Journal of Hydrology, 103 (1988) 127–137
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

[1]

THE THEORY OF RUNS WITH APPLICATIONS TO DROUGHT PREDICTION

LEMUEL A. MOYÉ, ASHA S. KAPADIA, IRINA M. CECH and ROBERT J. HARDY
The University of Texas School of Public Health, P.O. Box 20186, Houston, TX 77225 (U.S.A.)

(Received August 7, 1987; revised and accepted December 29, 1987)

ABSTRACT

Moyé, L.A., Kapadia, A.S., Cech, I.M. and Hardy, R.J., 1988. The theory of runs with applications to drought prediction. J. Hydrol., 103: 127–137.

Although statistical run theory is a useful tool, its application to the prediction of drought likelihood has met with limited success. Beginning from the rudiments of run theory, we have developed a pertinent probability distribution based on difference equations. This distribution allows the hydrologist to estimate the expected number of droughts of a prespecified duration, and the average drought length over the desired time period. The applicability of this new mathematical approach is demonstrated using precipitation records for different climatic regions of Texas.

INTRODUCTION

There has been interest in approaching the drought prediction problem as though droughts are entirely random occurrences. Yevjevich (1967) suggested that some elements of run theory may be applied to the estimation of drought likelihood, and his was among the first attempts at blending probabilistic aspects of run theory with the prediction of drought likelihood.

Review of previous work (Yevjevich, 1967; Saldariaga and Yevjevich, 1970) reveals that the estimation of drought likelihood remains pertinent, but the definitive mathematical model for its description has not yet been identified. Run theory holds promise, but the form of the theoretical results for the prediction of future runs (Cramer and Leadbetter, 1967; Schwager, 1983) require information that hydrologists are not likely to have. Thus the nature of its incorporation into drought likelihood models has been restricted, leading to models of limited usefulness in the prediction of drought likelihood (Yevjevich, 1972). We believe a further enhancement of statistical run theory is required to more fully employ its benefits in the context of drought likelihood assessment. Such an enhancement is presented here, and its applicability to drought likelihood estimation examined.

0022-1694/88/$03.50

METHODS

Definitions

Bernoulli trials play an important role in statistical theory. Consider a sequence of experiments in which only one of two possible outcomes (one considered a success, the other a failure) may occur. The probability that a success occurs in the experiment is p, and the probability of failure is $q = 1 - p$. If these probabilities remain the same from experiment to experiment, and knowledge of the result of a previous experiment gives no insight into the result of any succeeding experiment, then each experiment is a Bernoulli trial.

In any given year, water resource requirements in a given region are either met with the probability p or not met with the probability q, where $p + q = 1$. Therefore, if knowledge of a previous year's result (i.e., whether water requirements were met or not) do not aid in determining a succeeding year's result, and the probability of meeting water requirements remains unchanged across years, this sequence of years may be considered a sequence of Bernoulli trials.

In addition, if in k consecutive years, the water requirements are not met, then it is observed that a drought of length κ has occurred. A drought may be defined as being a run of consecutive years in which water requirements have not been met, i.e., a drought is a failure run. Furthermore, one may specify a minimum drought length of interest κ_d, focusing consideration on only failure runs of at least a given length κ_d. Therefore, any advances in run theory, in particular, any advances in the ability to predict the occurrence of runs of failures in a sequence of Bernoulli trials yet to be observed, may be applied to the prediction of drought likelihood.

If κ is the a-priori specified run length of interest and n the number of trials, then $R_{i,\kappa}(n)$ = Probability (there are exactly i failure runs of length κ in the next sequence of n trials, and all other run lengths are possible).

If i is the number of failure runs of length κ in the next sequence of n trials, then the range of i is finite, ranging from $i = 0$ to $i = I$ where $I = [(n + 1)/(\kappa + 1)]$, [ ] denoting the greatest integer function. Thus:

$$\sum_{i=0}^{I} R_{i,\kappa}(n) = 1$$

Using this information:

$$\varepsilon_{r,\kappa} = \sum_{i=0}^{I} iR_{i,\kappa}(n) = \text{the expected number of runs of length } \kappa$$

It follows that:

$$v_{r,\kappa} = \text{variance of the number of runs of length } \kappa \text{ in } n \text{ trials}$$

$$= \sum_{i=0}^{I} i^2 R_{i,\kappa}(n) - \varepsilon_{r,\kappa}^2$$

Furthermore:

$$\sum_{\kappa=1}^{n} \varepsilon_{r,\kappa}$$

is the expected number of droughts.

Another computation of interest is:

$$L(n) = \frac{\sum_{\kappa=1}^{n} \kappa \varepsilon_{r,\kappa}}{\sum_{\kappa=1}^{n} \varepsilon_{r,\kappa}}$$

which is an estimate of the average run length in the sequence.

Recalling that a drought may be defined as a run, we may apply these theoretical results to drought prediction. Note that:

$$L_{\kappa_d}(n) = \frac{\sum_{\kappa=\kappa_d}^{n} \kappa \varepsilon_{r,\kappa}}{\sum_{\kappa=\kappa_d}^{n} \varepsilon_{r,\kappa}}$$

is the average drought length in the next n years.

From the knowledge of this distribution, the capability now exists to compute (a) the expected number of droughts in the next n years, and (b) the expected duration of a drought over the next n years. Thus, by identifying $R_{i,\kappa}(n)$, the hydrologist will have access to quantities of importance in predicting drought likelihood. The goal of the remainder of this communication is to develop this distribution and to demonstrate the application of this development to the problem of drought prediction.

Theoretical development

As a first step, $R_{i,\kappa}(n)$ is derived. It is clear that:

$$R_{0,\kappa}(n) = 1, \quad \text{for } n < \kappa$$

$$R_{0,\kappa}(k) = 1 - q^{\kappa}$$

The difference equation for $R_{0,\kappa}(n)$ can be written as: Prob (there are no runs of length κ in n trials) =

$$\begin{aligned} R_{0,\kappa}(n) &= pR_{0,\kappa}(n-1) + qpR_{0,\kappa}(n-2) + q^2pR_{0,\kappa}(n-3) + q^3pR_{0,\kappa}(n-4) \\ &\quad + \ldots + q^{\kappa-1}pR_{0,\kappa}(n-\kappa) + q^{\kappa+1}pR_{0,\kappa}(n-\kappa-1) \\ &\quad + \ldots + q^{n-1}pR_{0,\kappa}(0) \\ &= \sum_{i=0}^{\kappa-1} pq^i R_{0,\kappa}(n-i-1) + \sum_{i=\kappa+1}^{n-1} pq^i R_{0,\kappa}(n-i-1) \end{aligned}$$

for $\kappa \leqq n < \infty$

The generating function approach (Goldberg, 1958) is then used to solve this equation resulting in:

$$R_{0,\kappa}(n) = 1, \quad \text{if } n < \kappa$$

$$R_{0,\kappa}(n) = 1 - q^{\kappa}, \quad \text{for } n = \kappa$$

$$\begin{aligned} R_{0,\kappa}(n) = {} & q^{\kappa} D(n-\kappa) + \sum_{h=\kappa}^{\min(n,2\kappa-1)} q^{h-\kappa}(1-q^{2\kappa-h})D(n-h) \\ & + \sum_{h=\min(n,\kappa+2)}^{\min(2\kappa+1,n)} q^{\kappa+1}(1-q^{n-\kappa-1})D(n-h) \\ & + \sum_{h=\min(n,2\kappa+2)}^{n} q^{h-\kappa}[1-q^{n-(h-\kappa)}]D(n-h) \\ & - q\left\{\sum_{h=\kappa}^{\min(n-1,2\kappa-1)} q^{h-\kappa}(1-q^{2\kappa-h})D(n-h-1)\right. \\ & + \sum_{h=\min(n,\kappa+2)}^{\min(2\kappa+1,\,n-1)} q^{\kappa+1}(1-q^{n-\kappa-1})D(n-h-1) \\ & \left. + \sum_{h=\min(n,2\kappa+2)}^{n-1} q^{h-\kappa}[1-q^{n-(h-\kappa)}]\,D(n-h-1)\right\}, \quad \text{for } n > \kappa \end{aligned}$$

where:

$$\begin{aligned} D(x) = {} & \sum_{m=0}^{[x/\kappa]} \sum_{h=0}^{[x/\kappa]-m} \binom{x-m\kappa-m+h}{m}\binom{m}{h} \\ & \times (-1)^{m+(m-h)(\kappa+1)} p^{m} q^{m(\kappa+1)-h} I(x-m\kappa-m+h \geqq m) \end{aligned}$$

Proceeding analogously, the difference equations and solutions are obtained for $R_{1,\kappa}(n)$, and $R_{i,\kappa}(n)$.

RESULTS

The derivation of $R_{i,\kappa}(n)$ allows, for a given q, κ and n the computation of the exact probabilities of the occurrence of runs of length κ. The applicability of these computations was tested using Texas precipitation records.

Rainfall data were obtained from the climatological records for Texas (Texas Almanac, 1986–1987). Data were available for the 93 years (1892–1984) for ten climatological regions of this state (Fig. 1). The Texas Almanac defined a drought as a period when annual precipitation was less than 75% of the thirty-year normal (1931–1960 period). Using this definition, the new method was tested in different regions.

Tables 1a, 1b and Tables 2a, 2b illustrate these tests for two contrasting climatological regions, Upper Coast (average annual precipitation equal

TABLE 1A

Application of model to recorded precipitation history, Upper Coast Climatic Division, Texas, 1892–1984

Year	Percent of normal rainfall	Year	Percent of normal rainfall	Year	Percent of normal rainfall
1892	73	1923		1954	57
1893	64	1924		1955	
1894		1925		1956	62
1895		1926		1957	
1896		1927	74	1958	
1897		1928		1959	
1898		1929		1960	
1899		1930		1961	
1900		1931		1962	
1901	70	1932		1963	73
1902		1933		1964	
1903		1934		1965	
1904		1935		1966	
1905		1936		1967	
1906		1937		1968	
1907		1938		1969	
1908		1939		1970	
1909		1940		1971	
1910	74	1941		1972	
1911		1942		1973	
1912		1943		1974	
1913		1944		1975	
1914		1945		1976	
1915		1946		1977	
1916		1947		1978	
1917		1948	67	1979	
1918		1949		1980	
1919		1950	68	1981	
1920		1951		1982	
1921		1952		1983	
1922		1953		1984	

TABLE 1B

Comparison of observed and expected droughts of various lengths

Length of drought (yr)	Observed droughts	Expected droughts
1	8	9.13
2	2	1.17
3	0	0.15
4	0	0.02
5	0	0.00

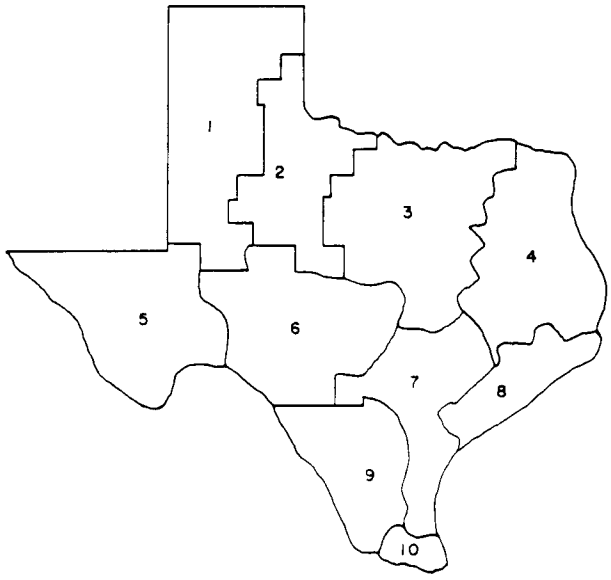

Fig. 1. Climatologic regions of Texas. *1* = High Plains; *2* = Low Rolling Plains; *3* = North Central; *4* = East Texas; *5* = Trans Pecos; *6* = Edwards Plateau; *7* = South Central; *8* = Upper Coast; *9* = Southern; *10* = Lower Valley.

1173 mm) and Southern (average annual precipitation equal 558 mm). Tables 1a and 2a show years during which annual precipitation was less than 75% of the normal. Since the average precipitation varies from region to region, the model's estimate of drought likelihood for a specific region uses information concerning average precipitation for that region alone.

In order to calculate the expected drought frequencies one must: (1) estimate q; (2) determine n, κ; (3) compute the probabilities for the distribution of interest; and (4) compute the expected number of runs for each value of κ.

For each climatic region, a value for q, the probability of the occurrence of insufficient rainfall in any given year is required. In the approach suggested by Yevjevich (1967), this value might be obtained by dividing the total number of years in the past when precipitation was less than 75% of the normal by the duration of the period of record. For Upper Coast and Southern regions, these estimates of q were, respectively, 0.13 and 0.14.

For different values of $\kappa = 1$ to 4, $R_{i,\kappa}(n)$ was computed. From these probabilities, the expected number of runs of length κ were computed using a computer program written with the Microsoft QuickBasic Compiler 3.0, and are reported in Table 1b and Table 2b. These tables report the expected and observed frequencies of droughts of various lengths, from one to five years, during the period from 1892 to 1984.

It is important to consider alternative drought definitions to investigate the model's robustness. In this regard, a drought may be alternatively defined as

TABLE 2A

Application of model to recorded precipitation history, Southern Climatic Division, Texas, 1892–1984

Year	Percent of normal rainfall	Year	Percent of normal rainfall	Year	Percent of normal rainfall
1892		1923		1954	71
1893	53	1924		1955	
1894		1925		1956	53
1895		1926		1957	
1896		1927		1958	
1897	72	1928		1959	
1898	69	1929		1960	
1899		1930		1961	
1900		1931		1962	67
1901	44	1932		1963	
1902	65	1933		1964	
1903		1934		1965	
1904		1935		1966	
1905		1936		1967	
1906		1937		1968	
1907		1938		1969	
1908		1939		1970	
1909		1940		1971	
1910	59	1941		1972	
1911		1942		1973	
1912		1943		1974	
1913		1944		1975	
1914		1945		1976	
1915		1946		1977	
1916		1947		1978	
1917	32	1948		1979	
1918		1949		1980	
1919		1950	74	1981	
1920		1951		1982	
1921		1952	56	1983	
1922		1953		1984	

TABLE 2B

Comparison of observed and expected droughts of various lengths

Length of drought (yr)	Observed droughts	Expected droughts
1	8	9.65
2	2	1.34
3	0	0.18
4	0	0.03
5	0	0.00

having a minimum length of two years. This application is illustrated using the entire state of Texas annual average precipitation for the same period 1892–1984. Each entry in the Table 3a represents the average rainfall in Texas for a given year. From these data the model's ability to predict the frequencies of droughts of various lengths may be examined (Table 3b).

In Table 3c we change the period for which q and the mean precipitation was computed to the first 30 years of the record. The projections of drought likelihood are then obtained for the remaining 69 years. Thus, the model's ability to project forward, making prospective drought likelihood assessments could be evaluated.

TABLE 3A

Average Texas rainfall, 1892–1984

Year	cm	Year	cm	Year	cm
1892	26.32	1923	37.24	1954	19.30
1893	18.50	1924	22.32	1955	23.59
1894	25.61	1925	25.37	1956	16.17
1895	29.83	1926	32.97	1957	36.93
1896	25.15	1927	24.32	1958	32.71
1897	24.21	1928	27.56	1959	31.29
1898	24.56	1929	29.47	1960	33.78
1899	27.57	1930	28.44	1961	30.20
1900	36.87	1931	28.37	1962	24.50
1901	20.13	1932	32.76	1963	20.95
1902	28.28	1933	26.15	1964	24.11
1903	29.64	1934	25.59	1965	28.68
1904	26.78	1935	35.80	1966	28.68
1905	35.98	1936	30.32	1967	28.44
1906	29.19	1937	25.89	1968	34.54
1907	28.51	1938	25.25	1969	29.85
1908	29.06	1939	23.52	1970	26.36
1909	21.58	1940	32.70	1971	29.58
1910	19.52	1941	42.62	1972	27.73
1911	26.83	1942	30.68	1973	38.37
1912	24.92	1943	24.28	1974	32.78
1913	33.25	1944	34.80	1975	28.70
1914	35.19	1945	30.60	1976	33.37
1915	28.79	1946	35.16	1977	24.40
1916	23.05	1947	24.75	1978	27.00
1917	14.30	1948	21.79	1979	31.43
1918	26.02	1949	35.80	1980	24.49
1919	42.15	1950	24.48	1981	32.65
1920	29.90	1951	21.99	1982	26.97
1921	25.18	1952	23.27	1983	25.75
1922	29.83	1953	24.76	1984	26.80

TABLE 3B

Model application to Texas statewide data

Length of drought (yr)	Observed droughts	Expected droughts
2	5	4.95
3	2	1.51
4	0	0.50
5	0	0.16
6	0	0.05
7	1	0.02
8	0	0.01
9	0	0.00
10	0	0.00

Drought definition: any year with less than 90% of the 1931–1960 statewide average; $q = 0.34$.

TABLE 3C

Prospective drought predictions

Length of drought (yr)	Observed droughts	Expected droughts
2	4	4.03
3	2	2.14
4	1	1.14
5	0	0.61
6	0	0.32
7	1	0.17
8	0	0.09
9	0	0.05
10	0	0.03

Drought definition: at least two consecutive years with less than normal (1892–1922) state average; $q = 0.525$.

DISCUSSION

Modern societies are less inclined to accept the conventional risks of drought, making the best possible estimate of their likelihood imperative. The estimation of drought likelihood will continue to occupy the attention of hydrologists and statisticians (Yevjevich, 1967). With this in mind, we studied the options that statistical run theory has to offer the practicing hydrologist. A review of run theory revealed that there have been many examinations of the issue of the future occurrence of runs in a sequence of Bernoulli trials. Mood (1940) and Schwager (1983) have developed expressions which theoretically predict future run behavior. However these results are not left in the form most

suitable to hydrologists, i.e., the offered solutions are functions of parameters to which hydrologists have no access. The same is true of the major contributions to crossing theory made by Cramer and Leadbetter (1967). Although these results on "exceedence measures" and "length of upward excursions" are precise, they offer only a theoretical solution to the problem of run prediction. Therefore, hydrologists formulate the problem of drought likelihood in manners similar to Mood or Cramer (Yevjevich, 1967), but are unable to use the advanced mathematical solutions offered by these theorists because of the difficulty in translating the abstract terms in which these solutions are formulated to a specific hydrologic context. A hydrologist attempting to implement these findings must estimate parameters such as "the probability of no upcrossings in the time interval 0 to infinity", clearly an endeavor full of risk and inaccuracy.

It therefore comes as no surprise that a review of the drought literature revealed that the application of run theory to drought prediction is restricted. Run theory has been identified as useful to hydrologists in the assessment of drought likelihood (Yevjevich, 1967, 1972). However, results of importance to hydrologists based on the established theory of runs rest of either simplistic probability models or require familiarity with asymptotic behavior of run length (Yevjevich, 1967).

The hydrologist has access to q (the probability of inadequate annual rainfall in the past), n (the number of years for which the predictions are to be made), and κ (the defined drought length). It is our contention that this information is sufficient to obtain accurate estimates of drought likelihood. The approach offered here permits this and is therefore of interest from a number of perspectives. First, it offers an exact solution for the prediction of the future occurrence of runs of an arbitrary length in a sequence of Bernoulli trials, with no approximations required. Second, the solution is in terms of parameters with which the hydrologist is familiar and has direct access. Third, the outcome measures of this model (expected number of droughts of an arbitrary length and average drought duration) are measures of hydrologic importance. Thus, starting from estimates of familiar parameters and using the presented model, the practicing hydrologist will gain a pertinent, accurate asessment of the likelihood of droughts in the region of interest.

The Bernoulli model represented by $R_{i,\kappa}(n)$ works in predicting droughts. Using alternative drought definitions, we observed that the model performed reasonably well in various climatologic regions of Texas.

These results are encouraging, but it must be emphasized that more theoretical development is required along the following lines. First, a rigorous goodness of fit test must be developed. Such a test is not presently available. The traditional chi-square test (Bickel, 1977) will not be entirely appropriate since it assumes that all expectations are derived from the same probability distribution. In fact, the expected frequency of each drought length has its own distribution.

Secondly, it was assumed that the value of q is region-specific and remains

constant over time. Initial observations suggest that the model is not overly sensitive to variations in q. Nevertheless, additional estimation techniques for q might need to be developed.

In addition, the underlying assumption of the Bernoulli model must be examined. The Bernoulli model assumes independence of rainfall over consecutive years. This assumption may (Friedman, 1957) or may not (Tannehill, 1947) be valid. The ability of this model to accurately assess drought likelihood prospectively (using the first thirty years of Texas data to assess drought likelihood for the remaining sixty-nine years) assumes no variability in q. There is no significant evidence of serial correlation in the data set ($\rho_1 = 0.03$). However, it is appropriate to develop a general model which would allow for such persistence. Efforts to develop this option suggest that this more sophisticated alternative would use as the probability of a run of length κ q_1 q_2 q_3 $q_4 \ldots q_\kappa$ and not q^κ. The ultimate model would incorporate these second or higher order Markovian dependencies into the statistical run theory.

The development presented in this paper is a first step to extend a potentially useful theory to a concept fitted for hydrological needs. We continue to work on the development of this model along the above outlined avenues. The limitations notwithstanding, the present developments offer an encouraging picture for the prospect of drought prediction.

REFERENCES

Bickel, P.J. and Doksum, K.A., 1977. Mathematical Statistics Basic Ideas and Selected Topics. Holden-Day, San Francisco, Calif., pp. 312–332.

Chiang, C.L., 1968. An Introduction to Stochastic Processes and their Application. Kreiger, New York, N.Y.

Cramer, H.L. and Leadbetter, M.R., 1967. Stationary and Related Stochastic Processes. Wiley, New York, N.Y.

Texas Almanac, 1986–1987. Sesquecentenial Edition. Dallas Morning News, Dallas, Tex.

Friedman, D.C., 1957. The prediction of long-continuing drought in South and Southwest Texas, 1957. The Travelers Insurance Company, Weather Res. Center, Occasional Res. Pap. Meteorol. No. 1, Hartford, Conn.

Goldberg, S., 1958. Introduction to Difference Equations. Wiley, New York, N.Y.

Mood, A.M., 1940. The distribution theory of runs. Ann. Math. Stat., 11: 367–392.

Saldariaga, J. and Yevjevich, V., 1970. Application of run-lengths to hydrologic series. Hydrol. Pap., Colorado State University, Fort Collins, Colo.

Schwager, S.J., 1983. Run probabilities in sequences of Markov-dependent trials. J. Am. Stat. Assoc., 78: 168–175.

Tannehill, I.R., 1947. Drought, its Causes and Effects. Princeton University Press, Princeton, N.J.

Yevjevich, V., 1967. An objective approach to definitions and investigations of continental hydrologic droughts. Hydrol. Pap., Colorado State University, Fort Collins, Colo.

Yevjevich, V., 1972. Stochastic Processes in Hydrology. Water Resour. Publ., Fort Collins, Colo.