International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 2 (2015) pp. 5301-5315

http://www.ripublication.com

Rainfall Pattern And Classification Of Districts In Tamilnadu Using Data Mining Techniques

G. Palanivel[1] , Dr.P.Vaishnavi[2] And Dr.K.Duraiswamy[3]

[1]Research Scholar, Department of Computer Science and Engineering,K.S.Rengasamy College of Technology,Tiruchengode,S.India.E-mail: palani01@gmail.com
[2] Assistant Professor, Department of Computer Applications,BIT Campus, AnnaUniversity-Chennai. Email: vaishmk@gmail.com
[3]Dean, Department of Computer Science and Engineering,K.S.Rengasamy College of Technology,Tiruchengode,S.India

Abstract

The present research is aimed at analyzing rainfall pattern and classification to evaluate in district wise data in Tamilnadu and make possible the result in various seasons to understand the climate change. The dataset relates to monthly rainfall from various districts of Tamilnadu in the period of January to December in the Indian Metrological Department database. The time frame of the data pertaining to the present study is 2004-2010. The salient feature of this study is the application of *Factor Analysis*, *K-means clustering and GIS (Geographical Information System) Map* as data mining tools to explore the hidden pattern present in the dataset for each of the study periods. Factor analysis is applied first and the factor scores of extracted factors are used to find initial groups by k-means clustering algorithm. Finally, data mining tools are applied and the groups are identified as rainfall belonging to **HR** (High Rainfall), **MR (**Moderate Rainfall**)** and **LR** (Low Rainfall). The results of the present study indicate that *Data Mining Tools* can be used as a feasible tool for the analysis of large set of rainfall data.

Keywords: Rainfall, Data mining, Factor Analysis, GIS Map and K-means Clustering.

1. Introduction

Rainfall is the key climatic inconsistent that governs the regional hydrologic cycle and accessibility of water resources. It is also one of the most composite and difficult

elements of the hydrological cycle to understand due to the great range of variation over a wide range of scales both in space and time (French *et al*.,1992).

In this section we describe the state climate, boundaries and average rainfall of Tamilnadu. In the state situated in the southern part of India, Tamil Nadu is one of the most well known and important states in India. Andhra Pradesh forms the northern border for it whereas Karnataka and Kerala lie on the North West and west respectively. Two water bodies enclose the state on the southern as well as the eastern sides, the Indian Ocean and the Bay of Bengal respectively. In fact, geometrically, Tamil Nadu touches the acute southern tip of the Indian Peninsula. The climate of Tamilnadu is generally wet subtropical climate and features fairly hot temperature over the year except during the monsoon season. The state has three distinct monsoon periods of rainfall. The *south west monsoon* starts from the period of June to September with strong southwest winds. The *north east monsoon* starts from the period of October to December with dominant northeast winds. Finally, *dry season* starts from January to May. The normal annual rainfall of the state is about 945 mm (37.2 in) of which 48% is through the North East monsoon, and 32% through the South West monsoon. Since the state is fully dependent on rains for recharging its water resources, monsoon failures lead to acute water scarcity and severe drought.

Moreover, factors like climate change and urbanization have also had an impact on the variation in rainfall. Recent studies have stated that any analysis of hydro-climatic variables should be done at the local scale rather than at a large or global scale (Sharma and Shakya, 2006; Barua et al., 2013).

2. Brief Review of Literature

Rainfall is key factor determining the sustainability and conservation of living species on the earth. In dry farming areas, where rainfall is the sources of water for crops, changes in both quantity and distribution of rainfall during the year could affect the economy an area (M.C.Ramos, 2001). Many researchers have applied MPL (Multi variables Polynomial regression) to implement the precipitation forecast model over Myanmar. The model output result in station wide monthly and annual rainfall amount during summer monsoon season. It is observed that MPR method achieves closer agreement between actual and estimated rainfall.

In this paper attempts have been made to study pattern in annual and classification of rainfall over Tamilnadu from 2004 to 2010. Long term trends of Indian monsoon rainfall for the country as well as for smaller subdivisions were studied by Pramanik and Jagannathan (1954), Parthasarathy and Dhar (1978), Parthasarathy (1984), Mooley and Parthasarathy (1983), Parthasarathy *et al.* (1993). Rao and Jagannathan (1963), Thapliyal and Kulshrestha (1991) and Srivatsava *et al*. (1992) also reported that All-India southwest monsoon/annual rainfall observed no significant trend. Long term trend in small spatial scale was reported by Koteswaram and Alvi (1969), Jagannathan and Bhalme (1973), Naidu et al. (1999) and Singh and Sontakke (1999). Rupa Kumar *et a*l. (1992) have found significant increasing trend in monsoon rainfall along the West Coast, north Andhra Pradesh and northwest India while significant decreasing trends over Madhya Pradesh and adjoining area, northeast India and parts

of Gujarat and Kerala. All these studies reveal that there is no similarity in rainfall trends at the regional level. In the present study, the main objective to all the season of rainfall data in Tamil Nadu was examined to identify the pattern and classification in the following statistical techniques:

(a) To identify the pattern of rainfall data in the study period using Factor Analysis and extracted the factor scores.

(b) To identify the final cluster centres and classification of rainfall data using k-mean clustering techniques.

3. Methodology

This section brings out the discussion of the database, the monthly wise rainfall data selected and the Data Mining Techniques.

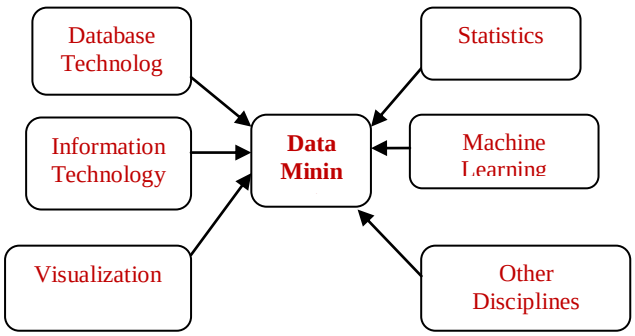
3.1. Database and Study Region

This section brings out the discussion of the database; the monthly wise rainfall parameters recorded various districts and in Tamilnadu. The district wise rainfall data were collected from secondary source from Meteorological Department in India, department during the period of 2004 to 2010 was considered as the database. In this study, three seasonal rainfall data were chosen that had been used in previous studies. The data mainly consists of three major categories, such as winter, summer and monsoon seasons.

3.2 Data Mining Techniques

Data Mining is an interdisciplinary field, the confluence of a set of disciplines in the following figure, including database systems, Statistics, machine learning, visualization and information science.

Although data mining is a new term, the technology is not. Data Mining or Knowledge Discovery in Databases (KDD) is the process of discovering previously unknown and potentially



useful information from the data in databases. In the present context data mining exhibits the patterns by applying few techniques namely, factor analysis, k-means clustering and discriminant rule. As such KDD is an iterative process, which mainly consist of the following steps;

Step 1: Data cleaning;
***Step 2*:** Data Integration;
Step 3: Data selection and transformation;
*S**tep 4***: Data Mining and
*S**tep 5***: Knowledge representation

Of these above iterative process Steps 4 and 5 are most important. If clever techniques are applied in Step 5, it provides potentially useful information that explains the hidden structure. This structure discovers knowledge that is represented visually to the user, which is the final phase of data mining.

3.3 Factor Analysis

The data to be compressed consist of N data vectors, from k-dimensions. Principal Component Analysis (PCA) searches for c k dimensional orthogonal vectors that can be best be used to represent the data $c \leq k$. Principal The original data set are projected onto a much smaller space, resulting in data compression. PCA can be used as a form of dimensionality deduction (J. Han M. Kamber, 2002). Factor analysis provides the tools for analyzing the structure of the interrelationships (correlations) among the large number of variables by defining sets of variables that are highly interrelated, known as factors. In the present study, factor analysis is initiated to uncover the patterns underlying monthly wise rainfall data. In factor extraction method the number of factors is decided based on the proportion of sample variance explained. Orthogonal rotations such as Varimax and Quartimax rotations are used to measure the similarity of a variable with a factor by its factor loading.

3.4 k-Means Clustering Methods

In unsupervised learning (a) no class label, (b) finding common patterns and (c) grouping similar objects. McQueen (1967) suggests the term k-means for describing an algorithm of his that assigns item to the cluster having the nearest centroid (mean). Generally this technique uses Euclidean distances measures computed by variables. Since the group labels are unknown for the data set, k-means clustering is one such technique in applied statistics that discovers acceptable meaningful classes or groups.

3.5 Geographical Information System

In the present study, GIS map is used to exhibit groups graphically and judge the nature of overall performance of the rainfall data.

4. Algorithms

A brief algorithm to classify the rainfall data during each of the study period based on their overall performance is described below:

Step 1: Factor analysis is initiated to find the structural pattern underlying the data set.

Step 2: k –means analysis partitioned the data set into k-clusters using year wise rainfall data as input matrix.

Step 3: Construct a Map using GIS Map and the final cluster centered (mean) values with appropriate hits of the district in the rainfall data set that are assigned group labels in step 2.

5. Results and Discussion

As mentioned in Section 3.3 Varimax and Quartimax criterion for orthogonal rotation have been used for the pruned data. Even though the results obtained by both the criterions were very similar, the varimax rotation provided relatively better clustering of financial ratios. Consequently, only the results of varimax rotation are reported here. We have decided to retain 75 to 80 percent of total variation in the data, and thus accounted consistently four factors for 2004, 2005 and 2010, the remaining years consistently three factors for each year with eigen values little less than or equal to unity. *Table 1* shows variance accounted for each factors

Table 1. Percentage of variance explained by factors

Factors	2004	2005	2006
1	30.734	27.235	30.194
2	22.359	23.128	23.477
3	17.806	13.725	16.552
4	8.822	9.361	0.000
Total	**79.72**	**75.49**	**70.22**

Table 2. Percentage of Variance explained by factors (Year-wise)

Factors	2007	2008	2009	2010
1	30.409	33.480	32.962	33.260
2	27.609	26.369	28.351	22.195
3	16.250	17.792	13.906	16.213
4	0.000	0.000	0.000	9.859
Total	**74.26**	**77.64**	**75.21**	**81.52**

From the above table we observe that the total variances explained by the extracted factors are over 75 percent, which are relatively higher. Also, for each factors the variability is more or less the same for the study period, though the number of rainfall in each year, after data cleaning and selection, kept varying owing to various reasons.

The monthly wise data loaded in the factors are presented in *Table 3 (1 to 10)*. Only those ratios with higher loadings are indicated by red color. From the *Table 4* it is clear that the clustering of rainfall is stable and unstable during the study period. We observed slight changes in factor loadings during the periods considered. The

differences in factor loadings may be due to statistical variations in the original data. In every year the factors are loaded depends upon the seasonal rainfall, we named the factors like, summer factor, winter factor and monsoon factors.

Table 3. Financial Ratios in Rotated Factors
(Yearwise from 2004 to 2010) Table-1(2004)

Months	Component			
	1	2	3	4
December	.859	-.013	-.219	.157
November	.851	.401	.103	-.074
October	.850	.220	.364	-.052
April	-.580	.406	-.254	.224
June	-.257	.876	.208	-.024
August	.217	.805	.289	-.114
January	.259	.717	-.097	.008
May	.116	.162	.843	.359
September	.336	.184	.790	.067
February	.280	.515	-.616	.335
July	-.506	.087	.614	.401
March	.029	.094	-.226	-.848

Table 2. (2005)

Months	Component			
	1	2	3	4
July	.923	.256	-.001	.060
June	.901	.130	-.140	-.026
August	.749	-.463	.238	-.133
April	.079	.794	-.028	-.056
January	.348	.726	-.107	-.167
February	-.306	.564	-.381	-.085
November	-.223	.018	.820	.148
March	-.080	.235	-.707	-.197
September	.434	-.361	.677	-.027
May	.115	-.318	-.431	-.370
October	.177	-.299	.056	.884
December	-.175	.003	.276	.858

Table 3. (2006)

Months	Component		
	1	2	3
June	.881	-.087	.072
July	.857	.220	.294
May	.716	.007	-.529
September	.705	-.335	.067
August	.648	-.461	.174
January	-.032	.909	.077
November	-.178	.773	.435
February	-.096	.698	.013
March	-.065	.675	-.432
December	.100	.021	.831
October	.127	.389	.771
April	-.073	.141	-.646

Table 4. (2007)

Months	Component		
	1	2	3
December	.896	.001	-.033
April	-.894	.201	.094
October	.702	.244	.378
March	-.673	.555	.319
May	-.531	.378	-.461
September	-.088	.878	-.079
July	.122	.859	-.056
June	-.266	.850	.093
August	.580	.631	-.357
November	.037	.072	.832
January	-.225	-.294	.615
February	.460	.403	.524

Table 5. (2008)

Months	Component		
	1	2	3
June	.910	.246	.034
July	.880	.204	-.291
September	.782	-.235	.130
August	.652	-.120	-.523
March	-.202	.914	.057
October	.234	.844	.039
February	.414	.725	-.152
April	-.074	.629	-.611
May	.422	-.569	-.399
November	-.174	-.051	.913
January	.098	-.007	.894
December	-.347	.442	.579

Table 6. (2009)

Months	Component		
	1	2	3
December	.950	-.147	.102
January	.933	-.106	-.041
November	.858	.235	-.052
March	.615	.575	-.202
June	-.219	.881	.215
July	-.040	.854	.198
October	.422	.755	-.254
May	-.526	.567	.233
August	.161	.474	.734
September	-.451	.265	.674
April	.257	.313	-.673
February	-.117	-.071	-.636

Table 7. (2010)

Months	Component			
	1	2	3	4
August	.887	-.117	.155	.242
June	.845	.296	.054	.199
July	.771	.337	-.269	.328
September	.702	-.137	.145	-.359
February	.679	.557	-.170	-.251
April	-.141	.891	-.138	.089
March	.357	.867	-.015	.061
October	.109	.645	.388	.057
December	.164	-.260	.901	.068
November	-.041	.020	.859	-.126
January	-.035	.204	.836	-.094
May	.157	.084	-.107	.936

Red color Indicates financial ratios highly loaded in respective factors

After performing factor analysis, the next stage is to assign initial group labels to each company. Step 2 of the algorithm is explored with rainfall data by Step 1, by conventional k-means clustering analysis. Formations of clusters are explored by considering 2-clusters, 3-clusters, 4-cluster and so on. Out of all the possible trials, 3-cluster exhibited meaningful interpretation than two, four and higher clusters. Having decided to consider only 3 clusters, it is possible to tempo a rainfall as group **HR**, group **MR** or group **LR** depending on whether the rainfall belonged to Cluster 1, Cluster 2 or Cluster 3 respectively. Cluster 1 (group **HR**) is a group of rainfall that have high values for the rainfall data, indicating that these years are performing well. The years with lower values for the rainfalls are grouped into Cluster 3 (Group **LR**). This suggested that Cluster 3 is a group of years with low rainfall. Cluster 2 (Group **MR**) are those years which perform moderately well as compared to the Cluster 1 and Cluster 3. Inspite of incorporating the results for each year, only the summary statistics are reported in *Table 4*. *Table 4* indicates that majority of rainfalls are in the monsoon seasons category except for the year 2004. The possible reasons that climate change, earthquake in Indian Ocean. High Rainfall in the year 2004 may be due to the reason for climate change in Tamilnadu coastal area. *Figure 1* through *7* shows the groupings of rainfalls into 3 clusters for each year of the study period. It is interesting to note that the mean vectors of these clusters can be arranged in the up and down order of magnitude as show in *Table 4 and Figure 1 to 7.*

Table 4. Number of Rainfall districts with Cluster Centers (Table 1 to 7)

Table 1

Factor Scores	2004		
	1	2	3
January	11	11	7
February	7	1	2
March	7	5	4
April	45	2	35
May	92	322	270
June	18	22	57
July	23	29	79
August	16	76	32
September	103	11	238
October	167	1	245
November	142	5	157
December	13	2	10
Total	**07**	**04**	**19**

Table 2

Factor Scores	2005		
	1	2	3
January	4	322	0
February	13	22	2
March	31	29	9
April	139	76	113
May	77	269	30
June	36	525	30
July	86	329	109
August	81	85	85
September	96	166	136
October	266	199	679
November	268	597	532
December	124	178	413
Total	**20**	**07**	**03**

Table 3

Factor Scores	2006		
	1	**2**	**3**
January	8	17	28
February	0	0	0
March	45	23	89
April	38	17	37
May	69	30	66
June	50	60	63
July	10	53	39
August	72	80	59
September	120	128	103
October	193	484	318
November	157	286	284
December	15	70	28
Total	**16**	**05**	**09**

Table 4

Factor Scores	2007		
	1	**2**	**3**
January	3	0	7
February	15	12	4
March	2	0	1
April	29	23	68
May	46	27	50
June	173	36	56
July	236	78	63
August	254	195	110
September	169	91	69
October	261	360	192
November	92	98	77
December	227	350	181
Total	**02**	**06**	**21**

Table 5

Factor Scores	2008		
	1	**2**	**3**
January	6	41	3
February	28	26	87
March	181	173	224
April	36	8	64
May	58	33	59
June	31	49	155
July	64	48	246
August	135	97	219
September	58	75	127
October	230	255	368
November	192	604	89
December	42	85	30
Total	**18**	**09**	**03**

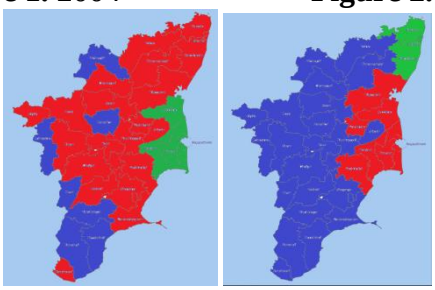
Table 6

Factor Scores	2009		
	1	**2**	**3**
January	2	26	5
February	0	0	0
March	98	61	23
April	46	40	37
May	143	30	68
June	184	16	34
July	636	25	44
August	234	117	87
September	212	87	124
October	162	70	57
November	658	539	272
December	73	325	75
Total	**01**	**06**	**23**

Table 7

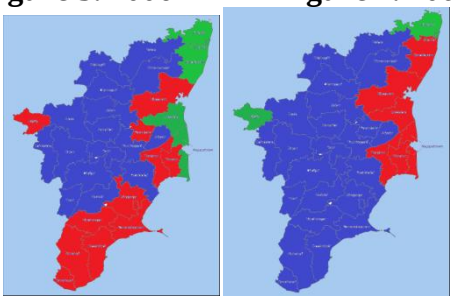
Factor Scores	2010		
	1	**2**	**3**
January	10	33	7
February	1	0	0
March	10	2	1
April	30	16	24
May	124	103	108
June	191	91	65
July	327	63	76
August	308	156	88
September	194	134	110
October	185	187	149
November	257	448	297
December	122	295	98
Total	**02**	**07**	**21**

Figure 1. 2004 **Figure 2.** 2005



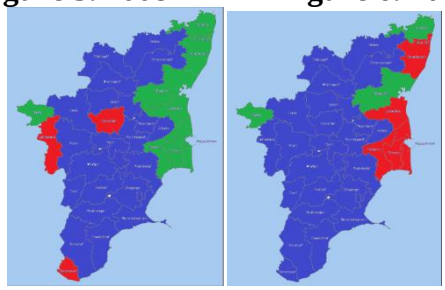
1 – HR 2 – MR 3-LR

Figure 3. 2006 **Figure 4.** 2007



1 – HR 2 – MR 3-LR

Figure 5. 2008 **Figure 6.** 2009



1 – HR 2 – MR 3-LR

Figure 7. 2010

1 – HR 2 – MR 3-LR

Clustered Groups (Figures 1 – 7)

6. Conclusion

The purpose of this paper was to identify the meaningful classification of rainfall data that are classified as best with respect to their rainfall in terms of monthly wise rainfall and data mining techniques. An attempt is made to analysis the rainfall data relating to various climate and monsoon period of seven years from 2004 to 2010. The present analysis has shown that only 3 groups could be meaningfully formed for each year. This indicates that only 3 types of rainfall existed over a period of seven years. Further, the year find them classified into *High Rainfall* (Group **HR**), *Medium Rainfall* (Group **MR**) and *Low Rainfall* (Group **LR**) categories depending on the climate and rainfall. A generalization of the results is under investigation to obtain an included class of 3 groups of rainfall for any given year in Tamilnadu.

7. References

[1] Anderson T W (1984), An Introduction To Multivariate Statistical Analysis 2/e,John Wiley and sons,Inc,New York.

[2] M.C.Ramos.2001.Rainfall distributionPattern and their over time in a

[3] Mediterranean area.Theoretical and Applied Climatology.69,163170.

[4] Pramanik, S.K., and Jagannathan, P., (1954), Climate change in India – 1: rainfall. *Indian Journal of Meteorology Geophysics* , 4, 291–309.

[5] Parthasarathy, B., (1984), Inter annual and long term variability of Indian summer monsoon rainfall. In: *Proceedings of the Indian Academy of Sciences (Earth Planetary Sciences),* vol. 93, pp. 371–385.

[6] Parthasarathy, B., and Dhar, O.N., (1978), Climate Fluctuations Over Indian Region – Rainfall: a Review, vol. 31. Indian Institute of Tropical Meteorology, Pune. Research Report No. RR-025.

[7] Parthasarathy, B., Rupakumar, K., Munot, A.A., (1993), Homogeneous Indian monsoon rainfall: variability and prediction. In: *Proceedings of the Indian Academy of Sciences (Earth Planetary Sciences),* vol. 102, pp. 121–155

[8] Mooley, D.A., Parthasarathy, B., (1983), Variability of Indian summer monsoon rainfall and tropical circulation features. *Monthly Weather Review*, 111, 967–968.

[9] Rao K, PR. and Jagannathan,P.(1953). A study of the northeast monsoon rainfall of Tamil 31) Nadu; *Indian journal of meteorology and Geophysics,* 4:22.

[10] Thapliyal, V., Kulshrestha, S.M., (1991), Climate changes and trends over India. *Mausam*, 42, 333–338.

[11] Srivatsava, H.N., Dewan, B.N., Dikshit, S.K., Prakasa Rao, G.S., Singh, S.S., Rao, K.R., (1992), Decadal trends in climate over India. *Mausam*,

[12] 43, 7–20.

[13] Koteswaram, P., Alvi, S.M.A., (1969), Secular trends and periodicities in rainfall at west coast stations in India. *Current Science*, 101, 371–375.

[14] Jagannathan, P., Bhalme, H.N., (1973), Changes in pattern of distribution of southwest monsoon rainfall over India associated with sunspots. *MonthlyWeather Review* , 101, 691–700.

[15] Naidu, C.V., Srinivasa Rao, B.R., Bhaskar Rao, D.V., (1999), Climatic trends and periodicities of annual rainfall over India. *Meteorological Application*, 6, 395–404.

[16] Singh,N. and Sontakke, NA.(1999): On the variability and prediction of rainfall in the postmonsoon season over India; *International journal of climatology*, 19: 309.

[17] Rupa Kumar, K., Pant, G.B., Parthasarathy, B., Sontakke, N.A., (1992), Spatial and sub-seasonal patterns of the long term trends of Indian summer monsoon rainfall. *International Journalof Climatology*, 12, 257–268.