Indian Journal of Radio & Space Physics
Vol. 39, February 2010, pp. 39-44

Higher order Markov chain models for monsoon rainfall over West Bengal, India

Avik Ghosh Dastidar[1], Deepanwita Ghosh[2], S Dasgupta[3] & U K De[1,$,*]

[1]School of Environmental Studies, Jadavpur University, Kolkata 700 032, India
[2]Department of Spectroscopy, Indian Association for the Cultivation of Science, Kolkata 700 032, India
[3]Department of Statistics, St Xavier's College, Kolkata 700 016, India
[$]Email: deutpal2003@yahoo.com

Received 21 October 2008; revised received and accepted 19 November 2009

Two-state Markov chain models of different orders have been used to simulate the pattern of rainfall during the monsoon season (June-September) over Gangetic West Bengal (India). The analysis is based on the relevant data for 31-year period (1970-2000) for four major meteorological stations in the region. The determination of the proper order of the Markov chain that best describes the rainfall pattern is an interesting problem and Bayesian information criterion has been used for the purpose. Bayesian information criterion (BIC) reveals that third order Markov chain model best describes the rainfall pattern in general except for one station. This is verified and it is found that third/fourth order chain simulates the observed data more closely than the chains of other orders using the classical goodness of fit test. The time independent behaviour of the chain has also been studied with the help of steady state probabilities. The theoretical and observed values of the mean recurrence time have been found to be in close agreement.

Keywords*:* Markov chain model, Bayesian information criterion, Rainfall probability, Steady state probability, Mean recurrence time

PACS No: 92.40.*eg*; 02.50.Ga

1 Introduction

A two-state Markov chain is a stochastic model for persistence of binary events. The occurrence / non-occurrence of precipitation on a given day is a simple meteorological example of a random binary event and a sequence of daily observations on precipitation / no precipitation for a particular location constitutes a time series of that variable. The use of a two-state Markov chain to analyze data of this kind dates back as early as 1962 when Gabriel & Neumann[1] fitted a two-state first order Markov chain to the daily rainfall occurrence at Tel Aviv. Several similar analyses have been attempted by many researchers.

Gates & Tong[2] used Akaike's information criterion (AIC)[3] to examine the proper order of the Markov chain that can be used to model weather data. They re-examined Tel Aviv data to show that a Markov chain of order not less than two should be used to model the distribution of dry/wet spells of weather. Azzalini & Bowman[4], Charantois & Liakatas[5] and Sung Euii Moon *et al.*[6] attempted similar probabilistic description of weather events in and outside India using Markov chains. Lana & Burgueno[7] have made use of Markov chains of second order with a large number of states to analyze daily precipitation occurrence at Catalonia in northeast Spain. Pant & Shivhare[8] used a two-state third order chain to analyze monsoon rain in India. Dasgupta & De[9] have shown that pre-monsoon (March-May) thunderstorm phenomenon over Kolkata (India) follows simple probabilistic considerations, being most appropriately described by a two-state Markov chain of the third order. The model also very realistically simulates the fact that a weather spell of shorter duration is more frequent than longer ones. Kulkarni *et al.*[10] have considered Markov chain models for pre-monsoon season thunderstorm over Pune in western India.

Drton *et al.*[11] developed a Markov Chain model of tornadic activity based on data on tornado counts. It has been shown that tornadic activity is affected mostly by the activity on the previous day. Ochola & Kerkides[12] developed a Markov chain simulation model for predicting wet and dry spells in Kenya analyzing events in the Kano plains. Emanuel *et al.*[13] have applied both stochastic and deterministic models for assessment of hurricane

and wind risk. Some more recent works in this field are by Zhao & Chu[14], Briggs & Ruppert[15], and Park *et al.*[16]

The main purpose of the current work is to find the order of two-state Markov chain suitable for describing the monsoon spells over Gangetic West Bengal. An extensive survey of the existing literature shows that the determination of the proper order of the Markov chain that can model the distribution of weather spells adequately is an interesting and important problem. Spoof & Pryor[17] have used the Bayesian information criterion (BIC) in the choice of the optimal model. Since the BIC did not adequately represent the distribution of the weather spells for all stations, they advocated the use of multiple criteria in the selection of the optimal model using BIC initially for parsimony. The present authors have also used BIC for determination of the optimal model order for simulating the distribution of weather spells in Gangetic West Bengal (India), but no second criterion has been implemented owing to the satisfactory results obtained from BIC. As a test for model skill, the classical goodness of fit test has been carried out.

2 Data and Methodology

The present work is based on data related to monsoon (June-September) rainfall over four major stations in Gangetic West Bengal, India, viz. Alipore (22.53°N, 88.33°E), Dum Dum (22.65°N, 88.45°E), Purulia (23.18°N, 86.22°E), and Midnapore (22.25°N, 87.65°E) for 31 years (1970-2000). The missing data are randomly distributed in terms of dry and wet days in the time series.

The original data which gives the daily precipitation at the stations over the said period has been transformed to a sequence of daily observations of the random, binary events: precipitation / no precipitation.

With the k^{th} day, a random variable X_k is associated such that

$$X_k = \begin{cases} \mathbf{1} \text{ if precipitation occurs on the } k^{th} \text{ day} \\ \mathbf{0} \text{ if precipitation does not occur on the } k^{th} \text{ day} \end{cases} \quad \ldots(1)$$

where, k = 1, 2,…etc

The data for each season of a particular year are considered as a separate sample of the above time series.

A two-state Markov chain is natural for dichotomous data, since each of the two states will pertain to one of the possible data values.

Initially, the process may be in any of the two states, which is referred to as the initial state of the chain. Thereafter, the process may either remain in the same state or move to the other state. In the second case, one can say that a transition has occurred from one state to the other. The probability of such a transition is known as the transition probability. A Markov chain is thus characterized by a set of such transition probabilities.

Let $\{X_t, t \varepsilon T\}$ be a Markov chain with index set T and state space S. In particular, if S = {0, 1}, then $\{X_t, t \varepsilon T\}$ is said to be a two-state Markov chain. The most common form of a two-state Markov chain is that of first order, where

$$P_r(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1 \ldots\ldots X_n = i_n) = \Pr(X_{n+1} = j \mid X_n = i_n) \text{ for all } (i_0, i_{1,\ldots\ldots} i_n) \varepsilon S \quad \ldots(2)$$

The two-state Markov chains of second, third and fourth orders satisfy the following conditions:

$$P_r(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1 \ldots\ldots X_n = i_n) = \Pr(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}) \text{ for all } (i_0, i_1, \ldots\ldots i_n) \varepsilon S \quad \ldots(3)$$

$$P_r(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1 \ldots\ldots X_n = i_n) = \Pr(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}) \text{ for all } (i_0, i_{1,\ldots\ldots\ldots} i_n) \varepsilon S \quad \ldots(4)$$

$$P_r(X_{n+1} = j \mid X_0 = i_0, X_1 = i_1 \ldots\ldots X_n = i_n) = \Pr(X_{n+1} = j \mid X_n = i_n, X_{n-1} = i_{n-1}, X_{n-2} = i_{n-2}, X_{n-3} = i_{n-3}) \text{ for all } (i_0, i_{1,\ldots\ldots\ldots} i_n) \varepsilon S \quad \ldots(5)$$

A two-state Markov chain of any order is completely determined by its initial state and a set of transition probabilities, p_{ij}, p_{ijk}, p_{ijkl} and p_{ijklm}. The transition probabilities are estimated using conditional relative frequencies.

The parameter estimates of Markov chain of different orders are obtained separately for each of the sample. The overall estimates are then obtained by averaging the estimates from all the samples. The suitability of Markov chains of different orders for modeling the given time series is examined using Bayesian information criterion[18]. According to this criterion, an 'S' state Markov chain of order 'm' is the most appropriate model, if it minimizes the function:

$$BIC(m) = -2L_m + s^m \cdot (\ln n) \quad \ldots(6)$$

where,

$$L_0 = \sum_{j=0}^{S-1} n_j \ln(p_j),$$

$$L_1 = \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} n_{ij} \ln(p_{ij}),$$

$$L_2 = \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} \sum_{k=0}^{S-1} n_{ijk} \ln(p_{ijk}),$$

$$L_3 = \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} \sum_{k=0}^{S-1} \sum_{l=0}^{S-1} n_{ijkl} \ln(p_{ijkl}),$$

$$L_4 = \sum_{i=0}^{S-1} \sum_{j=0}^{S-1} \sum_{k=0}^{S-1} \sum_{l=0}^{S-1} \sum_{m=0}^{S-1} n_{ijklm} \ln(p_{ijklm}).$$

Qualitatively, BIC gives a mathematical formulation of the principle of parsimony in model building. Quantitatively for eight or more observations, BIC leans towards lower dimension models. But for large number of observations, BIC perhaps selects the optimal model because of certain asymptotic properties which are yet to be proved for other available criteria like the AIC[18].

A comparison is made between the observed and expected number of wet and dry spells of different orders at the four stations for the arbitrarily chosen years 2001-2004 using the classical goodness of fit test. The data for the years 2001-2004 were not included in the development of the stochastic model but were reserved for cross validation of the results obtained. This appears more objective than the usual auto-verification scheme adopted by most authors.

In line with the approaches of some earlier works using Markov chains of first order, the n-step transition probabilities of the chain are, in fact, the elements of matrix P^n, where P is the one-step transition matrix of the chain. It is observed that after four to five steps, these probabilities become constant (up to five places after decimal) and hence become independent of the initial state. These steady state probabilities are:

π_0 = (steady state probability of "no precipitation")

π_1 = (steady state probability of "precipitation")

By using an alternative computational formula, based on the theory of conditional probability, it is found, from the Markov chain of first order:

$$\pi_1 = \frac{p_{01}}{1 + p_{01} - p_{11}} \quad \ldots (7)$$

$$\pi_0 = 1 - \pi_1 \quad \ldots (8)$$

Similar formulae for Markov chain of higher orders have also been derived. The expression for the second order chain is

$$\pi_0 = \frac{(p_{10}p_{100} + p_{11}p_{110})}{(1 - p_{00}p_{000} + p_{10}p_{100} - p_{01}p_{010} + p_{11}p_{110})} \quad \ldots (9)$$

$$\pi_1 = 1 - \pi_0 \quad \ldots (10)$$

The expressions for the chain of third and higher orders are not being presented because of their complicated nature.

3 Results and Discussions

Computation of the transition probabilities of the Markov chain of first, second, third and fourth orders reveal that the probability of a dry spell is maximum in all cases. In other words, the transition probabilities P_{00}, P_{000}, P_{0000} and P_{00000} are highest among all other transition probabilities. Having obtained the transition probabilities of the chains of different orders and taking into account the various transition counts, it is seen that a two-state third order chain minimizes the function in BIC in case of Midnapore, Purulia and Dum Dum stations, whereas the two-state fourth order chain minimizes the BIC in Alipore (Table 1).

Using the transition probabilities of the chains of different orders, the expected number of "wet" (precipitation) and "dry" (no precipitation) spells of various lengths have been calculated for the years 2001, 2002, 2003 and 2004. The closeness of the observed and expected values for these years have been judged using the classical goodness of fit test (Table 2). The test is accepted in 23 out of 24 available cases.

The steady state probabilities as obtained from the n-step transition matrix of the chain are the same as calculated from the first order Markov chain using the computational formula (Table 3). The mean recurrence times for "wet" and "dry" days have been calculated using the reciprocal of the steady state probabilities. The mean recurrence time for the wet days is obtained as 2.0346, 1.9818, 2.1580, and

2.2065 for Alipore, Dum Dum, Midnapore, and Purulia, respectively and that for dry days as 1.9666, 2.0186, 1.8636, and 1.8288 for Alipore, Dum Dum, Midnapore and Purulia, respectively. These mean recurrence times are then compared with the observed mean recurrence times for the test data [Tables 4(a) and (b)].

The steady state probabilities as derived using notion of conditional probabilities for the Markov chain of higher orders is found to yield values quite close to that obtained from both the N-step transition matrix as well as first order chain.

The observed and model derived counts for different length of dry and wet spells in the year 2003

Table 1 — BIC scores of the model of different order at four locations

	Alipore		Dum Dum		Midnapore		Purulia	
Order (i)	**L_i**	**BIC_i**	**L_i**	**BIC_i**	**L_i**	**BIC_i**	**L_i**	**BIC_i**
1	-2448.679	4906.967	-2517.083	5043.773	-2439.151	4887.910	-2491.015	4991.638
2	-2425.048	4869.312	-2492.347	5003.911	-2418.303	4855.822	-2467.244	4953.703
3	-2403.727	4845.885	-2470.475	4979.383	-2398.418	4835.269	-2445.629	4929.690
4	-2384.113	4845.091	-2465.094	5007.052	-2388.016	4852.896	-2432.232	4941.328
5	-2471.727	5097.182	-2552.174	5258.076	-2481.482	5116.694	-2545.718	5245.165
6	-2925.598	6158.653	-2999.737	6306.932	-2940.383	6188.224	-3008.717	6324.891
7	-3942.754	8500.423	-4028.700	8672.315	-3977.301	8569.516	-3988.431	8591.776
8	-5346.703	11923.235	-5476.769	12183.367	-5396.168	12022.166	-5427.770	12085.370
9	-6917.185	16294.029	-7091.861	16643.380	-6929.071	16317.801	-7004.057	16467.772
10	-8322.082	21563.481	-8613.897	22147.111	-8349.431	21618.179	-8417.442	21754.202

Table 2 — Goodness of fit test for observed and expected count (as obtained from the optimal model) of dry and wet spell for the test data

	Alipore		Dum Dum		Midnapore		Purulia	
Year	**Dry spell**	**Wet spell**	**Dry spell**	**Wet spell**	**Dry spell**	**Wet spell**	**Dry spell**	**Wet spell**
2001	2.0410	19.5905	5.3981	6.0130	4.8350	0.6749	2.6028	6.1877
2002	3.4731	1.8314	1.7097	5.5596	7.4075	5.5536	-	-
2003	5.2222	9.0409	3.6966	5.8705	-	-	-	-
2004	3.4500	4.2869	9.7894	3.7044	5.0263	6.8331	-	-

*Upper 5% value of chi-square distribution with 6 degrees of freedom is 12.592.

Table 3 —Comparison of stationary probabilities obtained from N-step transition matrix and computational formula based on a first order chain

	Stationary probability of no rainfall		Stationary probability of rainfall	
Station	**From matrix**	**From 1st order**	**From matrix**	**From 1st order**
Alipore	0.5085	0.5085	0.5085	0.4915
Dum Dum	0.4954	0.4954	0.5046	0.5046
Midnapore	0.5366	0.5366	0.4634	0.4634
Purulia	0.5468	0.5468	0.4532	0.4532

Table 4(a) — Mean recurrence time for dry days at different stations

	From Markov model		Observed mean recurrence time			
Station	**Stationary probability**	**Mean recurrence time**	**2001**	**2002**	**2003**	**2004**
Alipore	0.5085	1.9666	1.3077	0.8906	0.9516	1.0167
Dum Dum	0.4954	2.0186	1.3529	1.0000	0.9516	1.1607
Midnapore	0.5366	1.8636	1.1250	0.9180	—	0.7647
Purulia	0.5468	1.8288	1.1228	—	—	—

Table 4(b) — Mean recurrence time for wet days at different stations

	From Markov model		Observed mean recurrence time			
Station	**Stationary probability**	**Mean recurrence time**	**2001**	**2002**	**2003**	**2004**
Alipore	0.4915	2.0346	1.6618	2.0714	1.9655	1.9667
Dum Dum	0.5046	1.9818	1.5942	2.0000	2.0000	1.8281
Midnapore	0.4634	2.1580	1.8906	2.0169	—	2.3077
Purulia	0.4532	2.2065	1.7143	—	—	—

are presented in Figs 1[(a) and (b)], respectively for the station Dum Dum. One can conclude that the match, in general, is satisfactory. This conclusion holds for four years, i.e. 2001-2004, and four stations used for cross validation.

4 Conclusions

The daily rainfall in the monsoon season over major stations in Gangetic West Bengal is seen to follow simple, probabilistic considerations being most appropriately described by a two-state Markov chain of third or fourth order. The said models also very realistically simulate the fact that any weather spell (wet or dry) of shorter length is generally more frequent and longer spells being rare occurrences.

The overall climatological probability of precipitation on a given day in the season as obtained from the first order chain is seen to be

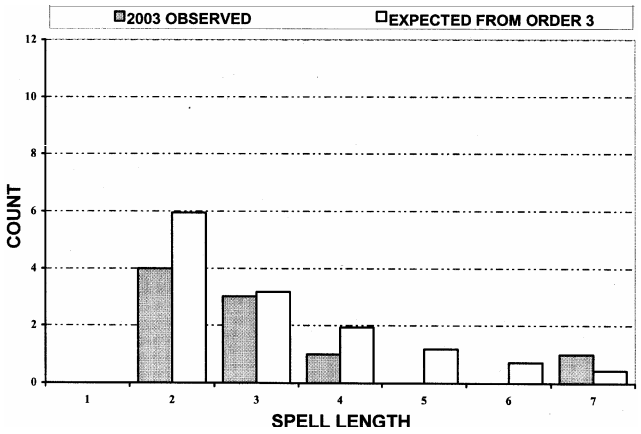

Fig. 1(a) — Observed and expected (from Markov chain of order 3) count of dry spells in Dum Dum for 2003

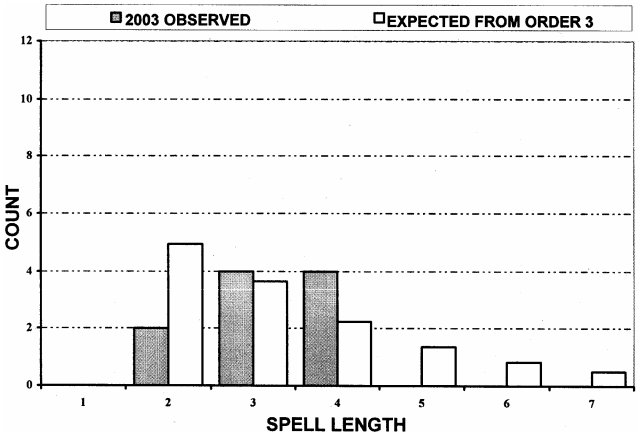

Fig. 1(b) — Observed and expected (from Markov chain of order 3) count of wet spells in Dum Dum for 2003

0.5085, 0.4954, 0.5366, and 0.5468 for Alipore, Dum Dum, Midnapore and Purulia, respectively.

Incidentally, the climatological probabilities as obtained from the chain of higher orders are almost identical to the value obtained from the first order chain. This indicates that the steady state probabilities are not only independent of the initial state of the chain but also of the order which is a significant finding in respect of the general behaviour of the Markov chain.

Interestingly, the first order Markov chain does satisfactorily describe the process of de-correlation over time as is revealed by the realistically close values of the observed and theoretical mean recurrence times for the test data.

Acknowledgements

The authors thank Department of Science & Technology, Govt. of India for the sanction of a research project. The present work is a part of the project.

References

1 Gabriel K R & Neumann J, A Markov chain model for daily rainfall occurrence at Tel-Aviv, *Q J R Meteorol Soc (UK)*, 8 (1962) pp 85-90.

2 Gates P & Tong H, On Markov chain modeling to some weather data, J *Appl Meteorol (USA)*, 15 (1976) pp 1145–1151.

3 Akaike H, A new look at the statistical model identification, *IEEE Trans Autom Control (USA),* 19 (1974) pp 716-723.

4 Azzalini A & Bowman A W, A look at some data on the old fatigue geyser, *Appl Stat (UK)*, 39 (1990) pp 357-365.

5 Charantois T & Liakatas A, Study of minimum temperatures employing Markov chain, *Mausam (India)*, 41 (1990) pp 69- 80.

6 Sung E M, Sang B R & Jai-gi-kuon, A Markov chain model for daily precipitation occurrence in South Korea, *Int J Climatol (UK)*, 14 (1994) 1009.

7 Lana X & Burgueno P, Daily dry-wet behaviour in Catalonia (NE Spain) from the viewpoint of Markov chains, *Int J Climatol (UK)*, 18 (1998) pp 793–815.

8 Pant B C & Shivhare R P, Markov chain model for study of wet/dry spells at AF station Sarsawa, *Vatavaran (India)*, 22 (1998) pp 37- 50.

9 Dasgupta S & De U K, Markov chain model for pre-monsoon thunderstorm in Calcutta, India, *Indian J Radio Space Phys*, 30 (2000) pp 138-142.

10 Kulkarni M K, Kandalgaonka S S, Tinmaker M R & Nath A, Markov chain models for pre-monsoon season thunderstorms over Pune, *Int J Climatol (UK)*, 22 (2002) pp 1415-1420.

11 Drton M, Marzban C & Guttorp P, A Markov chain model of tornadic activity, *Mon Weather Rev (USA)*, 131 (2003) pp 2941–2953.

12 Ochola W O & Kerkides P, A Markov chain simulation model for predicting critical wet and dry spells in Kenya: Analysing rainfall events in the Kano plains, *Irrig Drain (UK)*, 52 (2003) pp 327-342.

13 Emanuel K, Ravela S, Vivant E & Risi C, A statistical deterministic approach to Hurricane risk assessment, *Bull Am Meteorol Soc (USA)*, 87 (2006) pp 299–314.

14 Zhao X & Chu Pao-Shin, Bayesian multiple changepoint analysis of Hurricane activity in the eastern North Pacific: A Markov chain Monte Carlo approach, *J Clim* (*USA*), 19 (2006) pp 564–578.

15 Briggs W & Ruppert W, Assessing the skill of yes/no forecasts for Markov observations, *Mon Weather Rev (USA)*, 134 (2006) pp 2601–2611.

16 Park E, Elfeki A M M, Song Y & Kim K, Generalized coupled Markov chain model for characterizing categorical variables in soil mapping, *Soil Sci Soc Am J (USA)*, 71 (2007) pp 909-917.

17 Spoof J T & Pryor S C, On the proper order of Markov chain model for daily precipitation occurrence in the contiguous United States, *J Appl Meteorol (USA),* 47 (2008) pp 2477-2486.

18 Schwarz G, Estimating the dimension of a model, *Ann Stat (USA)*, 6 (1978) pp 461-464.