MAUSAM, **66**, 4 (October 2015), 809-814

551.579 : 551.577 : 551.509.314 (540.33)

Rainfall day and heavy rainfall day prediction of Kolkata (22.53° N, 88.33° E), India during June to October using linear discriminant analysis technique

SUKUMAR LALA (ROY) and NABAJIT CHAKRAVARTY*

Regional Meteorological Centre, Alipore, Kolkata – 700 027, India

**Positional Astronomy Centre, Salt Lake City, Kolkata – 700 091, India*

(*Received 16 August 2013*)

e mail : sukumarlala@gmail.com

सार – इस शोध पत्र में 25 वर्षों (1980-2005) के आँकड़ों का उपयोग करके जून से अक्तूबर के दौरान कोलकाता, (22.53° उ., 88.33° पू.), भारत में वर्षा के दिन और भारी वर्षा के दिन का पूर्वानुमान लगाने के लिए प्रभावी मौसम विज्ञानिक प्राचलों का उपयोग करके सांख्यिकीय सूचकांक तैयार किया गया है।

रैखिक विविक्तिकर विश्लेषण (एल डी ए), जो बहुचर सांख्यिकीय तकनीक है, का उपयोग वर्ष 2006-2008 के जून से अक्तूबर की अवधि के लिए वर्षा के दिनों और भारी वर्षा के दिनों का पूर्वानुमान लगाने के लिए सांख्यिकीय सूचकांक का पता लगाने के लिए 22 चयनित मौसम विज्ञानिक प्राचलों का उपयोग किया गया है और फिर इन वर्षों के वास्तविक वर्षा के दिनों और भारी वर्षा के दिनों की वैधता की तुलना की गई है । यह पाया गया है कि अगले तीन वर्षों (2006-2008) की अवधि में वर्षा नहीं होने के दिनों और वर्षा के दिनों में क्रमश: 60.34 प्रतिशत और 73.36 प्रतिशत का सही पूर्वानुमान रहा। अगले तीन वर्षों (2006-2008) के दौरान भारी वर्षा के दिनों में क्रमश: 90.98 प्रतिशत और 84.21 प्रतिशत सही पूर्वानुमान रहा। सभी 22 प्राचलों के लिए सक्षम स्किल स्कोर नामत: ट्रू स्किल स्कोर (टी एस एस), हेडके स्किल स्कोर (एच एस एस, क्रिटिकल सक्सेस इंडेक्स (सी एस आई) की गणना वर्षा के पूर्वानुमान और भारी वर्षा के पूर्वानुमान दोनो के लिए की गई। जाँच से पता चला कि एल डी ए तकनीक भारी वर्षा के पूर्वानुमान के लिए अधिक कुशल है जहाँ दो समूहों, नामत: भारी वर्षा नहीं होने वाले दिनों और भारी वर्षा के दिनों, के प्राचलों के मध्य तीव्र व्यतिरेक रहा।

ABSTRACT. In the present work, statistical index are formed using the effective meteorological parameters for predicting the rainfall day and heavy rainfall day of Kolkata (22.53° N, 88.33° E) India during June to October utilizing the data of 25 years (1980-2005).

Linear Discriminant Analysis (LDA), which is a multivariate statistical technique has been utilized to 22 selected meteorological parameters to find out the statistical index that has been utilized to predict the rainfall day and the heavy rainfall day for the period June to October for the year 2006-2008 and then validate with the actual rainfall day and heavy rainfall day of these years. It was found that it yielded 60.34% and 73.36% correct prediction for No rainfall days and rainfall days respectively during the period in the next three years (2006-2008). It yielded 90.98% and 84.21% correct prediction for No heavy rainfall days and Heavy rainfall days respectively during the period in the next three years (2006-2008). For all 22 parameters the efficient skill scores namely, True Skill Score (TSS), Heidke Skill Score (HSS), Critical Success Index (CSI) are computed for both rainfall prediction and heavy rainfall prediction. The investigation revealed that LDA technique is more efficient in prediction of heavy rainfall where there were sharp contrast between the parameters of the two groups, namely no heavy rainfall days and heavy rainfall days.

Key words – Linear discriminant analysis (LDA), Rainfall day, Heavy rainfall day, Discriminant function and centroid.

1. Introduction

The prediction of rainfall day and heavy rainfall day during the period June to October in Kolkata is of much concern to the forecasters due to it's erratic nature. Also the inhabitants of this area are interested to know whether there will be rain on a particular day and if it occurs whether it will be heavy, during the period June to October because this period falls in monsoon season and initial parts of post monsoon season where forecasting of a rainy day or heavy rainy day is difficult. Thus prediction of these events are always of ultimate interest for the

TABLE 1

Data size for the period 1980-2008 for event rainfall (No rainy and rainy days) and for event heavy rainfall (No heavy rainy day and heavy rainy day)

Event	No. of parameters	Category	For construction of discriminant index (1980-2005) (Total 3825 cases out of which 811 missing data of one or more variables not taken)		For validation test (2006-2008) (Total 459 cases out of which 11 missing data of one or more variables not taken)	
			No. of observations	Total	No. of observations	Total
Rainfall	22	No rainy day (NR)	1245	3014	174	448
		Rainy day (R)	1769		274	
Heavy rainfall	22	No heavy rainy day (NHR)	1690	1769	255	274
		Heavy rainy Day (HR)	79		19	

researchers. In fact, many previous researchers utilized different multivariate techniques in different situations of atmosphere to predict the events which is reflected from the works of Brier and Allen, 1952; Agresti, 1996 and Asnani, 2005. To study the principal anomaly in winter temperature, eigenvector methods have been applied by Diaz and Fulbright, 1981. To describe a multivariate statistical model for forecasting anomalies of surface pressure over Europe and North America, Cluster analysis (CL) and LDA have been comprehensively used (Maryon and Storey, 1985). A composite Empirical orthogonal function (EOF) of monthly sea surface temperature (SST) and also of precipitation in the tropical Pacific ocean region was performed (Weare, 1987). Ward and Folland, 1991, utilized both multiple linear regressions and LDA to forecast rainfall and SST in north-east Brazil. Works on objective evaluator of techniques for prediction of severe weather events has been done by Donaldson *et al*., 1975.

Several multivariate statistical methods have been used by several researchers to establish different phenomenon in India. A good number of attempts have been made to predict the occurrence of rainfall by two-state Markov-chains (Dasgupta and De, 2001; Pant and Shivhare, 1998 and Thiagarajan *et al*., 1995). PCA has been applied by several scientists to understand the monsoon rainfall (Iyenger and Basak, 1994; Sengupta and Basak, 1998) where they have identified specific regions of India with respect to rainfall. The relationship between the frequency of rain and various meteorological parameters has been studied by Hanssen and Kuippers, 1965.

In this paper, an attempt has been made to find the statistical index by assigning 22 meteorological parameters which prevail during the occurrence of the events, *i.e*., rainfall and heavy rainfall of this region so that prediction can be made about the rainfall days and the heavy rainfall days during the most active period, *i.e.*, June to October of the event. This can be an added guiding tool along with the existing tools that the forecasters are equipped with for forecasting rainy and heavy rainy days Table 1.

2. Data

The meteorological parameters responsible for occurrence of rainfall and heavy rainfall within the next 24 hours of the 0830 hrs IST observation of the station after incorporating the Bright sunshine hours and Total radiation received by it's nearby station (Dumdum : 22.65° N, 88.45° E) as these two parameters are also an effective parameter for the cause of rainfall of the station which has been emphasized by Wong and Chow, 2001 was taken for Linear Discriminant Analysis (LDA) in same line as per the works of Basak, 2012. Total 22 parameters, P_i (i =1, 2,…, 22) are formed from the surface and radiation observations of 25 years (1980-2005) and those have been utilized for LDA.

Total 22 parameters, P_i (i =1, 2,…, 22) are constructed as follows:

P_1 = Bright sunshine hours,

P_2 = Dry bulb temperature in ° C,

P_3 = Direction of High cloud in meteorological code,

P_4 = Direction of Low cloud in meteorological code,

P_5 = Direction of Medium cloud in meteorological code,

P_6 = Dew point in °C,

P_7 = Form of High cloud in meteorological code,

P_8 = Amount of High cloud in okta,

P_9 = Form of Low cloud in meteorological code,

P_{10} = Amount of Low cloud in okta,

P_{11} = Height of Low cloud in meteorological code,

P_{12} = Maximum temperature in °C,

P_{13} = Form of Medium cloud in meteorological code,

P_{14} = Amount of Medium cloud in okta,

P_{15} = Minimum temperature in °C,

P_{16} = Daily Total Radiation in watt per square metre,

P_{17} = Relative Humidity in percent,

P_{18} = Sea Level Pressure in hPa,

P_{19} = Saturated Vapor Pressure in hPa,

P_{20} = Total amount of Cloud in okta,

P_{21} = Wet Bulb temperature in °C and

P_{22} = Wind Direction in meteorological code.

In the first stage the 22 meteorological parameters were grouped into two categories, one set contained the met. parameters related to rainfall below and equal to 0.4 mm which was taken as a No rainy day and the other set contained the met. Parameters related to rainfall above 0.4 mm which was taken as a rainy day.

In the second stage the 22 meteorological parameters among the set of rainy days were grouped into two categories, one set contained the meteorological parameters related to rainfall below and equal to 60 mm which was taken as a No heavy rainy day and the other set contained the met. parameters related to rainfall above 60 mm which was taken as a Heavy rainy day Table 2.

TABLE 2

Discriminant functions for 22 parameters for no rainy day and rainy day and no heavy rainy day and heavy rainy day

Events	Nature of days for auto-verification	Number of parameters	Number of days involved	Discriminant function
Rainfall	NR	22	1245	RD_x 0.50474
	R	22	1769	RD_y -0.35523
Heavy rainfall	NHR	22	1690	HD_x 0.05875
	HR	22	79	HD_y -1.25685

3. Methodology

In this section the basics of the multivariate of the technique, namely, Linear discriminant analysis (LDA) has been discussed in short. It is followed by the result of the analysis.

3.1. *Linear discriminant analysis (LDA)*

We consider the two sets of observations $X = [X_{ij}]$, (i =1, 2,…, k and j = 1, 2,…., m) and $Y = [Y_{ij}]$, (i = 1, 2,…, k and j=1, 2,….n) where i, j stands for number of parameters (k in each case) and number of days (m and n in two cases) respectively. X and Y are the group-symbols of no rainy day and rainy day for one set and no heavy rainy day and heavy rainy day for another set.

The groups X and Y are arranged as follows :

$$X = [\,\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k\,]\;;\; \overline{X}i = \left\langle \frac{1}{m} \right\rangle \sum_{j=1}^{m} X_{ij}$$

$$Y = [\,\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k\,]\;;\; \overline{Y}i = \left\langle \frac{1}{n} \right\rangle \sum_{j=1}^{n} Y_{ij}$$

The covariance matrices of each group are as follows:

$$S_x = [S_x(i, j)]_{kxk}$$

$$\text{where, } S_x(i,j) = [\,1/(m\text{-}1)] \sum_{p=1}^{m} (X_{ip} - \overline{X_i})\,(X_{jp} - \overline{X}_j)$$

$$S_y = [S_y(i,j)]_{kxk}$$

$$\text{where, } S_y(i, j) = [\,1/(n\text{-}1)] \sum_{p=1}^{n} (Y_{ip} - \overline{Y_i})\,(Y_{jp} - \overline{Y}_j)$$

TABLE 3(a)

LDA analysis for 22 parameter combination for set 1 (No rainy day and Rainy day)

Event	Nature of days	Number of variables	Number of days involved	Number of correct results	Percentage of success	Average percentage of success
Rainfall	NR	22	174	105	60.34	68.30
	R	22	274	201	73.36	

TABLE 3(b)

LDA analysis for 22 parameter combination for set 2 (No heavy rainy day and Heavy rainy day)

Event	Nature of days	Number of variables	Number of days involved	Number of correct results	Percentage of success	Average percentage of success
Heavy Rainfall	NHR	22	255	232	90.98	90.51
	HR	22	19	16	84.21	

In the analysis, without any loss of generosity, it is assumed that the population in each of the groups have same covariance matrix, the pooled estimate of the dispersion of data around their means are

$$S = [1/(m - n - 2)\ ].\ [(m - 1)S_x + (n - 1)\ S_y]$$

We now verify the nature of the unknown group $U = [U_{ij}]$ (i = 1, 2,……, k and j = 1, 2,……l); i, j stand for number of parameters and number of days respectively. The discriminant functions of X, Y and U are

$$D_x = \overline{X}\,'\ S^{-1}(\overline{X} - \overline{Y})$$

$$D_y = \overline{Y}\,'\ S^{-1}(\overline{X} - \overline{Y})$$

$$D_u = \overline{U}\,'\ S^{-1}(\overline{X} - \overline{Y})$$

where, dash denotes the transpose of the matrix.

If $|Dx - Du| < |Dy - Du|$, then U belongs to the X – group, that is the nature of unknown days to be of the nature of No rainy day or No heavy rainy day.

If $|Dx - Du| > |Dy - Du|$, then U belongs to the Y – group, that is the nature of unknown days to be of the nature of Rainy day or Heavy rainy day [Tables 3(a&b) and Table 4].

TABLE 4

Contingency table of skill scores

Observation	Prediction	
	Events predicted	Events not predicted
Events observed	A (Hits)	B (Misses)
Events not observed	C (False Alarm)	D (Non-events Hits)

4. Analysis

The analysis has been performed for 2 sets. In the first set, LDA technique is applied to the matrices X and Y where X and Y contain 22 parameters P_i (i = 1, 2,…, 22) as mentioned in the earlier section X consists of the parameters of No Rainy day and Y consists of those of Rainy days. The discriminant function for the No Rainy day is denoted by RD_x and for Rainy day by RD_y (Table 2). The indices RD_x and RD_y are constructed utilizing surface data of 0830 hrs IST for Kolkata and radiation data of 0830 hrs IST for Dumdum (near Kolkata) for June to October of 25 years (1980-2005). The dimensions of X and Y matrices are 1245 × 22 and 1769 X 22 (*e.g.*, x, y ; x = number of days ; y = number of parameters (Table 1). These RD_x and RD_y are used to predict the nature of days of Unknown system (US) from the set of 3 years (2006 - 2008). The results are presented in Table 3(a).

The same procedure has been applied for the data set of Rainfall days from which two categories have been made, one set comprising of no heavy rainy day and the other set comprising of Heavy rainy day. The data matrix sizes are 1690 × 22 and 79 × 22 representing No heavy rainy day and heavy rainy day respectively. The corresponding discriminant functions have been denoted by HD_x and HD_y for No heavy rainy day and heavy rainy day respectively (Table 2). Utilizing these HD_x and HD_y, the result of the prediction from the nature of days of the US are presented in Table 3(b).

5. Results and discussion

5.1. *No rainy day and rainy day*

With the 22 parameters, the LDA analysis yields 64.4 % correct prediction for No rainy days and 62.0 % for rainy days [Table 3(a)] for verification of 3 years (2006-2008). With 22 parameters, the LDA analysis yields 91.0 % correct prediction for No heavy rainy days and 85.7 % for heavy rainy days [Table 3(b)] for verification of 3 years (2006-2008) (Tables 5 and 6).

TABLE 5

Description of different skill scores

Skill score	Code	References	Equation	Limits
Probability	POD	Donaldson *et al*. (1975)	POD = A/(A+B)	$0 \leq POD \leq 1$
False alarm ratio	FAR	Donaldson *et al*. (1975)	FAR = C/(A+C)	$0 \leq FAR \leq 1$
Critical success index	CSI	Donaldson *et al*. (1975)	CSI = A/(A+B+C)	$0 \leq CSI \leq 1$
True skill statistics	TSS	Hanssen and Kuippers (1965)	TSS = (A/A+B)-(C/C+D) = (AD-BC)/(A+B)(C+D)	$-1 \leq TSS \leq 1$
Hiedke skill score	HSS	Brier and Allen (1952)	HSS = (CFE)/(N-E) = 2(AD-BC)/(A+B) (B+D) + (A+C)(C+D)	$-1 \leq HSS \leq 1$
Miss Rate	MR	-	B / (B+A)	$0 \leq MR \leq 1$
Correct Non-Occurrence	C-Non	Dhawan *et al*. (2008)	D / (D+C)	$0 \leq C\text{-}Non \leq 1$
Bias	BIAS	Dhawan *et al*. (2008)	(A+C) / (A+B)	-
Percent Correct	PC	-	[(A+D) / (A+B+C+D)] × 100	$0 \leq PC \leq 100$

6. Skill scores

For each stage of the analysis the results are presented in the form of a 2 × 2 contingency table. The entries of the table are 'correctly forecasted events (A)', 'events not correctly forecasted (B)', 'events forecasted but not observed (C)' and 'events not forecasted and also not observed (D)'. The presentation is shown in Table 4. Based on these, nine skill scores, namely Probability of Detection (POD), False Alarm Ration (FAR), Critical Success Index (CSI), True Skill Score (TSS), Heidke Skill Score (HSS), Percentage of Correct result (PC) and others are computed (Table 5). Brief description of the skill scores are presented in Table 5.

It may be stressed that perfect forecast will show a HSS score of 1, a set of random forecast will be 0 and a lesser hits compared to the forecast by chance will have negative score. TSS and HSS both are being used in literature as Rain and Heavy rain forecast skill parameters (Mukhopadhyay *et al*., 2003; Tyagi *et al*., 2010), however there seems to be a quite difference between their characteristics namely, TSS pursues a high POD, HSS attempts to reduce FAR to reasonable rate (Haklander and Deldon, 2003). The limitations of TSS and HSS is that, if the number of correct forecast (A) and number of correct non event forecast (D) are interchanged and number of misses (B) and also number of False alarms (C) are interchanged, scores remain unchanged. But, CSI would change. Thus, no single forecast would give complete picture. However it is desirable to include CSI, POD, FAR, MR, C-NON, BIAS and PC in addition to HSS for broader and useful forecast.

TABLE 6

Table of different skill scores

Skill scores	Rainfall	Heavy rainfall
POD	0.7444	0.4103
FAR	0.2664	0.1579
MR	0.2556	0.5897
C-NON	0.5899	0.9872
CSI	0.5860	0.3810
TSS	0.3343	0.3975
HSS	0.3356	0.5056
BIAS	1.0148	0.4872
Per cent correct	68.30%	90.51%

6.1. *Results of skill score*

For each stage of the analysis and also for each set of rainy and heavy rainy days, the skill scores are presented in Table 6.

6.1.1. *Rainy day*

The overall forecast skills are almost consistent. The TSS and HSS for stage (22 parameter) are also consistent. The highest score occurred is case of 22 parameter combination (*i.e.*, TSS is 0.33 approx. and HSS is 0.33 approx.). The POD, PC and C-NON are 0.74, 68.3 and 0.59 approx. respectively. The CSI is 0.59 approx. whereas FAR and MR is 0.27 and 0.26 approx. respectively.

6.1.2. *Heavy rainy day*

The overall forecast skills are also consistent and shows relatively high percentage of accuracy than that of Rainy day set. The TSS and HSS for stage (22 parameters) are also consistent. The highest score occurred is case of 22 parameter combination (*i.e.*, TSS is 0.40 approx. and HSS is 0.51 approx.). The POD, PC and C-NON is 0.41, 90.51 and 0.99 approx. respectively. The CSI is 0.38 approx., whereas FAR and MR is 0.16 and 0.59 approx.).

7. Conclusions

The above analysis reveals that in case of predicting heavy rainfall day the LDA technique was more efficacious probably due to sharp distinction between the two group of 22 parameters due to some strong system in addition to the normal monsoon and initial post monsoon condition that prevailed during the occurrence of this event (*i.e.*, heavy rainfall) . Hence we can incorporate the use of the technique of LDA for prediction of heavy rainfall for the next 24 hours by observing the 0830 hrs IST observations of the station of date. Also the technique can be used optimally by adding other parameters that are responsible for the occurrence of the events.

It also reveals that in case of predicting rainfall day the accuracy is relatively low, hence it can be used as an added feature with the existing practice. The optimality of this feature can be increased by incorporating other parameters responsible for the occurrence of this event and is the avenue that can be explored by the researchers.

References

Agresti, A. 1996, "An Introduction to Categorical Data Analysis".

Asnani G. C., 2005, "Tropical Meteorology", **2**, 829-833.

Basak, P., 2012, "Convective development at Kolkata (22.53° N, 88.33° E), India during pre-monsoon season using linear discriminant analysis technique" *Mausam*, **63**, 3, 423-432.

Brier, G. W. and Allen, R. A., 1952, "Verification of weather forecasts", Compendium of Meteorology, *Amer. Meteor. Soc.*, 841-848.

Dasgupta, S. and De, U. K., 2001, "Markov chain models for pre-monsoon thunderstorm in Calcutta", *Ind. J. Radio & Space Phys.*, **31**, 138-142.

Dhawan, V. B., Tyagi, A. and Bansal, M.C., 2008, "Forecasting of thunderstorms in pre-monsoon season over north-west India", *Mausam*, **63**, 3, 423-432.

Diaz, H. F. and Fulbright, D. C., 1981, "Eigenvector analysis of seasonal temperature, precipitation and synoptic scale system frequency over contiguous United States, Part.1 (winter)", *Mon. Wea. Rev.,* **109**, 1267-1284.

Donaldson, R., Dyer, R. and Kraus, M., 1975, "An objective evaluator of techniques for prediction of severe weather events, Reprints, ninth Conf. on Severe Local Storms, Norman, OK", *Amer. Meteor. Soc.*, 321-326.

Haklander, A. J. and Delden, A. V., 2003, "Thunderstorm predictors and their forecast skill for the Netherlands", *Atmos. Res.*, **67-68**, 273-299.

Hanssen, A. W. and Kuippers, W. J. A., 1965, "On the relationship between the frequency of rain and various meteorological parameters", *Verhand. K. Nederlands. Meteor. Inst*., **81**, 2-15.

Iyenger, R. N. and Basak, P., 1994, "Regionalization of Indian Monsoon rainfall and long term variability signals", *Int. J. Climatol*., **14**, 1095-1114.

Maryon, R. H. and Storey, A. H., 1985, "A multivariate statistical model for forecasting anomalies of half-monthly mean surface pressure", *Int. J. Climatol.,* **5**, 561-578.

Mukhopadhyay, P., Sanjay, J. and Singh, S. S., 2003, "Objective forecast of thundery/non-thundery days using conventional indices over three northeast Indian stations", *Mausam*, **54**, 4, 867-880.

Pant, B. C. and Shivhare, R. P., 1998, "Markov chain model for study of wet/dry spells at A. F. station, Sarsawa during SW monsoon season", *Vatavaran*, **22**, 37-50.

Sengupta, P. R. and Basak, P., 1998, "Some studies of southwest monsoon rainfall", *Proc. Ind. Nat. Sci. Acad*., **64 A**, 737-745.

Thiagarajan, R., Ramadoss and Ramaraj, 1995, "Markov chain model for daily rainfall occurrences at east Thanjavur district", *Mausam*, **46**, 383-388.

Tyagi, B., Naresh Krishna,V. and Satyanarayana, A. N. V., 2010, "Study of thermodynamic indices in forecasting pre-monsoon thunderstorm over Kolkata during STORM pilot phase 2006-2008", *Nat. Hazards, online Pub.*, August, 2010.

Ward, M. N. and Folland, C. K., 1991, "Prediction of seasonal rainfall in the North east of Brazil using eigenvector of sea-surface temperature", *Int. J. Climatol.*, **11**, 711-743.

Weare, B. C., 1987, "Relationship between monthly precipitation and SST variation in the Tropical Pacific region", *Mon. Wea. Rev.,* **115**, 2687-2698.

Wong, L.T. and Chow, W. K., 2001, "Solar radiation model", *Applied Energy*, **69**, 191-224.