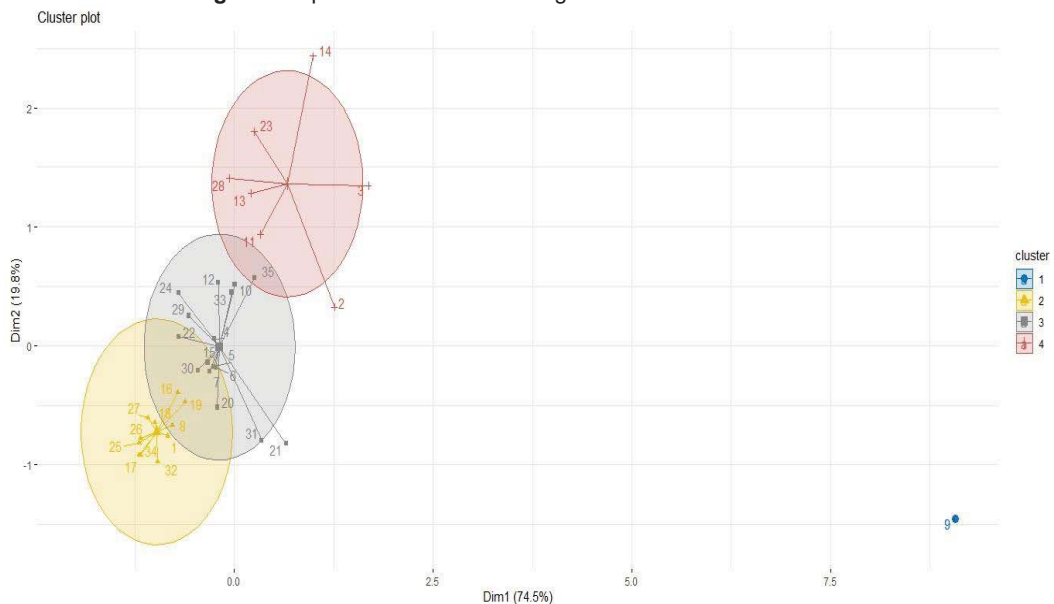


Figure 5. Number of cluster using k-means

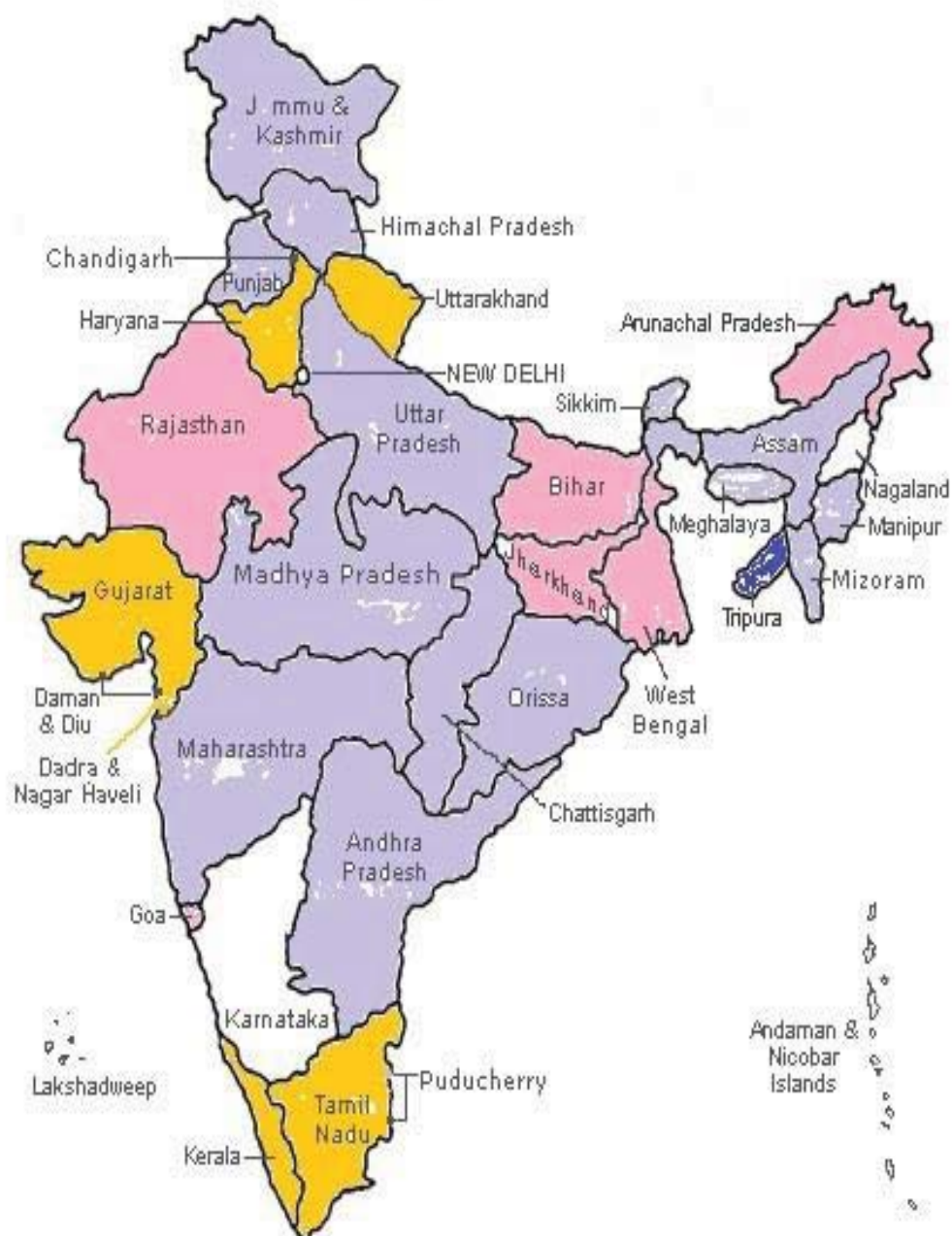


Figure 6. Clustering of rainfall states of India using k-means

In case of model based clustering we are getting 6 clusters of rainfall data using BIC and is given in figure (7). The obtained cluster is presented diagrammatically in figure (8). This shows that the best model using finite mixture normal model fitted by EM-algorithm is VEV which is of varying volume, orientation and equal shape. For this fitted model, the BIC value is - 181.2515. We are getting 6 cluster which includes 11, 4, 12, 1, 3, 4 rainfall states with mixing probability 0.31357755, 0.11428571, 0.34365848, 0.02857143, 0.08571520, 0.11419163 respectively. The spatial distribution of rainfall cluster of India using model based clustering has shown in figure (9).

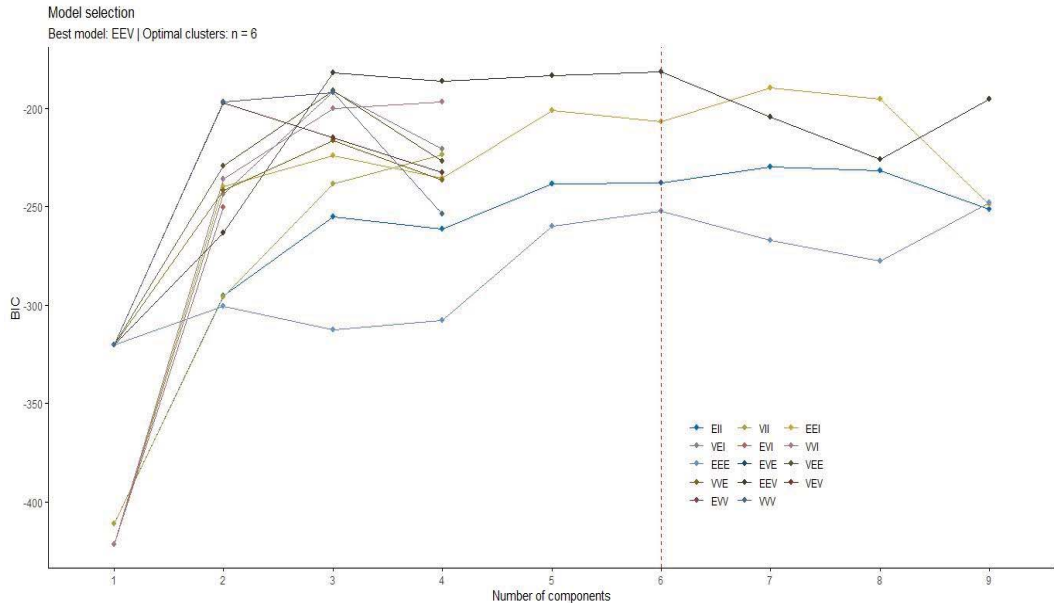


Figure 7. BIC plot for fitted model and no of clusters

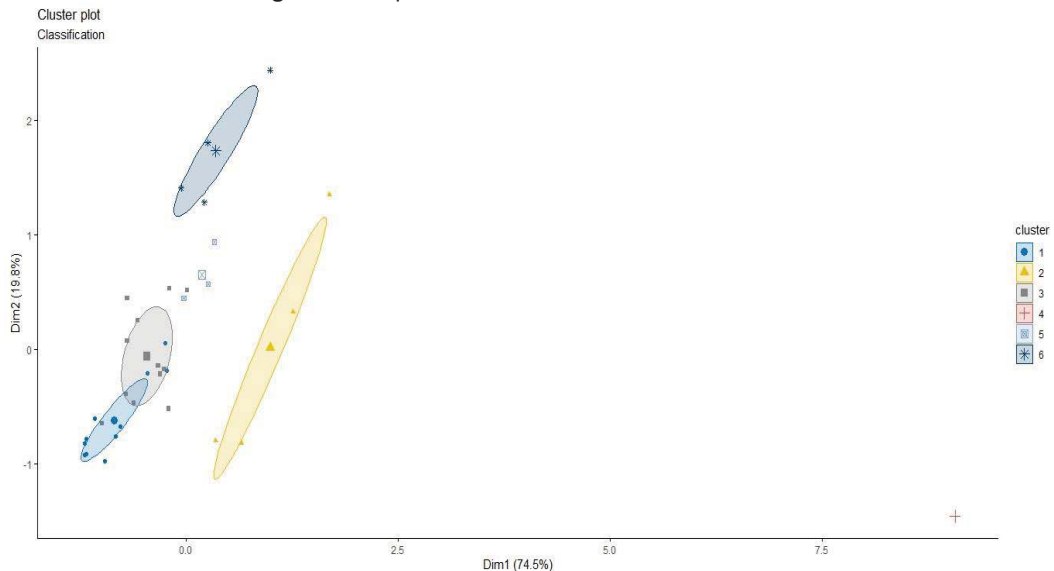


Figure 8. No. of cluster and data assignment in clusters using the VEV model

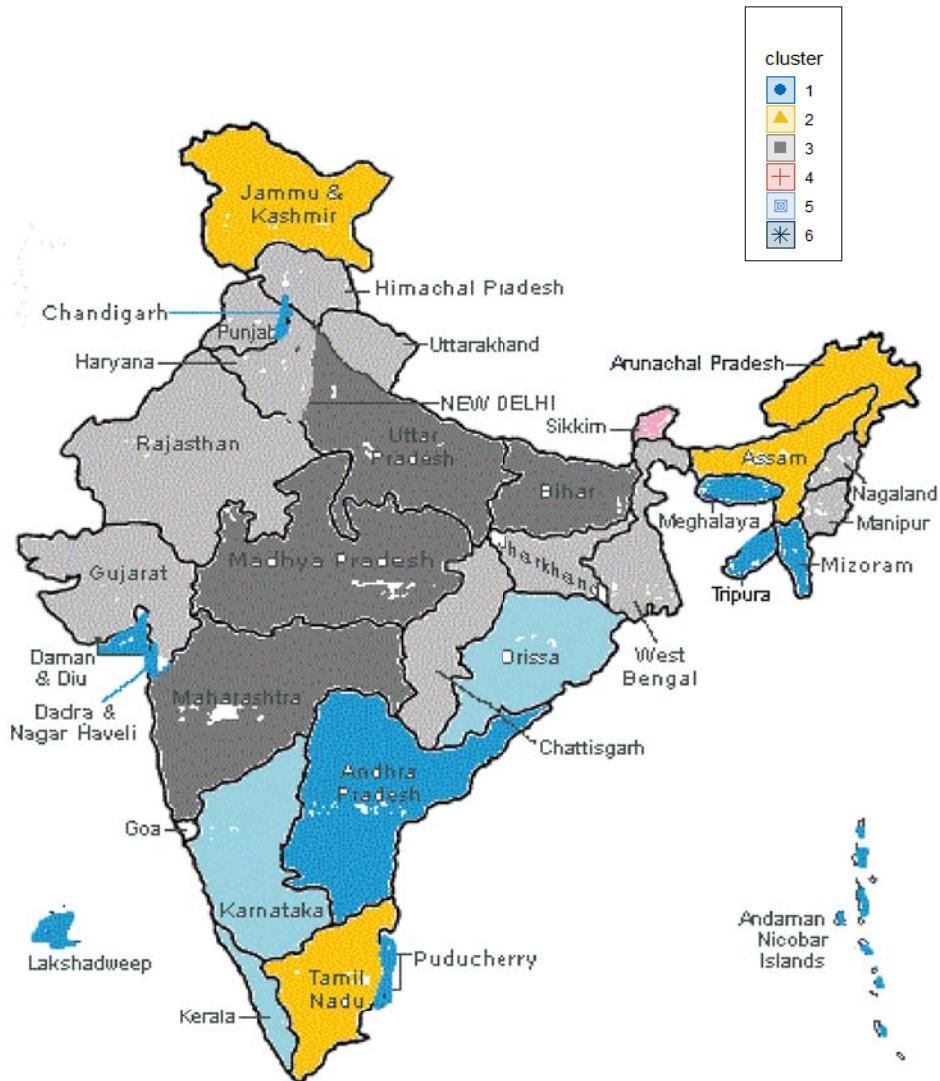


Figure 9. Clustering of rainfall states of India using model based clustering

We observed that the figures (6 & 9) clearly representing how the rainfall is distributed over the states of India using cluster analysis of K-means and model based clustering.

8. CONCLUSION

This study identifies clusters of rainfall for 35 states situated in the spatial province of India. The identified clusters along with spatial distribution will help policymakers to observe the rainfall inequality in India. This study contributes in identifying heavy rainfall/drought regions. It is nevertheless to say that rainfall plays a crucial role in human life. Therefore, planning agencies might consider findings of this

study while framing the policies related to agriculture, management of flood and drinking water supply & storage etc.

8. REFERENCES

- Banfield, J.D., Raftery, A.E., 1993, Model Based Gaussian and Non Gaussian Clustering. *Biometrics*, **49**, 803-821.
- Dasgupta, A., Raftery, A. E., 1998, Detecting Features in Spatial Point Processes with Clutter via Model-Based Clustering, *Journal of the American Statistical Association*, **93**, 294-302.
- Dempster, A.P., Laird, N.M., Rubin, D.B., 1977, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B.* **39**, 1–38.
- Dikbas, F., Firat, M., Koc, A.C., Gungor, M., 2012, Classification of precipitation series using fuzzy cluster method, *Int. J. Climatol.* **32**, 1596–1603.
- Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011, *Cluster Analysis*. John Wiley and Sons.
- Fraley, C., Raftery, AE, 1998, How many clusters? Which clustering method? Answer via model-based cluster analysis. *The Computer Journal.* **41**(8):578–588.
- Fraley, C., Raftery, AE, 1999, MCLUST: Software for model-based cluster analysis. *Journal of Classification*, **16**, 297–206.
- Goyal, M.K., Gupta, V., 2014. Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis, *Water Resour. Manag.* **28**, 4491–4511.
- Hartigan, J.A., 1975, *Clustering Algorithms*, Wiley, New York.
- Hosking, J.R.M., Wallis, J.R., 1993, Some statistics useful in regional frequency analysis, *J. Clim.* **29**, 271–281.
- Jackson, I. J., Weinand, H., 1995, Classification of tropical rainfall stations: A comparison of clustering techniques. *Int. J. Climatol.* **15**, 985-994.
- Johnson, F., Green, J., 2018, A comprehensive continent-wide regionalization investigation for daily design rainfall, *J. Hydrol. Reg. Stud.* **16**, 67–79.
- Jobson, J. D., 1992, *Applied Multivariate Data Analysis, Vol. II: Categorical and Multivariate Methods*. Springer-Verlag, 731.
- Johnson, R.A., Wichern, D.W., 2015, *Applied Multivariate Statistical Analysis*. Pearson education Inc.
- Kaufman, L., Rousseeuw, PJ, 1990, *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons, Inc.; New York.
- Kulkarni, A., Kripalani, R.H., 1998, Rainfall patterns over India: Classification with fuzzy c-means method, *Theor. Appl. Climatol.* **59**, 137–146.
- Kyseláží, J., Pícek, J., Huth, R., 2007, Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech Republic, *Studia Geophysica et Geodaetica*, **51**, 327–344.

- McLachlan, G., Basford, K.E., 1988, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.
- McLachlan, G., Peel, D., 2004, *Finite Mixture Models*. John Wiley & Sons: Hoboken, NJ, USA.
- Melnykov, V., Maitra, R., 2010, Finite mixture models and model-based clustering. *Statistics surv*, **4**, 80-116.
- Meddi, M., Meddi, H., Toumi, S., Mehaiguen, M., 2013, Regionalization of rainfall in North-western Algeria, *Geographia Technica*, **17**.
- Milligan, G. W., 1980. An examination of the effects of six types of errors perturbation on fifteen clustering algorithms. *Physcometrika*, **45**, 325-342.
- Modarres, R., 2006, Regional precipitation climates of Iran, *J. Hydrol. (NZ)* **45**, 15.
- Polymenis, A., 2019, Some Comments on the Performance of Information Ratios for Assessing Numbers of Components in Finite Mixtures. *International Journal of Applied Mathematics and Statistics*, **58**, 36-43.
- R Core Team, 2018, R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
- Ramos, M. C., 2001. Divisive and hierarchical clustering techniques to analyze variability of rainfall distribution patterns in a Mediterranean region. *J. hydrol.* **57**, 123-138.
- Raftery, AE., Dean, N., 2006, Variable selection for model based clustering. *Journal of American Statistical Association*, **101** (473), 168-178.
- Rainfall data for the year 2016-17, Indian Meteorological Department (IMD)
- Raziei, T., Bordi, I., Pereira, L., 2008, A precipitation-based regionalization for Western Iran and regional drought variability, *Hydrol. Earth Syst. Sci.* **12**, 1309–1321.
- Sarhadi, A., Heydarizadeh, M., 2014, Regional frequency analysis and spatial pattern characterization of dry spells in Iran, *Int. J. Climatol.* **34**, 835–848.
- Schwarz, G., 1978, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.
- Singh, V. and Kumar, M. (2020). Cluster detection through model based clustering approach. *International Journal of Applied Mathematics and Statistics*, **59(1)**, 55-62.
- Sharma, S., 1996, *Applied Multivariate Techniques*. John wiley & Sons, Inc. USA.
- Shirin, A.S., Thomas, R., 2016, Regionalization of rainfall in Kerala state, *Proc. Technol.* **24**, 15–22.
- Stahl, D., Sallis, H., 2012, Model-based cluster analysis. *WIREs Comput Stat.* **4**, 341-358.
- Yue, S., Wang, P., Wang, J., Huang, T., 2013, Extension of the gap statistics index to fuzzy clustering, *Soft. Comput.* **17**, 1833–1846.