

MODIFIED MARKOV PROBABILITY MODELS OF SEQUENCES OF PRECIPITATION EVENTS¹

E. H. WISLER

North Carolina State University, Raleigh, N.C.

ABSTRACT

A number of modifications to the Markov chain probability model are proposed for cases in which the simple model does not fit sequences of wet or dry days. The modified models are shown to fit the observed records in most cases, and can also be applied to sequences of wet or dry hours. A solution to the problem of counting sequences within a limited time period is also given.

1. INTRODUCTION

Several authors have found that sequences in daily rainfall occurrences could be described by a simple Markov chain model. The first explicit application of the model seems to have been made by Gabriel and Neumann [8] in a study of sequences of data at Tel Aviv. More recently, Caskey [3] and Weiss [13] have been successful in describing sequences at a number of different locations, indicating that application of the model might be quite general.

There are, however, several sets of data which have been reported, which are not described properly by the simple Markov chain model. Among these may be mentioned several of the results given by Newnham [12] for the British Isles, sequences of wet days at Montsouris given by Besson [2], sequences of dry days at San Francisco given by Jorgensen [10], the results cited as fitting the logarithmic distribution given by Williams [14] for Harpenden and by Cooke [5] for Moncton, and the results cited as fitting a higher order Markov chain given by Feyerherm and Bark [6] in the Midwestern States. Green [9] stated that the fit of the simple model was unsatisfactory for several of the cases cited by Weiss [13]. A comparison of the observed counts with those calculated according to the simple Markov chain model shows that there were observed more very short, fewer intermediate, and more very long sequences than would be expected.

The consistency of the manner in which the discrepancies occur is indicative that there may be a more general probability model, of which the Markov chain model is a special case, which will describe in a more suitable way the behavior of wet and dry sequences. Several such models have been developed and will be presented herein with some comments on the estimation of parameters and on the problem of counting sequences within a limited time period.

2. ELEMENTARY URN MODELS

In order to clarify the discussion that follows, it may be worthwhile to describe three elementary urn models which may be used to choose between two alternatives in a specified manner.

The Bernoulli urn model consists of a single urn containing (say) red and black balls. A ball is drawn at random from the urn, its color is noted, and it is then replaced in the urn. In a sequence of such draws, the probability of drawing a red ball is a constant independent of whatever event occurred at any previous draw. This urn model matches the case of independence which Jorgensen [10] tried to use in fitting his San Francisco data.

The Polya urn model also consists of a single urn containing red and black balls. After a ball is drawn at random from the urn and its color noted, it is replaced in the urn, together with D balls of the same color. In a sequence of such draws, the probability of drawing a red ball changes after each draw and is related to the number of red balls drawn previously in the sequence.

The simple Markov urn model consists of three urns containing red and black balls. From one urn (the "First" urn) only the first draw is made. If the ball drawn is red, the next draw is made from the "Red" urn. If the ball drawn is black, the next draw is made from the "Black" urn. After any draw, the ball is replaced in the urn from which it was drawn. Thus, each of the three urns acts as a Bernoulli urn and the probability of drawing a red ball from any of the urns is constant. Since each urn may have a different proportion of red balls, however, the probability of drawing a red ball at any draw may depend on the outcome of the previous draw.

These three urn models illustrate three types of dependence of an event on previous events. The Bernoulli urn model, representing the case of independence, has been applied (Beer et al. [1]) to sequences of months of above and below average rainfall. The model is sufficient for cases of no persistence. When persistence occurs, the

¹ Contribution from the Agricultural Engineering Department, North Carolina Agricultural Experiment Station, Raleigh, N.C. Published with the approval of the Director of Research as Paper No. 1961 of the Journal Series.

simple Markov urn model is satisfactory provided that the persistence is limited to the effect of one time interval on the next.

There may also be cases in which persistence extends over several time intervals. Such was true in the data reported by Newnham [12] where the probability of a rainy day was influenced by the number of consecutive rainy days preceding it. This type of contagious behavior might be expected to be matched by the Polya urn model. However, the data show a marked change when a sequence of dry days is broken by a single wet one. The probability is clearly influenced more by the previous outcome than by the entire previous history of outcomes.

3. MODIFIED MARKOV PROBABILITY MODELS

Four modified urn models have been developed with the desired characteristic of reducing to the simple Markov model as a special case, but in a more general manner being able to account for the contagious cases. Each of these models uses the three urns of the simple Markov model in the usual manner. The modifications arise by allowing the content of an urn to vary in a specified way during a sequence of draws of balls of the same color from the urn. When the sequence is terminated by drawing a ball of the opposite color, the content of the urn is restored to its initial state.

In order to simplify our description, we consider only the case of black runs. A black run of length m consists of one black ball drawn from either the "First" or "Red" urns, and $(m-1)$ black balls drawn from the "Black" urn. If the run is exactly of length m , we must further specify the drawing of one red ball from the "Black" urn following the $(m-1)$ black balls. We are thus led naturally to the definitions

p_m = the conditional probability that a run of black draws will last for *exactly* m draws, given that a black ball has been drawn.

$F(m)$ = the conditional probability that a run of black draws will last for *at least* m draws, given that a black ball has been drawn.

$$\bar{m} = p_m + p_{m+1} + p_{m+2} + \dots$$

\bar{m} = the mean length of black run.

Since draws from the "Black" urn *must* be preceded by the draw of a black ball from one of the other urns, we see for example that the conditional probabilities $F(1)$, $F(2)$, $F(3)$, . . . are associated respectively with the events of drawing 0, 1, 2, . . . black balls from the "Black" urn, nothing being said about succeeding draws. The event having probability $F(1)$ is always satisfied, so that

$$F(1) = 1.$$

Since the probabilities as defined are affected only by properties of the "Black" urn, we need direct our attention only to this urn. Suppose in its initial state it contains N balls, of which R are red and S are black.

Then for the simple Markov chain model

$$p_m = xy^{m-1}$$

$$F(m) = y^{m-1}$$

$$\bar{m} = 1/x$$

where

$$y = S/N$$

$$x = R/N = 1 - y.$$

In the First Modification, an urn acts as a Polya urn, D balls being added after each draw. In this case

$$p_m = x \frac{(b)_{m-1}}{(c+1)_{m-1}}$$

$$F(m) = \frac{(b)_{m-1}}{(c)_{m-1}}$$

$$\bar{m} = 1 + (y/x - z)2 \quad x > z$$

$$= \infty \quad z \geq x$$

where

$$z = D/N$$

$$b = y/z$$

$$c = 1/z$$

$$(b)_r = b(b+1)(b+2) \dots (b+r-1)$$

$$(b)_0 = 1.$$

In the Second Modification, an urn acts as a Friedman urn. This is a generalized Polya urn proposed by Friedman [7] in which D balls of the same color and H balls of the opposite color are added after each draw. In this case

$$p_m = [x + (m-1)v] \frac{(b)_{m-1}}{(f+1)_{m-1}} g^{m-1}$$

$$F(m) = \frac{(b)_{m-1}}{(f)_{m-1}} g^{m-1}$$

$$\bar{m} = F(b, 1; f; g) \quad v > 0$$

where

$$v = H/N$$

$$f = 1/(z+v)$$

$$g = z/(z+v).$$

In the Third Modification, an urn acts as a Polya urn, D balls being added after each draw, for only a specified number of draws w . Thereafter the urn acts as a Bernoulli urn, no balls being added. In this case

$$p_m = x \frac{(b)_{m-1}}{(c+1)_{m-1}} \quad m \leq w+1$$

$$= p_{w+1} \left(\frac{b+w}{c+w} \right)^{m-w-1} \quad m > w+1$$

$$F(m) = \frac{(b)_{m-1}}{(c)_{m-1}} \quad m \leq w+1$$

$$= F(w+1) \left(\frac{b+w}{c+w} \right)^{m-w-1} \quad m > w+1$$

$$m = \frac{(b+w)(c+w)}{a^2(1-a)} p_{w+1} + \frac{(1-c)}{(1-a)}$$

where

$$a = x/z.$$

In the Fourth Modification, an urn acts as a Polya urn, but the contagion parameter D varies in some manner, converging toward zero for a large number of draws. Suppose that after each successive draw, the total number of balls that have been added equals the contagion parameter D times the successive terms in the sequence

$$A = [A_1, A_2, A_3, A_4, \dots]$$

Then

$$p_m = x \frac{b(b+A_1)(b+A_2) \dots (b+A_{m-2})}{(c+A_1)(c+A_2) \dots (c+A_{m-1})}$$

$$F(m) = \frac{b(b+A_1)(b+A_2) \dots (b+A_{m-2})}{c(c+A_1)(c+A_2) \dots (c+A_{m-2})}$$

A general solution for the average length of run has not been found.

The expected number of black runs of length m for any of these models is

$$M(m) = T_s p_m$$

where T_s is the total number of black runs.

Analogous equations apply for red runs.

4. THE EXPECTED NUMBER OF RUNS IN A SPECIFIED NUMBER OF DRAWS

Equations derived in the previous section for the expected number of runs are valid for what may be termed a normally terminated series, i.e., the number of complete runs in the data are counted, a run never being truncated. In some cases, it may be desirable to count runs in a rigidly specified time period such as a given month. This is equivalent to specifying the number of draws from an urn model and may be termed an artificially terminated series. More short runs and fewer long runs will occur than in the normally terminated series.

Suppose that a series of exactly n draws is to be made from some urn model containing only red and black balls. A general solution has been obtained for the expected number of runs of specified length under the sole restriction that the probability of obtaining such a run is dependent only on its length and is independent of preceding runs or of its location in the series. The solution thus applies to all models previously referred to except for the Polya model.

Let the probabilities that the first ball drawn will be black or red be y_1 and x_1 respectively, and let $M(m)$ be the expected number of black runs of length m . Then

$$M(n) = y_1 F(n)$$

$$M(n-1) = y_1 p_{n-1} + x_1 X_0 F(n-1)$$

$$M(n-2) = y_1 p_{n-2} + (y_1 Z_1 + x_1 X_1) F(n-2) - x_1 X_0 F(n-1)$$

$$M(n-i) = y_1 p_{n-i} + (y_1 Z_{i-1} + x_1 X_{i-1}) F(n-i) - (y_1 Z_{i-2} + x_1 X_{i-2}) F(n-i+1) \quad 2 < i < n$$

where

$$X_i = \sum_{j=1}^{i+1} q_j + \sum_{j=1}^{i-1} q_j Z_{i-j}$$

$$Y_i = \sum_{j=1}^{i-1} p_j q_{i-j}$$

$$Z_i = \sum_{j=1}^i Y_{j+1} + \sum_{j=1}^{i-2} Y_{j+1} Z_{i-j-1}.$$

The q_j are defined in the same way for red runs as the p_m are defined for black runs, and together with the $F(m)$ can be calculated directly for any specific model.

The expected total number of black runs

$$T_s = y_1 + y_1 Z_{n-2} + x_1 X_{n-2}.$$

The expected number of black balls drawn

$$B = y_1 [F^* + Z^*] + x_1 X^*$$

where

$$F^* = \sum_{m=1}^n F(m)$$

$$X^* = \sum_{m=1}^{n-1} X_{n-1-m} F(m)$$

$$Z^* = \sum_{m=1}^{n-2} Z_{n-1-m} F(m).$$

Analysis shows that the initial probabilities y_1 and x_1 must satisfy certain restrictions. The restriction on y_1 is satisfied by the particular value

$$y_1 = B/n,$$

which is the expected average probability of drawing a black ball.

The special cases of the Bernoulli and Markov urn models can be solved readily. For the Bernoulli model let y and x be the respective probabilities of drawing a black or red ball. Then

$$X_r = y(1+rx)$$

$$Z_r = rxy$$

$$y_1 = y,$$

and

$$M(r) = 2xy^r + x^2 y^r (n-r-1) \quad r < n$$

$$= y^r \quad r = n$$

$$= 0 \quad r > n$$

$$T_s = y + (n-1)xy.$$

These results were given by Cochran [4] and Mahalanobis [11].

For the Markov chain model let y_R be the probability of drawing a black ball from the red urn, x_S the probability of drawing a red ball from the black urn,

$$y_S = 1 - x_S$$

and

$$h = y_S - y_R.$$

Then

$$X_r = (r+1)y_R - \frac{y_R^2}{1-h} \left[r - h \frac{(1-h^r)}{(1-h)} \right]$$

$$Z_r = \frac{x_S y_R}{1-h} \left[r - h \frac{(1-h^r)}{(1-h)} \right]$$

$$y_1 = \frac{y_R}{x_S - y_R},$$

and

$$M(r) = 2 \frac{x_S y_R}{x_S + y_R} y_S^{r-1} + \frac{x_S^2 y_R}{x_S + y_R} y_S^{r-1} (n-r-1) \quad r < n$$

$$= \frac{y_R}{x_S + y_R} y_S^{r-1} \quad r = n$$

$$= 0 \quad r > n$$

$$T_S = \frac{y_R}{x_S + y_R} [1 + (n-1)x_S].$$

5. APPLICATIONS

The data for dry runs at San Francisco, given by Jorgensen [10] have already been cited as one instance for which the simple Markov model did not provide a satisfactory fit. In figure 1 the data have been plotted together with expected values according to five theoretical models. The length of run is plotted against the logarithm of the number of times that the given length was equaled or exceeded. It can easily be shown that expected values for both the Bernoulli and Markov models will plot as straight lines, for a normally terminated series. For an artificially terminated series, the plots will exhibit a slight downward curvature, but in most practical uses this is negligible and may be ignored.

The data for San Francisco illustrate the upward curvature that is characteristic of situations in which persistence extends over a prolonged period. The straight lines predicted by the Bernoulli and Markov models are clearly unsatisfactory. On the other hand, the First Modification with its assumption of unlimited persistence seems to have overcompensated. The Third Modification, with persistence limited in this case to five days, seems satisfactory. The Second Modification with decreasing persistence seems equally satisfactory.

Each of these models has a certain number of parameters which must be estimated from the data, and the more general the model, the more parameters there are to be estimated. At this time there is not sufficient information to

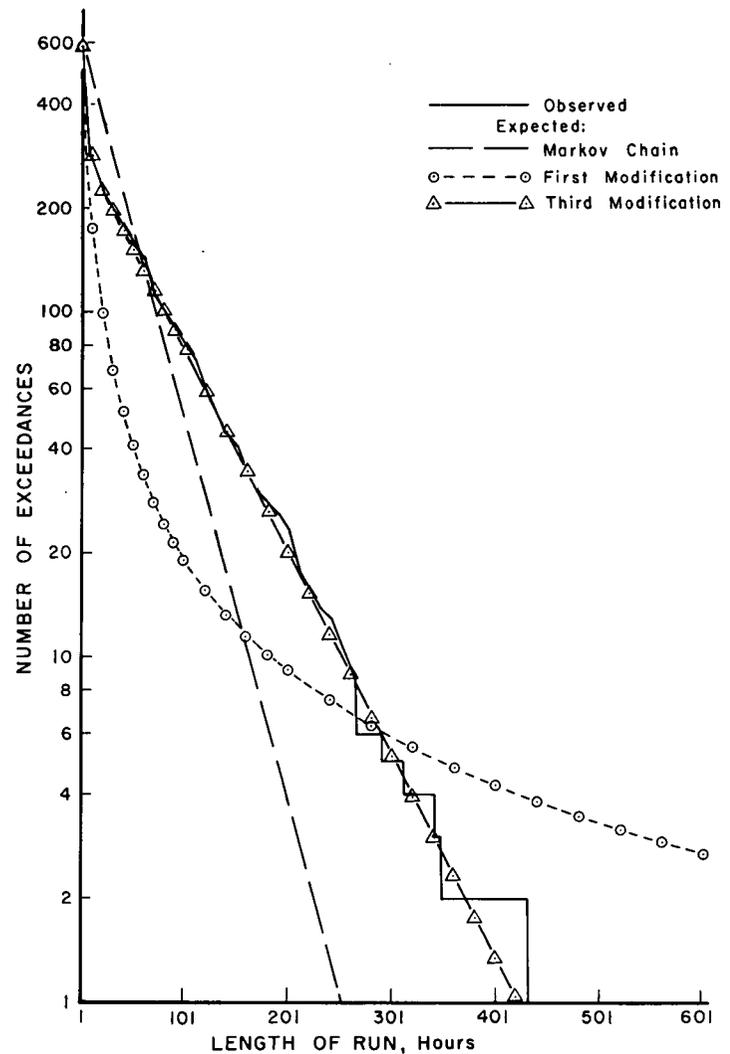


FIGURE 1.—The number of times a given length of run of dry days was equaled or exceeded at San Francisco, observed values given by Jorgensen [10] being compared with expected values for several probability models.

determine which estimators are most efficient, and the choice is made either for other reasons or for ease of calculation. For example, the purpose for which these models was developed required that the relative number of wet events should be controlled. This can be done by estimating the average length of run from the data, so that all models are fitted with this restriction.

The total number of runs may also be estimated from the data, a restriction which is usually easy to apply. These two restrictions are sufficient to fit the parameters of the simple Markov model. For the modified models additional parameters must be estimated. The estimator of the number of runs of length one has been used in the present example, this being sufficient to fit the parameters of the First Modification. In addition, the estimator of the number of runs of length two has been used to fit the Second Modification, and the slope of the straight line

section has been used to fit the Third Modification. No use of the Fourth Modification has been made at this time because the residual disagreement between the expected and observed values for the other models is not of a type that could be reduced by application of this model.

Another example is given in figure 2 in which several models are fitted to length of runs of dry hours in April for the 35-yr. period 1929-63 at Greensboro, N.C. The data were obtained as an artificially terminated series of exactly 720 hr., a period which is sufficiently long that equations derived for a normally terminated series are sufficiently accurate. The shape of the curve of the data is rather striking, and shows a major reason for the development of the Third Modification.

In addition to the results given in the figures, several other sets of data have been fitted to some of the models, these tests generally being restricted to cases for which the Markov model had proved unsatisfactory. While these results are too voluminous to report here, the Chi-square values from goodness of fit tests are given in table 1 with associated significance levels.

TABLE 1.—Summary of goodness-of-fit tests of several probability models, giving the Chi-square value and associated significance level.

Location	Type	Markov	Modifications		
			First	Second	Third
Aberdeen[12].....	Wet.....	22.6 .05	20.8 .06		19.0 .06
	Dry.....	25.2 .001	4.3 .74		
Kew[12].....	Wet.....	24.6 .003	8.8 .36		
	Dry.....	80.8 <.00001	11.8 .29		
Valentia[12].....	Wet.....	48.9 .0003	28.1 .08		8.4 .97
	Dry.....	60.5 <.00001	10.6 .16		10.4 .11
Greenwich[12].....	Wet.....	19.4 .05	7.7 .65		
	Dry.....	196.8 <.00001	36.5 .0005		20.1 .06
Montsouris[2].....	Wet.....	41.6 .0004	8.1 .92		
San Francisco[10].....	Wet.....	6.6 .58	4.5 .72	3.5 .74	
	Dry.....	40.6 .0004	25.5 .03	15.7 .26	13.2 .43
Harpenden[14].....	Wet.....	35.1 .0001	8.1 .52		
	Dry.....	77.6 <.00001	20.3 .03		16.8 .05
Moncton (Fall)[5].....	Wet.....	22.1 .0005	6.7 .15		
Greensboro.....	Wet.....	40.9 .00003	10.2 .42		9.0 .43
	Dry.....	456.5 <.00001	456.4 <.00001		9.5 .30

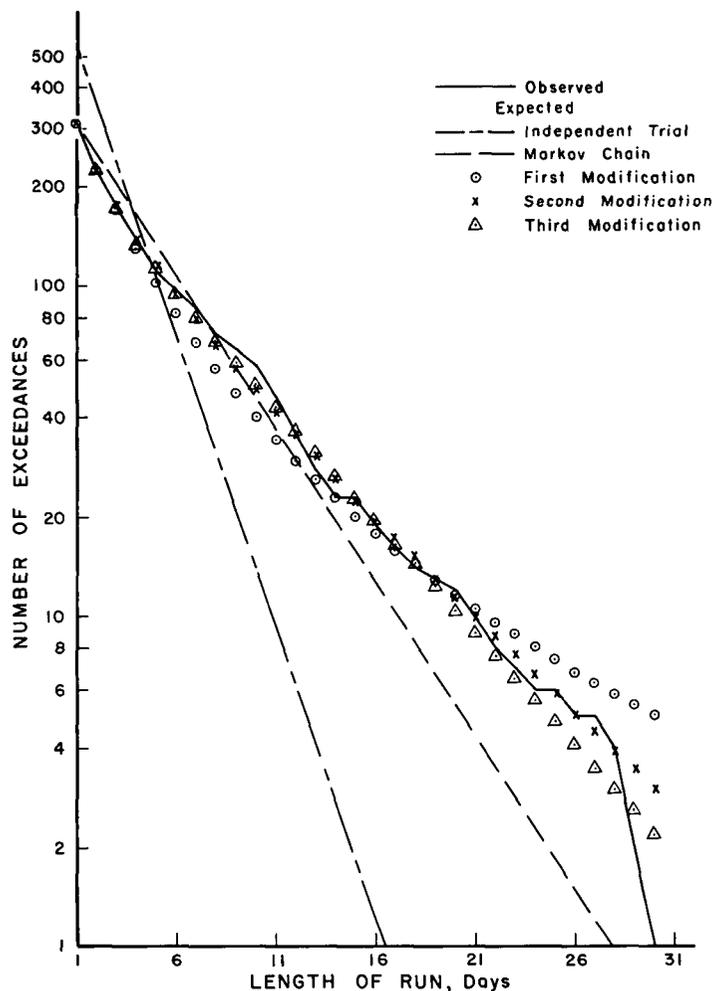


FIGURE 2.—The number of times a given length of run of dry hours was equaled or exceeded at Greensboro, N.C., observed values for April 1929-63 being compared with expected values for several probability models.

It may be seen that the data may be fitted in a satisfactory manner using one of the modified models. We cannot say from this that one of the models is superior to any other. The Second Modification has not been tested sufficiently, largely because the mean length of run is in the form of a hypergeometric function which is not readily computed. It may also be pointed out that the major part of the data curve which seems to be fitted more satisfactorily by the Third than by the First Modification is for the longer runs, for which there are insufficient data to test by the goodness of fit test except as a group. Therefore, although a plot of the data often showed the Third Modification to be superior, the difference did not appear in the Chi-square values. The tendency of the First Modification to overestimate the likelihood of long runs is emphasized by the fact that, for certain relevant parameter sets, the expected mean length of run is infinite.

We conclude that the Second and Third Modifications

are superior to the First, the only reason for using the First being its ease of application. We cannot choose between the Second and Third—the Second seems more rational, while the Third is easier to apply. The Fourth Modification does not appear to offer enough improvement to be worth the extra effort.

6. SUMMARY

Several probability models have been developed to fit sequences of wet or dry periods of various lengths; these models were obtained as generalizations of the simple Markov chain model which has often been used for this purpose. Strictly speaking, any of these models can be matched by higher-order Markov chain models, and parameters could be fitted by a procedure such as that given by Feyerherm and Bark [6]. What the modified models do, in effect, is to specify relations between the parameters of the higher-order chains, thus reducing the number of parameters that must be estimated.

Equations for the expected number of runs of any length are presented. These are available both for a normally terminated and for an artificially terminated series. Some remarks are made concerning estimation of the parameters necessary for fitting the models, and applications are demonstrated for runs of dry days at San Francisco and for runs of dry hours at Greensboro.

The modified models are clearly superior to the simple Markov model in some cases which reflect conditions when persistence plays a more important role than that assumed by the simple Markov model. While these cases seem relatively limited for daily intervals, they are probably general for shorter time intervals such as hours. Although various of the modified models may be superior in different cases, it appears that the Second and Third Modifications may be the most satisfactory for additional applications.

REFERENCES

1. A. Beer, A. J. Drummond and R. Fürth, "Sequences of Wet and Dry Months and the Theory of Probability," *Quarterly Journal of the Royal Meteorological Society*, vol. 72, 1946, pp. 74–86.
2. L. Besson, "Sur la probabilité de la Pluie," *Comptes Rendus*, t. 178, (19 Mai) 1924, pp. 1743–1745.
3. J. E. Caskey, Jr., "A Markov Chain Model for the Probability of Precipitation Occurrence in Intervals of Various Length," *Monthly Weather Review*, vol. 91, No. 6, June 1963, pp. 298–301.
4. W. G. Cochran, "The Statistical Analysis of Field Counts of Diseased Plants," *Journal of the Royal Statistical Society Supplement*, vol. 3, 1936, pp. 49–67.
5. D. S. Cooke, "The Duration of Wet and Dry Spells at Moncton, New Brunswick," *Quarterly Journal of the Royal Meteorological Society*, vol. 79, No. 342, Oct. 1953, pp. 536–538.
6. A. M. Feyerherm and L. D. Bark, "Statistical Methods for Persistent Precipitation Patterns," Sixth National Conference on Agricultural Meteorology, Lincoln, Nebr., Oct. 1964.
7. B. Friedman, "A Simple Urn Model," *Communications on Pure and Applied Mathematics*, vol. 2, 1949, pp. 59–70.
8. K. R. Gabriel and J. Neumann, "A Markov Chain Model for Daily Rainfall Occurrence at Tel Aviv," *Quarterly Journal of the Royal Meteorological Society*, vol. 88, No. 375, Jan. 1962, pp. 90–95.
9. J. R. Green, "Two Probability Models for Sequences of Wet or Dry Days," *Monthly Weather Review*, vol. 93, No. 3, Mar. 1965, pp. 155–156.
10. D. L. Jorgensen, "Persistence of Rain and No-Rain Periods During the Winter at San Francisco," *Monthly Weather Review*, vol. 77, No. 9, Nov. 1949, pp. 303–307.
11. P. C. Mahalanobis, "On Large-Scale Sample Surveys," *Philosophical Transactions of the Royal Society of London, Series B*, vol. 231, 1944, pp. 329–451.
12. E. V. Newnham, "The Persistence of Wet and Dry Weather," *Quarterly Journal of the Royal Meteorological Society*, vol. 42, No. 179, July 1916, pp. 153–162.
13. L. L. Weiss, "Sequences of Wet or Dry Days Described by a Markov Chain Probability Model," *Monthly Weather Review*, vol. 92, No. 4, Apr. 1964, pp. 169–176.
14. C. B. Williams, "Sequences of Wet and Dry Days Considered in Relation to the Logarithmic Series," *Quarterly Journal of the Royal Meteorological Society*, vol. 78, No. 335, Jan. 1952, pp. 91–96.

[Received April 12, 1965; revised June 7, 1965]