ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 5, Issue 10, October 2020

Spatial variation in months of June and July in Northeast India using genetic algorithm based clustering

[1] Debbarma Nilotpal, [2] Choudhury Parthsarathi, [3]Roy Parthajit, [4] Agarwal Shivam
[1] PhD Scholar, Civil Engineering Department, NIT Silchar, Silchar, Assam, India.
[2] Professor, Civil Engineering Department, NIT Silchar, Silchar, Assam, India.
[3] Associate Professor, Civil Engineering Department, NIT Silchar, Silchar, Assam, India.
[4] PhD Scholar, Civil Engineering Department, NIT Silchar, Silchar, Assam, India.

***Abstract:*-- Study of spatial variation of rainfall in monthly time scale for any region can serve as a crucial information in many climatological and water related development activities. The present study therefore investigates the spatial variation for months of June and July rainfall in Northeast India using genetic algorithm-based clustering. Seven station attributes of 33 ground stations are considered and the algorithm performance is assessed with eight cluster validation measures. L-moments technique was applied to determine regional rainfall quantiles for different return periods. Results obtained found that the region can be classified into three and five homogenous rainfall regions for months of June and July. The July month needed further rearrangement of stations to obtain homogeneity of the regions. The June month rainfall was found to defined by Gumbel Extreme Value (GEV) and Pearson type III(PE3) distribution, whereas the July month rainfall was defined by four probability distributions namely GEV, PE3, Generalized Pareto (GPA), and Generalized Logistic (GLO) distributions.**

Index Terms:- genetic algorithm, cluster, L-moments, probability distribution, rainfall quantile

1. INTRODUCTION

Mitigation of hazards and reliable water management strategies require rainfall information with long historical records in any area. When the area to be applied is ungauged and is situated in a mountainous region with difficult terrain, regionalization studies works as a trade-off for acquiring spatial information. The northeast region of India is chosen in the study for determining spatial variation of monthly rainfall during south west monsoon season. The Northeast region of India is severely affected by heavy rainfall occur every year with lots of casualties and devastating damages. The information available on preparedness for heavy rainfall occurrence in the region is insufficient and limited. And studies related to spatial variation of rainfall in each months of monsoon season in the region is also very scarce. The study thus takes up two most heavy rainfall months i.e. June and July to add to information studies on spatial distribution of rainfall during monsoon months. The regionalization studies on monthly time scale may vary and be different from that of annual or seasonal time scale (Wu et al. 2013). A study by Mahanta et al. 2013 found the heavy rainfall events occuring during the months of June and July to be more widespread. Therefore, the monthly rainfall variation for the months of June and July are studied and its spatial variation in the region. The northeast region receives heaviest rain in whole of India and Cherrapunjee in northeast region is also considered wettest place on earth. The region is highly undular, mountainous and is hindered with sparsity of rain gauge network for accurate and reliable information. And so in such cases regionalization studies employing clustering technique for spatial rainfall analysis is most suitable. Clustering analysis is a non-supervised method where the grouping of objects is done using some similarity criterion of the object attributes. Widely used cluster methods like K-means, fuzzy c-means (Goyal and Gupta, 2014) are available in plenty in literature but the application of evolutionary algorithms in clustering climatic attributes is very limited. So, in the present study, genetic algorithm as an evolutionary algorithm is considered in clustering 33 ground stations in the study area utilizing seven station attributes. The L-moments technique is used for parameter estimation, homogeneity testing of delineated regions and selection of regional distributions. Finally mapping of determine rainfall quantile estimates for different return time are presented and discussed.

2. MATERIALS AND METHODS

2.1 Data and area of study

The study area covers the South Bank of Brahmaputra and whole of Barak valley with location lying between 22^0 N to 28^0 N and 89^0 E to 96^0 E which covers the north eastern states comprising of Assam, Manipur, Nagaland,

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 5, Issue10, October 2020

Meghalyaya, Mizoram and Tripura. Records of monthly rainfall of 33 rain gauging stations for a period of 20 years (1997 – 2016) were collected from Regional Meteorological Centre Guwahati, India. The rainfall received in the region is maximum in India and is largely influenced by the Himalayas in the North, Meghalaya Hills in the south, Manipur and Nagaland hills to the east. Heavy precipitation results mostly due to orographic process in the region. The altitude in the study region is found to vary from 16 to 1598 meters with the range of maximum monthly total lying between 302 mm to 5222 mm. In literature, most of the studies were confined to only seasonal and annual assessments in the Northeast region. However, studies on spatial variability studies on monthly basis has been scarcely done for the region.

2.2 Methodology

Regional frequency analysis for rainfall parameter for any region starts with the determination of homogenous regions. Clustering technique to form homogenous regions using station characteristics has been found as most popular method. Seven station attributes viz. average monthly total, maximum monthly total, minimum monthly total, latitude, longitude, altitude and monthly coefficient of variation are normalized using min-max normalization. The attributes are used as real-valued chromosomes into genetic algorithm-based clustering. The output from the algorithm is not the final clustering and may often need further subjective adjustments to increase physical coherence of clustered regions (Hosking and Wallis, 1997).

2.2.1 Cluster Validation indices

A total of eight cluster validation indices each for crisp type clustering is considered. The indices used for assessing performance include Dunn, Calinski-Harbasz(CH), Kraznowski and Lai(KL), C-index, Davies-Bouldin(DB), Pakhira, Bandopadhyay and Maulik(PBM), SD indices and Xie-Beni(XB) index (Liu et al. 2013 ; Hubert and Levin, 1976; Kraznowski and Lai, 1988). For a dataset, X = {x_1, x_2 , x_3, …… x_N} where each x_i is a vector in an D-dimensional space i.e. $X \subseteq R^D$. The fitness in genetic algorithm is taken as Davies-Bouldin index that represents the ratio of total within-cluster scatter and between-cluster separation. The intra cluster separation density is defined by the average distance of each member to the center of cluster.

$$R_{ij} = \frac{S_i + S_j}{d_{ij}}, \; i \neq j \qquad \text{where,} \qquad d_{ij} = \|c_i - c_j\|^2 \quad (1)$$

$$S_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \|x_j - c_i\|^2 \quad (2)$$

$$R_i = \max_{j=1,\ldots M} R_{ij}, i = 1, \ldots\ldots, M \quad (3)$$

$$DB = \frac{1}{M} \sum_{i=1}^{M} R_i \quad (4)$$

And a smaller of the DB index indicates a better clustering result. The choice of selecting the optimum from a set of evaluated indices can best be achieved as a multi criteria decision problem. For strengthening the decision a total of three MCDM techniques viz. VIKOR, WASPAS and TOPSIS (Blanca and Ceballos, 2016) are considered and the agreement of any two technique is final ranking. Weights are assigned to the eight cluster validation indices using Shannon entropy method and taken constant in all three MCDM methods. Considering the number of clusters as alternatives and cluster validation indices as selection criteria, MCDM technique is utilized to decide the optimum number of homogeneous regions.

2.2.3 Heterogeneity measure, regional probability distribution and quantile estimation

The L-moments are alternative ways of determining the parameters of a probability distribution from a data sample. The steps involved in evaluating the discordancy measure and L-moments are completely described and can be found in Hosking and Wallis, 1997 where they proposed heterogeneity statistic for computing the degree of heterogeneity in a group of sites. For heterogeneity test of a group, a four parameter kappa distribution is fitted to the regional data set generated from series of 500 equivalent region data using Monte-Carlo simulation. For each region, the regional L-moment ratios and V-statistic are computed and based on the vector of V statistic, mean (μ_{vi}) and standard deviation σ_{vi}, the heterogeneity measure can be calculated as :

$$H_i = \frac{V_i - \mu_{vi}}{\sigma_{vi}}, \quad i = 1,2,3 \quad (5)$$

On the basis of homogeneity measurements as suggested by Hosking and Wallis, 1997, a region or group of sites is interpreted as “acceptably homogenous” if H_i is less than 1, possibly heterogeneous if H_i lies between 1 and 2 and definitely heterogeneous if H_i exceeds 2. Large values of H2 indicates larger regional and at-site estimate difference while larger H3 implies deviation between estimates of regional and at-site. However, H1 is considered prinicipal measure as H2 and H3 are not explicitly distinctive for grossly heterogenous regions (Hosking and Wallis 1997).

Choice of a single regional distribution for each individual homogenous regions formed after regionalization is the next

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 5, Issue10, October 2020

most important assessment in the regional frequency analysis. Every individual station may follow its own independent distribution but by regional frequency analysis such a distribution is fitted to the whole region such that it gives the closest outcome to the summation of individual distribution of all the stations in the region if applied individually. The goodness of fit criterion based on L-kurtosis as proposed by Hosking and Wallis, 1997 is considered for selecting the regional distribution. Five different types of distribution namely Generalized logistic (GLO), Generalized Extreme value (GEV), General Normal (GNO), Pearson type-III (PE3) and Generalized Pareto (GPA) are considered in the study. On the basis of the best fitted distribution and its parameters a growth curve is generated from Monte Carlo simulation. Then the approach of index flood method is employed to find the quantile of extreme rainfall (Hosking and Wallis, 1997).

3. RESULTS AND DISCUSSION

The data used were recorded monthly total precipitation for 33 rain gauge stations in genetic algorithm based clustering. Real valued chromosomes based on normalized attributes were performed in clustering with other criteria as: number of population = 140, heuristic crossover with cross-over percentage = 0.8, mutation percentage = 0.4, number of generations = 3000. The optimum number of clusters were determined for from trial range of 2 to 7 clusters and the performance determined using eight cluster validation measures. Four indices of maximization type(CH, KL, PBM and Dunn indices) and four minimization type (DB, C-index, SD and XB indices) were used altogether and the weights of indices determined using Shannon Entropy method. The study aimed at finding out the pattern of rainfall distribution during June and July. The delineation of homogenous regions for the study area for the months of June and July are found to be 3 and 5 regions respectively. The clustering results were further subjected to heterogeneity test and some stations had to be dropped for attaining homogeneity of the regions. The H1 values is considered as principal measure for determination of homogeneity of a region. H1 values for June month lies between -0.09 and 0.36 for the three regions thereby suggesting all regions to be definitely homogenous. And the value for July lies between -0.29 to 0.95, indicating grouped stations to have more difference and variation in similarity among each other. For June month, Gharmura station was highly discordant and the delineated homogenous regions for month of June comprised of 2, 10 and 20 stations. The July month rainfall was found to be highly complex and so the clustering required further adjustments to attain homogeneity for each regions. The grouping for July was done into five homogenous groups with the removal of one station. Large values of H2 and H3 obtained for both months and presented in table 1 for regions of July month suggests the agreement between at-site and regional estimates for the region are lesser homogenous compared to month of June. The spatial variation of rainfall is more for the month of July in the region and is evident from the analysis. The regional distributions and the parameters of each distribution are calculated for each regions for the months and presented in Table 1 and 2. The best fitted distribution for each homogenous regions is selected based on Z distribution statistic given by Hosking and Wallis, 1997. Minimum value of the distribution is selected as the most fit distribution. Rainfall quantile estimates corresponding to various return periods for both the months are also mentioned in Table 2. Further, it is observed that the maximum and minimum predicted rainfall in regions of the study area for month of June is 4592.14 and 486.15 mm respectively. While the range is slightly more for July with maximum and minimum rainfall values as 5576.7 and 448.91 mm respectively. The maximum rainfall region prediction observed for both months is a common region comprising of Cherrapunjee and Mawsynram. These two sites are found to be separate distinct cluster from other clusters in the genetic algorithm clustering. The reason may be obvious as both the stations are located near to each other and record the heaviest rainfall in entire of India. Finally, location of the gauge stations and spatial interpolated map of the estimated quantiles in the study region is done using inverse distance weighting method and presented in Figure 1.

4.CONCLUSION

The spatial variability of monthly precipitation in Northeast India was taken up and the study indicated different climatic homogenous sub zones. The 33 number of ground rain gauge stations formed different rainfall homogenous regions for the months of June and July. July month depicted maximum complexity in occurrence of rainfall and was divided into five regions with the removal of one station. Whereas the June month rainfall in the study area was grouped into three homogenous regions, and were found to follow GEV, PE3 for June month and GLO, GEV, PE3 and GPA for July month. H1 values obtained suggested the clustered rainfall regions are more homogenous for the month of June. The H2 and H3 values obtained indicate lesser stability and agreement between regional and at-site estimates of rainfall. As major amount of the monsoon rainfall occurs during these two months in Northeast India, prediction of rainfall has to be done judiciously during these months. Further, rainfall quantile estimates for various return periods are presented that may be useful in the flood control structures or other water resources applications. The present study may thus

serve as a useful information in augmenting the variability study of rainfall during the months of June and July in Northeast India.

ACKNOWLEDGMENTS

Authors want to thank Regional Meteorological Centre, Guwahati, India for providing with the data required in the study.

Conflict of Interest
The authors declare no conflict of interest.

REFERENCES

1. Blanca, A., Ceballos, M., 2016 MCDM: Multi-Criteria Decision Making Methods for Crisp Data. R Software Package, Version 1.2.
2. Y. Liu, Z. Li, H. Xiong, X. Gao, J. Wu and S. Wu 2013, "Understanding and Enhancement of Internal Clustering Validation Measures," IEEE Transactions on CYBERNETICS, Vol.43, No.3, pp. 982-994.
3. Goyal, M.K., and Gupta V. 2014 Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis. Water Resources Management, 28(13), 4491-4511.
4. Hosking, J.R.M. 1990 L-Moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc Ser B 52:105–124
5. Hosking, J.R.M., Wallis J.R. 1997 Regional frequency analysis: an approach based on L-moments. Cambridge Univ. Press, Cambridge, UK.
6. Hubert, L., and Levin, J. 1976. A general statistical framework for assessing categorical clustering in free recall. Psychologica Bulletin 1072–1080.
7. Krzanowski,W., and Lai, Y. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44 23–34.
8. Mahanta, R, Sarma, D and Choudhury, A. 2013 Heavy rainfall occurrences in northeast India. International Journal of Climatology, 33:1456-1469.
9. Wu, X., Zurita-Milla, R. and Kraak, M.J. 2013 Visual discovery of synchronization in weather data ast multi-temporal resolutions. The Cartographic Journal 50 (3), 247-256.

FIGURES AND TABLES

Table 1. Heterogeneity fit results of June and July months and results of candidate distributions

Month	Region	Number of stations	Heterogeneity measures			Probability Distributions (Z – statistic value)				
			H1	H2	H3	GLO	GEV	GNO	PE3	GPA
June	1	10	-0.09	-0.35	-0.28	2.4	1.14	0.83	0.18*	-1.8
	2	2	-0.59	-0.82	-0.19	1.26	0.29*	0.49	0.48	-1.55
	3	20	0.36	0.61	0.42	3.99	1.69	1.57	1.01*	-3.25
July	1	5	0.62	0.19	0.39	1.87	0.99	0.77	0.32*	-1.06
	2	4	0.95	0.89	-0.04	2.7	1.18*	1.53	1.52	-1.63
	3	9	-0.29	0.98	1.36	3.42	1.45*	1.68	1.56	-2.44
	4	2	0.20	-0.99	-1.01	3.01	2.04	2.15	2.09	0.12*
	5	12	0.18	-1.33	-2.00	0.84*	-0.97	-0.85	-1.05	-4.65

Values marked with asterisk(*) indicate best distribution

Table 2. Regional parameter of selected distributions and estimated rainfall quantiles

Month	Region	Distribution fitted	Parameters			Rainfall (mm)					
			Location	Scale	Shape	F=0.8 RT=5	F=0.9 RT=10	F=0.98 RT=50	F=0.99 RT=100	F=0.995 RT=200	F=0.998 RT=500
	1	PE3	1.00	0.34	1.22	486.15	566.39	736.59	806.11	873.99	962.02

ISSN (Online) 2456-1290

International Journal of Engineering Research in Mechanical and Civil Engineering (IJERMCE)
Vol 5, Issue10, October 2020

June	2	GEV	0.90	0.24	0.22	3228.17	3551.37	4097.9	4274.83	4425.73	4592.14
	3	PE3	1.00	0.42	0.90	570.37	674.08	885.09	968.69	1049.47	1229.15
	1	PE3	1.00	0.54	1.23	467.72	578.18	812.72	908.38	1001.98	1123.28
	2	GEV	0.88	0.31	0.24	448.91	500.85	586.23	613.06	635.56	659.89
July	3	GEV	0.88	0.24	0.15	574.17	641.03	764.86	808.92	848.59	895.15
	4	GPA	0.34	1.11	0.69	4117.69	4698.14	5330.8	5448.46	5521.24	5576.71
	5	GLO	0.96	0.19	-0.10	549.93	632.18	828.06	919.99	1018.13	1158.64

F = non-exceedence probability, RT = Return time in years

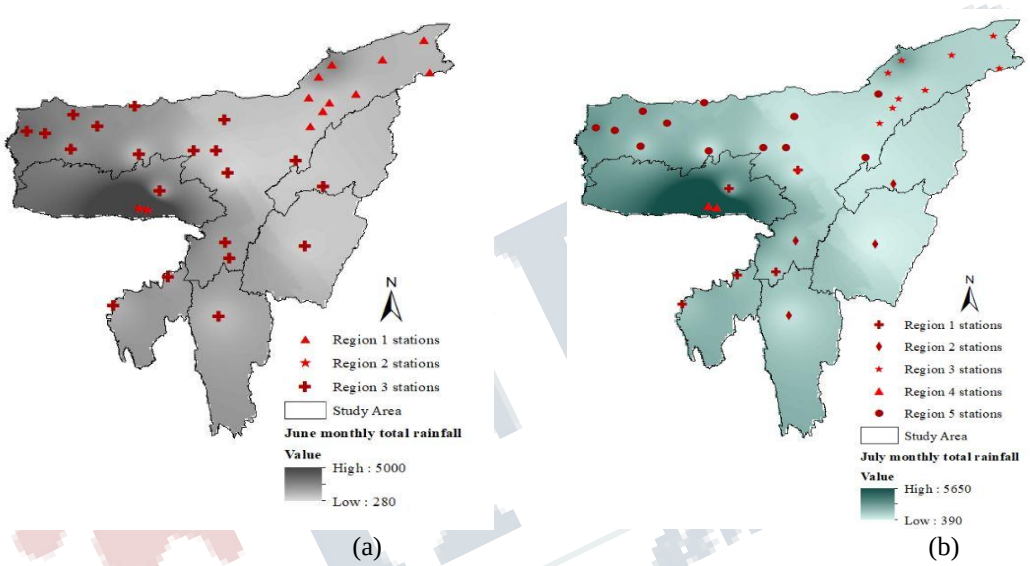

Figure 1. Monthly total rainfall map corresponding to return time years 50 for months of (a) June and (b) July