JOURNAL OF GEOPHYSICAL RESEARCH, VOL 92, NO D8, PAGES 9679–9682, AUGUST 20, 1987

Assessment of a Class of Neyman-Scott Models for Temporal Rainfall

EFI FOUFOULA-GEORGIOU

Department of Civil Engineering, Iowa State University

PETER GUTTORP

Department of Statistics, University of Washington

Neyman-Scott type cluster point processes have been used in several studies to model temporal rainfall at a single location. In this paper we study the applicability of such models with the rainfall thought of as instantaneous bursts at the points of the Neyman-Scott process. We find that this class of models does not provide adequate fit to some observed rainfall series. We also discuss some estimation problems associated with the fitting procedure, and examine the importance and appropriateness of the distributional assumptions made in the modeling.

1. INTRODUCTION

Hobbs and Locatelli [1978] describe mesoscale rainfall activity in cyclonic storms roughly as follows. Synoptic-scale weather fronts contain in them large mesoscale regions or rainfall bands, where precipitation activity is possible. In turn, these bands contain moving rain cells, which are points of higher rainfall rates. On the basis of a similar physical description, *LeCam* [1961] suggested modeling rainfall at a location by a cluster point process having as its underlying primary process the arrival of cyclonic storm systems, with the secondary process, the one actually observed, corresponding to the mesoscale precipitation phenomena. *Kavvas and Delleur* [1981] suggested a Neyman-Scott Poisson cluster process, in which the primary process is a nonhomogeneous Poisson process and where the secondary points are laid down in an independent and identically distributed fashion around the cluster centers of the primary process. Rodriguez-Iturbe and coworkers [*Rodriguez-Iturbe et al.*, 1984; *Valdes et al.*, 1985; *Rodriguez-Iturbe et al.*, this issue] have studied different versions of this model, usually made stationary by considering only a short time period each year, such as a month, and taking rainfall into account by somehow modeling the relation between the point process of rainfall occurrences and the precipitation amounts per event. The main approach of the previous studies to judging the goodness of the fit of the model has been to study the consistency of parameter estimates based on rainfall amounts subjected to varying amounts of aggregation.

In this paper we will concentrate on the point process of event occurrences for short, stationary periods. On the basis of some of the numerical results in the work by *Foufoula-Georgiou and Guttorp* [1986], henceforth referred to as FGG, we will present a critical assessment of the validity of a class of Neyman-Scott models for temporal rainfall. We will look at the consistency of parameter estimates based on different time scales. The problems associated with different estimation procedures will be discussed. In particular, we give an explicit recursive formula for computing the likelihood of an observed zero-one process indicating presence or absence of precipitation in each time interval. We will highlight the sensitivity of one estimation method to different distributional assumptions. Finally, we discuss some ways one may, directly or indirectly, check the assumptions made in the Neyman-Scott model. As an illustration, lid-opening data from the MAP3S (Multistate Atmospheric Power Product Pollution Study) acid rain monitoring network are used to study the distribution of cluster size.

2. TIME SCALE CONSISTENCY OF PARAMETER ESTIMATES FOR INSTANTANEOUS NEYMAN-SCOTT MODELS

We view the rainfall process in a simplistic fashion as instantaneous bursts of precipitation at the times of events. The amounts of precipitation corresponding to each event are called marks, and the overall process is a marked point process. No specification of the actual relation between rainfall amounts and the process of rainfall events will be attempted here, and, in fact, our results are valid for any such specification. We will follow *Rodriguez-Iturbe et al.* [1984], henceforth referred to as RGW, in assuming that the primary process follows a homogeneous Poisson process with intensity λ, whereas the cluster of secondary events associated with a given primary event has a size described by a geometric distribution of parameter p and occurs at temporal locations which have exponentially distributed distances from the location of the primary event. This exponential distribution has parameter β. This model is applicable to relatively short, and therefore approximately stationary, segments of precipitation data. In our work we use a separate model for each month.

Since we usually only have access to hourly data, we do not know the exact times of events. However, we can observe whether or not there was precipitation (corresponding to at least one event) in each time period. Using results from *Guttorp* [1986a], we base our inference on this binary occurrence series. We fit the model parameters using the method of moments. The observed mean and lag one and lag two autocorrelations of the binary series are set equal to their theoretical counterparts. A detailed analysis of two data sets, together with a Monte Carlo study of the accuracy of the estimation procedure, can be found in the work by FGG. We used six different aggregation scales, ranging from 1 to 24 hours. Here we show one example of the results. Data from Sea-Tac airport in Washington for the month of December, 1965–1982, yielded the parameter estimates in Table 1.

There is systematic variation of the parameter estimates with the discretization interval Δ. This was found for all months in data from both Sea-Tac and Denver. Since the problem could be associated with the fitting procedure, we applied the procedure to

Paper number 7D0322.
0148-0227/87/007D-0322$05.00

TABLE 1. Parameters of an Instantaneous Neyman-Scott Model Fitted to Sea-Tac Data for December, 1965–1982.

	Δ in days:					
	1/24	1/12	1/6	1/4	1/2	1
$\hat{\lambda}$	1.25	0.98	0.81	0.75	0.63	0.68
$\hat{p}$	0.025	0.055	0.083	0.098	0.202	0.316
$\hat{\beta}$	18.790	8.452	5.374	4.224	1.934	1.772

simulated data from Neyman-Scott processes with a wide range of parameter values. An example is shown in Table 2.

Since there is no indication in the simulation study of the systematic variation of parameter estimates with Δ seen in the rainfall data, we conclude that the instantaneous-burst Neyman-Scott model is not describing the data well. Furthermore, the appealing physical interpretation of the components of the model, identifying the primary process with the frontal systems and the secondary process with the rainbands, is untenable, since these physical processes do not change with the discretization scale. *Valdes et al.* [1985] found similar systematic variation of parameter estimates when fitting a particular case of this model to data simulated from the space-time model of *Waymire et al.* [1984].

3. ESTIMATION PROBLEMS FOR INSTANTANEOUS NEYMAN-SCOTT MODELS

The general problem of parameter estimation for point processes has been studied in the statistical literature [e. g., *Brillinger*, 1978; *Ogata*, 1978; *Kutoyants*, 1984]. However, the application to particular models is often far from straightforward. In this section we will discuss two estimation methods, maximum likelihood and the method of moments, each of which has its own problems.

3.1 *Maximum Likelihood Estimation*

The method of maximum likelihood is well respected in classical statistical theory, since it yields asymptotically efficient estimators. Similar results have been obtained for maximum likelihood estimation in stochastic processes [*Heyde*, 1978]. It is in principle possible to fit the binary occurrence series using maximum likelihood. In order to do so, we need two simple results. Let $\zeta(A)$ denote the probability of no points of the continuous time Neyman-Scott process N falling in the set A. Suppose that A is of the form $\cup_1^n(a_i, b_i)$ where $b_i \le a_{i+1}$.

$$\log\zeta(A) = -\lambda\{b_n + q\sum_{j=1}^{n}(b_j-a_j) - \frac{1}{\beta}\sum_{j=0}^{n-1}\log\left[\frac{qC_{j+1}+pd_j}{qC_{j+1}+pf_{j+1}}\right]$$

$$-\frac{1}{\beta}\sum_{j=1}^{n}\log\left[\frac{q(C_{j+1}-d_j)+f_j}{q(C_{j+1}-d_j)+d_j}\right] - \frac{q}{\beta}\sum_{j=1}^{n}\log\left[\frac{1+q(C_{j+1}-d_j)/d_j}{1+q(C_{j+1}-d_j)/f_j}\right]\}$$

where $q=1-p$, $d_j=\exp(-\beta b_j)$, $f_j=\exp(-\beta a_j)$, $C_j=\sum_{i=j}^{n}(f_i-d_i)$, and $C_{n+1}=0$. Furthermore, the likelihood $L(\lambda, p, \beta)$ from the observed binary series can be computed recursively using the fact that

$$\mathbf{P}(y\ 1\ x) = \mathbf{P}(y \bullet x) - \mathbf{P}(y\ 0\ x)$$

where y and x are arbitrary sequences of zeros and ones, the dot (•) stands for an arbitrary symbol (a zero or a one), and **P** is the probability measure of the binary series for given values of the parameters λ, p and β. In other words, the probability of the sequence y, followed by a one, and then followed by the sequence x, can be computed when one knows the probability of the sequence y, followed by any single event, and then the sequence x, as well as the probability of the sequence y, followed by a zero, and then followed by x. The procedure is to take the given sequence of zeros and ones, transform its probability into a linear combination of probabilities of sequences only involving the zero and the dot, which subsequently can be expressed in terms of ζ evaluated at sets of the form $\cup(k_i\Delta, l_i\Delta)$ for integers k_i and l_i. An example of this procedure is given in the appendix of FGG.

The resulting likelihood will now have to be optimized numerically. For a large data set the resulting likelihood function (which, of course, is just the probability of the observed sequence, viewed as a function of the unknown parameters) is a complicated expression which may require considerable computing time to optimize. One has to evaluate a linear combination of $2^{\nu(1)}$ terms of the form given in the expression for log $\zeta(A)$ above, where $\nu(1)$ is the number of rainy time intervals (e. g., hours or days) in the data. For small data sets the method of maximum likelihood should be used because of its higher efficiency, rather than the method of moments discussed below. In the data set previously discussed in Table 1, with 13,392 hourly observations, we were unable to compute the maximum likelihood estimates.

3.2 *Numerical Problems With the Method of Moments*

When the form of the likelihood is too complicated, one is forced to try a different estimation method which yields simpler expressions for the estimates. The method of moments often has this property. One simply equates expressions for theoretical moments to estimates of these moments until there are enough equations to solve for the unknown parameters. Generally, one

TABLE 2. Mean and Standard Deviation of Parameter Estimates From 500 Replicates of a Neyman-Scott Model

	Δ in days:					
	1/24	1/12	1/6	1/4	1/2	1
mean, $\hat{\lambda}$	0.105	0.103	0.102	0.102	0.102	0.101
sd, $\hat{\lambda}$	0.029	0.012	0.009	0.009	0.009	0.009
mean, $\hat{p}$	0.050	0.051	0.051	0.052	0.052	0.080
sd, $\hat{p}$	0.013	0.007	0.009	0.011	0.017	0.056
mean, $\hat{\beta}$	5.144	5.081	5.027	5.027	5.083	4.601
sd, $\hat{\beta}$	1.671	0.701	0.478	0.479	0.690	1.824

Parameters are: λ=0.10; p=0.05; β=5.0.

tries to use low-order moments or product moments for the equations, since high-order moments are difficult to estimate accurately. Naturally, the simplicity of the method has a price: the method of moments estimates typically have larger variance than the maximum likelihood estimates. In large data sets this may not matter too much, since then the variance of the moment estimates would be relatively small.

Since we have three parameters, λ, β, and p, in the instantaneous Neyman-Scott model, we use the mean and the lag one and lag two autocorrelations for the fitting. A major problem with using the lag two autocorrelation (r_2) is that it often is quite small, especially when Δ is large. Consequently, one may frequently obtain negative estimates of this parameter. Negative parameter values are impossible under the model (cf. FGG), and the resulting equations have no admissible solution. Furthermore, the derivative of r_2 with respect to λ is found to be very small for a large range of values of β. Thus the estimates may be quite unstable, in that a small change in an empirical moment estimate may lead to a large change in the estimated parameter value. This is one reason for the high variability of the method of moment estimator of β, even for the simulated data in Table 2.

The method of moments is sometimes praised for the quality of preserving the observed low-order moments, something that presumably is desirable for a model which is intended as the input to a hydrological simulation program. On the other hand, empirical moments have large variability, and it is more important to validate the structure of the model against data than to choose some particular empirical moments and preserve these. An advantage with the maximum likelihood estimates is that they automatically provide estimates of the low-order moments by plugging the parameter estimates into the theoretical formulae. Serious discrepancies of the maximum likelihood estimates of low-order moments from corresponding sample moments cast doubt over the validity of the model.

The sensitivity of the equations defining the method of moments estimates is also illustrated in a discussion by FGG. Two methods of fitting an instantaneous bursts Neyman-Scott model are compared, yielding models with very similar moments but with quite different estimated parameter values. This puts severe strains on the physical interpretability of the parameters and illustrates further the need for independent checking of the components of the stochastic model that is being fitted.

3.3 *Sensitivity to Choice of Cluster Size Distribution*

The choice of distribution of the number of events in a cluster has largely been a matter of mathematical convenience. In the hydrological literature, either the geometric or the Poisson distribution has been assumed. The Poisson assumption allows *Smith and Karr* [1985] to develop maximum likelihood estimates for a continuously observed Neyman-Scott process. The geometric assumption, which is more common, is used in this paper to obtain a closed form for the likelihood of the binary series discussed previously. We will relate these assumptions later in this section.

An example given by FGG shows how extremely sensitive parameters like the expected rainfall amounts in each cluster member and the rate of events, are relative to the specific choice of distribution for the cluster size. Ideally, observations of the cluster members should be used to determine the distribution of cluster size. Such observations have not usually been available to the hydrological community. However, in acid rain research it is often the case that observations are made based on events, corresponding to storm fronts passing over a station. Because chemical analyses need to be made on the samples, the data are collected in samplers which open only when it rains. Data from the MAP3S network station at Whiteface Mountain, New York, (described by *Guttorp*, [1986b]) contain for each event the number of lid openings. Taking these to correspond (at least roughly) to cluster events, we observed an average of 11.7 openings per event, with a standard deviation of 14.6. It is interesting to note that this average is similar in size to those reported by FGG and RGW, although the data of the latter are hourly observations rather than event-based observations. The size of the standard deviation immediately rules out the Poisson distribution, which has a variance equal to its mean. The estimated standard deviation from a geometric distribution is 11.2, and a χ^2-test rejects the geometric distribution. There are too many extreme events with either very few or very many openings. Hence a Neyman-Scott-type model for these data will have to use a more complicated cluster size distribution than Poisson or geometric.

A natural candidate is the negative binomial distribution [cf. *Johnson and Kotz*, 1969, chap. 5]. This is a two-parameter distribution, which allows for added flexibility in fitting. The geometric is a special case. There is a natural link between the Poisson, geometric, and negative binomial distributions. The latter two are derived from the former by suitable randomization of the Poisson mean. Thus if one regards the formation of secondary events as generated by a Poisson mechanism, one obtains a geometric distribution of cluster size by assuming that each storm draws its mean cluster size from an exponential distribution. The negative binomial is obtained by instead drawing the mean cluster size from a gamma distribution. Fitting a truncated negative binomial (excluding zeros) to the lid-opening data gave good agreement (the P-value is 0.17 for a χ^2-test). The estimated gamma shape parameter was 0.47, corresponding to a density with a tendency to give more small events than the exponential distribution (corresponding to a geometric cluster size). In addition, the fitted gamma density was more spread out more than the fitted exponential, so that more large events would be expected as well.

4. DISCUSSION

A stochastic model that is based on a conceptualization of the physical mechanism governing the phenomenon under study should have some components, at least, with physical interpretation. This is particularly important if the model is to be used as input to studies of complex systems, as is often the case in hydrology, since deviations from the physical structure of the input to the system may have quite unforeseen effects on the output. It is not enough to simply preserve a few empirical moments. Rather, the model and its conceptual basis must be validated.

In the case of the instantaneous Neyman-Scott model discussed in this paper, the temporal distribution of storm front arrivals and the size of clusters are parameters of the model, which could, at least in principle, be determined from data such as satellite and/or radar measurements. Efforts in this directions would be beneficial in uncovering structural properties of the actual precipitation processes that can be very useful in the stochastic model building. The results discussed in section 2 indicate that the particular version of the stochastic model studied does not have a physical basis. There are at least three different possible explanations. First, the structure of the conceptual model for the underlying unobserved mechanism may be incorrect. Second, the simplistic view of instantaneous bursts of rainfall may be incorrect. Third, the particular choices of components (Poisson primary process, geometric- or Poisson-distributed cluster sizes, exponential cluster dispersion function) may be incorrect. In this section we will discuss ways of assessing these explanations.

We mentioned briefly in section 1 the meteorological basis for the cluster process model. It is of course quite simplistic, e. g., by not distinguishing between convective and stratiform precipitation, by not taking rain cells into account, and by ignoring meteorological measurements on winds, etc., that may be available. Nevertheless, we feel that it provides a conceptually sound model and therefore can form the basis for a simple stochastic model of temporal precipitation.

By looking at the binary series of precipitation occurrence we have avoided making any specific assumptions about the relation between precipitation amounts and the point process of events. Our conclusions therefore apply to any such specification. The key simplifying assumption is that rainfall events are instantaneous bursts. There is little doubt, in view of the results of *Rodriguez-Iturbe et al.* [this issue], that a more complex description of the marked point process allowing for positive duration of events will yield improvements in the model fit. Since the binary process we study carries direct information about the underlying point process only for the instantaneous model and not for a model where precipitation events have a duration and may overlap, our results have no bearing on the validity of the model used by Rodriguez-Iturbe *et al.*. Further work is needed to map the correspondence between the physical system and the stochastic description given by Rodriguez-Iturbe and coworkers.

Kavvas and Herd [1985] used radar data to develop a model for spatial rainfall. These data may also be used to check the validity of the assumptions regarding the primary process, since the arrival of frontal systems can be identified from the radar pictures [*Browning*, 1985; *Austin*, 1985]. Research along these lines seems very promising, in particular for space-time model development. *Guttorp* [1986b], using event-based data, found that a homogeneous Poisson process is a reasonable model for the arrival of storm fronts to at least one of the stations studied in that paper. On the other hand, *Guttorp and Thompson* [1986] found that the arrival process of severe tropical cyclonic storms at the Bay of Bengal exhibits some evidence of clustering. As to the frequency of rainbands in a given frontal system, radar or satellite pictures may provide valuable information. Acid rain data of the type used in section 3.3 can be utilized to complement radar and satellite data. The results of *Rice* [1975] can be used to estimate nonparametrically the dispersion function for the cluster events if one can estimate the underlying covariance density. This method is computationally somewhat intensive but yields (pointwise) standard errors for the estimates, enabling validation of the exponential assumption for dispersion.

Acknowledgments. This work had partial support from the Environmental Protection Agency of the United States, through a cooperative agreement with the SIAM Institute for Mathematics and Society (SIMS).

REFERENCES

Austin, G. L., Application of pattern-recognition and extrapolation techniques to forecasting, *ESA J.*, *9*, 147–155, 1985.

Brillinger, D. R., Comparative aspects of the study of ordinary time series and of points processes, in *Developments in Statistics*, vol. 1, edited by P. N. Krishnayah, Academic, Orlando, Fla., 1978.

Browning, K. A., Conceptual models for precipitation systems, *ESA J.*, *9*, 157–180, 1985.

Foufoula-Georgiou, E., and P. Guttorp, Compatibility of continuous rainfall occurrence models with discrete rainfall observations, *Water Resour. Res.*, *22*, 1316–1322, 1986.

Guttorp, P., On binary time series obtained from continuous time point processes describing rainfall, *Water Resour. Res.*, *22*, 897–904, 1986a.

Guttorp, P., Modelling rainfall events using event-based data, *SIMS Tech. Rep. 99*, Dept. Statist., Univ. British Columbia, Vancouver, B. C., 1986b.

Guttorp, P., and M. L. Thompson, A probability model for severe cyclonic storms striking the coast around the Bay of Bengal, *Mon. Weather Rev.*, *114*, 2267–2271, 1986.

Heyde, C. C., An optimum property of the maximum likelihood estimator for stochastic processes, *Stochastic Process. Appl.*, *8*, 1–9, 1978.

Hobbs, P. V., and J. D. Locatelli, Rainbands, precipitation cores and generating cells in a cyclonic storm, *J. Atmos. Sci.*, *35*, 230–241, 1978.

Johnson, N. L., and S. Kotz, *Distributions In Statistics: Discrete Distributions*, Houghton Mifflin, Boston, Mass., 1969.

Kavvas, M. L., and J. W. Delleur, A stochastic cluster model for daily rainfall sequences, *Water Resour. Res.*, *17*, 1151–1160, 1981.

Kavvas, M. L., and K. R. Herd, A radar-based stochastic model for short-time-increment rainfall, *Water Resour. Res.*, *21*, 1437–1455, 1985.

Kutoyants, Yu. A., *Parameter Estimation for Stochastic Processes*, Heldermann, Berlin, 1984.

Le Cam, L. M., A stochastic description of precipitation, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 3, edited by J. Neyman, pp. 165–186, California, Berkeley, Calif., 1961.

Ogata, Y., The asymptotic behaviour of maximum likelihood estimators for stationary point processes, *Ann. Inst. Statist. Math.*, *30*, 243–261, 1978.

Rice, J., Estimated factorisation of the spectral density of a stationary point process, *Adv. Appl. Probab.*, *7*, 801–817, 1975.

Rodriguez-Iturbe, I., V. K. Gupta, and E. Waymire, Scale considerations in the modeling of temporal rainfall, *Water Resour. Res.*, *20*, 1611–1619, 1984.

Rodriguez-Iturbe, I., B. Febres de Power, and J. B. Valdés, Rectangular pulses point process models for rainfall: Analysis of empirical data, *J. Geophys. Res.*, this issue.

Smith, J. A., and A. F. Karr, Statistical inference for point process models of rainfall, *Water Resour. Res.*, *21*, 73–79, 1985.

Valdes, J. B., I. Rodriguez-Iturbe, and V. K. Gupta, Approximations of temporal rainfall from a multidimensional model, *Water Resour. Res.*, *21*, 1259–1270, 1985.

Waymire, E., V. K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-β scale, *Water Resour. Res.*, *20*, 1453–1465, 1984.

E. Foufoula-Georgiou, Department of Civil Engineering, Iowa State University, Ames, IA 50011.

P. Guttorp, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195.

(Received July 10, 1986;
revised March 23, 1987;
accepted March 27, 1987.)