Journal of Applied and Natural Science 2(1): 42-47 (2010)

JANS

Statistical modeling of wet and dry spell frequencies over North-East India

S. Deka[1]*, M. Borah[1] and S.C. Kakaty[2]

[1] Department of Mathematical Sciences, Tezpur University, Napaam,Tezpur- 784028 (Assam), INDIA
[2] Department of Statistics, Dibrugarh University, Dibrugarh-786004 (Assam), INDIA
*Corresponding author. E-mail: surobhi@tezu.ernet.in

Abstract: In this paper an attempt has been made to develop a discrete precipitation model for the daily series of precipitation occurrences over North East India. The point of approach is to model the duration of consecutive dry and wet days i.e. spell, instead of individual wet and dry days. Various distributions viz. uniform, geometric, logarithmic, negative binomial, Poisson and Markov chain of order one and two, Eggenberger-Polya distribution have been fitted to describe the wet and dry spell frequencies of occurrences. The models are fitted to the observed data of seven stations namely Imphal, Mohanbari, Guwahati, Cherrapunji, Silcoorie, North Bank and Tocklai (Jorhat) of North-East India with pronounced attention to summer monsoon season. The goodness of fit of the proposed model has been tested using Kolmogorov-Smirnov test. It is observed that Eggenberger-Polya distribution fairly fits wet and dry spell frequencies and can be used in the future for an estimation of the wet and dry spells in the area under study.

Keywords: Geometric Distribution, Logarithmic Series, Negative binomial distribution, Poisson distribution, Markov Chain, Kolmogorov-Smirnov test

INTRODUCTION

The definition of spell is based on the duration of consecutive wet and dry days. A wet spell is a sequence of wet days and it begins and ends the day after and the day before a dry day. In this study a wet day (W) is considered as one where the precipitation is = 1mm and, obviously, dry day (D) the one where there is not precipitation or is not > 1mm.

According to Fisher (1924) crop yield during a season mainly influenced by the distribution of rainfall rather than season total amount of rainfall. Again the distribution of rainfall depends on the wet and dry spells over a period of time, so it is very important to investigate the pattern of occurrence of such spells during the Indian Summer Monsoon Season. The main objective of the present study is to find the best fitting model to describe the wet and dry spell frequencies of occurrences considering the climatic features of the different parts of North-East India. Among the possible statistical models,we have used Discrete uniform distribution; Geometric distribution; Logarithmic series; Negative binomial distribution; Poisson distribution; Markov chain of order one and two; Eggenberger-Polya distribution.The Kolmogorov-Smirnov test for goodness of fit was employed as the significance test for every model, assuming the level of significance as 5% (a =.05).

In order to put our discussion into proper perspective we relate our work with the existing literature. The most frequently used model for generating consecutive series of dry and wet days is the first order, two state, homogeneous Markov chain that has been applied by several authors (cf. Gabriel and Neumann (1962), Katz (1974), Bruhn *et al.* (1980), Richardson (1981), Geng (1986), Matyasovszky and Dobi (1986), Wilks (1992), Dubrovsky (1997). The major disadvantage of this model is that it overestimates the very short, but underestimates the very long dry sequences. An essential improvement to reproduce the short and long spells were made by Berger and Goossens (1983) and Nobilis (1986) using higher order Markov chain and Eggenberger-Polya distribution. They found that short spells were best fitted by fourth order Markov chain, where as the Eggenberger-Polya distribution gave the best fit to the long series. Later, Racsko *et al.* (1991) proposed a model constituting two different geometric distributions. In the referred study, both the geometric distributions were separated according to the length of dry spells. Results of the works suggested that mixed distribution, including geometric one, could be promising in reproduction of long dry periods. For wet spells, it was also observed that simple geometric distribution could be promising. Recently, following the idea of Racsko *et al.* (1991) a mixture distribution based on a weighted sum of two geometric distributions, as well as that of one geometric and one poisson distribution have been applied by Wantuch *et al.* (2000). The first model exhibits good fitting for the dry spells and the latter one can be advised to employ

Table 1. Results of the Kolmogorov-Smirnov (K-S) tests for North Bank (1986-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.3636	1	Discrete Uniform	0.3750
2	Geometric	0.4056	2	Geometric	0.4866
3	Logarithmic	0.4208	3	Logarithmic	0.4922
4	Neg. Binomial	0.5633	4	Neg. Binomial	0.5096
5	Poisson	0.2100	5	Poisson	0.2817
6	M.C of order one	0.0402	6	M.C of order one	0.0661
7	M.C of order two	0.0306	7	M.C of order two	0.0226
8	Eggenberger-Polya	0.0178	8	Eggenberger-Polya	0.0121
Critical value at $a = .05$		0.0545	Critical value at $a = .05$		0.0545

Table 2. Results of the Kolmogorov-Smirnov (K-S) tests for Tocklai (1986-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.3333	1	Discrete Uniform	0.3333
2	Geometric	0.4439	2	Geometric	0.5385
3	Logarithmic	0.4526	3	Logarithmic	0.5484
4	Neg. Binomial	0.3313	4	Neg. Binomial	0.4692
5	Poisson	0.2095	5	Poisson	0.3749
6	M.C of order one	0.0235	6	M.C of order one	0.0730
7	M.C of order two	0.0152	7	M.C of order two	0.0096
8	Eggenberger-Polya	0.0178	8	Eggenberger-Polya	0.0186
Critical value at $a = .05$		0.0505	Critical value at $a = .05$		0.0504

for the wet periods. More recently, while Tolika and Maheras (2005) have found that both Markov chain of order two and Negative Binomial distribution can be used to estimate the wet spells in Greece, Eggenberger-Polya and Truncated Negative Binomial were found to be more efficient in fitting observed data both for wet/dry spells by Giuseppe *et al.* (2005). Although a good number of literatures are available describing the model for daily precipitation round the globe, no rigorous work barring the work by Medhi (1976) pursued in the North East region of India.

A brief outline of this paper is as follows. Section 2, introduces a brief specification of data set and the statistical methods used in this work. In section 3, a discussion is carried out on the results obtained from different statistical models applied to analyze the wet and dry spells frequencies. Finally, section 4 is devoted to a critical assessment of the results obtained in section 3.

MATERIALS AND METHODS

In this study series of daily rainfall data of seven stations in North East India viz. Imphal (2001-2005), Mohanbari (1993-2006), Guwahati (2001-2005), Cherrapunji (2001-2005), Silcoorie (1986-2005), North Bank (1986-2005), Tocklai (1986-2005) have been selected. The locations of these seven stations of North East India are shown in Fig 1. The series of daily rainfall are taken from Regional Meteorological Centre, Guwahati and Tocklai Experimental Station, Jorhat involving the aforesaid seven stations for the summer season (April to September) in each year.

When a spell overlaps a seasonal change (that is, it includes the 31st of march and 1st of April or 30th September and the 1st of October) it is considered in its whole up to its modality change even if it reaches the following season and we include it in the season in which it develop longer. The sample gives the observed frequency of wet/dry spell of i length (where i goes from 1 to the longest spell). The i length spell can be considered as a casual variable and its probability density can be calculated with theoretical models.

The models that have been used to describe the empirical data are uniform, geometric, logarithmic, negative binomial, Poisson, defined by [Eq. 1-Eq. 5] respectively. Further, following the trend of Berger *et al.* (1983) the spell frequencies have also been analyzed by Eggenberger-Polya distribution [Eq. 6] and Markov chain of order one and two defined by [Eq. 9] and [Eq. 10] respectively.

$$P(X = k) = \frac{1}{b-a}, \qquad a \le k \le b \tag{1}$$

$$P_1(X = k) = p_1.(1-p_1)^{k-1}, \quad 0 < p_1 < 1 \tag{2}$$

$$P_2(X = k) = \frac{-q^k}{x\ln(1-q)}, \quad 0 < q < 1 \tag{3}$$

Table 3. Results of the Kolmogorov-Smirnov (K-S) Tests for Silcoorie (1986-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.3333	1	Discrete Uniform	0.4285
2	Geometric	0.3499	2	Geometric	0.5256
3	Logarithmic	0.3795	3	Logarithmic	0.5334
4	Neg. Binomial	0.4249	4	Neg. Binomial	0.4704
5	Poisson	0.2567	5	Poisson	0.3513
6	M.C of order one	0.0618	6	M.C of order one	0.0763
7	M.C of order two	0.0232	7	M.C of order two	0.0215
8	Eggenberger-Polya	0.0176	8	Eggenberger-Polya	0.0190
Critical value at $\alpha = .05$		0.0597	Critical value at $\alpha = .05$		0.0601

$$P_3(X=k)=\binom{n+k-1}{k}\frac{p^n}{1-p^n}(1-p)^k$$

$$0 < p < 1, \text{ n>0} \qquad (4)$$

$$P_4(X=k)=\frac{e^{-l}\,l^k}{(1-e^{-l})k!},\quad l>0 \qquad (5)$$

where k (=1,2,3,…) is defined as the number of consecutive days of which a spell is composed. The Eggemberger-Polya distribution is:

$$P_5(X=k)=\frac{d^k}{(1+d)^{h/d+k}}\frac{\Gamma(h/d+k)}{k!\Gamma(h/d)} \qquad (6)$$

where Γ is the Gamma function.

It follows from the argument of Giuseppe *et al.* (2005) that the above distribution maintains the following recursive relation:

$$P_5(1)=\frac{1}{(1+d)^{m/d}} \qquad (7)$$

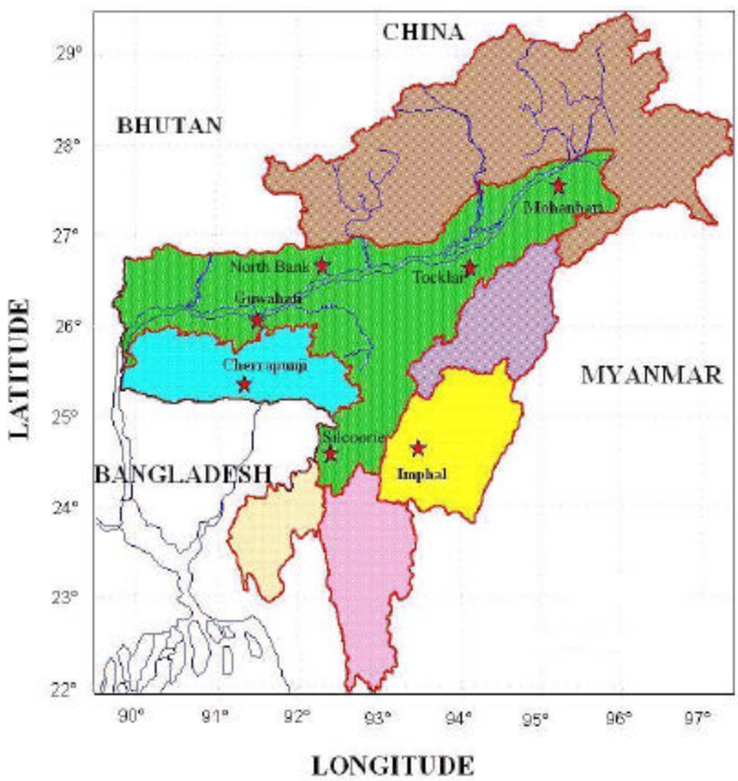

Fig. 1. *Locations of rain gauge stations used in this study.*

Table 4. Results of the Kolmogorov-Smirnov (K-S) Tests for Mohanbari (1993-2006).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.3333	1	Discrete Uniform	0.3571
2	Geometric	0.3951	2	Geometric	0.3922
3	Logarithmic	0.4126	3	Logarithmic	0.4104
4	Neg. Binomial	0.4439	4	Neg. Binomial	0.4343
5	Poisson	0.2237	5	Poisson	0.2292
6	M.C of order one	0.0574	6	M.C of order one	0.0368
7	M.C of order two	0.0287	7	M.C of order two	0.0105
8	Eggenberger-Polya	0.0321	8	Eggenberger-Polya	0.0325
Critical value at $a = .05$		0.0695	Critical value at $a = .05$		0.0694

$$P_5(k) = \frac{m + (k-2)d}{(k-1)(1+d)} P_5(k-1), \quad k \geq 2, \qquad (8)$$

where (m+1) is the mean length of a spell, d is given by $s^2/m - 1$, s^2 being the variance of sequences' length.

In the case of first order Markov chain the probability that a dry spell will last exactly n days is given by

$$Q_n = p_{00}^{n-1}.p_{01} = p_{00}^{n-1}.(1 - p_{00}) \quad \text{for } n \geq 1 \qquad (9)$$

where p_{00} is the probability of a dry day following a dry day and p_{01} the probability of a rainy day following a rainy day. The two parameters p_{01} and p_{11} are required to be estimated for describing the Markov Chain of order one. One can estimate these parameters according to the principle of maximum likelihood estimation. The maximum likelihood estimate of p_{ij} (i,j=0,1) is given by

$$p_{ij} = \frac{n_{ij}}{\sum_{j=0}^{1} n_{ij}} = \frac{n_{ij}}{n_{i.}}, \qquad (10)$$

n_{ij} is the number of direct transition from the state i to the state j.

In the second order Markov chain the probability Q_n is expressed as $Q_n = p_{100} p_{000}^{n-2}.p_{001}$, for $n \geq 2$ (11)

$$Q_1 = p_{101} \qquad (12)$$

and the maximum likelihood estimate of p_{ijk} (i,j,k=0,1) is given by

$$p_{ijk} = \frac{n_{ijk}}{\sum_{k=0}^{1} n_{ijk}} = \frac{n_{ijk}}{n_{ij\cdot}}, \qquad (13)$$

n_{ijk} is the number of transition from the state i to the state k through j. The first order Markov Chain only takes into account the state-wet or dry-of the day preceding a given one. In the same way, the second-order considers the states of the two preceding days. Raising the order of Markov chain does not necessarily do away the imperfections of the model. On the other hand, the number of parameters to estimate increases with 2^k for two state, k order Markov chain which may rapidly enhance the uncertainty of the estimation. Therefore the present study is confined to the Markov chain of order one and two.

The Kolmogorov-Smirnov test for goodness of fit is then employed as the significance test for each model which is one of the most powerful non parametric tests for differences between two cumulative frequency

Table 5. Results of the Kolmogorov-Smirnov (K-S) Tests for Cherrapunji (2001-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.2963	1	Discrete Uniform	0.4000
2	Geometric	0.2365	2	Geometric	0.5643
3	Logarithmic	0.3066	3	Logarithmic	0.5807
4	Neg. Binomial	0.2176	4	Neg. Binomial	0.4495
5	Poisson	0.3510	5	Poisson	0.4220
6	M.C of order one	0.0777	6	M.C of order one	0.0485
7	M.C of order two	0.0582	7	M.C of order two	0.0388
8	Eggenberger-Polya	0.0541	8	Eggenberger-Polya	0.0317
Critical value at $a = .05$		0.1338	Critical value at $a = .05$		0.1338

Table 6. Results of the Kolmogorov-Smirnov (K-S) Tests for Guwahati (2001-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.3750	1	Discrete Uniform	0.3333
2	Geometric	0.4558	2	Geometric	0.5019
3	Logarithmic	0.4631	3	Logarithmic	0.5077
4	Neg. Binomial	0.5086	4	Neg. Binomial	0.4213
5	Poisson	0.2290	5	Poisson	0.3087
6	M.C of order one	0.0399	6	M.C of order one	0.0787
7	M.C of order two	0.0341	7	M.C of order two	0.0112
8	Eggenberger-Polya	0.0382	8	Eggenberger-Polya	0.0396
Critical value at $a = .05$		0.1024	Critical value at $a = .05$		0.1018

Table 7. Results of the Kolmogorov-Smirnov (K-S) Tests for Imphal (2001-2005).

Summer Wet Spells			Summer Dry Spells		
Serial No.	**Distributions**	**K-S Statistic**	**Serial No.**	**Distributions**	**K-S Statistic**
1	Discrete Uniform	0.4167	1	Discrete Uniform	0.3333
2	Geometric	0.4416	2	Geometric	0.4663
3	Logarithmic	0.4505	3	Logarithmic	0.4727
4	Neg. Binomial	0.2187	4	Neg. Binomial	0.6678
5	Poisson	0.2415	5	Poisson	0.2466
6	M.C of order one	0.1056	6	M.C of order one	0.0736
7	M.C of order two	0.0435	7	M.C of order two	0.0307
8	Eggenberger-Polya	0.0491	8	Eggenberger-Polya	0.0152
Critical value at $a = .05$		0.1070	Critical value at $a = .05$		0.1064

distributions of the observed and estimated ones. Massey and Frank (1951) showed that Kolmogorov-Smirnov test treats individual observation separately leading to no loss of information in grouping while loss of information in chi-square procedure is large. Pal (1998) mentioned that the Chi square test's sensitivity to very small cell frequencies make itself unsuitable when expected frequencies work out at less than 5 in 20 percent of the cells. In this study, we have also observed that more than 20% of the cell frequencies are less than 5 and therefore the Kolmogorov-Smirnov test is applied to test the goodness of fit. The test statistics used is

$$D_n = \max\left|S_n(x) - F(x)\right|$$

where $S_n(x)$ and $F(x)$ are empirical and theoretical distribution functions, respectively. The distribution of D_n is independent of $F(x)$. The theoretical distribution function however, has to be completely specified. In this study the theoretical distribution function have been calculated by using the estimated parameters of the distribution in each case. The significance of a critical value of D_n depends on the no. of observations. For example, if n is over 35, the critical values of D at .05 level of significance can be determined by the formula $1.36/\sqrt{n}$. Any D_n equal to or greater than $1.36/\sqrt{n}$ will be significant at .05 levels (two tailed test). In the second phase, the goodness of fit has been tested by Kolmogorov-Smirnov statistics and results are summarized in Table 1 to Table 7.

RESULTS

This section deals with the comparative results obtained from different statistical models applied to analyze the wet and dry spells frequencies over North East India. In the first phase of this work we have calculated the empirical frequencies of wet and dry spells according to their length. Then the same frequencies have been estimated for each station using the aforesaid theoretical distribution models.

Results of Kolmogorov-Smirnov tests presented in the Table1 to Table 7 clearly indicate that apart from the M.C of order two (in some cases order 1 also) and Eggenberger-Polya distribution, the rest of the distributions work poorly to represent the spell frequencies.

In case of dry series, Eggenberger-Polya distribution and Markov Chain of order two shows better results in all seven stations where as Markov chain of order one shows good fit for the stations Mohanbari, Cherrapunji, Guwahati and Imphal. While Eggenberger-Polya distribution gives best fit for the stations North-Bank, Silcoorie, Cherrapunji and Imphal, Markov Chain of order two shows best fit for the stations Tocklai, Mohanbari and Guwahati. Summarizing the above experiences, we may conclude that Eggenberger-Polya distribution and Markov Chain of order two are competing each other in case of dry spells.

In comparison to dry series Markov Chain of order two shows better performance in case of wet series. Results of the Kolmogorov-Smirnov tests for Markov Chain of order one shows good fit to the observed data in most of the investigated cases. Like dry spells, Eggenberger-Polya and Markov Chain of order two are the best fitting models in case of wet spells also. Markov Chain of order two gives best fit to the observed data for four stations and Eggenberger-Polya distribution works better than Markov chain of order two for the rest three stations.

Conclusion

This section concerns with the critical evaluation of the work carried out. These are listed below:

Eggenberger-Polya distribution and Markov Chain of order two (in some cases Markov Chain of order one also) models are efficient in fitting the observed data. The other models do not fit at all.

In case of dry spells (wet spells) Eggenberger-Polya distribution (Markov Chain of order two) shows best fit in four stations out of seven stations.

Markov Chain of order two needs four parameters while Eggenberger-Polya needs only two parameters.

Considering the above discussions it can be concluded that Eggenberger-Polya is better than Markov Chain of order two and can be more easily used as a theoretical model to estimate the seasonal climatic characterization of precipitation over North-East India.

REFERENCES

Berger, A. and Goossens, Chr. (1983). Persistence of wet and dry spells at Uccle. (Belgium). *J. Climatol*., 3: 21-34.

Bruhn, J. A., Fry, W. E. and Fick, G. W. (1980). Simulation of daily weather data using theoretical probability distributions, *J. of Appl. Met.*, 19 :1029-1036.

Dubrovsky, M. (1997). Creating Daily Weather Series With Use of the Weather Generator. *Environmetrics*, 8: 409-424.

Fisher, R. A. (1924). The influence of rainfall on the yield of wheat at Rothamsted. *Phil.Trans. Roy. Stat. Soc*. London. B, 213:89-142.

Gabriel. K. R. and Neumann, J. (1962). A Markov Chain model for daily occurrence at Tel Aviv, *Quart. J.R. met. Soc.*, 88: 90-95.

Geng, S.(1986). A simple method for generating daily rainfall data. *Agricultural and Forest Meteorology*, 36:363-376.

Giuseppe, E. D., Vento, D., Epifani, C. and Esposito, S. (2005). Analysis of dry and wet spells from 1870 to 2000 in four Italian sites, *Geophysical Research Abstracts*, 7: 1-6.

Katz, R. W. (1974). Computing probabilities associated with the Markov chain model for precipitation, *J. Appl. Meteorol*., 13:953-954.

Massey, J. and Frank, J. (1951). The Kolmogorov test for goodness of fit. *JASA*, 46:68-78.

Matyasovszky, I. and Dobi, I. (1989). Methods for analysis of time series of precipitation data using Markov chains (in Hungarian) , *IdoÈjaÂraÂs*, 93: 276-288.

Medhi, J. (1976). A Markov Chain for the occurrence of wet and dry days, *Ind. J. Met. Hydro. & Geophys*. 27: 431-435.

Nobilis, F. (1986). Dry spells in the Alpine country Austria, *J. Hydrol.*, 88: 235-251.

Pal, S. K.(1998). *Statistics for Geoscientists: Techniques and Applications*, Concept Publishing Company, New Delhi.

Racsko, P., Szeidl, L and Semenov, L.(1991). A serial approach to local stochastic weather models. *Ecological Modelling*, 57: 27-41.

Richardson, C. W.(1981). Stochastic simulation of daily precipitation, temperature, and solar radiation, *Water Resources Research*, 17: 182-190.

Tolika, K. and Maheras, P.(2005), Spatial and temporal characteristics of wet spells in Greece, *Theor. Appl. Climatol.* 81: 71–85.

Wilks, D.(1992). Adapting stochastic weather generation algorithm for climate change studies. *Climate Change*, 22: 67-84.

Wantuch, W., Mika, J. and Szeidl, L.(2000). Modelling Wet and Dry Spells with Mixture Distributions, *Meteorology and Atmospheric Physics*, 73: 245-256.