Stochastic Hydrology and Hydraulics 9, 207-213 © Springer-Verlag 1995

Trend analysis: binary-valued and point cases

D. R. Brillinger
Dept. of Statistics, University of California, Berkeley, CA 94720, USA

Abstract: A sequence of occurrence times of floods may be considered to be part of a realization of a binary-valued time series or of a stochastic point process. In this paper a criterion for detecting the presence of a monotonic trend in the rate of the process is considered. The criterion is based on linear functions of the data with the coefficients chosen to emphasize a monotonic rate. In the case that the process is stationary and mixing, the null distribution of the test statistic is approximately standard normal.

Key words: Abelson-Tukey coefficients, binary time series, detection, floods, monotone trend, point process, Rio Negro, time series, trend.

1 Introduction

Floods are an important hazard. There is concern with description of their character and possible changes of behavior. This paper studies data consisting of the years of occurrence of floods of the Rio Negro River at Manaus in Brazil. In this case there is the possibility that upriver deforestation is leading to an increase in the number of floods, see Sternberg (1987). Data on the height of the Rio Negro have been recorded at Manaus, daily since 1903. Following Sternberg, a flood is defined as the river level exceeding 28.5m sometime in the year. This allows a record to be constructed for the years 1903 on. Early floods of 1892, 1895, 1898 were recorded by Le Cointe, see Sternberg. This allows the series to be extended backwards somewhat and thus there is a motivation to reduce the data from 1903 on to simply whether or not flooding occurred in a given year. The years of floods are found to be: 1892, 1895, 1898, 1904, 1908, 1909, 1913, 1918, 1920, 1921, 1922, 1944, 1953, 1955, 1971, 1972, 1973, 1975, 1976, 1982, 1989. These values allow construction of a 0-1 time series with Y_t taking the value 1 if there is a flood in year 1892 + t and the value 0 otherwise for, t = 0, ..., 100. Figure 1 graphs the cumulative count of floods and the corresponding flood years. The dashed line corresponds to a constant rate of flooding. The step function fluctuates away from the line and a question is whether the departure is significant.

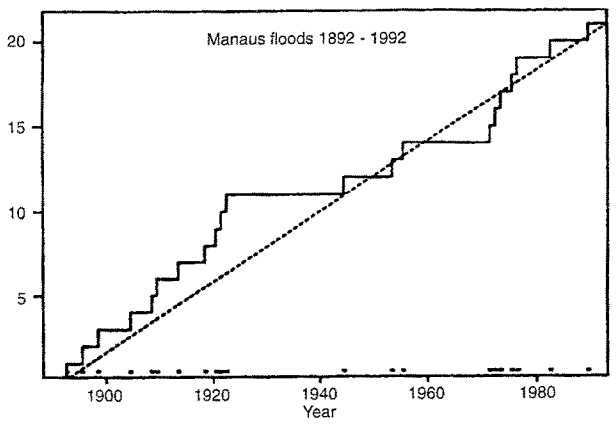

Figure 1. Flood years and cumulative count of years with floods

The paper has sections as follows: trend detection for time series, the binary case, results for the Rio Negro data, an extension to the point process case, discussion and finally an Appendix indicating the derivation of the results.

2 Techniques for detecting trend

In this work a trend will be taken to refer to the monotonic increase (or decrease) of some mean level or rate. In the case of independent observations a variety of procedures have been set down for testing for the presence of trends of other types. The papers Cuzick (1988), Lee (1988), Margolin (1988) may be mentioned. For the time series case there are: Cox (1981), Dagum and Dagum (1988), Harvey (1989).

Consider the problem of detecting the presence of a monotonic trend when a time series has the form

$$Y_t = S_t + \mathcal{E}_t \tag{1}$$

and data are available for t = 0, 1, 2, ... Here $S_t = E\{Y_t\}$, the mean level, is supposed either constant or monotonic in t and $\mathcal{E}$ is a stationary noise process with mean 0 and power spectrum $f_2(\lambda)$.

A naive statistic on which, to base a trend detection procedure is $\sum(t-\bar{t})Y_t$, see Lee (1988). This statistic can be anticipated to be particularly effective in the presence of a linear trend. Abelson and Tukey (1963) considered instead a linear statistic $\sum c_t Y_t$, involving a set of coefficients $\{c_t\}$, summing to 0 which maximized the minimum correlation coefficient squared between $\{c_t\}$ and all $\{S_t\}$ subject to $S_0 \leq S_1 \leq S_2 \leq \ldots \leq S_{T-1}$. The values found were

$$c_t = \left\{t\left(1 - \frac{t}{T}\right)\right\}^{1/2} - \left\{(t+1)\left(1 - \frac{t+1}{T}\right)\right\}^{1/2} \tag{2}$$

for t = 0, 1, ..., T-1. Schaafsma and Smid (1966) found the same coefficients in deriving a most stringent somewhere most powerful test.

To deal with the serial dependence, generally present in time series data, Brillinger (1989) considered the detection statistic

$$\sum c_t Y_t / \left\{ 2\pi \hat{f}_2(0) \sum c_t^2 \right\}^{1/2} \tag{3}$$

with $\hat{f}_2(0)$ an estimate of $f_2(0)$. Because $\sum c_t = 0$ the numerator of (3) has expected value 0 for constant mean level. Following the Abelson and Tukey (1963) result, the numerator is larger if the S_t are monotonic increasing. The denominator of (3) is an estimate of the standard error of the numerator. For stationary mixing $\mathcal{E}$ the statistic (3) was found to be approximately standard normal and the test procedure consistent. The value $\hat{f}_2(0)$ could be obtained by smoothing the Y values to estimate S_t and then basing a spectrum estimate on the residuals, $Y_t - \hat{S}_t$.

3 The binary time series case

Consider now a series representing the years of floods, specifically a 0-1 time series Y_t. Binary series are discussed in Cox (1970) for example. The model (1) with the series $\mathcal{E}$ stationary, is unreasonable, for suppose that

$$\text{Prob}\{Y_t = 1\} = \pi_t$$

then $\text{Prob}\{Y_t = 0\} = 1 - \pi_t$, $E\{Y_t\} = \pi_t$ and $\text{var}\{Y_t\} = \pi_t[1 - \pi_t]$, and so the series Y is nonstationary in variance.

The binary case was studied in Brillinger (1994), through a probit model,

$$\text{Prob}\{Y_t = 1 | Y_{t-1}, Y_{t-2}, \ldots\} = \Phi\left(S_t + \sum_{u=1}^{U} a_u Y_{t-u} \right)$$

with $S_t = \alpha + \beta t$ and Φ the normal cumulative. This is a fairly strong assumption. The hypothesis of no trend may be formalized here as $\beta = 0$. In the present work the statistic (3) will be employed, but an estimate of $f_2(0)$ specific to the binary case needs to be developed.

A model is required for a 0-1 valued Y_t allowing the presence of a trend. To that end consider the model of random (time dependent) deletion of the 1's of a stationary 0-1 process. Specifically suppose

$$Y_t = I_t X_t$$

with X_t a 0-1 stationary process having mean μ and autocovariance function, $\text{cov}\{X_{t+u}, X_t\} = \gamma(u)$ and with I_t a sequence of independent 0-1 variates having $E\{I_t\} = \nu_t$. Then, see Appendix,

$$\pi_t = E\{Y_t\} = \nu_t \mu$$

and

$$\text{cov}\{Y_s, Y_t\} = \delta_t^s \nu_t \mu (1 - \nu_t \mu) + (1 - \delta_t^s) \nu_s \nu_t \gamma(s - t)$$

where δ_t^s is 1 if $s = t$ and is 0 otherwise. Nonstationarity of the rate π_t comes from nonconstancy of the ν_t. Now

$$E\left\{ \sum c_t Y_t \right\} = \sum c_t \nu_t \mu \tag{4}$$

and

$$\text{var}\left\{\sum c_t Y_t\right\} = \sum c_t^2 \nu_t \mu(1-\nu_t\mu) + \sum_{s\neq t} c_s c_t \nu_s \nu_t \gamma(s-t) \tag{5}$$

For an estimate of this last, one needs estimates of $E\{Y_t\} = \nu_t\mu$ and of $\nu_s\nu_t\gamma(s\text{-}t)$. An estimate of $\pi_t = E\{Y_t\} = \nu_t\mu$ may be obtained by smoothing the Y values in some fashion. An estimate of $\gamma(u)\mu^{-2}$ is provided by

$$g(u) = \frac{1}{T}\sum_{s-t=u}(Z_s - \overline{Z})(Z_t - \overline{Z}) \tag{6}$$

where $Z_t = Y_t/\hat{\pi}_t$. Now (4) may be estimated by

$$\sum c_t^2 \hat{\pi}_t(1-\pi_t) + \sum_{s\neq t, |s-t|\leq U_T} c_s c_t \hat{\pi}_s \hat{\pi}_t g(s-t) \tag{7}$$

for some U_t.

Under a mixing condition on the series X, the statistic $\sum c_t Y_t$ will be asymptotically normal and the variance estimate (7), will be consistent, see Appendix. One might damp down the terms in the second sum of (7) by inserting a window function, as is usual in spectrum estimation.

Such estimates as (6), with terms weighted inversely to probabilities, are considered in Brillinger (1979), Lee and Brillinger (1979), Guttorp and Thompson (1991).

4 Results for the Rio Negro

Figure 2 presents an estimate of π_t obtained via the procedure gam() of Hastie (1991) for the Rio Negro data. The dashed line is at the overall level. Of course there is uncertainty in the estimates. (In gam() the options family = binomial (link=probit) and loess(span=.75) were employed.)

For the Rio Negro data

$$\sum c_t Y_t = -1.143$$

and taking $U_T = 10$ the standard error estimate is .732 . The value of the statistic (3) is -1.563, so there is no substantial evidence of a monotonic increasing (or even decreasing) trend.

5 The point process case

Consider a stochastic point process $N \equiv \{\tau_j\}$. As in the binary case one can ask if the rate function of N is monotonic increasing as opposed to constant. In this section a procedure paralleling the 0-1 case will be indicated.

Chapter 3 of Cox and Lewis (1966) is devoted to trend analysis of point processes. Cox and Lewis consider the Poisson and also renewal models. The Poisson is a common model for times of floods, see Todorovic (1979), and may be motivated by the idea that floods are rare. Ogata and Katsura (1986) proceed to model point processes incorporating trends by linearly parametrizing the conditional intensity function of the point process. (This function completely defines a broad class of point processes, as indicated in the reference.) These authors then obtain maximum likelihood estimates.

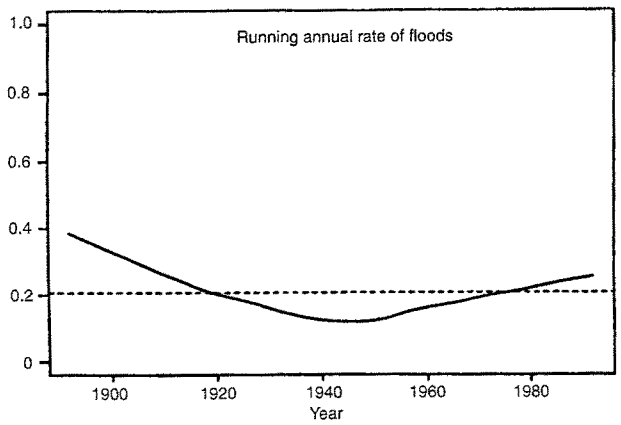

Figure 2. Estimate of $\hat{\pi}_t$

Turning to the development of a procedure of the character of (3), suppose that the times of points are available for the interval (0, T]. Let N_t denote the cumulative count, that is the number of points in (0, t]. An analog of the statistic (3) is based on

$$\int_1^{T-1} c_t dN_t = \sum_{1 \le \tau_j \le T-1} c_{\tau_j}$$

with the c_t given by (2). In studying this case and since T is large it is convenient to replace c_t by b_t where

$$c_t \approx b_t = \frac{2t/T - 1}{2[t(1 - t/T)]^{1/2}}$$

for $1 \le t \le$ T-1. The intervals (0,1) and (T-1,T] are avoided in the integral to keep quantities involved finite.

If the process has rate π_t, defined via $E\{dN_t\} = \pi_t dt$, then

$$E\left\{\int_1^{T-1} b_t dN_t\right\} = \int_1^{T-1} b_t \pi_t dt$$

which will be 0 for π_t constant. Supposing $\mathrm{cov}\{dM_s, dM_t\} = q(s\text{-}t)dsdt$ for $s \neq t$, one has

$$\mathrm{cov}\{dN_s, dN_t\} = [\delta(t - s)\nu_t\mu + \nu_s\nu_t q(s - t)]dsdt$$

and so

$$\mathrm{var}\left\{\int_1^{T-1} b_t dN_t\right\} = \int_1^{T-1} b_t^2 \nu_t \mu dt + \int_1^{T-1}\int_1^{T-1} \nu_s \nu_t q(s - t)dsdt \qquad (8)$$

The estimation of q(u), given the π_t, is considered in Brillinger (1979) and Lee and Brillinger (1979). The rate function π_t can be estimated by a kernel or some such

procedure, see Boneva et al (1971), Guttorp and Thompson (1991) and so one can construct an estimate of the variance and thereby form a standardized test statistic.

An advantage of the present approach is simplicity, specifically the linearity of the numerator. Another advantage is the broad assumption of stationarity instead of the narrow Poisson or renewal. If a full specification is available, as in Ogata and Katsura (1986), then one would expect a likelihood approach to be more efficient.

The point process problem of this section may sometimes be conveniently reduced to the corresponding 0-1 case. One simply breaks the time interval up into small equiwidth cells and sets Y_t equal to 1 if there is a point in the t-th cell and to 0 otherwise.

6 Discussion

Little evidence was found for an increasing trend in the rate of flooding of the Rio Negro near Manaus.

If it is felt reasonable to assume that the observed process is Bernoulli or Poisson, then there would be a simpler denominator for (3), specifically the second terms in (4) and (8) are not needed. In terms of inadequacies of the analysis, one thing to be mentioned is that the data analyzed started with a flood rather than at an arbitrary time origin. The statistical derivations assumed an arbitrary origin.

In the future it seems worth developing test statistics for other nonstationary 0-1 valued and point processes based on some stationary process.

Acknowledgements

The Rio Negro data were provided by Hilgard O'Reilly Sternberg. The paper was prepared with the partial support of the NSF Grant DMS-9300002 and the ONR Grant N00014-94-1-0042.

Appendix

This Appendix refers to some of the analytic details. Basically the methods and assumptions of Brillinger (1972) and Brillinger (1979) are employed. Re the rate ν_t it is assumed that $\nu_t = \nu(t/T)$ with $\nu(.)$ satisfying Assumption A.2 of Brillinger (1989). To obtain the expressions (4) and (5) for the mean and variance one uses the results

$$E\{Y_t\} = E_x\{E\{Y_t|X\}\}$$

and

$$\mathrm{cov}\{Y_s, Y_t\} = E_X\{\mathrm{cov}\{Y_s, Y_t|X\}\} + \mathrm{cov}_X\{E\{Y_s,|X\}, E\{Y_t|X\}\}$$

To develop the asymptotic normality of the numerator of the criterion, one simply evaluates cumulants, as in Brillinger (1989). The estimate of the variance tends to the variance, in probability, and the statistic is consequently asymptotically normal.

The consistency of the test follows from the fact that

$$\frac{|\sum c_t S_t|^2}{\sum c_t^2} \geq \frac{2\sum(S_t - \bar{S})^2}{\log T}(1 + o(1))$$

for monotonic S_t, here $\nu_t \mu$, and the coefficients in question.

References

Abelson, R.P.; Tukey, J.W. 1963: Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, Ann. Math. Statist. 34, 1347-1369

Boneva, L.I.; Kendall, D.G.; Stefanov, I. 1971: Spline transformations: Three new diagnostic aids for the statistical data-analyst. J. Roy. Statist. Soc. B, 1-37.

Brillinger, D.R. 1972: The spectral analysis of stationary interval functions. In Proc. Sixth Berkeley Symp. Math. Star. Prob., pp. 483-513. Univ. Calif. Press, Berkeley

Brillinger, D.R. 1979: Analyzing point processes subjected to random deletions. Canadian J. Statist. 7, 21-27

Brillinger, D.R. 1989: Consistent detection of a monotonic trend superposed on a stationary time series. Biometrika 76, 23-30

Brillinger, D.R. 1994: Trend analysis: Time series and point process problems. Environmetrics 5, 1-19

Cox, D.R. 1970: Analysis of Binary Data. Methuen, London

Cox, D.R. 1981: Statistical analysis of time series: Some recent developments, Scand. J. Statist. 8, 93-108

Cox, D.R.; Lewis, P.A.W. 1966: The Statistical Analysis of Series of Events. Methuen, London

Cuzick, J. 1988: Trend tests. Encyclopedia of Statistical Sciences 9, 336-342

Dagum, C.; Dagum, E.B. 1988: Trend. Encyclopedia of Statistical Sciences 9, 321-324

Guttorp, P.; Thompson, M.L. 1991: Estimating second-order parameters of volcanicity from historical data. J. Amer. Statist. Assoc. 86, 578-583

Harvey, A.C. 1989: Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge Univ. Press, Cambridge

Hastie, T.J. 1991: Generalized additive models. In Statistical Models in S (eds. J.M. Chambers and T.J. Hastie). Wadsworth, Pacific Grove, pp. 249-308

Lee, W.H.K.; Brillinger, D.R. 1979: On Chinese earthquake history - an attempt to model an incomplete data set by point process analysis. Pageog 117, 1229-1257

Lee, Y.J. 1988: Tests for trend in count data. Encyclopedia of Statistical Sciences 9, 328-334

Margolin, B.H. 1988: Test for trend in proportions. Encyclopedia of Statistical Sciences 9, 334-336

Ogata, Y.; Katsura, K. 1986: Point-process models with linearly parametrized intensity for application to earthquake data. In Essays in Time Series and Allied Processes (eds. J. Gani and M.B. Priestley). Applied Probability Trust, Sheffield, pp. 291-310

Schaafsma, W.; Smid, L.J. 1966: Most stringent somewhere most powerful test against alternatives restricted by a number of linear inequalities. Ann. Math. Statist. 37, 1161-1172

Sternberg, H. O'R. 1987: Aggravation of floods in the Amazon River as a consequence of deforestration? Geografiska Annaler 69A, 201-219

Todorovic, T. 1979: A probabilistic approach to analysis and prediction of floods. Bull. Internat. Statist. Inst. 48, Book 1, 113-124