TECHNICAL RESEARCH REPORT

Exceedances and Moments in Data Containing Zeros

by K. Fokianos, B. Kedem, and D.A. Short

T.R. 96-47

INSTITUTE FOR SYSTEMS RESEARCH

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Exceedances and Moments in Data Containing Zeros

by

Konstantinos Fokianos[1], Benjamin Kedem[1], and David A. Short[2]

[1]Mathematics Department and Institute for Systems Research
University of Maryland, College Park
Maryland 20742, USA

and

[2]Laboratory of Atmospheres
NASA/GSFC, Greenbelt
Maryland 20771, USA

March 1996 [1]

[1]Work supported by NSF grant EEC-940-2384, and NASA grant NAG52783

Abstract

Rain rate–the speed of rain in mm/hr–assumes zero values when it is not raining, and positive values on a continuum otherwise. For large areas, empirical evidence points to a high correlation between the instantaneous area average rain rate and the instantaneous fractional area where rain rate exceeds a fixed positive threshold. An explanation is provided by appealing to the mean value theorem for integrals, in conjunction with the mixed nature of the probability distribution of rain rate. Using a multinomial logits model, the fractional area also is shown useful as a time dependent covariate in categorical prediction of the area average.

Keywords: Rain rate, area average, fractional area, correlation, categorical time series, partial likelihood.

1 Introduction

The quantity of interest in this article is *rain rate* in mm/hr observed instantaneously, at every location over a given large area in the tropics, by means of a precipitation radar. So, imagine large instantaneous maps or snapshots of rain rate taken periodically over a fixed area. Such snapshots consist of zeros when it is not raining, and positive values otherwise when it is raining. It follows the probability distribution of rain rate is of a *mixed type*: a mixture of a discrete component supported at 0, and a continuous component. We use this fact in explaining a certain remarkable linearity observed in nature.

Instantaneous tropical rain rate snapshots obtained from a radar over a given large area show that *the instantaneous fraction of the area (FA) where rain rate exceeds a threshold and the instantaneous area average (AA) are highly correlated quantities.* An example of this fact is shown in Figure 3400 pairs (FA,AA) from snapshots over a large area 290 km in diameter in the tropics–near 2°S and 156°E–obtained every 10 minutes. The data are described briefly in the enclosed appendix. We see for a threshold in the range of 3 to 9 mm/hr the correlation is quite high exceeding 97%. The same has been observed time and again using different radar data sets of tropical rainfall from different parts of the globe, yielding correlations well above 95% and at times reaching 99%. Short et al. (1993) have observed the same high degree of linearity from raingauge data. This seemingly puzzling phenomenon has attracted a fair amount of attention in the scientific community and several attempts to explain it have been suggested including Atlas et al. (1990), Braud et al. (1993), Kedem et al. (1990), Kedem and Pavlopoulos (1991), Mase (1994), Morrissey (1994), Rosenfeld et al. (1990). Shimizu et al. (1993) have observed the high linearity persists for higher order moments as well–it necessitates higher thresholds.

We argue the observed linearity can be explained very simply by appealing to the mean value theorem for integrals in conjunction with the fact rain rate is nonnegative and its distribution is of mixed type.

We further illustrate the fractional area is a useful covariate in categorical prediction of the area average rain rate. This has an application in rainfall measurement from imprecise satellite-borne instruments, for it is still possible to classify reliably rain rate as being above or below a threshold and determine accordingly the fractional area and from it the area average.

2 Variation in the Discrete Component

Let X be a random variable of a mixed type

$$X = \begin{cases} Y, & \text{With prob. } p \\ 0, & \text{With prob. } 1-p \end{cases}$$

where Y is a positive continuous random variable with absolutely continuous distribution function F. It follows

$$P(X \leq x) = (1-p)H(x) + pF(x),$$

where $H(x)$ is a step function, $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$. Therefore for $x \geq 0$,

$$P(X > x) = [1 - F(x)]p$$

For a positive b the mean value theorem gives for $\tau \in [0, b]$,

$$E(X) = bP(X > \tau) + \int_b^\infty P(X > x)dx$$

Taking b sufficiently large we obtain

$$E(X) \approx bP(X > \tau) = b[1 - F(\tau)]p \tag{1}$$

Clearly, since $E(X) = pE(Y)$, b depends on F only. The relationship (1) points to a close association between the first moment and the exceedance probability for sufficiently large threshold τ, since τ grows monotonically with b.

Let $X_1, ..., X_n$ be a random sample from X. Then, in general, the sample average $\overline{X}$ and sample proportion of positive values $\hat{p}$ are positively correlated, since a larger proportion of positive values–at the expense of the 0's–tends to increase the sample average. However, the correlation cannot be too high, because pairs $(\overline{X}, \hat{p})$ from different samples from X tend to cluster around the point $(E(X), p)$.

If on the other hand each sample $X_1, ..., X_n$ comes from the *same* F but *different* p, as p varies in $[0, 1]$ from sample to sample while *holding* F *fixed (and hence also* b*)*, we see from (1) the pairs $(\overline{X}, \hat{p})$ tend to vary jointly along a straight line with slope $b[1 - F(\tau)]$, thus producing high correlations. Empirical evidence shows τ may even be only (relatively) moderately large.

This is illustrated in Figures corresponding to thresholds $\tau = 5$ and $\tau = 10$, respectively. The figures show scattergrams from 400 samples from a mixed lognormal–that is $\log Y$ is $\mathrm{N}(1,1)$–produced with different p values as indicated, while holding $b[1 - F(\tau)]$ fixed. In Figure 2-(a) $p = 0.07$ is held constant in all the 400 samples giving a correlation between $(\overline{X}, \hat{p})$ of 0.758. In 2-(b) p is allowed to vary in the set (.07,.08,.09,.1,.12) and the correlation increases to 0.878. In Figure yielding a correlation of 0.994. Similar remarks hold for Figure in p is more pronounced, as it "stretches" over the range $[0,1]$ it produces a high correlation between $(\overline{X}, \hat{p})$ in excess of 99%.

This can help to explain the high correlations in Figure as a sample (not necessarily random) from X. In this case $\hat{p}$ at time t is the percent of the area covered with precipitation, and the fractional area is approximated by $[1-F(\tau)]\hat{p}$. One can argue that over a period of about three weeks, rainfall is sufficiently homogeneous *conditional on rain*, meaning that F hardly varies throughout the period–some empirical evidence of this is given in Kedem and Pavlopoulos (1991). But p certainly varies in $[0,1]$ as is evident from Figure 1 with $\tau = 0$: some snapshots are devoid of rainfall while quite a few are more than 50% covered with precipitation. Thus for a sufficiently large threshold τ, for a snapshot at time t, (1) is approximated by

$$\overline{X}_t \approx b\{[1 - F(\tau)]\hat{p}_t\} = \{b[1 - F(\tau)]\}\hat{p}_t \tag{2}$$

That is, the pairs (FA_t, AA_t) tend to fall along a straight line as $\hat{p}_t$ varies from snapshot to snapshot, while F does not vary.

2.1 Optimal Thresholds

The requirement that b and hence τ be relatively high means $\tau > 0$. However, the threshold cannot be too large since then the fractional area is mostly zero across all snapshots regardless of the area average, leading to degraded correlation. Figure 1 suggests $\tau = 3$ is a reasonable threshold as it maximizes the correlation between AA and FA. It is interesting to note the same conclusion is reached by a goodness of fit procedure discussed next.

3 The Fractional Area as a Covariate

The close connection between the area average and the fractional area also reveals itself in categorical prediction in time of the first using the second as a covariate.

Let AA_t denote the area average rain rate in the TOGA/COARE data at time t, and let AA_t be categorized in m categories. Define a longitudinal random variable Y_t by,

$$Y_t = j \text{ if the j'th category is observed at time t}, \quad j = 1, \ldots, m$$

A widely used model is the multinomial logits model defined as,

$$P[Y_t = j \mid \mathcal{F}_{t-1}] = \frac{\exp(\beta_j' z_{t-1})}{1 + \sum_{i=1}^{q} \exp(\beta_i' z_{t-1})}, \quad j = 1, \ldots, q = m - 1 \tag{3}$$

where z_t is a vector of random covariates. Let

$$z_t = (FA(\tau)_t, FA(\tau)_{t-3})'$$

where $FA(\tau)_t$ is the fractional area at time t corresponding to threshold τ. As in Figure 1, attention is restricted to $\tau = 0, 1, 3, 5, 7, 9$.

For $m = 3$, we define

$$Y_t = \begin{cases} 1 & \text{if } 0 \leq AA_t < .004 \\ 2 & \text{if } .004 \leq AA_t < .2 \\ 3 & \text{if } AA_t \geq .2 \end{cases} \tag{4}$$

The results of partial likelihood (PL) analysis of this case are reported in Table 1, where χ_4^2 refers to a goodness of fit statistic described in Slud and Kedem (1994) and Fokianos and Kedem (1995).

Table 1. Three categories: Multinomial logits model diagnostics.

τ	$-2 \log \mathrm{PL}$	χ_4^2	Probabilities of Misclassification			
			1st Cat.	2nd Cat.	3rd Cat.	Total
0	1040.26	140.61	6.9%	39.1%	52.4%	30.8%
1	408.27	13.99	3.8%	4.4%	29.8%	8.9%
3	197.99	1.97	5.1%	5.7%	6.1%	5.2%
5	242.28	13.31	3.8%	8%	3.9%	5.8%
7	308.98	3.45	3.9%	7.8%	4.5%	5.9%
9	392.43	7.28	6.4%	12.1%	5.5%	9.2%

As in Figure 1, $\tau = 3$ gives the best fit, though it does not minimize the probability of misclassification.

The same analysis was repeated with 4 categories,

$$Y_t = \begin{cases} 1 & \text{if } 0 \leq AA_t < .004 \\ 2 & \text{if } .002 \leq AA_t < 0.03 \\ 3 & \text{if } 0.03 \leq AA_t < .2 \\ 3 & \text{if } AA_t \geq .2 \end{cases} \tag{5}$$

Table 2 gives the corresponding diagnostics. Once again, $\tau = 3$ gives the best fit. Similar results were obtained for other categorical models including proportional odds.

Table 2. Four categories: Multinomial logits model diagnostics.

τ	$-2\log \mathrm{PL}$	χ^2_6	Probabilities of Misclassification				
			1st Cat.	2nd Cat.	3rd Cat.	4th Cat.	Total
0	1492.85	258.36	1.5%	93.6%	53.9%	49.8%	43.7%
1	547.75	23.69	3.7%	14.6%	9.6%	29.4%	12.2%
3	269.28	1.09	4.6%	20.7%	5.2%	6.1%	7.8%
5	367.81	13.58	3.4%	30.6%	8.3%	3.9%	9.75%
7	448.24	15.49	3.7%	31.0%	8.8%	4.5%	10.1%
9	496.14	16.13	4.6%	37.8%	11.2%	5.5%	12.5%

4 Exceedances in Time Series

We end by noting our discussion regarding the high correlation between the area average and the fractional area carries over to time series as well. Figure average and the proportion exceeding a threshold in a rain rate time series obtained from 25 time series from TOGA/COARE. Each time series is a 2 km by 2 km area average rain rate observed every 10 minutes, and is over 3020 long (about 3 weeks). The proportion of 0's ranged from 76.7% to 78.7%. The correlation between the average and the proportion exceeding a threshold is maximized at 0.83 for threshold $\tau = 20$. This corresponds roughly to an exceedance probability of about 0.001. That is, 3 or 4 exceedances determine the average of the whole time series to a fair degree.

In this connection Barnett et al. (1995) show in Gaussian and monotonically transformed Gaussian stationary processes the variance can be estimated from the magnitude of high peaks exceeding a high threshold. Examples are given where this gives a smaller mean square error than the commonly used sum of squares estimator.

5 Appendix: The TOGA/COARE Data

Radar rainfall measurements over the western Pacific warm pool were collected by two shipboard-Doppler-radars as part of the Tropical Oceans Global Atmosphere (TOGA) Coupled Ocean Atmosphere Response Experiment (COARE) during the Intensive Observing Period (IOP), November 1992 - February 1993. The MIT and TOGA radars were carried by the Research Vessel (R/V) John V. Vickers

(U.S.A.) and R/V Xiang Yang-Hong #5 (People's Republic of China) in a series of three cruises during which the ships maintained station with the Intensive Flux Array (IFA), near 2°S, 156°E. Merged and single radar rainfields having a spatial resolution of 2 km × 2 km and a temporal resolution of 10 minutes have been produced from these observations. In order to obtain a constant sized averaging domain under a wide variety of meteorological conditions, the gridded rainfall data used in this study are from the MIT radar, covering a circle of diameter 290 km, during cruise 3, January 29 to February 23, 1993. For more details see Short et al. (1996).

References

Atlas, D., D. Rosenfeld, and D.A. Short (1990), "The estimation of convective rainfall by area integrals 1. The theoretical and empirical basis," *Journal of Grophysical Research*, 95, No. D3, 2153-2160.

Barnett, J., R. Clough, and B. Kedem (1995), "Power considerations in acoustic emission," *Jour. Acoust. Soc. Amer.*, 98, pp. 2071-2081.

Braud, I., J. D. Creutin, and C. Barancourt (1993), "The relationship between the mean areal rainfall and the fractional area where it rains above a given threshold," *Jour. Appl. Meteorology.*, 32, pp. 193-202.

Fokianos, K. and B. Kedem (1995), " Partial likelihood analysis of categorical time series models." Report T.R. 95-86, Institute for Systems Research, Univ. of Maryland, College Park.

Kedem, B., L.S. Chiu and Z. Karni (1990), "An analysis of the threshold method for measuring area average rainfall," *Jour. Appl. Meteorology*, 29, pp. 3-20.

Kedem, B. and H. Pavlopoulos (1991), "On the threshold method for rainfall estimation: Choosing the optimal threshold level," *Jour. Amer. Stat. Assoc.*, 86, pp. 626-633.

Mase, S. (1994), "On the threshold method for estimating total rainfall." To appear.

Morrissey, M.L. (1994), "The effect of data resolution on the area threshold method," *Jour. Appl. Meteorology*, 33, pp. 1263-1270.

Rosenfeld, D., D. Atlas, and D. Short (1990), "The estimation of convective rainfall by area integrals 2. The height-area threshold (HART) method," *Journal of Geophysical Research*, 95, No. D3, pp. 2161-2176.

Shimizu, K., D.A. Short, and B. Kedem (1993), "Single and double threshold methods for estimating the variance of area rain rate," *Jour. Meteor. Soc. Japan*, 71, pp. 673-683.

Short, D.A., D.B Wolff, D. Rosenfeld and D. Atlas (1993), "A study of the threshold method utilizing rain-gauge data," *J. Appl. Meteor.*, 32, pp. 1379-1387.

Short, D. A., P. A. Kucera, B. S. Ferrier, J. C. Gerlach, S. A. Rutledge and O. W. Thiele (1996), "Shipboard Radar Rainfall Patterns Within the TOGA/COARE IFA". To be submitted to *Bulletin of the American Meteorological Society.*

Slud, E. and B. Kedem (1994), "Partial likelihood analysis of logistic regression and autoregression," *Statistica Sinica*, 4, pp. 89-106.

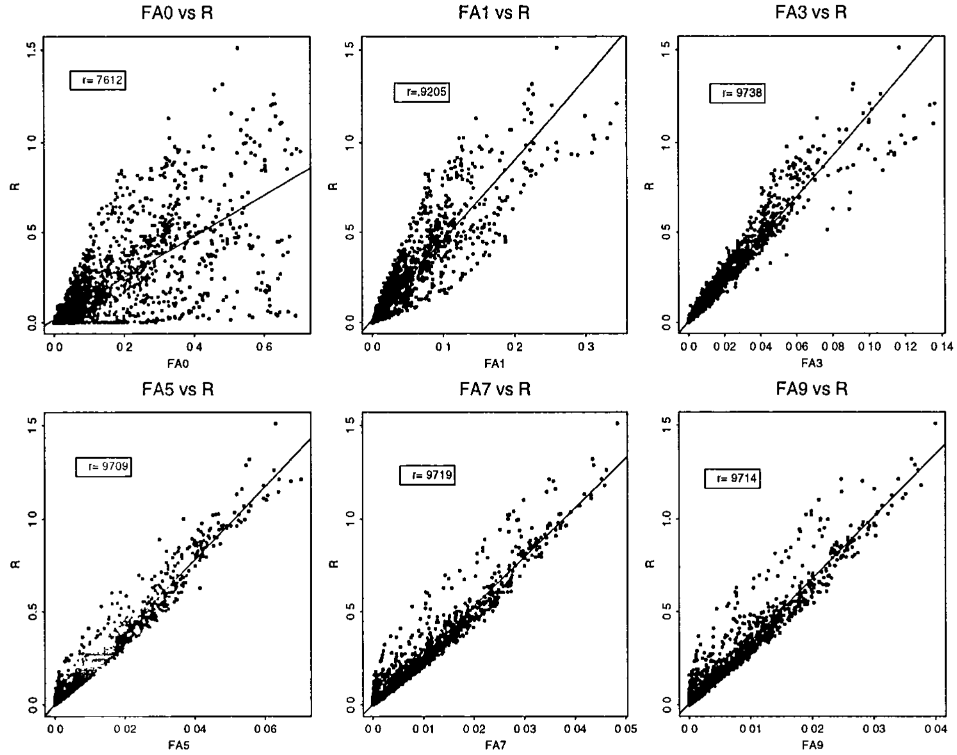

Figure 1: Area average rain rate versus fractional area for TOGA/COARE data. $R = AA$.

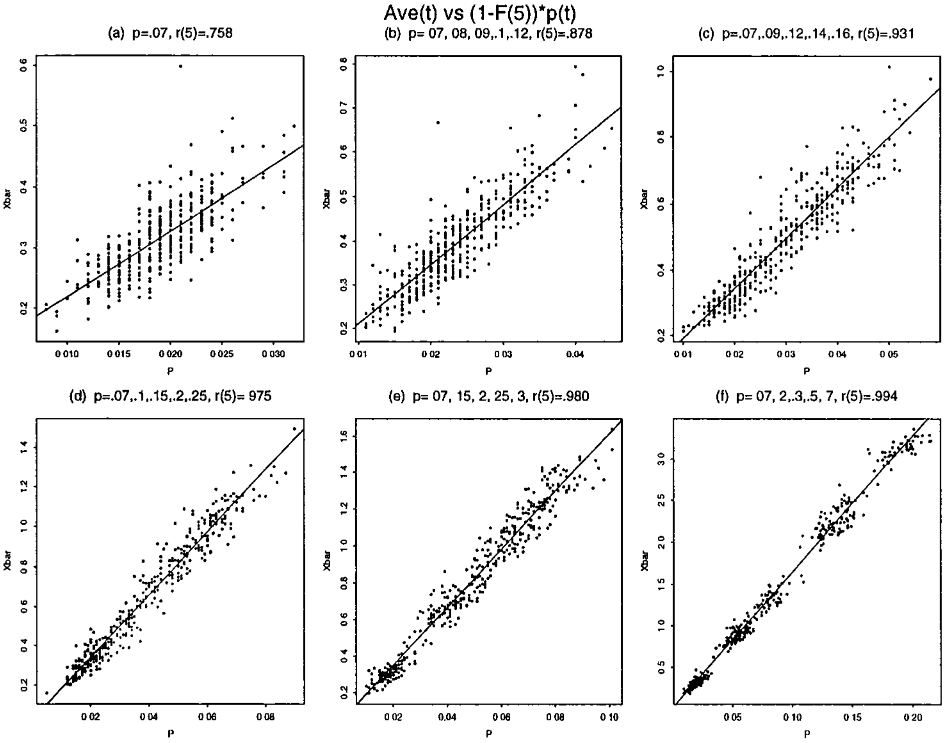

Figure 2: $\overline{X}_i$ vs $(1-F(5))\hat{p}_i$, from 400 mixed lognormal samples of size 1000 each. In (a) $p = 0.07$ is constant. In (b)-(f) p varies

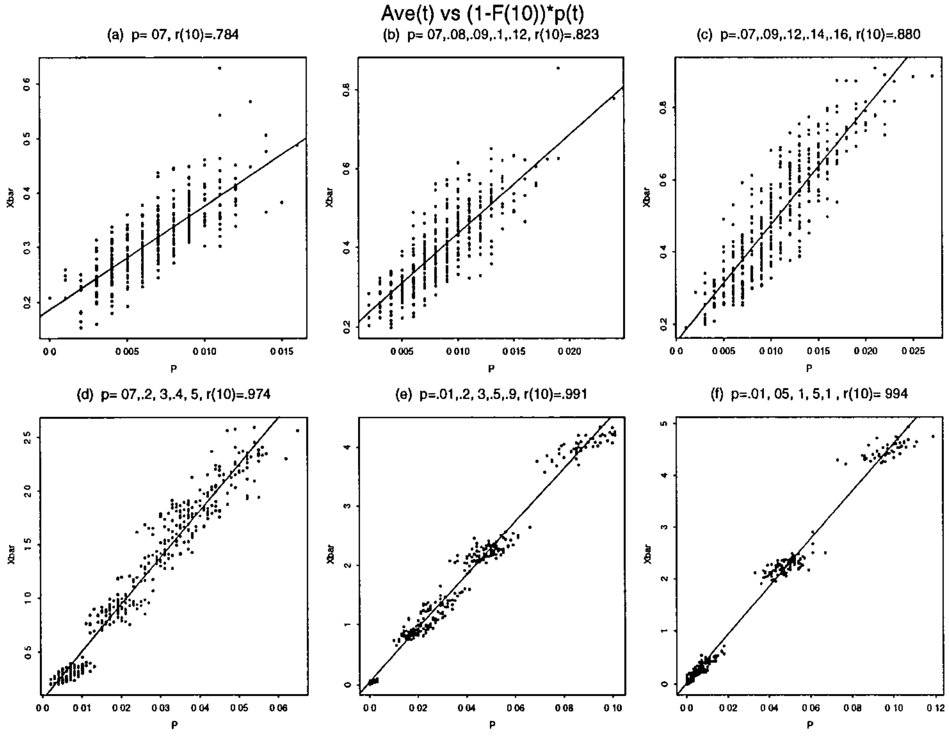

Figure 3: $\overline{X}_i$ vs $(1 - F(10))\hat{p}_i$, from 400 mixed lognormal samples of size 1000 each. In (a) $p = 0.07$ is constant. In (b)-(f) p varies.

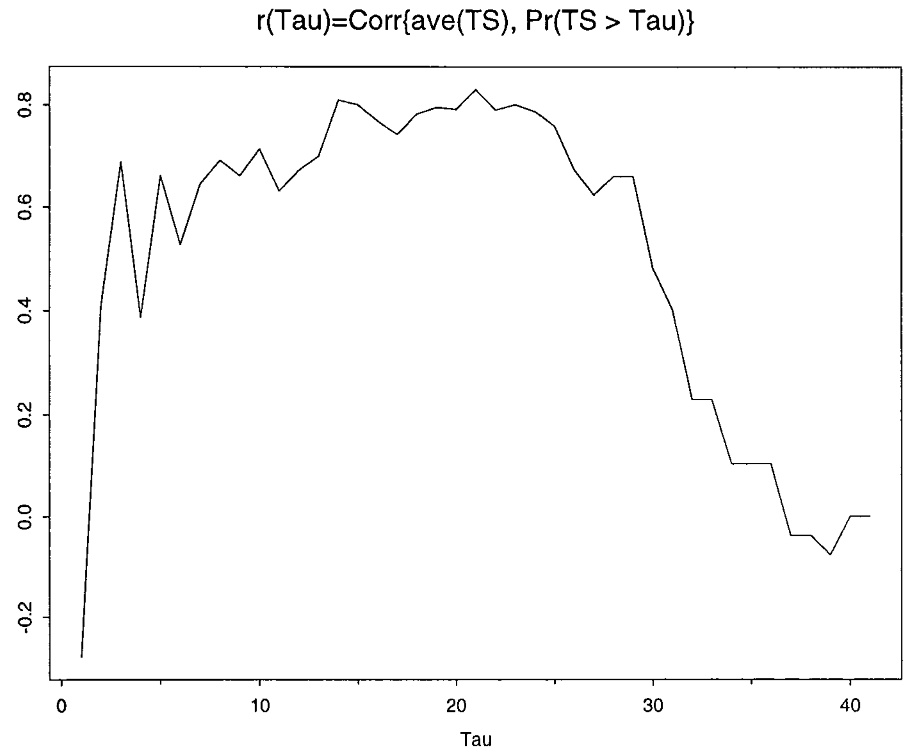

Figure 4: Time series average versus an estimated probability of exceeding τ.