AN ANALYSIS OF DAILY PEAK STREAM DISCHARGE USING A NON-GAUSSIAN TIME SERIES MODEL

S. Perera

AN ANALYSIS OF DAILY PEAK STREAM DISCHARGE USING A NON-GAUSSIAN TIME SERIES MODEL

S. Perera
Assistant Professor
Center on Aging and Department of Preventive Medicine
University of Kansas Medical Center
3901 Rainbow Boulevard
Kansas City, Kansas 66160-7117

ABSTRACT

Daily peak stream discharge data, collected over time, are typically characterized by a few large peaks separated by runs of small values, where peaks correspond to the occurrence of storms. Furthermore, the peak discharge on the first day of a storm has little or no relationship to the previous day's discharge. These characteristics are not present in standard Gaussian time series models in which a zig-zag behavior not conducive to runs of small values is observed and the present value always depends on the previous value. However, they can be successfully captured with non-Gaussian time series models.

Daily peak stream discharge between 1926 and 1953 of Kaukonahua Stream, Hawaii is analyzed using a new exponential autoregressive (NEAR) time series model. The distribution of the length of contiguous periods in which the stream discharge stays below a fixed percentage of the average is estimated. This estimate is shown to be closer to the actual distribution than that obtained using standard Gaussian time series models, with data from the same stream obtained during two disjoint time periods 1926-1952 and 1960-1996 .

Keywords: stochastic modeling of streamflow; NEAR; exponential autoregressive; runs; non-Gaussian time series.

1. INTRODUCTION

The daily peak discharge (in ft^3/sec) of Kaukonahua Stream, Hawaii, during the period 1926 to 1953 is presented in Figure 1. This type of data resulting from a natural phenomenon has special features such as being bounded below by zero and a heavily positively skewed distribution caused by the presence of a few very high peaks separated by runs of relatively small values. These characteristics make the shape of the distribution of such data very different from that of a Gaussian distribution. Furthermore, transformation to normality, using a logarithmic or a similar transformation, can be difficult due to a large number of possible zero or near-zero values. These special features of this type of data suggest that they may have been generated from a mechanism based on an exponential or a similar distribution. Therefore, the possibility of using a non-Gaussian model, perhaps one based on the exponential distribution, should be explored (see Perera, 2000).

Some of the non-Gaussian time series models that have appeared in the literature include the product autoregressive (PAR) model proposed by McKenzie (1982), mixed exponential autoregressive (MEAR) model and a mixed exponential moving-average model by Jevremović (1990), the NMEAR(1) model by Lawrance and Lewis (1982), the AREX(1) model by Mališić (1987), the exponential moving average model (EMA) and the exponential autoregressive moving average model (EARMA) by Lawrance and Lewis (1980), and the new exponential autoregressive models (NEAR) by Lawrance (1981). MEAR(1) and MEAR(1) models are special cases of the AREX(1) model. The exponential autoregressive (EAR) model discussed by Billard and Mohamed (1991) and Sim (1987) is a special case of the NEAR(1) model. All of these models are based on the exponential distribution, and except for NEAR models, little follow up work on parameter estimation or assessing fit has appeared in the literature.

The NEAR models were discussed in detail and applied to a series of wind velocity data by Lawrance and Lewis (1985). There are many difficulties in using NEAR models with data, including parameter estimation and assessing fit, as indicated by Raftery (1985), Rao (1985), Chatfield (1985), and others. However, the work by Karlen and Tjøstheim (1988), and Smith (1986) has eliminated some of the difficulties in parameter estimation.

In the present paper, a non-Gaussian model is used to model the data presented in Figure 1, and estimate the distribution of length of a contiguous period of low stream discharge. A NEAR models is chosen, due to the availability of parameter estimation methods. Also, similar models based on the exponential distribution have been used Lewis and Hugus (1982) to model river-flow. However, the above reference did not consider estimation of the distribution of related quantities of interest.

The NEAR(1) model is given by:

$$X_t = \begin{cases} \beta X_{t-1} & \text{w.p. } \alpha \\ 0 & \text{w.p. } 1-\alpha \end{cases} + \epsilon_t, \tag{1.1}$$

where $\{\epsilon_t\}$ is the residual sequence defined as

$$\epsilon_t = \begin{cases} E_t & \text{w.p. } p \\ bE_t & \text{w.p. } 1-p \end{cases}. \tag{1.2}$$

Here, $\{E_t\}$ is an independent and identically distributed sequence of standard exponential variates, $p=(1-\beta)/\{1-(1-\alpha)\beta\}$, and $b=(1-\alpha)\beta$. Chan (1988) showed that the conditions $0\le\alpha\le1$, $0\le\beta\le1$, and $\alpha\beta<1$ are both necessary and sufficient for the existence of a stationary and ergodic NEAR(1) process. These conditions also define the parameter space Ω for a NEAR(1) process. The marginal distribution of X_t is standard exponential, and the autocorrelation structure of a NEAR(1) model is identical to that of an AR(1) Gaussian autoregressive process given by $Z_t=aZ_{t-1}+e_t$, where $e_t \sim N(0,\sigma^2)$, $a=\alpha\beta$. Thus the NEAR(1) model establishes a dependence structure among exponential random variables analogous to that in the AR(1) model for Gaussian random variables. These properties make the NEAR(1) model a good alternative to modeling heavily skewed data such as that presented in Figure 1.

As seen from (1.1), another characteristic of the NEAR processes is that the current value of the time series is allowed to depend on the previous observation only with some probability less than unity, whereas in Gaussian ARMA processes, the current value always partly depend on the previous observation. This characteristic makes NEAR models particularly appealing in situations similar to the present, because the peak stream discharge after a storm has little or no dependence on the previous day's peak discharge. On other days, the peak discharge presumably depends on the previous day's peak discharge to a larger extent.

In Section 2, the deseasonalization of data, required before fitting a stationary time series model, is performed. In Sections 2 and 3, a discussion about selecting, fitting and estimation of NEAR and ARMA models is given. Section 4 compares estimates of distributions obtained using the two approaches, and shows that estimates obtained using a NEAR model is more accurate when compared to both the observed data set, and another from the same stream obtained during 1960-1996. Section 5 includes a summary.

2. DESEASONALIZATION OF DATA

Strong seasonal components with periods 365 and 122 days can be easily identified from the tall peaks in the periodogram $I_n(\lambda)=n^{-1}|\Sigma_{t=1}^{n} \log(Y_t) e^{-it/\lambda}|^2$ of log-transformed data in Figure 2, where the sequence $\{Y_t\}$ denotes the original data. These two seasonal components can be removed by fitting the regression model,

$$Y_t=\exp\{\beta_0+\beta_1\sin(2\pi t/365)+\beta_2\cos(2\pi t/365)+\beta_3\sin(2\pi t/122)+\beta_4\cos(2\pi t/122)\} \cdot \epsilon_t$$

For this data set, the least squares estimates were $\hat{\beta}_0=1.998$, $\hat{\beta}_1=0.148$, $\hat{\beta}_2=0.325$, $\hat{\beta}_3=0.146$ and $\hat{\beta}_4=0.081$. The estimated residual sequence,

$$\hat{\epsilon}_t=\log Y_t-\hat{\beta}_0-\hat{\beta}_1\sin(2\pi t/365)-\hat{\beta}_2\cos(2\pi t/365)-\hat{\beta}_3\sin(2\pi t/122)-\hat{\beta}_4\cos(2\pi t/122),$$

referred to as deseasonalized stream data, is void of any seasonal components and therefore suitable for stationary time series modeling.

3. FITTING A NEAR MODEL

Since NEAR models have standard exponential marginal distributions, the data must be first power transformed to make the mean and the standard deviation equal, and then re-scaled to make them equal to unity. For this data set, the transformation $x=\hat{\epsilon}^{0.635}/1.357$ is appropriate, and the selection of this transformation was data-driven. Figures 3 and 4 respectively contain the autocorrelation and partial autocorrelation plots of power transformed deseasonalized data. The tailing off in the autocorrelation plot and the sudden drop at lag 1 in the partial autocorrelation plot suggest an AR(1) autocorrelation structure, and therefore a NEAR(1) model was selected. Using the two-stage conditional least squares estimation method proposed by Nicholls and Quinn (1982) for random coefficient autoregressive models and used by Karlsen and Tjøstheim (1988) for NEAR(2) models, the parameter estimates of the NEAR(1) model were obtained to be $\hat{\alpha}=0.749$ and $\hat{\beta}=0.739$. Since the true innovation sequence $\{\epsilon_t\}$ defined in (1.2) cannot be estimated even with the knowledge of the true values of parameters, model checking was based on the AR(1)-type residuals, as was done by Lawrance and Lewis (1985). The AR(1)-type

residuals $R_t=X_t-\alpha\beta X_{t-1}$ are the conditional expectations of the true residuals ϵ_t and are estimated by $\hat{R}_t=X_t-\hat{\alpha}\hat{\beta}X_{t-1}$. The uncorrelated nature of AR(1)-type residuals, as evidenced by the autocorrelation plot in Figure 5, indicates a satisfactory fit.

3. FITTING AN ARMA MODEL

A transformation suggested by Box and Cox (1964), given by

$$x=\begin{cases}(\epsilon^{\lambda}-1)/\lambda & \text{if } \lambda\neq 0\\ \log(\epsilon) & \text{if } \lambda=0\end{cases}$$

is typically applied to the data before fitting a Gaussian model, where the parameter λ is chosen by using a data-driven mechanism so that the transformed data is approximately normally distributed. For this data set, $\lambda=-0.025$ was chosen since it makes the skewness of the transformed data approximately zero and the normal probability plot almost linear along the diagonal. The autocorrelation plot of Box-Cox transformed deseasonalized stream data in Figure 6 gradually tails off and the partial autocorrelation plot in Figure 7 has a sudden drop at lag 1, suggesting that an AR(1) model may be appropriate. Maximum likelihood estimation yields $\hat{\mu}=-0.079$, $\hat{\phi}=0.754$, and $\hat{\sigma}^2=0.260$ respectively for the overall mean, serial correlation, and variance of the innovation sequence. The autocorrelation plot of the residuals in Figure 8 indicates a satisfactory fit.

4. ESTIMATING THE DISTRIBUTION OF LENGTHS OF DROUGHTS

A drought can be interpreted as a contiguous run of days in which the stream discharge stays below an arbitrary but fixed percentage (e.g. 40% or 20%) of the average for the season under consideration, and it corresponds to a run of values in the deseasonalized series which are less than the same percentage. The distribution of lengths of droughts (runs below 40% of the average) is approximated via simulation for both AR(1) and NEAR(1) processes using parameter combinations similar to those estimated using transformed deseasonalized data from the period 1926-1953. This approximation is then compared to the drought length distribution from the same stream for the "future" period 1960-1996.

Figure 9 presents a comparison of

(i) a simulated distribution of the lengths of droughts obtained by simulating an AR(1) process of length 1,000,000 defined by $X_t-0.754X_{t-1}=-0.079+Z_t$, where $\{Z_t\}$ is a sequence of independently and identically distributed variates with mean 0 and variance 0.260 (dotted line),

(ii) a simulated distribution of the lengths of droughts obtained by simulating a NEAR(1) process of length 1,000,000 defined by (1.1) and (1.2), where $\alpha=0.749$ and $\beta=0.739$ (broken line),

(iii) observed distribution of lengths of droughts during the period 1926-1952 (thick solid line), and

(iv) the distribution of lengths of droughts observed at the same location of the same stream

during the "future" time period 1960-1996 (thin solid line).

The distribution of lengths of droughts appear to have changed during the later period compared to the earlier period, as indicated by (iii) and (iv). However, both of the observed distributions (iii) and (iv) are closer to distribution (ii) than they are to distribution (i), indicating that in this situation, estimate of the distribution of lengths of droughts obtained by using a NEAR(1) model is more accurate compared to those obtained using an AR(1) model. Moreover, curve (ii) intersects with curves (iii) and (iv), indicating that average length of a drought estimated using a NEAR(1) model is close to its observed value. In contrast, curve (i) stays completely to the left of (iii), consistently underestimating the length of a drought. This phenomenon is probably due to the zig-zag type behavior of Gaussian ARMA models, which is not conducive to runs of small values. Similar results were obtained when 20% was used to define a drought, instead of 40%.

Table 1 contains the statistics for measuring discrepancy between curves (i) and (iii), (ii) and (iii), (i) and (iv), and (ii) and (iv) in Figure 9, and quantitatively confirms the above observations. Based on absolute difference and squared difference metrics, using a NEAR(1) model in this instance has reduced the "distance" between the estimated and observed distributions of length of drought by 30-57% for the "current" period 1926-1952, and by 13-14% for the "future" period 1960-1996, compared to using a Gaussian AR(1) model.

5. SUMMARY

In order to satisfactorily model certain types of real data, non-Gaussian time series models are needed. Even with a transformation to normality, the Gaussian models may not perform as well as some non-Gaussian models, as illustrated with the present data set. For this particular data set, the NEAR(1) model is shown to fit well and produce a more accurate estimate of the distribution of length of droughts, compared to a Gaussian AR(1) model. This emphasizes that additional difficulties encountered in using these models are sometimes worth the effort.

ACKNOWLEDGEMENTS

I wish to thank Professor Paul Nelson for his guidance. This work was supported in part by NSF-SCREMS grant DMS-9628643.

REFERENCES

Billard, L. and Mohamed, F.Y. (1991). Estimation of the parameters of an EAR(p) process. *Journal of Time Series Analysis*, **12**, 179-192.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society-Series B*, **26**, 211-243.

Chan, Kung-Sik (1988). On the existence of the stationary and ergodic NEAR(p) model. *Journal of Time Series Analysis*, **9**, 319-328.

Chatfield, C. (1985). Discussion after "Modelling and residual analysis of nonlinear

autoregressive time series in exponential variables." *Journal of the Royal Statistical Society-Series B*, **47**, 185-189.

Jevremović, V. (1990) Two examples of nonlinear processes with a mixed exponential marginal distribution. *Statistics & Probability Letters*, **10**, 221-224.

Karlsen, H. and Tjøstheim, D. (1988). Consistent estimates for the NEAR(2) and NLAR(2) time series models. *Journal of the Royal Statistical Society-Series B*, **50**, 313-320.

Lawrance, A. J. (1981). A new autoregressive time series in exponential variables (NEAR(1)). *Advances in Applied Probability*, **13**, 826-845.

Lawrance, A. J. and Lewis, P. A. W. (1980). The exponential autoregressive-moving average EARMA(p, q) process. *Journal of the Royal Statistical Society-Series B*, **42**, 150-161.

Lawrance, A. J. and Lewis, P. A. W. (1982). A mixed exponential time series model. *Management Science, Journal of the Institute of Management Sciences*, **28**, 1045-1053.

Lawrance, A. J. and Lewis, P. A. W. (1985). Modelling and residual analysis of nonlinear autoregressive time series in exponential variables. *Journal of the Royal Statistical Society-Series B*, **47**, 165-202.

Lewis P. A. W. and Hugus, D. K. (1982). An analysis of 15 years of wind velocity data from ship PAPA. In D. K. Hugus' Ph. D. thesis, *Extensions of Some Models for Positive Valued Time Series*. Naval Postgraduate School, Monterey.

Mališić, J. D. (1987). On exponential autoregressive time series models, in: P. Bauer et al., eds., *Mathematical Statistics and Probability Theory, Vol. B*, 147-153.

McKenzie, E. (1982). Product autoregression: a time series characterization of the gamma distribution. *Journal of Applied Probability*, **19**, 463-468.

Nicholls, D. F. and Quinn, B. G. (1982). *Random Coefficient Autoregressive Models: An Introduction*. Heidelberg: Springer-Verlag.

Perera, S. (2000). Ph.D. thesis, *Autoregressive Time Series Models with Exponential Marginal Distributions*. Kansas State University, Manhattan.

Raftery A. E. (1985). Discussion after "Modelling and residual analysis of nonlinear autoregressive time series in exponential variables." *Journal of the Royal Statistical Society-Series B*, **47**, 185-187.

Rao, T. S. (1985). Discussion after "Modelling and residual analysis of nonlinear autoregressive time series in exponential variables." *Journal of the Royal Statistical Society-Series B*, **47**, 187.

Sim, C. H. (1987). A stochastic bivariate process associated with the EAR(1) model. *IEEE Transactions on Information Theory*, **33**, 47-51.

Smith, R. L. (1986). Maximum likelihood estimation for the NEAR(2) model. *Journal of the Royal Statistical Society-Series A*, **48**, 251-257.

Figure 1: Daily peak discharge of Kaukonahua Stream, Hawaii during 1926-1953

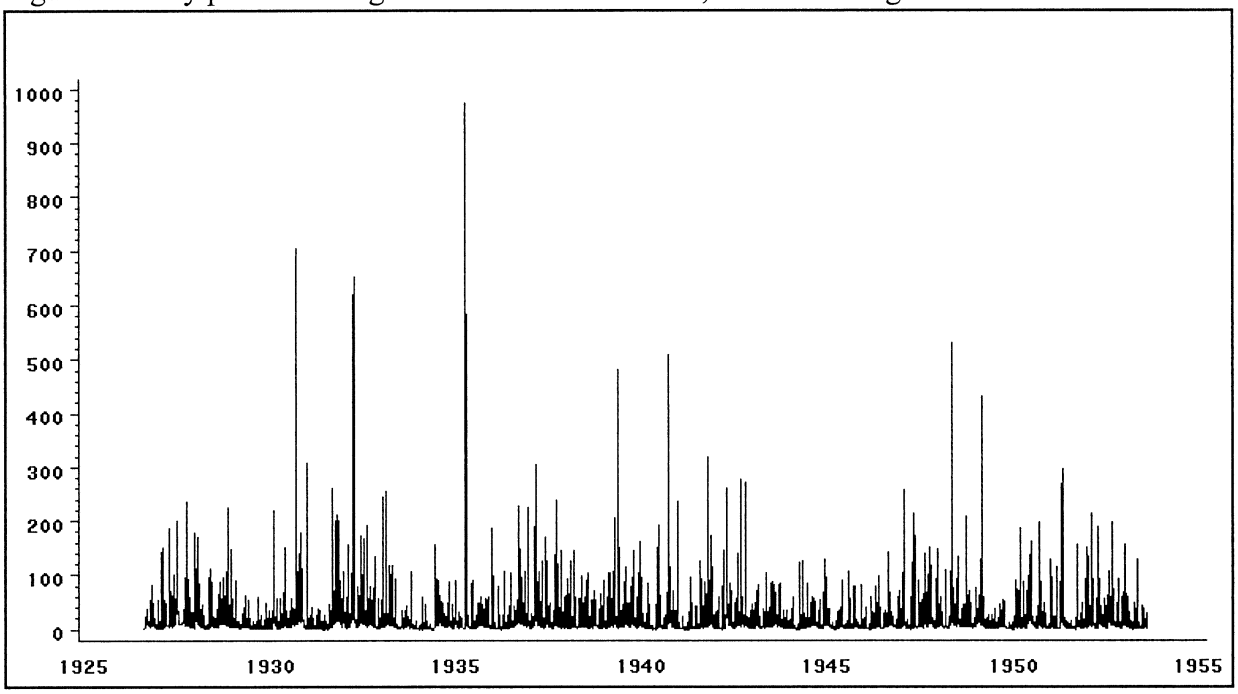

Figure 2: Periodogram of the log-transformed stream data

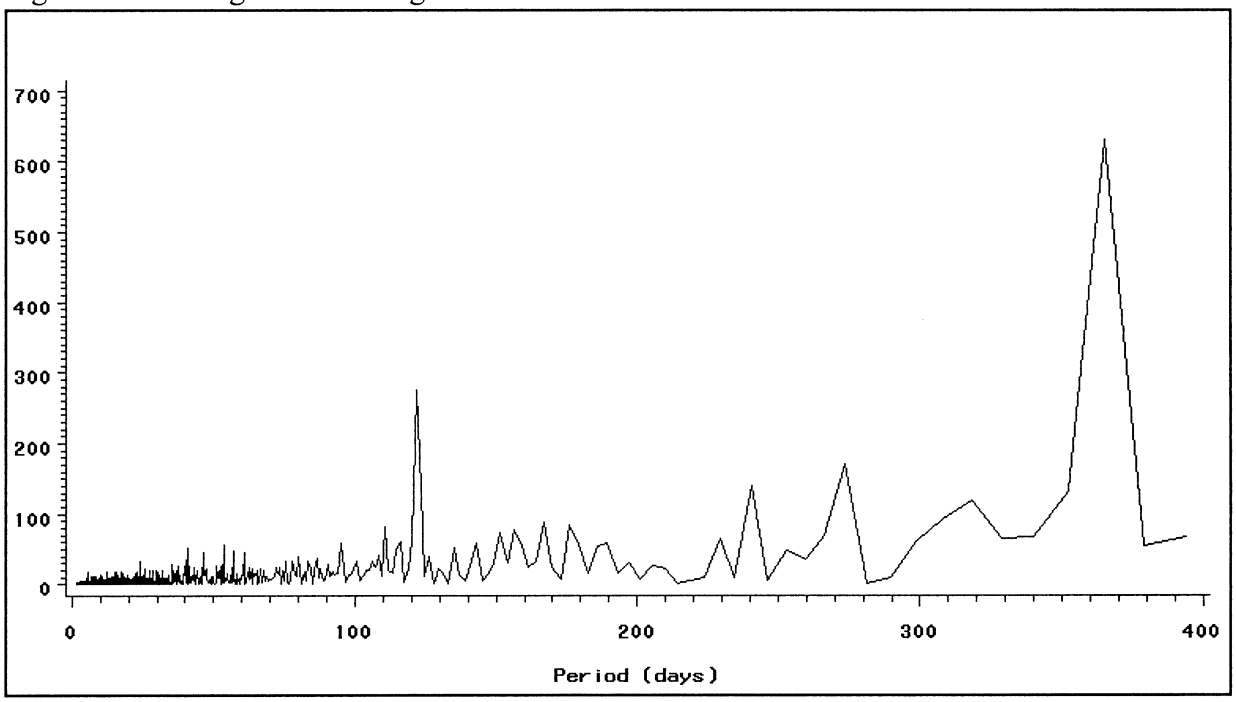

Figure 3: Autocorrelation plot of the power transformed deseasonalized data.

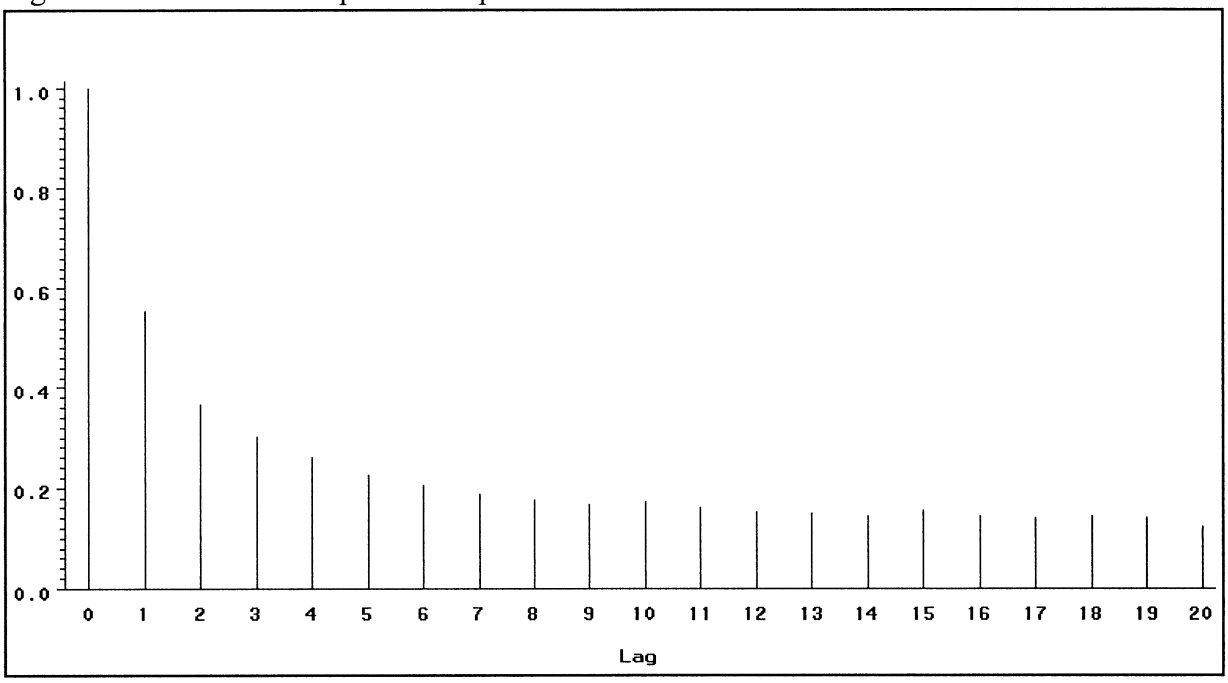

Figure 4: Partial autocorrelation plot of the power transformed deseasonalized data

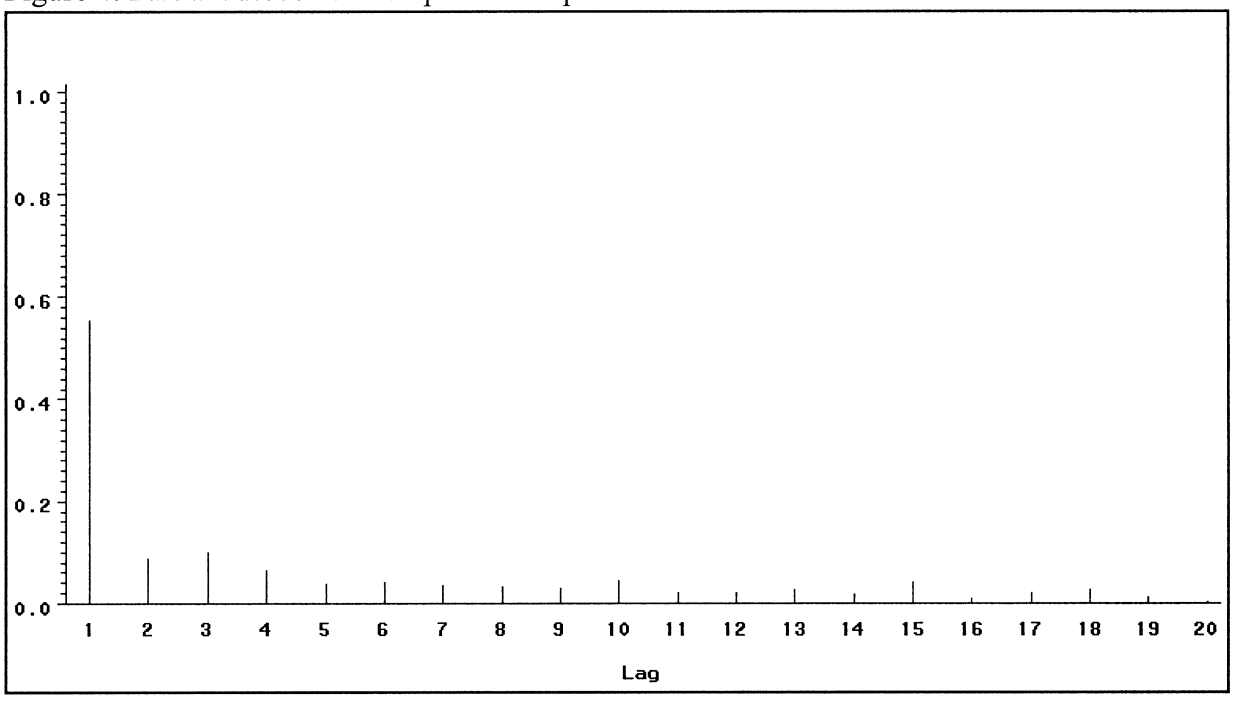

Figure 5: Autocorrelation plot of the AR(1)-type residuals obtained from a NEAR(1) model

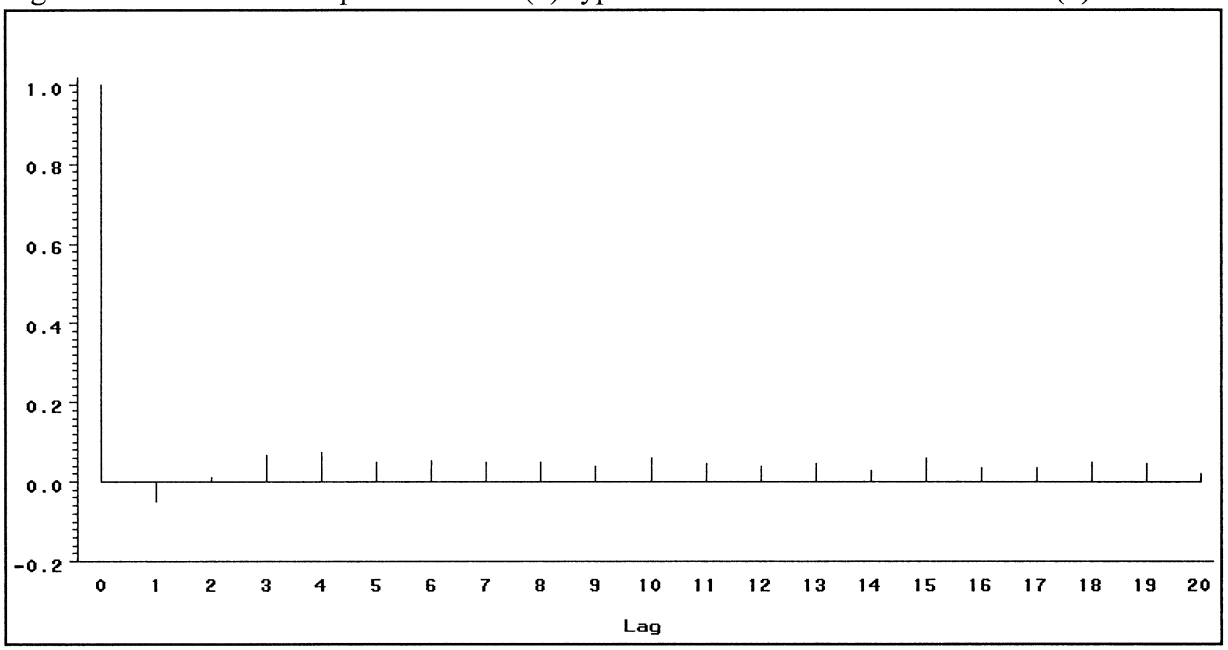

Figure 6: Autocorrelation plot of the Box-Cox transformed deseasonalized data

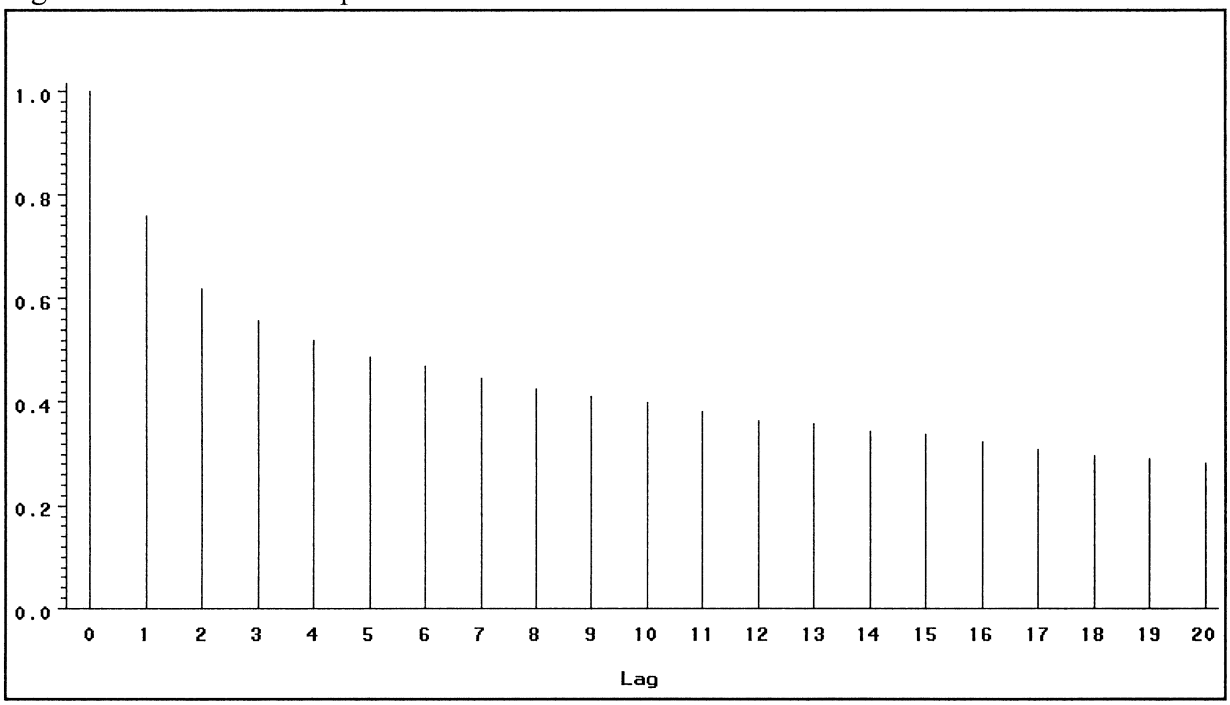

Figure 7: Partial autocorrelation plot of the power transformed deseasonalized data

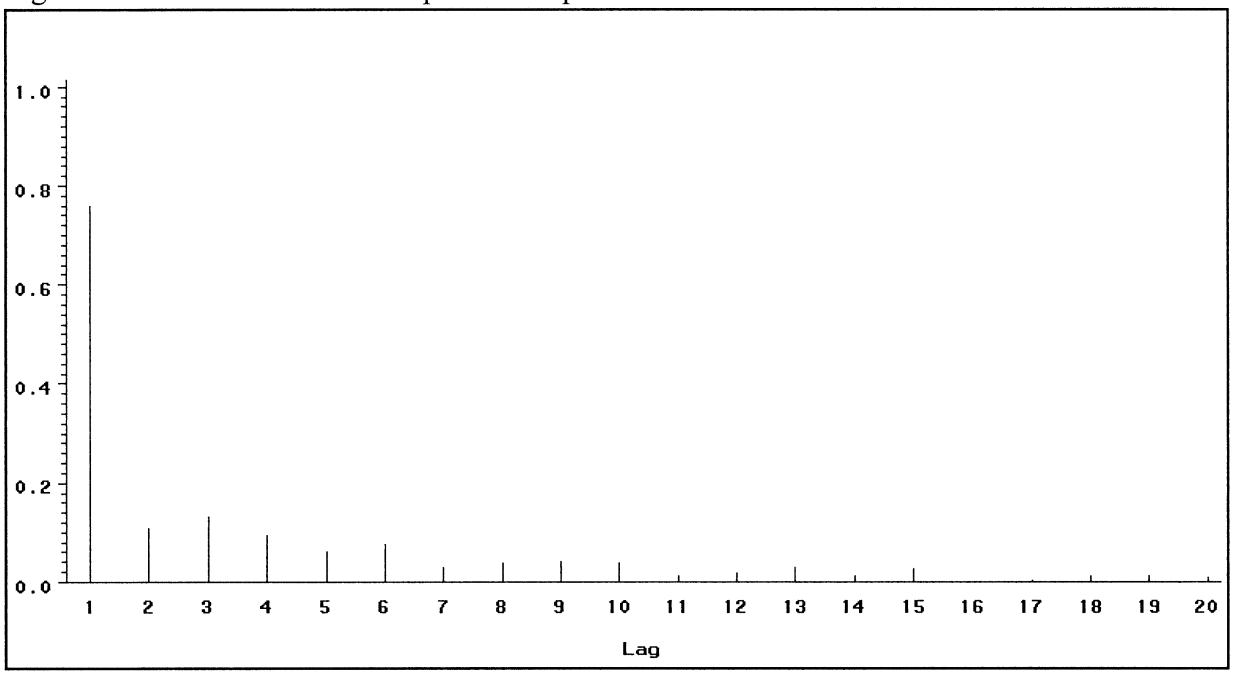

Figure 8: Autocorrelation plot of the residuals obtained from fitting an AR(1) model

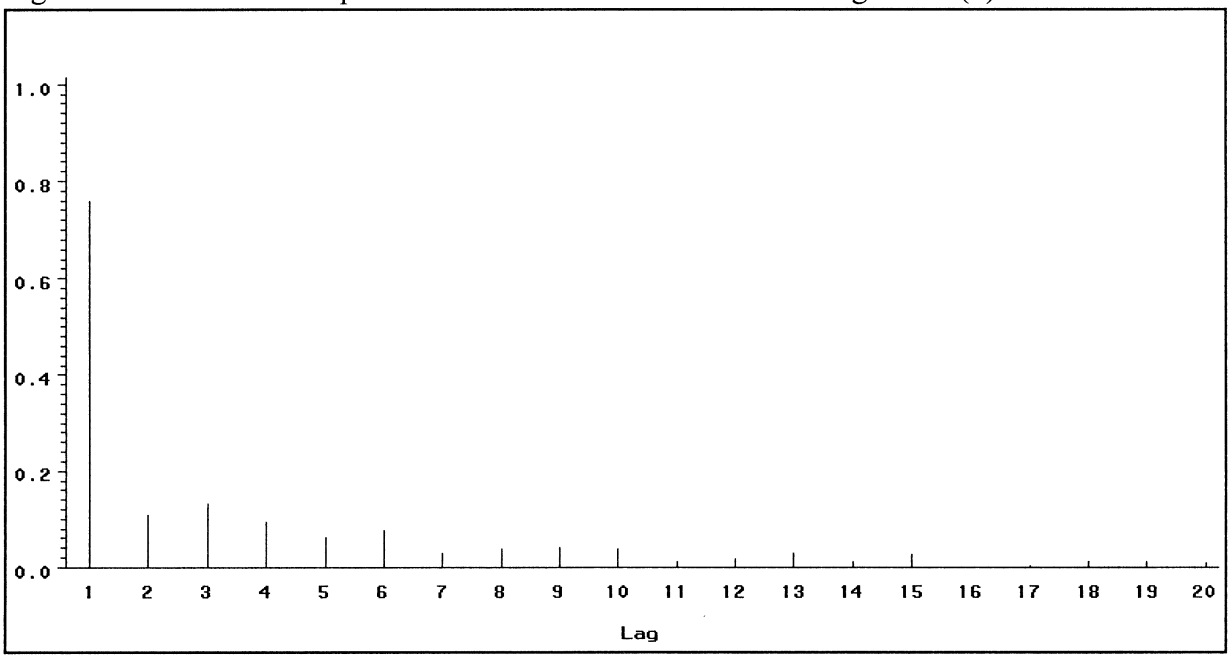

Figure 9: Comparison of cumulative distributions of lengths of droughts (i) estimated using an AR(1) model, (ii) estimated using a NEAR(1) model, (iii) observed during the "current" period 1926-1952, and (iv) observed during the "future" period 1960-1996.

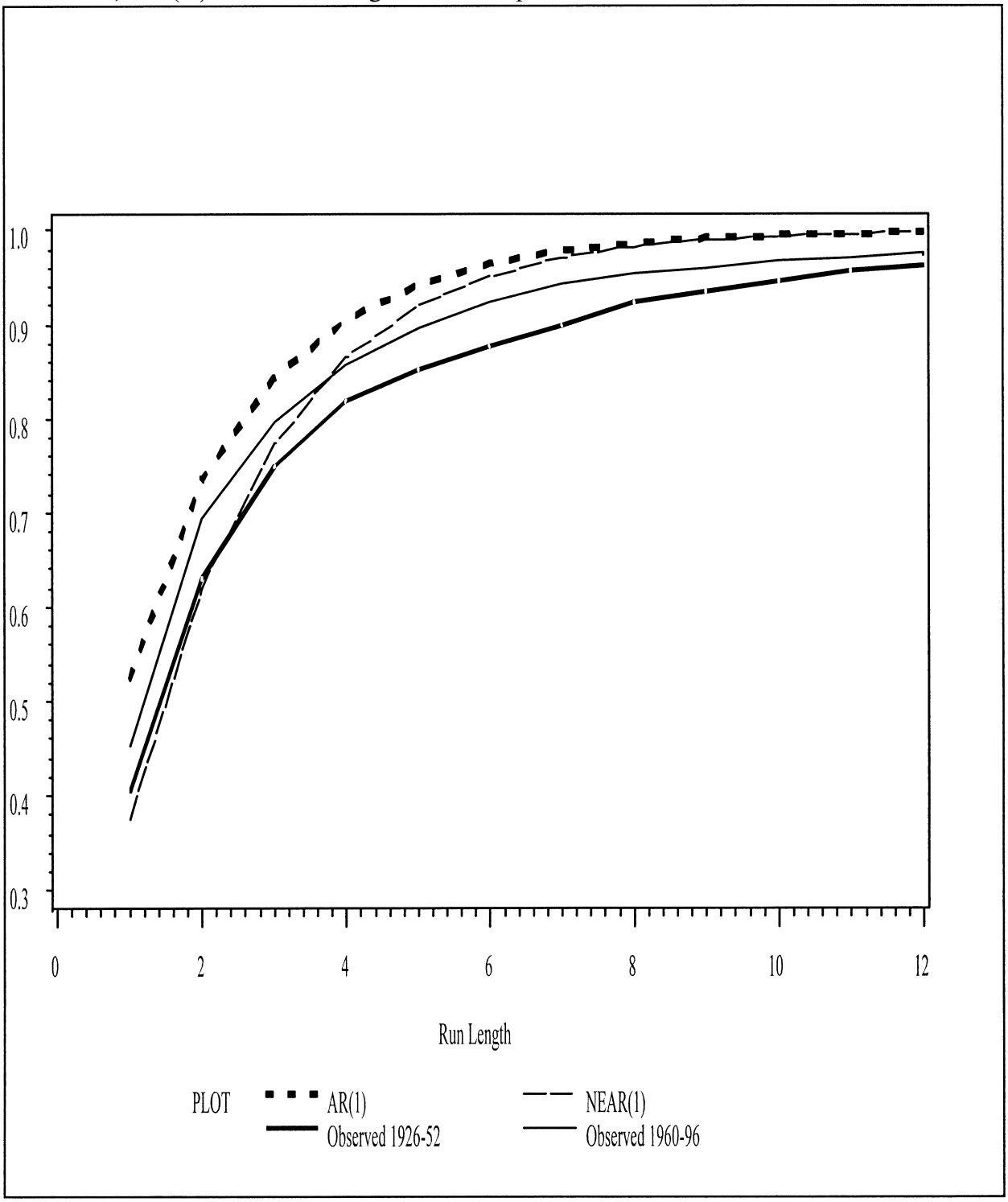

Table 1: Discrepancy statistics between predicted and observed distributions of lengths of droughts

Metric	Between Estimated and Observed During the "Current" Period 1926-1952			Between Estimated and Observed During the "Future" Period 1960-1996		
	AR(1)	NEAR(1)	Reduction Obtained Using NEAR(1)	AR(1)	NEAR(1)	Reduction Obtained Using NEAR(1)
Sum of absolute differences	1.132	0.792	30.0%	0.586	0.510	13.0%
Sum of squared differences	0.079	0.034	57.0%	0.022	0.019	13.6%