MAUSAM, **69**, 3 (July 2018), 433-436

551.577.3 (540.69)

Determination of rainfall regions among the districts of Kerala state

P. A. AKHISHA, V. SRINIVASA RAO, S. K. NAFEEZ UMAR and K. UMA DEVI

Department of Statistics and Mathematics, Agricultural College, Bapatla – 522 101, India

(*Received 14 August 2017 Accepted 8 June 2018*)

e mail : Akhishapa377@gmail.com

सार - इस शोध पत्र में केरल राज्य के वर्षा वाले क्षेत्रों की पहचान केरल के चौदह जिलों में 2004 से 2016 (156 महीने) के दौरान हुई मासिक वर्षा समय श्रृंखला की विशेषताओं के आधार पर की गई है क्षेत्रीय वर्षा के पैटर्नो का निर्धारण करने के लिए नॉन हिरारिकल समूह विशेलषण जैसे K क्लस्टरिंग अलॉगरिथ्म के अभिप्राय वाले का प्रयोग विभिन्न लैग और वर्षा के चार समूह समय श्रृंखला मॉडलों के आधार पर पाए गए आॉटोकोरिलेशन सहसंबंध पर किया गया। समय श्रृंखला मॉडलिग के परिणामों से केरल में मासिक वर्षा के स्थानिक पैटर्नों में अत्यधिक परिवर्तन शीलता का पता चला।

ABSTRACT. In this study, rainfall regions of Kerala State were identified based on the properties of monthly rainfall time series of fourteen districts of Kerala from 2004 to 2016 (156 months). To determine regional rainfall pattern, a non hierarchical cluster analysis, *i.e.*, K means Clustering Algorithm, was applied on autocorrelation coefficients at different lags and four rainfall groups were found based on the time series models. The results of the time series modeling showed a high variation of temporal pattern of the monthly rainfall over Kerala.

Key words – Autocorrelation, Principal component analysis, Cluster analysis.

1. Introduction

Nature has bestowed Kerala with abundant rainfall. The average annual rainfall of the State is about 3000 mm, which is about three times the average for the whole of India. Even though the state does not suffer large inter annual variations in annual or seasonal rainfall, there is large spatial variation in the rainfall distribution. A thorough investigation on the rainfall characteristics both on spatial and temporal scales with emphasis on the influence of geography is needed.

Linda (2001) applied a number of statistical techniques to assess the appropriateness of the rainfall districts of Western Australia with agglomerative hierarchical method and classified South West into six largely non overlapping regions.

Hierarchical methods like average method and Ward method were applied by Saed Soltani *et al.* (2006) to classify 28 capitals of the provinces of Iran to eight clusters which cover more than 95% of rainfall variance over Iran.

2. Materials and method

The present study is confined to the Kerala State, having fourteen districts namely Kasargod, Kannur, Kozhikkode, Wayanad, Malappuram, Palakkad, Thrissur, Ernakulam, Idukki, Kottayam, Alappuzha, Pathanamthitta, Kollam and Thiruvananthapuram. The secondary data was collected from the Regional Meteorological Centre, Thiruvananthapuram for monthly (156 months) rainfall for each district covering the years 2004 - 2016.

Multivariate techniques are common methods for classifying meteorological data such as rainfall. Principal components and cluster techniques were used in this study to classify the districts based on the rainfall by using autocorrelation coefficients up to a certain lag.

2.1. *Autocorrelation*

It is defined as correlation between the members of the series of observations ordered in time or space. Autocorrelation function (ACF) at lag k, denoted by $\rho_{(k)}$ is defined as:

$$\rho_k = \frac{\varphi_k}{\varphi_o} = \frac{\text{covariance at lag k}}{\text{variance}}$$

$$\varphi(k) = \frac{1}{N}\sum_{t=1}^{N-k}(Y_t - \bar{Y})(Y_{t-k} - \bar{Y})$$

and

TABLE 1

Autocorrelation coefficients matrix X (14 × 12)

Lags	1	2	3	4	5	6	7	8	9	10	11	12
KGD	.57	.20	-.11	-.35	-.48	-.52	-.45	-.31	-.09	.21	.54	.77
KNR	.58	.17	-.10	-.32	-.48	-.54	-.45	-.29	-.09	.18	.51	.72
CLT	.55	.16	-.05	-.27	-.48	-.55	-.44	-.26	-.02	.17	.49	.70
WYD	.53	.20	-.07	-.30	-.46	-.49	-.42	-.27	-.04	.17	.47	.61
MLP	.56	.22	-.04	-.27	-.49	-.57	-.47	-.24	-.03	.20	.50	.68
PKD	.55	.23	-.03	-.29	-.51	-.58	-.48	-.25	-.02	.20	.49	.64
TCR	.60	.24	-.03	-.30	-.53	-.62	-.53	-.26	-.03	.24	.53	.69
EKM	.57	.23	-.03	-.31	-.51	-.61	-.50	-.23	-.02	.24	.49	.62
IDK	.60	.33	-.02	-.32	-.51	-.58	-.51	-.28	-.03	.26	.53	.66
KTM	.57	.24	-.02	-.28	-.49	-.60	-.49	-.22	-.01	.24	.48	.63
ALP	.55	.18	-.05	-.28	-.46	-.57	-.47	-.20	-.04	.21	.50	.60
PTM	.50	.21	-.04	-.26	-.43	-.51	-.46	-.21	-.05	.20	0.43	.56
KLM	.51	.14	-.06	-.26	-.39	-.49	-.43	-.18	-.04	.14	.44	.55
TVM	.42	.00	-.24	-.22	-.16	-.13	-.12	-.12	-.12	.01	.33	.40

$$\varphi_o = \frac{1}{N}\sum_{t=1}^{N}(Y_t - \bar{Y})^2$$

Let the matrix X (m × k) consist of autocorrelation coefficients at lags k = 1, ..., 12 of m districts. K = 12 was chosen as the autocorrelation; coefficients of higher lags were not significant or they had similar seasonal fluctuations as the first k = 12. This means a matrix of 14 rows of districts and 12 columns of autocorrelation coefficients of rainfall series. As the variables must not be correlated with each other, PCA was carried out and thus data dimension was reduced.

2.2. *Principal component analysis*

The general objective is data dimensionality reduction. Principal Components (PCs) are special kinds of transformations that transform the original vector of X variables to a new vector of Z variables which are mutually independent and each of the Z variables are linear combinations of the original vector of X variables. The first PC captures as much of the variation in the original data as possible. The second component captures the maximum variation that is uncorrelated with the first component and so on.

Here we had X = 12 variables, *i.e.*, 12 autocorrelation coefficients for each district which are correlated. Using the scree plot of PC number *vs* Eigen value, the required number of uncorrelated PCs was obtained and the corresponding Eigen vectors by using SAS 9.3. Later, the PC scores for each district were obtained, thus forming the data for cluster analysis.

2.3. *Cluster analysis*

Cluster analysis procedures are used for classifying the objects on the basis of their observational vectors into homogeneous groups, referred as clusters. Here objects are classified on the basis of similarities between them.

Let the new matrix be X′(m × p) where m is the number of districts and p is the number of PCs selected. Hence, the objects to be classified are represented by districts and variables by PC scores. The similarity is measured in the form of inter object distances. A commonly used similarity measure is the Euclidean distance (d^2_{rs}) which is written as follows.

$$d^2_{rs} = \sum_{j=1}^{k}(x_{rj} - x_{sj})^2$$

where, the r^{th} and s^{th} rows of the data matrix X′ was denoted by (x_{r1}, x_{r2},...,x_{rk}) and (x_{s1},x_{s2},...,x_{sk}) respectively. In the present study, a non hierarchical method of clustering namely, *k* means clustering algorithm, was followed in SAS 9.3.

2.4. *K means clustering*

k means clustering aims to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This method of clustering was chosen because of its iterative procedure. The optimum number of clusters to be formed was decided by the Elbow method in *r*

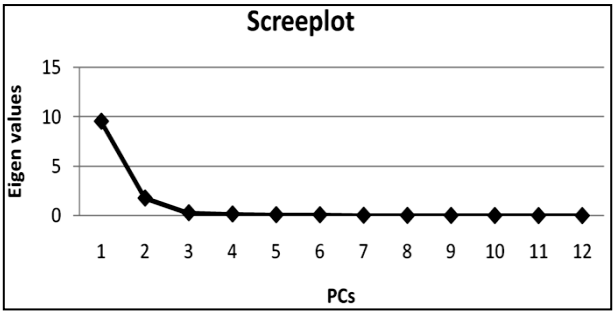

Fig. 1. Scree plot of Eigen values *vs* principal components

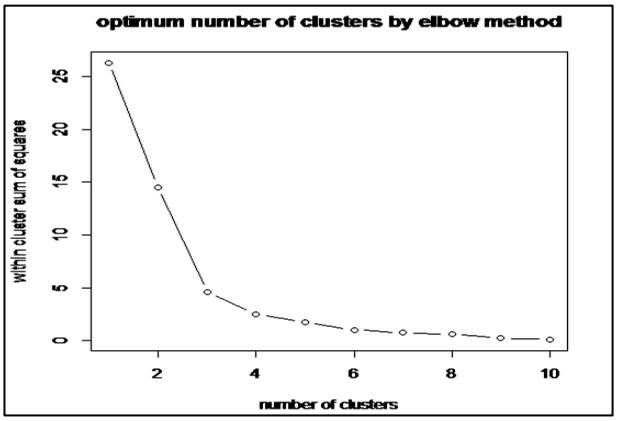

Fig. 2. Optimum number of clusters by Elbow method

language. *k* means clustering algorithm consisted of three major steps.

(*i*) The items were partitioned into *k* initial clusters

(*ii*) Proceeded through the list of items to the clusters, whose centroid was nearest, *i.e*., distance was usually computed using distance with either standardised or non standardised observations. Recalculated the centroid for the cluster receiving the new item and for the cluster loosing the item.

(*iii*) Step 2 was repeated until no more reassignment took place.

3. Results and discussion

Each time series was subjected to autocorrelation such that the autocorrelation coefficients for each district at lag = 1, 2, ..., 12 could be obtained. Hence, the X (14 × 12) matrix was made as given in Table 1.

The X matrix was used for carrying out Principal Component Analysis in SAS 9.3 and 12 Eigen values were obtained. The scree plot showed that the 12 original variables could be reduced to 3 variables and it was given in Fig. 1.

TABLE 2

Output of PCA

PC	Eigen value	Percentage variation	Cumulative percentage variation
PC1	9.5323	79.44	79.44
PC2	1.7667	14.72	94.16
PC3	0.2904	2.42	96.58
PC4	0.1622	1.35	97.93
PC5	0.0909	0.76	98.69
PC6	0.0855	0.71	99.40
PC7	0.0305	0.25	99.66
PC8	0.0174	0.15	99.80
PC9	0.0161	0.13	99.94
PC10	0.0062	0.05	99.99
PC11	0.0011	0.01	100.0
PC12	0.0000	0.00	100.0

TABLE 3

PC scores of each district

District	PC1 Score	PC2 Score	District	PC1 Score	PC2 Score
KGD	0.45	-2.54	EKM	0.55	0.51
KNR	0.23	-1.68	IDK	0.93	-0.08
CLT	0.11	0.11	KTM	0.41	1.03
WYD	-0.07	-0.31	ALP	0.05	0.56
MLP	0.29	0.29	PTM	-0.34	0.97
PKD	0.39	0.47	KLM	-0.67	0.93
TCR	0.80	0.09	TVM	-3.14	-0.45

The first principal component explained 79.44 per cent of total variations; second principal component explained 14.72 per cent of total variation. Since the first two PCs together explained a total of 94.16 per cent of total variations, those two PCs which were linearly independent to each other were selected. The total data dimensionality was reduced from twelve to two. The Eigen values, percentage of total variation explained by each PC and cumulative percentage variation explained were given in Table 2. The principal component scores of each district corresponding to two PCs were used to form the clusters and shown in Table 3.

Optimum number of clusters was determined by the Elbow method using R language. The plot of Within Sum of Squares *vs* number of clusters obtained by Elbow method was given as in Fig. 2

TABLE 4

Cluster listing

District	Cluster	Distance from seed
Kasargod	1	0.3992
Kannur	1	0.3992
Kozhikkode	4	0.3152
Wayanad	4	0.6945
Malappuram	4	0.1653
Palakkad	4	0.3183
Thrissur	4	0.3737
Ernakulam	4	0.3785
Idukki	4	0.5596
Kottayam	3	0.5716
Alappuzha	3	0.3669
Pathanamthitta	3	0.2304
Kollam	3	0.5358
Thiruvananthapuram	2	0.0000

TABLE 5

Cluster summary

Cluster	Frequency	RMS Std Deviation	Maximum distance from Seed to Observation	Nearest cluster
1	2	0.3992	0.3992	4
2	1	-	0.0000	3
3	4	0.3655	0.5716	4
4	7	0.3313	0.6945	3

TABLE 6

Distance between cluster centroids

Nearest cluster	1	2	3	4
1	-	3.8486	2.9884	2.2283
2	3.8486	-	3.2907	3.6304
3	2.9884	3.2907	-	0.9202
4	2.2283	3.6304	0.9202	-

The approach was based on within cluster sum of squares. Since the kink was observed at the point corresponding to the 4 number of clusters, the optimum number of clusters to be formed out of fourteen districts was four.

Once the number of clusters was decided, K means clustering algorithm was followed using SAS 9.3. The analysis showed that the convergence criterion was satisfied. The cluster was listing obtained as given in Table 4. The northernmost districts of the State, namely, Kasargod and Kannur were listed under cluster 1 while the southernmost district namely, Thiruvananthapuram was listed under Cluster 2 alone. Districts namely, Kottayam, Alappuzha, Pathanamthitta and Kollam were listed under Cluster 3 while, districts namely, Kozhikkode, Wayanad, Malappuram, Palakkad, Thrissur, Ernakulam and Idukki were listed under Cluster 4. Pseudo F statistic was found to be 32.22. Cluster Summary and Distance between cluster centroids were given in Tables (5&6).

4. Summary and conclusions

Using Autocorrelation coefficients of each time series up to twelve lags, Principal Component Analysis was carried out and based on the Principal Component Scores of each districts for two Principal Components which accounted for 94.16% of total variation, the districts were grouped into four clusters by following K-means Clustering Algorithm, a non hierarchical method of clustering. This grouped the districts having similar monthly rainfall pattern into one cluster through an iterative procedure. Kasargod and Kannur districts were listed under cluster 1 while Thiruvananthapuram was listed under Cluster 2 alone. Kottayam, Alappuzha, Pathanamthitta and Kollam were listed under Cluster 3 while Kozhikkode, Wayanad, Malappuram, Palakkad, Thrissur, Ernakulam and Idukki were listed under Cluster 4.

The contents and views expressed in this research paper/article are the views of the authors and do not necessarily reflect the views of the organizations they belong to.

References

Linda, E. Chambers, 2001, "Classifying Rainfall Districts : A South Western Australian Study", *Australian Meteorological Magazine*, **50**, 2, June, 2001.

Soltani, S., Modarres, R. and Eslamian, S. S., 2006, "The use of time series modeling for the determination of rainfall climates of Iran", *International Journal of Climatology*, **27**, 819-829.