WATER RESOURCES
research center

Publication No. 67

ESTIMATING MISSING VALUES IN MONTHLY RAINFALL SERIES

By

EFI FOUFOULA-GEORGIOU

A Thesis Presented to the Graduate Council of
the University of Florida
in Partial Fulfillment of the Requirements for the
Degree of Master of Engineering

University of Florida
Gainesville

UNIVERSITY OF FLORIDA

ESTIMATING MISSING VALUES IN MONTHLY RAINFALL SERIES

By

EFI FOUFOULA-GEORGIOU

Publication No. 67

FLORIDA WATER RESOURCES RESEARCH CENTER

Research Project Technical Completion Report

Sponsored by

South Florida Water Management District

A THESIS PRESENTED TO THE GRADUATE COUNCIL OF
THE UNIVERSITY OF FLORIDA
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF ENGINEERING

UNIVERSITY OF FLORIDA

1982

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to all those who contributed towards making this work possible.

I am particularly indebted to the chairman of my supervisory committee, Professor Wayne C. Huber. Through the many constructive discussions along the course of this research, he provided an invaluable guidance. It was his technical and moral support that brought this work into completion.

I would like to express my sincere appreciation to the other members of my supervisory committee: Professors J. P. Heaney, D. L. Harris, and M. C. K. Yang, for their helpful suggestions and their thoughtful and critical evaluation of this work.

Special thanks are also given to my fellow students and friends, Khlifa, Dave D., Bob, Terrie, Richard, Dave M., and Mike, for their cheerful help and the pleasant environment for work they have created.

Finally my deepest appreciation and love go to my husband, Tryphon, who has been a constant source of encouragement and inspiration for creative work. Many invaluable discussions with him helped a great deal in gaining an understanding of some problems considered in this thesis.

The research was supported in part by the South Florida Water Management District.

Computations were performed at the Northeast Regional Data Center on the University of Florida campus, Gainesville.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table | Page

LIST OF FIGURES

Figure Page

Abstract of Thesis Presented to the Graduate
Council of the University of Florida in Partial
Fulfillment of the Requirements for the Degree of
Master of Engineering

ESTIMATION OF MISSING OBSERVATIONS IN MONTHLY RAINFALL SERIES

By

Efstathia Foufoula-Georgiou

December, 1982

Chairman: Wayne C. Huber
Cochairman: James P. Heaney
Major Department: Environmental Engineering Sciences

This study compares and evaluates different methods for the estimation of missing observations in monthly rainfall series. The estimation methods studied reflect three basic ideas:

(1) the use of regional-statistical information in four simple techniques:
   - mean value method (MV),
   - reciprocal distance method (RD),
   - normal ratio method (NR),
   - modified weighted average method (MWA);

(2) the use of a univariate autoregressive moving average (ARMA) model which describes the time correlation of the series;

(3) the use of a multivariate ARMA model which describes the time and space correlation of the series.

An algorithm for the recursive estimation of the missing values in a series by a parallel updating of the univariate or multivariate ARMA model is proposed and demonstrated. All methods are illustrated in a case study using 55 years of monthly rainfall data from four south Florida stations.

Chairman

CHAPTER 1

INTRODUCTION

Rainfall Records

Rainfall is the source component of the hydrologic cycle. As such it regulates water availability and thus land use, agricultural and urban expansion, maintenance of environmental quality and even population growth and human habitation. As Hamrick (1972) points out, water may be transported for considerable distances from where it fell as rain and may be stored for long periods of time, but with very few exceptions it originates as rainfall. Consequently, the measurement and study of rainfall is in actuality the measurement and study of our potential water supply.

Rainfall studies attempt to derive models, both probabilistic and physical, to describe and forecast the rainfall process. Since the quality of every study is immediately related to the quality of the data used, the need for "good quality" rainfall data has been expressed by all hydrologists. By "good quality" is meant accurate, long and uninterrupted series of rainfall measurements at a range of different time intervals (e.g., hourly, daily, monthly, and yearly data) and for a dense raingage network. Missing values in the series (due, for example, to failure of the recording instruments or to deletion of a station) is a real handicap to the hydrologic data users. The estimation of these missing values is often desirable prior to the use of the data.

For instance, the South Florida Water Management District prepared a magnetic tape with monthly rainfall data for all rainfall stations in south Florida for use in this study (T. MacVicar, SFWMD, personal communication, May, 1982). The data included values for the period of record at each station, ranging from over 100 years (at Key West) to only a few months at several temporary stations. Approximately one month was required to preprocess these data prior to performing routine statistical and time series analyses. The preprocessing included tasks such as manipulations of the magnetic tape, selection of stations with desirable characteristics (e.g., long period of record, proximity to other stations of interest, few missing values) and a major effort at replacement of missing values that did exist. This effort, in fact, was the motivation for this thesis.

Many different kinds of statistical analyses may be performed on a given data set, e.g., determination of elementary statistical parameters, auto- and cross-correlation analysis, spectral analysis, frequency analysis, fitting time series models. For routine statistics (e.g., calculation of mean, variance and skewness) missing values are seldom a problem. But for techniques as common as autocorrelation and spectral analysis missing values can cause difficulties. In multivariate analysis missing values result in "wasted information" when only the overlapping period of the series can be used in the analysis, and in inconsistencies (Fiering, 1968, and Chapter 4 of this thesis) when the incomplete series are used.

In general, two approaches to the problem of missing observations exist. The first consists of developing methods of analysis that use only the available data, the second in developing methods of estimation of the missing observations followed by application of classical methods of analysis.

Monthly rainfall totals are usually calculated as the sum of daily recorded values. Thus, if one or more daily observations are missing the monthly total is not reported for that month. An investigation conducted by the Weather Bureau in 1950 (Paulhus and Kohler, 1952), showed that almost one third of the stations for which monthly and yearly totals were not published had only a few (less than five) days missing. Furthermore, for some of these missing days there was apparently no rainfall in the area as concluded by the rainfall observations at nearby stations. Therefore, in many cases estimation of a few missing daily rainfall values can provide a means for the estimation of the monthly totals.

Statisticians have been most concerned with the problem of handling short record multivariate data with missing observations in some or all of the variables, but no explicit and simple solutions have been given, apart from a few special cases in which the missing data follow certain patterns. A review of these methods is given by Afifi and Elashoff (1956). In the time domain, "the analysis of time series, when missing observations occur has not received a great deal of attention" as Marshall (1980, p. 567) comments, and he proposes a method for the estimation of the autocorrelations using only the observed values. Jones (1980) attempts to fit an ARMA model to a stationary time series which has missing observations using Akaike's Markovian representation and Kalman's recursive algorithm. In the frequency domain, spectral analysis with randomly missing observations has been examined by Jones (1962), Parzen (1963), Scheinok (1965), Neave (1970) and Bloomfield (1970).

In hydrology, the problem of missing observations has not been studied much as Salas *et al*. (1980) state:

> The filling-in or extension of a data series is a topic which has not received a great deal of attention either in this book or elsewhere. Because of its importance, the subject is expected to be paid more attention in the future. (Salas *et al*., 1980, p. 464)

Simple and "practicable" methods for the estimation of missing rainfall values for large scale application were proposed by Paulhus and Kohler (1952), for the completion of the rainfall data published by the Weather Bureau. The study was initiated after numerous requests of the climatological data users. Beard (1973) adopted a multisite stochastic generation technique to fill-in missing streamflow data, and Kottegoda and Elgy (1977) compared a weighted average scheme and a multivariate method for the estimation of missing data in monthly flow series. Hashino (1977) introduced the "concept of similar storm" for the estimation of missing rainfall sequences. Although the same methods of estimation can be applied to both rainfall and runoff series, a specific method is not expected to perform equally well when applied to the two different series due mainly to the different underlying processes. This is true even for rainfall series from different geographical regions, since their distributions may vary greatly as shown in Fig. 1.1.

This analysis will use monthly rainfall data from four south Florida stations. First, a frequency analysis of the missing observations has been performed and their typical pattern has been identified. In this work the term "missing observations" is used for a sequence of missing monthly values restricted to less than twelve, so that unusual cases of lengthy gaps (a year or more of missing values) is avoided since they do not reflect the general situation.

Frequency Analysis of Missing Observations in the South Florida Monthly Rainfall Records

An analysis of the monthly rainfall series of 213 stations of the South Florida Water Management District

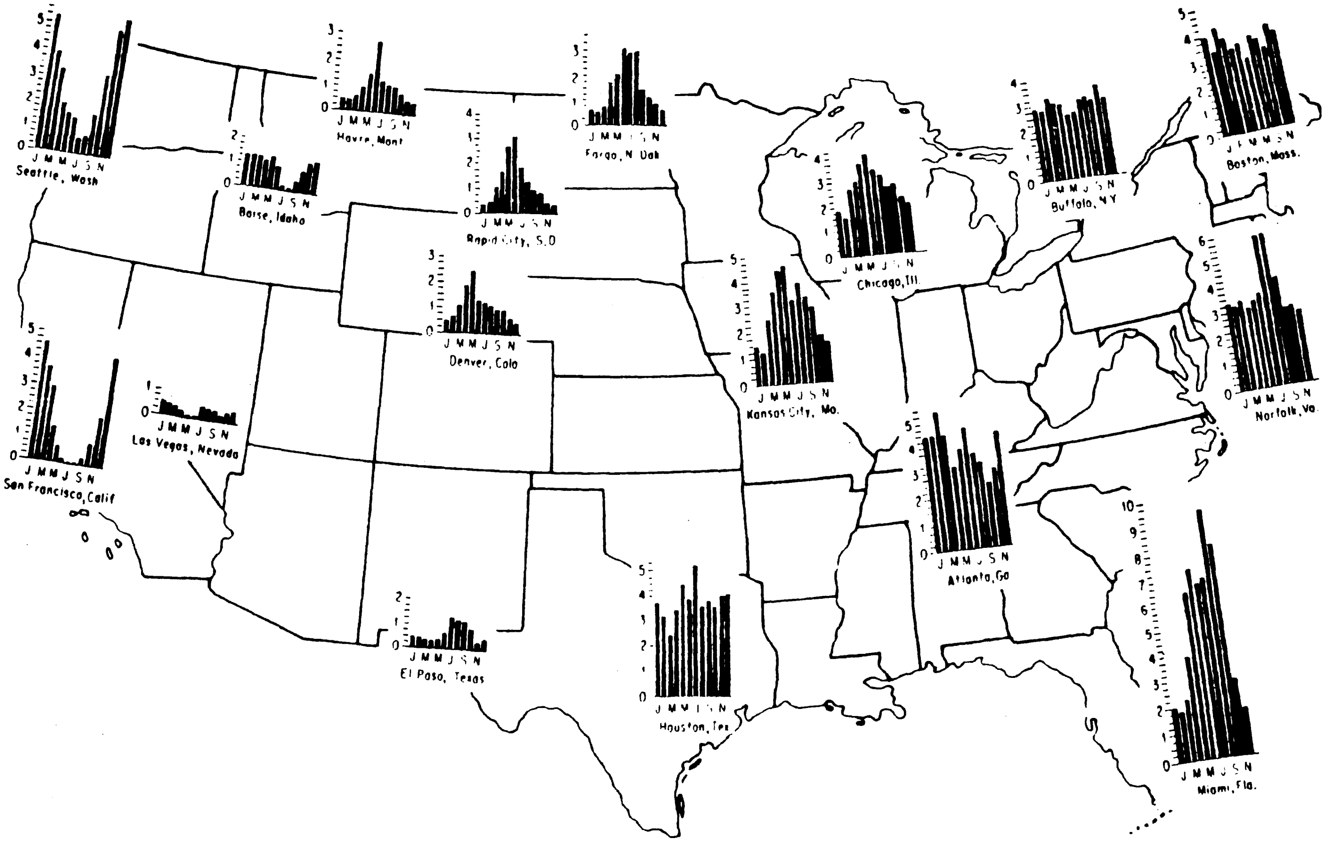


Fig. 1.1. Monthly distribution of fainfall in the United States (after Linsley R.K., Kohler M.A. and Paulhus J.L., Hydrology for Engineers, 1975, McGraw-Hill, 2nd. edition p. 90)

(SFWMD) gave the results shown on Table 1.1. Figure 1.2 shows the probability density function (pdf) plot of the percent m of missing values, f(m), which is defined as the ratio of the probability of occurrence over an interval to the length of that interval (column 4 of Table 1.1). The shape of the pdf f(m) suggests the fit by an exponential distribution

$$f(m) = \lambda e^{-\lambda m} \tag{1.1}$$

where λ is the parameter of the distribution calculated as the inverse of the expected value of m, E(m);

$$E(m) = \Sigma p(m_i)\ m_i \tag{1.2}$$

where $p(m_i)$ is the probability of having m_i percent of missing values. The mean value of the percentage of missing values is $\bar{m} = E(m) = 13.663$, and therefore the fitted exponential pdf is

$$f(m) = 0.073\ e^{-0.073m} \tag{1.3}$$

which gives an interesting and unexpectedly good fit as shown by Fig. 1.2 and column 5 of Table 1.1

The question now arises as to whether the missing values within a record follow a certain pattern. In

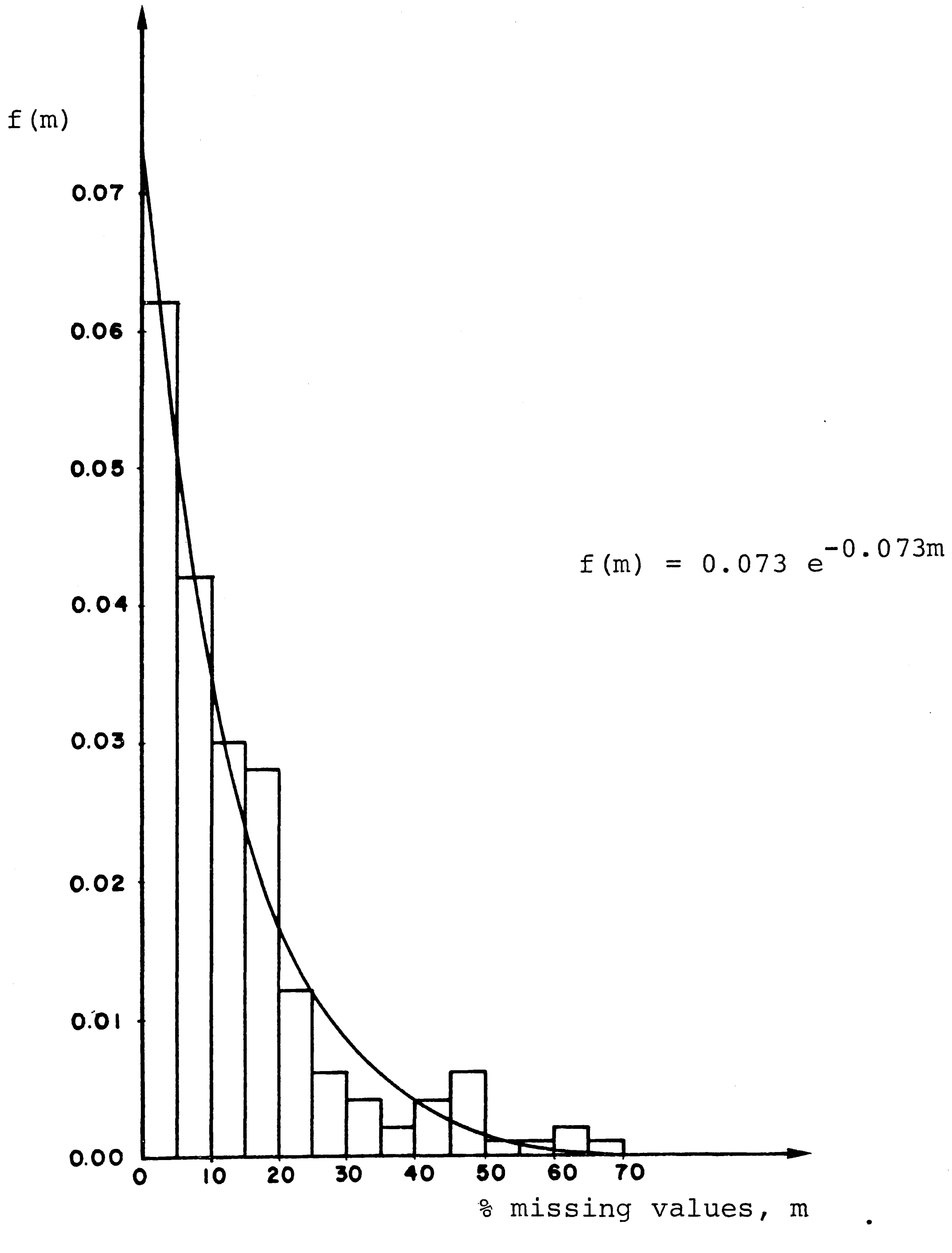


Fig. 1.2. Probability density function, f(m), of the percentage of missing values. Based on 213 stations, $\bar{m} = 13.663\%$.

Table 1.1. Frequency Distribution of the Percent of Missing Values in 213 South Florida Monthly Rainfall Records.

1	2	3	4	5
% of Missing Values	% of Stations	Cumulative % of Stations	Empirical pdf	Fitted Exponential pdf
0-5	30.52	30.52	0.061	0.061
5-10	21.12	51.64	0.042	0.042
10-15	14.55	66.19	0.029	0.029
15-20	13.61	79.80	0.027	0.020
20-25	6.10	85.90	0.012	0.014
25-30	3.29	89.10	0.007	0.010
30-35	1.88	91.70	0.004	0.007
35-40	0.94	92.01	0.002	0.005
40-45	2.35	94.36	0.005	0.003
45-50	2.82	97.18	0.006	0.002
50-55	0.47	97.65	0.001	0.002
55-60	0.47	98.12	0.001	0.001
60-65	1.41	99.53	0.003	0.001
65-70	0.47	100.00	0.001	0.001

particular, if the occurrence of a gap is viewed as an "event" then the distribution of the interevent times (sizes of the interevents) and of the durations of the events (sizes of the gaps) may be examined.

The probability distribution of the size of the interevents (number of values between two successive gaps) has been studied for four "typical" stations of the SFWMD, as far as length of the record, distribution and percent of missing values is concerned. These four stations are:

MRF 6018, Titusville 2W, 1901-1981, 7.5% missing
MRF 6021, Fellsmere 4W, 1911-1979, 9.3% missing
MRF 6029, Ocala, 1900-1981, 4.4% missing
MRF 6005, Plant City, 1892-1981, 8.6% missing

A derived pdf for the four stations combined and the fitted exponential pdf are shown in Fig. 1.3. The mean size of the interevent, $\bar{\tau}$, is 19.03 months; therefore, the fitted exponential distribution is

$$f(\tau) = 0.053\ e^{-0.053\tau} \tag{1.4}$$

Also, the probability distribution of the size of the gaps (number of values missing in each gap) has also been studied for the same four stations. These have been treated as discrete distributions since the size of the gap ($k = 1, 2, \ldots, 11$) is small as compared to the interevent times. A probability distribution for the four stations combined is then derived, which is also the discrete probability mass function (pmf). This plot is shown in Fig. 1.4 and suggests either a Poisson distribution or a discretized exponential.

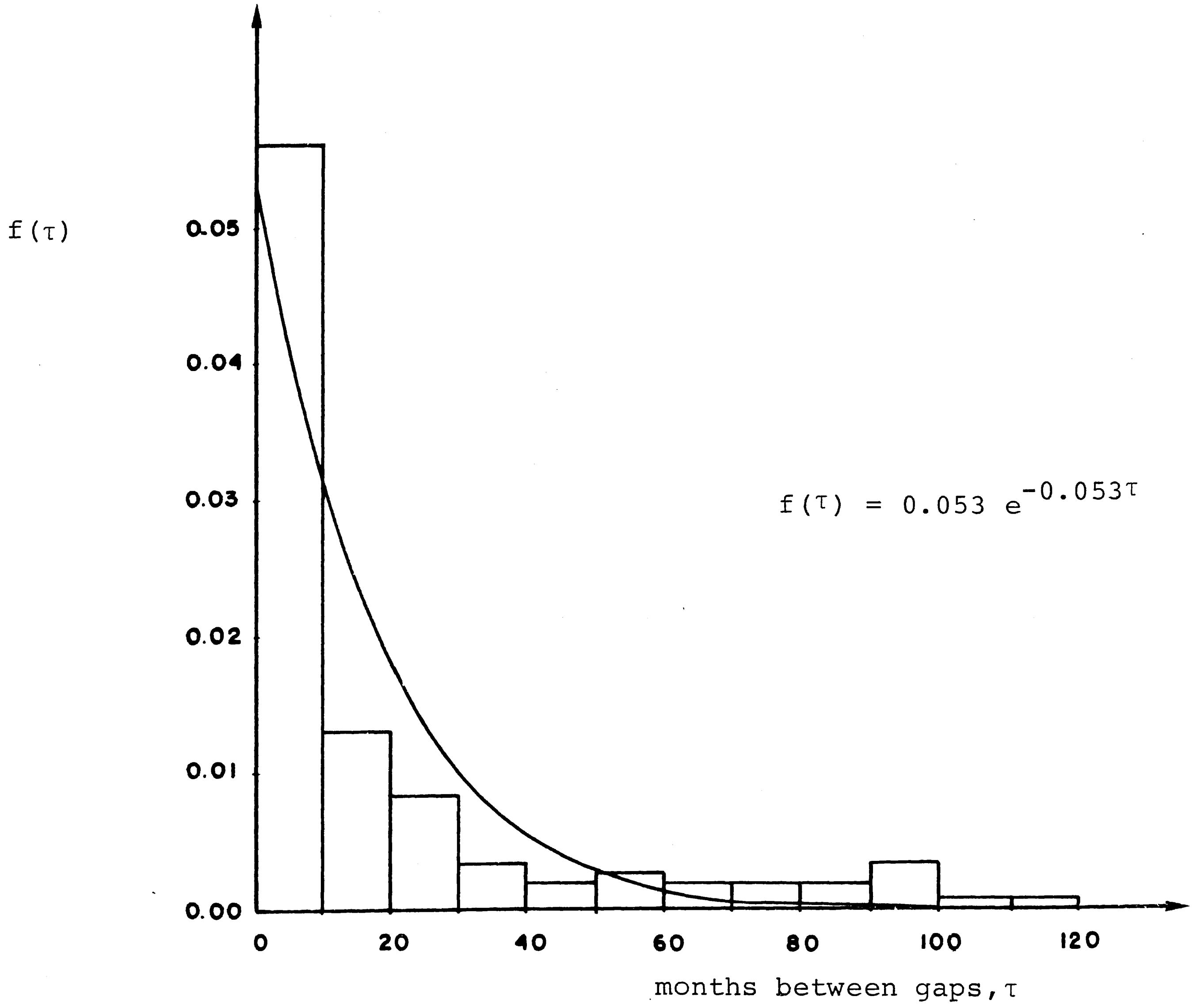


Fig. 1.3. Probability density function, $f(\tau)$, of the interevent size. Based on four stations.

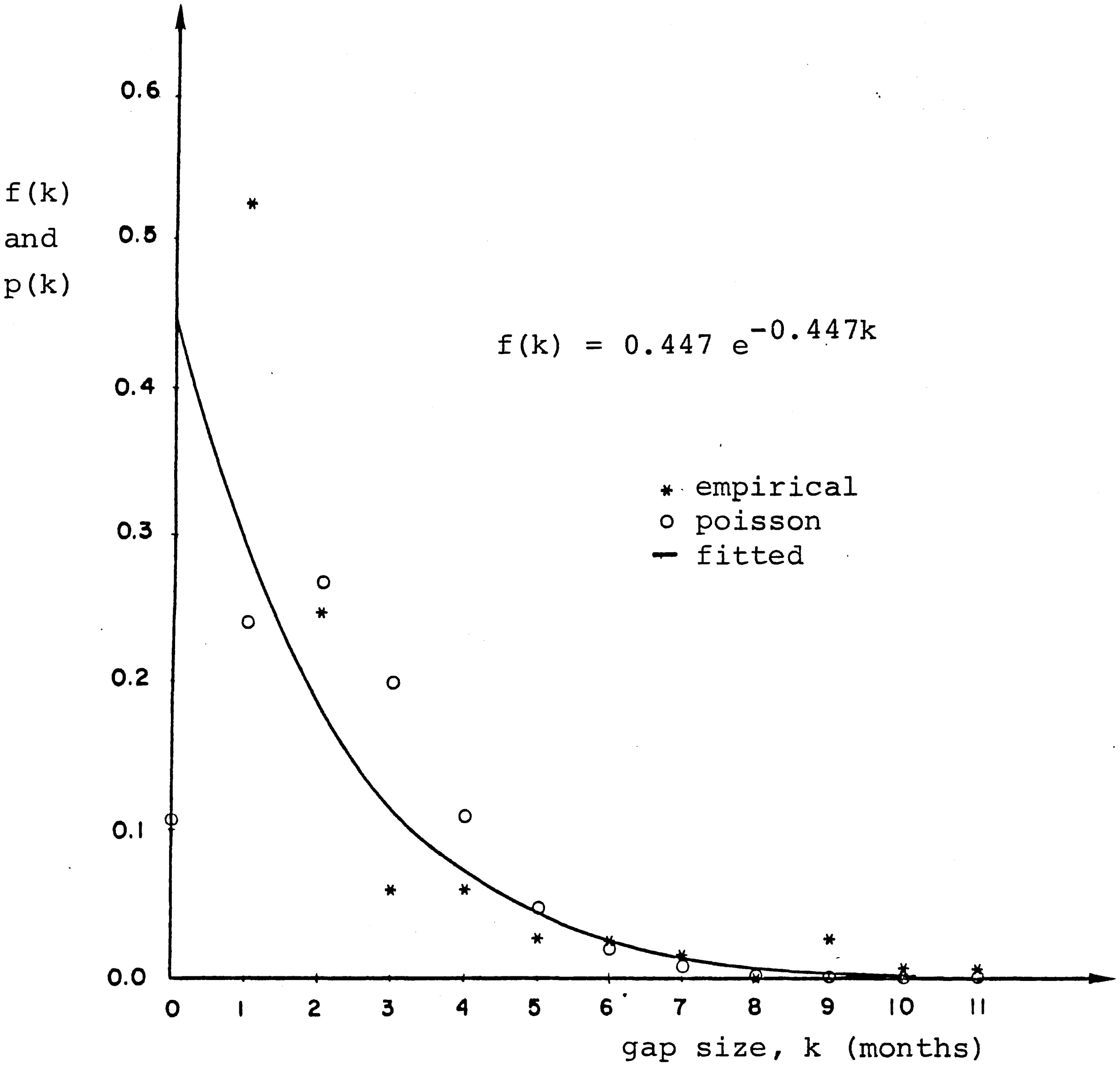


Fig. 1.4. Probability density, f(k), and mass function, p(k), of the gap size. Based on four stations.

The mean value $\bar{k}$ is 2.237, which is also the parameter λ of the Poisson distribution. The Poisson distribution

$$f(k) = \frac{e^{-\lambda} \lambda^k}{k!} \tag{1.5}$$

is nonzero at k = 0 and does not fit the peak of the empirical point very well at k = 1 (it gives a value of 0.24 instead of the actual 0.53). The fitted continuous exponential pdf shown in Fig. 1.4 gives a better fit in general but also implies a nonzero probability for a gap size near zero. To overcome this problem and to discretize the continuous exponential pdf, the area (probability) under the exponential curve between zero and 1.5 is assigned to k = 1, ensuring a zero probability at k = 0. Areas (probabilities) assigned to values of $k > 1$ are centered around those points. The fitted discretized exponential and the Poisson are also shown in Fig. 1.4.

The distributions of the size of the gaps (k) and of the size of interevents (τ) will be used to generate randomly distributed gaps in a complete record. Suppose that we have a complete record and desire to remove randomly m percent missing values. If the mean size of the gap ($\bar{k}$) is assumed constant, the mean size of interevent ($\bar{\tau}$) must vary, decreasing as the percent of missing values increases. Let N denote the total number of values in the record, m the where

$$R^2 = \rho_1\phi_1 + \rho_2\phi_2 + . . . + \rho_p\phi_p \tag{3.8}$$

is called the multiple coefficient of determination and represents the fraction of the variance of the series that has been explained through the regression.

If we denote by ϕ_{kj} the jth coefficient in an autoregressive process of order k, then the last coefficient ϕ_{kk} of the model is called the partial autocorrelation coefficient. Estimates of the partial autocorrelation coefficients $\phi_{11}, \phi_{22}, . . ., \phi_{pp}$ may be obtained by fitting to the series autoregressive processes of successively higher order, and solving the corresponding Yule-Walker equations. The partial autocorrelation function ϕ_{kk}, k = 1, 2, . . ., p may also be obtained recursively by means of Durbin's relations (Durbin, 1960)

$$\phi_{k+1,k+1} = [r_{k+1} - \sum_{j=1}^{k} \phi_{k,j}\ r_{k+1-j}]/[1 - \sum_{j=1}^{k} \phi_{k,j}\ r_j] \tag{3.9}$$

$$\phi_{k+1,j} = \phi_{k,j} - \phi_{k+1,k+1}\ \phi_{k,k-j+1} \quad j = 1, 2, . . ., k$$

It can be shown (Box and Jenkins, 1976, p. 55) that the autocorrelation function of a stationary AR(p) process is a mixture of damped exponential and damped sine waves, infinite in extent. On the other hand, the partial autocorrelation function ϕ_{kk} is nonzero for $k \leq p$ and zero for $k > p$. The plot of autocorrelation and partial autocorrelation functions of the series may be used to identify the kind and the order of the model that may have generated it (identification of the model).

Moving Average Models

In a moving average model the deviation of the current value of the process from the mean is expressed as a finite sum of weighted previous shocks a's. Thus a moving average process of order q can be written as:

$$\tilde{z}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q} \tag{3.10}$$

or

$$\tilde{z}_t = \theta(B) a_t \tag{3.11}$$

where

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \tag{3.12}$$

is the moving average operator of order q. An MA(q) model contains (q+2) parameters, μ, θ_1, θ_2, ..., θ_q, σ_a^2 to be estimated from the data.

From the definition of stationarity (see Appendix A) it follows that an MA(q) process is always stationary since $\theta(B)$ is finite and thus converges for $|B|<1$. But for an MA(q) process to be invertible the q moving average coefficients $\theta_1, \theta_2, \ldots, \theta_q$ must be chosen so that $\theta^{-1}(B)$ converges on or within the unit circle, in other words the characteristic equation $\theta(B) = 0$ must have its roots outside the unit circle.

By multiplying equation (3.10) by $\tilde{z}_{t-k}$ and taking expected values on both sides we define the autocovariance at lag k:

$$\gamma_k = E\,[(a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q})\ (a_{t-k} - \theta_1 a_{t-k-1} - \ldots - \theta_q a_{t-k-q})] \qquad (3.13)$$

which gives

$$\gamma_o = (1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2)\ \sigma_a^2 \qquad k = 0 \qquad (3.14)$$

$$\gamma_k = (-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \ldots + \theta_{q-k}\theta_q)\sigma_a^2\ , \quad k = 1, 2, \ldots, q \qquad (3.15)$$

$$\gamma_k = 0 \qquad k > q \qquad (3.16)$$

By substituting in equation (3.15) the value of σ_a^2 from equation (3.14) we obtain a set of q nonlinear equations for $\theta_1, \theta_2, \ldots, \theta_q$ in terms of $\rho_1, \rho_2, \ldots, \rho_q$.

$$\rho_k = \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2}, \quad k=1,2,\ldots,q \tag{3.17}$$

These equations are analogous to the Yule-Walker equations for an autoregressive process, but they are not linear and so must be solved iteratively for the estimation of the moving average parameters θ, resulting in estimates that may not have high statistical efficiency. Again it was shown by Wold (1938) that these parameters may need corrections (e.g., to fit better the correlogram as a whole and not only the first q correlation coefficients), and that there may exist several, at most 2^q solutions, for the parameters of the moving average scheme corresponding to an assigned correlogram $\rho_1, \rho_2, \ldots, \rho_q$. However, only those θ's are acceptable which satisfy the invertibility conditions.

From equation (3.14) an estimate for the white noise variance σ_a^2 may be obtained

$$\sigma_a^2 = \frac{\sigma_z^2}{1 + \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2} \tag{3.18}$$

According to the duality principle (see Appendix A) an invertible MA(q) process can be represented as an AR process of infinite order. This implies that the partial autocorrelation function ϕ_{kk} of an MA(q) process is infinite in extent. It can be estimated after tedious algebraic manipulations from the Yule-Walker equations by substituting ρ_k as functions of θ's for $k < q$ and $\rho_k = 0$ for $k > q$. So, in contrast to a stationary AR(p) process, the autocorrelation function of an invertible MA(q) process is finite and cuts off after lag q, and the partial autocorrelation function is infinite in extent, dominated by damped exponentials and damped sine waves (Box and Jenkins, 1976).

Mixed Autoregressive-Moving Average Models

In practice, to obtain a parsimonious parameterization, it will sometimes be necessary to include both autoregressive and moving average terms in the model. A mixed autoregressive-moving average process of order (p,q), ARMA(p,q), can be written as

$$\tilde{z}_t = \phi_1 \tilde{z}_{t-1} + \cdots + \phi_p \tilde{z}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q} \tag{3.19}$$

or

$$\phi(B)\ \tilde{z}_t = \theta(B)\ a_t \tag{3.20}$$

with (p+q+2) parameters, μ, $\theta_1, \ldots, \theta_q, \phi_1, \ldots, \phi_p, \sigma_a^2$ to be estimated from the data.

An ARMA(p,q) process will be stationary provided that the characteristic equation $\phi(B) = 0$ has all its roots outside the unit circle. Similarly, the roots of $\theta(B) = 0$ must lie outside the unit circle for the process to be invertible.

By multiplying equation (3.19) by $\tilde{z}_{t-k}$ and taking expectations we obtain

$$\gamma_k = \phi_1 \gamma_{k-1} + \dots + \phi_p \gamma_{k-p} + \gamma_{za}(k) - \theta_1 \gamma_{za}(k-1) - \dots - \theta_q \gamma_{za}(k-q) \qquad (3.21)$$

where $\gamma_{za}(k)$ is the cross covariance function between z and a, defined by $\gamma_{za}(k) = E[\tilde{z}_{t-k} a_t]$. Since z_{t-k} depends only on shocks which have occurred up to time t-k, it follows that

$$\begin{aligned} \gamma_{za}(k) &= 0 \qquad && k > 0 \\ \gamma_{za}(k) &\neq 0 \qquad && k \leq 0 \end{aligned} \qquad (3.22)$$

and (3.21) implies

$$\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2} + \dots + \phi_p \rho_{k-p} \qquad k \geq q + 1 \quad (3.23)$$

or

$$\phi(B) \rho_k = 0 \qquad k \geq q + 1 \quad (3.24)$$

Thus, for the ARMA(p,q) process the first q autocorrelations $\rho_1, \rho_2, \ldots, \rho_q$ depend directly on the choice of the q moving average paramaters θ, as well as on the p autoregressive parameters ϕ through (3.21). The autocorrelations of higher lags ρ_k, $k \geq q + 1$ are determined through the difference equation (3.24) after providing the p starting values $\rho_{q-p+1}, \ldots, \rho_q$. So, the autocorrelation function of an ARMA(p,q) model is infinite in extent, with the first q-p values $\rho_1, \ldots, \rho_{q-p}$ irregular and the others consisting of damped exponentials and/or damped sine waves (Box and Jenkins, 1976; Salas et al., 1980).

Autoregressive Integrated Moving Average Models

An ARMA(p,q) process is stationary if the roots of $\phi(B) = 0$ lie outside the unit circle and "explosive nonstationary" if they lie inside. For example, an explosive nonstationary AR(1) model is $z_t = 2z_{t-1} + a_t$ (the plot of z_t vs. t is an exponential growth) in which $\phi(B) = 1 - 2B$ has its root B = 0.5 inside the unit circle. The special case of homogeneous nonstationarity is when one or more of the roots lie on the unit circle. By introducing a generalized autoregressive operator $\phi_0(B)$, which has d of its roots on the unit circle, the general model can be written as

$$\phi_0(B) = \phi(B)\ (1-B)^d\ \tilde{z}_t = \theta(B)\ a_t \qquad (3.25)$$

that is

$$\phi(B)\ w_t = \theta(B)\ a_t \qquad (3.26)$$

where

$$w_t = \nabla^d\ \tilde{z}_t = \nabla^d\ z_t \qquad (3.27)$$

and $\nabla = 1 - B$ is the difference operator. This model corresponds to assuming that the dth difference of the series can be represented by a stationary, invertible ARMA process. By inverting (3.27)

$$z_t = \nabla^{-d} w_t = S^d w_t \tag{3.28}$$

where S is the infinite summation operator

$$S = 1 + B + B^2 + \dots = (1-B)^{-1} = \nabla^{-1} \tag{3.29}$$

Equation (3.28) implies that the nonstationary process z_t can be obtained by summing or "integrating" the stationary process w_t, d times. Therefore, this process is called a simple autoregressive integrated moving average process, ARIMA(p,d,q).

It is also possible to take periodic or seasonal differences at lag's of the series, e.g., the 12th difference of monthly series, introducing the differencing operator ∇_s^D with the meaning that seasonal differencing ∇_s is applied D times on the series. This periodic ARIMA$(P,D,Q)_s$ model can be written as

$$\Phi(B^s)\ \nabla_s^D\ z_t = \Theta(B^s)\ a_t \tag{3.30}$$

The combination of nonperiodic and periodic models leads to the multiplicative ARIMA(p,d,q) x ARIMA$(P,D,Q)_s$ model which can be written as

$$\phi(B)\ \Phi(B^s)\ \nabla^d\ \nabla_s^D\ z_t = \theta(B)\ \Theta(B^s)\ a_t \tag{3.31}$$

After the model has been fitted to the differenced series an integration should be performed to retrieve the original process. But such an integrated series would lack a mean value since a constant of integration has been lost through the differencing. This is the reason that the ARIMA models cannot be used for synthetic generation of time series, although they are useful in forecasting the deviations of a process (Box and Jenkins, 1976; Salas _et al._, 1980).

Transformation of the Original Series

Transformation to Normality

Most probability theory and statistical techniques have been developed for normally distributed variables. Hydrologic variables are usually assymetrically distributed or bounded by zero (positive variables), and so a transformation to normality is often applied before modeling. Another approach would be to model the original skewed series and then find the probability distribution of the uncorrelated residuals. Care must then be taken to assess the errors of applying methods developed for normal variables to skewed variables, especially when the series are highly skewed, e.g., hourly or daily series. On the other hand, when transforming the original series into normal, biases in the mean and standard deviation of the generated series may occur. In other words, the statistical properties of the transformed series may be reproduced in the generated but not in the original series. An alternative for avoiding biases in the moments of the generated series would be to estimate the moments of the transformed series through the derived relationships between the moments of the skewed and normal series. Matalas (1967) and Fiering and Jackson (1971) describe how to estimate the first two moments of the log-transformed series so as to reproduce the ones of the original series. Mejia et al. (1974) present another approach in order to preserve the correlation structure of the original series.

However, the most widely used approach is to transform the original skewed series to normal and then model the normal series. Several transformations may be applied to the original series, and the transformed series then tested for normality, e.g. the graph of their cumulative distribution should appear as a straight line when it is plotted on normal probability paper. The transformation will be finally chosen that gives the best approximation to normality, e.g., the best fit to a straight line.

Another advantage of transforming the series to normal is that the maximum likelihood estimates of the model parameters are essentially the same as the least squares estimates, provided that the residuals are normally distributed (Box and Jenkins, 1976, Ch. 7). This facilitates the calculation of the final estimates since they are those values that minimize the sum of squares of the residuals.

Box and Cox (1964) showed how a maximum likelihood and a parallel Bayesian analysis can be applied to any type of transformation family to obtain the "best" choice of transformation from that family. They illustrated those methods for the popular power families in which the observation x is replaced by y, where

$$y = \begin{cases} \frac{x^{\lambda}-1}{\lambda} , & \lambda\neq 0 \\ \log x , & \lambda=0 \end{cases} \tag{3.32}$$

The fundamental assumption was that for some λ the transformed observations y can be treated as independently normally distributed with constant variance σ^2 and with expectations defined by a linear model

$$E[y] = A\ L \tag{3.33}$$

where A is a known constant matrix and L is a vector of unknown parameters associated with the transformed observations (Box and Cox, 1964).

This transformation has the advantage over the simple power transformation proposed by Tukey (1957)

$$y = \begin{cases} x^{\lambda} , & \lambda \neq 0 \\ \log x , & \lambda = 0 \end{cases} \tag{3.34}$$

of being continuous at $\lambda=0$. Otherwise the two transformations are identical provided, as has been shown by Schlesselman (1971), that the linear model of (3.33) contains a constant term.

Further, Draper and Cox (1969), showed that the value of λ obtained from this family of transformations can be useful even in cases where no power transformation can produce normality exactly. Also, John and Draper (1980) suggested an alternative one-parameter family of transformations when the power transformation fails to produce satisfactory distributional properties as in the case of a symmetric distribution with long tails.

The selection of the exact transformation to normality (zero skewness) is not an easy task, and over-transformation, i.e., transformation of the original data with a large positive (negative) skewness to data with a small negative (positive) skewness, or under-transformation, i.e., transformation of the original data with a large positive (negative) skewness to data with a small positive (negative) skewness, may result in unsatisfactory modeling of the series or in forecasts that are in error. This was the case for the data used by Chatfield and Prothero (1973a), who applied the Box-Jenkins forecasting approach and were dissatisfied with the results, concluding that the Box-Jenkins forecasting procedure is less efficient than other forecasting methods. They applied a log transform to the data which evidently over-transformed the data, as shown by Box and Jenkins (1973) who finally suggested the approximate transformation $y = x^{0.25}$, even though the complicated but precise Box-Cox procedure gave an estimate of $\lambda = 0.37$ [Wilson (1973)].

Thus, the selection of the normality transformation greatly affects the forecasts, as Chatfield and Prothero (1973b) experienced with their data. They concluded that

> . . . We have seen that a "small" change in λ from 0 to 0.25 has a substantial effect on the resulting forecasts from model A [ARIMA(1,1,1) x ARIMA$(1,1,1)_{12}$] even though the goodness of fit does not seem to be much affected. This reminds us that a model which fits well does not necessarily forecast well. Since small changes in λ close to zero produce marked changes in forecasts, it is obviously advisable to avoid "low" values of λ, since a procedure which depends critically on distinguishing between fourth-root and logarithmic transformation is fraught with peril. On the other hand a "large" change in λ from 0.25 to 1 appears to have relatively little effect on forecasts. So we conjecture that Box-Jenkins forecasts are robust to changes in the transformation parameter away from zero. . . .[Chatfield and Prothero (1973b) p. 347]

Stationarity

Most time series occurring in practice exhibit nonstationarity in the form of trends or periodicities. The physical knowledge of the phenomenon being studied and a visual inspection of the plot of the original data may give the first insight into the problem. Usually the length of the series is not long enough, and the detection of trends or cycles only through the plot of the series is ambiguous. Useful tools for the detection of periodicities are the autocorrelation function and the spectral density function of the series (which is the Fourier transform of the autocorrelation function). If a seasonal pattern is present in the series then the correlogram (plot of the autocorrelation function) will exhibit a sinusoidal appearance and the periodogram (plot of the spectral density function) will show peaks. The period of the sinusoidal function of the correlogram, or the frequency where the peaks occur in the periodogram, can determine the periodic component exactly (Jenkins and Watts, 1968). Another device for the detection of trends and periodicities is to fit some definite mathematical function, such as exponentials, Fourier series or polynomials to the series and then model the residual series, which is assumed to be stationary. More details on the treatment of nonstationary data as well as on the interpretation of the correlogram and periodogram of a time series can be found in textbooks such as Bendat and Piersol (1958), Jenkins and Watts (1968), Wastler (1969), Yevjevich (1972), and Chatfield (1980).

Apart from the approach of removing the nonstationarity of the original series and modeling the residual series with a stationary ARMA(p,q) model, the original nonstationary series can be modeled directly with a simple or seasonally integrated ARIMA model. Actually, the second approach can be viewed as an extension of the first one, e.g., the nonstationarity is removed through the simple (∇) or seasonal (∇_s) differencing. However, the integrated model cannot be used for generation of data, as has already been discussed.

For many hydrologic applications, one is satisfied with second order or weak stationarity, e.g., stationarity in the mean and variance. Furthermore, weak stationarity and the assumption of normality imply strict stationarity (see Appendix A).

Monthly Rainfall Series

Normalization and Stationarization

Stidd (1953, 1968) suggested that rainfall data have a cube root normal distribution because they are product functions of three variables: vertical motion in the atmosphere, moisture, and duration time. Synthetic rainfall data generated using processes analogous to those operating in nature showed that the exponent required to normalize the distribution is between 0.5 (square root) and 0.33 (cubic root) for different types of rainfall (Stidd, 1970).

The square root transformation has been extensively used for the approximate normalization of monthly rainfall series (see Table C12 of Appendix C) with satisfactory results: Delleur and Kavvas (1978), Salas *et al*. (1980), Ch. 5, Roesner and Yevjevich (1966). However, Hinkley (1977) used the exact Box-Cox transformation for monthly rainfall series. Although, Asley _et_ _al_. (1977) have developed an efficient algorithm for the estimation of λ along with other parameters in an ARIMA model, it seems that the exact value of λ is not more reliable than the approximate one $\lambda = 0.5$ (Chatfield and Prothero, 1973b). The reasons for this follow.

First, Chatfield and Prothero (1973b) used the Box-Cox procedure to evaluate the exact transformation of their data. They obtained estimates $\hat{\lambda} = 0.24$ using all the data (77 observations), $\hat{\lambda} = 0.34$ using the first 60 observations and $\hat{\lambda} = 0.16$ excluding the first year's data. Therefore, it is logical to infer that even if the complicated Box-Cox procedure for the incomplete rainfall record is used, the missing values may be enough to give a spurious λ, which is not "more exact" than the value of 0.5 used in practice.

Second, we may also notice that the use of either $\lambda = 0.33$ (cubic root) or $\lambda = 0.5$ (square root) is not expected to greatly affect the forecasts since, according to Chatfield and Prothero (1973b), the Box-Jenkins forecasts are not too sensitive to changes of λ for $\lambda > 0.25$.

Monthly rainfall series are nonstationary. The variation in the mean is obvious since generally the expected monthly rainfall value for January is not the same as that of July. Although the variation of the standard deviation is not so easy to visualize, calculations show that months with higher mean usually have higher standard deviation. Thus, each month has its own probability distribution and its own statistical parameters resulting in monthly series that are nonstationary.

By introducing the concept of circular stationarity as developed by Hannan (1960) and others (see Appendix A for definition), the periodic monthly rainfall series can be considered not as nonstationary but circular stationary, since circular stationarity suggests that the probability distribution of rainfall in a particular month is the same for the different years. Then, the monthly rainfall series is composed of a circularly stationary (periodic) component and a stationary random component.

The time-series models currently used in hydrology are fitted to the stationary random component, so the circularly stationary component must be removed before modeling. This last component appears as a sinusoidal component in the autocorrelation function (with a 12-month period) or as a discrete spectral component in the spectrum (peak at the frequency 1/12 cycle per month). Usually several subharmonics of the fundamental 12-month period are needed to describe all the irregularities present in the autocorrelation function and spectral density function, since in nature the periodicity does not follow an ideal cosine function with a 12-month period. The use of a Fourier series approach for the approximation of the periodic component of monthly rainfall and monthly runoff series has been illustrated by Roesner and Yevjevich (1966).

Kavvas and Delleur (1975) investigated three methods of removal of periodicities in the monthly rainfall series: nonseasonal (first-lag) differencing, seasonal differencing (12-month difference), and removal of monthly means. They worked both analytically and empirically using the rescaled (divided by the monthly standard deviation) monthly rainfall square roots for fifteen Indiana watersheds. They concluded that "all the above transformations yield hydrologic series which satisfy the classical second-order weak stationarity conditions. Both seasonal and nonseasonal differencing reduce the periodicity in the covariance function but distort the original spectrum, thus making it impractical or impossible to fit an ARMA model for generation of synthetic monthly series. The subtraction of monthly means removes the periodicity in the covariance and the amount of nonstationarity introduced is negligible for practical purposes." (Kavvas and Delleur, 1975, p. 349.) In other words, they concluded that the best way for modeling monthly rainfall series is to remove the seasonality (by subtracting the monthly means and dividing by the standard deviations of the normalized series) and then use a stationary ARMA(p,q) model to model the stationary normal residuals.

Modeling of Normalized Series

It is assumed that the nonstationarities due to long-term trends are removed before any operation. Then the appropriate transformation is applied to the data in order to obtain an approximately normal distribution. For monthly rainfall series experience has shown that the best practical transformation is the square root transformation, as has already been discussed. What remains is the modeling of the normalized series with one of the following models: stationary ARMA(p,q), simple nonstationary ARIMA(p,d,q), seasonal nonstationary ARIMA$(P,D,Q)_s$, or multiplicative ARIMA$(p,d,q)x(P,D,Q)_s$ model.

Delleur and Kavvas (1978) fitted different models to the monthly rainfall series of 15 basins in Indiana and compared the results. They studied the models: ARIMA (0,0,0), ARIMA(1,0,1), ARIMA(1,1,1), ARIMA$(1,1,1)_{12}$, and ARIMA$(1,0,0)x(1,1,1)_{12}$ on the square-root transformed series. They concluded that from the nonseasonal ARIMA models, ARMA(1,1) "emerged as the most suitable for the generation and forecasting of monthly rainfall series." The goodness-of-fit tests applied on the residuals were the portemanteau lack of fit test (see Appendix A) of Box and Pierce (1970) and the cumulative periodogram test (Box and Jenkins, 1976, p. 294). The ARMA(1,1) model passed both tests in all cases studied. From the nonseasonal models, ARIMA$(1,0,0)x(1,1,1)_{12}$ also passed the goodness-of-fit tests in all cases, but they stress that this model "has only limited use in the forecasting of monthly rainfall series since it does not preserve the monthly standard deviations." As far as forecasts are concerned, they showed that "the forecasts by the several models follow each other very closely and the forecasts rapidly tend to the mean of the observed rainfall square roots (which is the forecast of the white noise model)."

CHAPTER 4

MULTIVARIATE STOCHASTIC MODELS

Introduction

For univariate stochastic models the sequence of observations under study is assumed independent of other sequences of observations and so is studied by itself (single or univariate time series). However, in practice there is always an interdependence among such sequences of observations, and their simultaneous study leads to the concept of multivariate statistical analysis. For example, a rainfall series of one station may be better modeled if its correlation with concurrent rainfall series at other nearby stations is incorporated into the model. Multiple time series can be divided into two groups: (1) multiple time series at several points (e.g., rainfall series at different stations, streamflow series at various points of a river), and (2) multiple series of different kinds at one point (e.g., rainfall and runoff series at the same station). In general, both kinds of multiple time series are studied simultaneously, and their correlation and cross-correlation structure is used for the construction of a model that better describes all these series. The parameters of this so called multivariate stochastic model are calculated such that the correlation and cross-correlation structure of the multiple measured series are preserved in the multiple series generated by the model.

The multivariate models that will be presented in this chapter have been developed and extensively used for the generation of synthetic series. How these models can be adapted and used for filling in missing values will be discussed in chapter 5.

General Multivariate Regression Model

The general form of a multivariate regression model is

$$Y = A X + B H \tag{4.1}$$

where Y is the vector of dependent variables, X the vector of independent variables, A and B matrices of regression coefficients, and H a vector of random components. The vectors Y and X may consist of either the same variable at different points (or at different times) or different variables at the same or different points (or at different times).

For convenience and without loss of generality all the variables are assumed second order stationary and normally distributed with zero mean and unit variance. Transformations to accomplish normality have been discussed in Chapter 3. A random component is superimposed on the model to account for the nondeterministic fluctuations.

In the above model, the dependent and independent variables must be selected carefully so that the most information is extracted from the existing data. A good summary of the methods for the selection of independent variables for use in the model is given in Draper and Smith (1966). Most popular is the stepwise regression procedure in which the independent variables are ranked as a function of their partial correlation coefficients with the dependent variable and are added to the model, in that order, if they pass a sequential F test.

The parameter matrices A and B are calculated from the existing data in such a way that important statistical characteristics of the historical series are preserved in the generated series. This estimation procedure becomes cumbersome when too many dependent and independent variables are involved in the model, and several simplifications are often made in practice. On the other hand, restrictions have to be imposed on the form of the data, as we shall see later, to ensure the existence of real solutions for the matrices A and B.

Multivariate Lag-One Autoregressive Model

If only one variable (e.g., rainfall at different stations) is used in the analysis then the model of equation (4.1) becomes a multivariate autoregressive model. Since in the rest of this chapter we will be dealing only with one variable (rainfall) which has been transformed to normal and second order stationary, the vectors Y and X are replaced by the vector Z for a notation consistent with the univariate models. Matalas (1967) suggested the multivariate lag-one autoregressive model

$$Z_t = A\ Z_{t-1} + B\ H_t \tag{4.3}$$

where Z_t is an (mxl) vector whose ith element z_{it} is the observed rainfall value at station i and at time t, and the other variables have been described previously.

Such a model can be used for the simultaneous generation of rainfall series at m different stations. The correlation and cross-correlation of the series is incorporated in the model through the parameters A and B.

The matrices A and B are estimated from the historical series so that the means, standard deviations and autocorrelation coefficients of lag-one for all the series, as well as the cross-correlations of lag-zero and lag-one between pairs of series are maintained.

Let M_0 denote the lag-zero correlation matrix which is defined as

$$M_0 = E[Z_t\ Z_t^T] \tag{4.4}$$

Then a diagonal element of M_0 is $E[z_{i,t}\ z_{i,t}] = \rho_{ii}(0) = 1$ (since Z_t is standardized) and an off diagonal element (i,j) is $E[z_{i,t}\ z_{j,t}] = \rho_{ij}(0)$ which is the lag-zero cross correlation between series $\{z_i\}$ and $\{z_j\}$. The matrix M_0 is symmetric since $\rho_{ij}(0) = \rho_{ji}(0)$ for every i, j.

Let M_1 denote the lag-one correlation matrix defined as

$$M_1 = E[Z_t \; Z_{t-1}^T] \tag{4.5}$$

A diagonal element of M_1 is $E[z_{i,t} \; z_{i,t-1}] = \rho_{ii}(1)$ which is the lag-one serial correlation coefficient of the series $\{z_i\}$, and an off-diagonal element (i,j) is $E(z_{i,t} \; z_{j,t-1}) = \rho_{ij}(1)$ which is the lag-one cross-correlation between the $\{z_i\}$ and $\{z_j\}$ series, the latter lagged behind the former. Since in general $\rho_{ij}(1) \neq \rho_{ji}(1)$ for $i \neq j$ the matrix M_1 is not symmetric.

After some algebraic manipulations (see Appendix B) the coefficient matrices A and B are obtained as solutions to the equations

$$A = M_1 \; M_0^{-1} \tag{4.6}$$

$$B \; B^T = M_0 - M_1 \; M_0^{-1} \; M_1^T \tag{4.7}$$

where M_0^{-1} is the inverse of M_0, and M_1^T the transpose of M_1. The correlation matrices M_0 and M_1 are calculated from the data. Then an estimate of the matrix A is given directly by equation (4.6), and an estimate for B is found by solving equation (4.7) by using a technique of principal component analysis (Fiering, 1964) or upper triangularization (Young, 1968). For more details on the solution of equation (4.7) see Appendix B.

Comments on Multivariate AR(1) Model

Assumption of Normality and Stationarity

We have assumed that all random variables involved in the model are normal. The assumption of a multivariate normal distribution is convenient but not necessary. It has been shown (Valencia and Schaake, 1973) that the multivariate AR(1) model preserves first and second order statistics regardless of the underlying probability distributions.

Several studies have been done using directly the original skewed series. Matalas (1967) worked with log-normal series and constructed the generation model so that it preserves the historical statistics of the log-normal process. Mejia _et_ _al_. (1974) showed a procedure for multivariate generation of mixtures of normal and log-normal variables. Moran (1970) indicated how a multivariate gamma process may be applied, and Kahan (1974) presented a method for the preservation of skewness in a linear bivariate regression model. But in general, the normalization of the series prior to modeling is more convenient, especially when the series have different underlying probability distributions. In such cases different transformations are applied on the series, and that combination of transformations is kept which yields minimum average skewness. Average skewness is the sum of the skewness of each series divided by the number of series or number of stations used. This operation is called finding the MST (Minimum Skewness

Transformation) and results in an approximately multivariate normal distribution (Young and Pisano, 1968).

We have also assumed that all variables are standardized, e.g., have zero mean and unit variance. This assumption is made without loss of generality since the linear transformations are preserved through the model. On the other hand this transformation becomes necessary when modeling periodic series since by subtracting the periodic means and dividing by the standard deviations we remove almost all of the periodicity.

If the data are not standardized, M_0 and M_1 represent the lag-zero and lag-one covariance matrices (instead of correlation matrices), respectively. If S denotes the diagonal matrix of the standard deviations and R_0, R_1 the lag-zero and lag-one correlation matrices then

$$M_0 = S\ R_0\ S \tag{4.8}$$

and

$$M_1 = S\ R_1\ S \tag{4.9}$$

When we standardize the data the matrix S is an identity matrix and M_0, M_1 become the correlation matrices R_0 and R_1 respectively. Thus, one other advantage of standardization is that we work with correlation matrices whose elements are less than unity and the computations are likely to be more stable (Pegram and James, 1972).

Cross-Correlation Matrix M_1

Notice that the lag-one correlation matrix M_1 has been defined as $M_1 = E[Z_t \; Z_{t-1}^T]$ which contains the lag-one cross-correlations between pairs of series but having the second series lagged behind the first one. Following this definition the lag-minus-one correlation matrix will be

$$M_{-1} = E[Z_{t-1} \; Z_t^T] \tag{4.10}$$

and it will contain the lag-one correlations having now the second series lagged ahead of the first one. It is easy to show that M_{-1} is actually the transpose of M_1:

$$M_{-1} = E[Z_{t-1} \; Z_t^T] = E[(Z_t \; Z_{t-1}^T)^T] = M_1^T \tag{4.11}$$

Care then must be taken so that there is a consistency between the equation used to calculate matrix A and the way that the cross-correlation coefficients have been calculated. Such an inconsistency was present in the numerical multisite package (NMP) developed by Young and Pisano (1968) and was first corrected by O'Connell (1973) and completely corrected and improved by Finzi *et al*. (1974, 1975).

Incomplete Data Sets

In practice, hydrologic series at different stations are unlikely to be concurrent and of equal length. With lag-zero auto- and cross-correlation coefficients calculated from the incomplete data sets, the lag-zero correlation matrix M_0 obtained may not be positive semidefinite, and, its inverse M_0^{-1} needed for the calculation of matrix A thus may have elements that are complex numbers. Also, a necessary and sufficient condition for a real solution of matrix B is that $C = M_0 - M_1 M_0^{-1} M_1^T$ is a positive semidefinite matrix (see Appendix B).

When all of the series are concurrent and complete then M_0 and C are both semidefinite matrices [Valencia and Schaake, 1973], and the generated synthetic series are real numbers. When the series are incomplete there is no guarantee that real solutions for the matrices A and B exist causing the model of Matalas (1967) to be conditional on M_0 and C being positive semidefinite [Slack, 1973].

Several techniques have been proposed which use the incomplete data sets but guarantee the posite semidefiniteness of the correlation matrices. Fiering (1968) suggested a technique that can be used to produce a positive semidefinite correlation matrix M_0. If M_0 is not positive semidefinite then negative eigenvalues may occur and hence negative variables, since the eigenvalues are variances in the principal component system. In this technique, the eigenvalues of the original correlation matrix are calculated. If negative eigenvalues are encountered, an adjustment procedure is used to eliminate them (thereby altering the correlation matrix, M_0 [Fiering, 1968]).

A correlation matrix is called consistent if all its eigenvalues are positive. But consistent estimates of the correlation matrices M_0 and M_1 do not guarantee that C will also be consistent.

Crosby and Maddock (1970) proposed a technique that is suitable only for monotone data (data continuous in collection to the present but having different starting times). This technique produces a consistent estimate of the matrix M_0 as well as of the matrix C, and is based on the maximum likelihood technique developed by Anderson (1957).

Valencia and Schaake (1973) developed another technique. They estimate matrices A and B from the equations

$$A = M_1 \, M_{01}^{-1} \tag{4.12}$$

$$B \, B^T = M_{02} - M_1 \, M_{01}^{-1} \, M_1^T \tag{4.13}$$

where M_{01} is the lag-zero correlation matrix M_0 computed from the first (N-1) vectors of the data, and M_{02} is computed from the last (N-1) vectors, where N is the number of data points (number of times sampled) in each of the n series.

Further Simplification

Sometimes in practice, the preservation of the lag-zero and lag-one autocorrelations and the lag-zero cross-correlations is enough. In such cases, i.e., when the lag-one cross-correlations are of no interest, a nice simplification can be made due to Matalas (1967, 1974). He defined matrix A as a diagonal matrix whose diagonal elements are the lag-one auto-correlation coefficients. With A defined as above, the lag-one cross-correlation of the generated series ($\rho'_{ij}(1)$) can be shown to be the product of the lag-zero cross-correlation ($\rho_{ij}(0)$) and the lag-one auto-correlation of the series ($\rho_{ii}(1)$), but of course different than the actual lag-one cross-correlation ($\rho_{ij}(1)$).

$$\rho'_{ij}(1) = \rho_{ij}(0)\ \rho_{ii}(1) \qquad (4.14)$$

By using $\rho'_{ij}(1)$ of equation (4.14) in place of the actual $\rho_{ij}(1)$, thus avoiding the actual computation of $\rho_{ij}(1)$ from the data, the desired statistical properties of the series are still preserved.

Higher Order Multivariate Models

The order p of a multivariate autoregressive model could be estimated from the plots of the autocorrelation and partial autocorrelation functions of the series (Salas *et al.*, 1980) as an extension of the univariate model identification, which is already a difficult and ambiguous task. However, in practice first and second order models are usually adequate and higher order models should be avoided (Box and Jenkins, 1976).

In any case, the multivariate multilag autoregressive model of order p takes the form

$$Z_t = \sum_{k=1}^{p} A_k Z_{t-k} + B H_t \tag{4.15}$$

and the matrices $A_1, A_2, \ldots A_p$, B are the solutions of the equations

$$M_i = \sum_{k=1}^{p} A_k M_{i-k} \quad , i = 1, 2, \ldots, p \tag{4.16}$$

$$B B^T = M_0 - \sum_{k=1}^{p} A_k M_k^T \tag{4.17}$$

where M_ℓ is the lag-ℓ correlation matrix. Equation (4.16) is a set of p matrix equations to be solved for the matrices $A_1, A_2, \ldots, A_p$, and matrix B is obtained from (4.17) using techniques already discussed. Here, the assumption of diagonal A matrices becomes even more attractive. For a multivariate second-order AR process the above simplification is illustrated in Salas and Pegram (1977) where the case of periodic (not constant) matrix parameters is also considered.

O'Connell (1974) studied the multivariate ARMA(1,1) model

$$Z_t = A Z_{t-1} + B H_t - C H_{t-1} \tag{4.18}$$

where A, B, and C are coefficient matrices to be determined from the data. Specifically they are solutions of the system of matrix equations

$$\begin{aligned} B\,B^T + C\,C^T &= S \\ C\,B^T &= T \end{aligned} \qquad (4.19)$$

where S and T are functions of the correlation matrices M_0, M_1 and M_2. Methods for solving this system are proposed by O'Connell (1974).

Explicit solutions for higher order multivariate ARMA models are not available and Salas _et_ _al_. (1980) propose an approximate multivariate ARMA(p,q) model.

CHAPTER 5

ESTIMATION OF MISSING MONTHLY RAINFALL VALUES-- A CASE STUDY

Introduction

This section compares and evaluates different methods for the estimation of missing values in hydrological time series. A case study is presented in which four of the simplified methods presented in Chapter 2 have been applied to a set of four concurrent 55 year monthly rainfall series from south Florida and the results compared. Also a recursive method for the estimation of missing values by the use of a univariate or multivariate stochastic model has been proposed and demonstrated. The theory already presented in Chapters 2, 3 and 4 is supplemented whenever needed.

Set Up of the Problem

The monthly rainfall series of four stations in the South Florida Water Management District (SFWMD) have been used in the analysis. These stations are:

Station A	:	MRF6038, Moore Haven Lock 1
Station 1	:	MRF6013, Avon Park
Station 2	:	MRF6093, Fort Myers WSO Ap.
Station 3	:	MRF6042, Canal point USDA.

For convenience the four stations will sometimes be addressed as A, 1, 2, 3 instead of their SFWMD identification numbers 6038, 6013, 6093 and 6042, respectively. Their locations are shown in the map of Fig. 5.1. Station A in the center is considered as the interpolation station (whose missing values are to be estimated) and the other three stations 1, 2 and 3 as the index stations. Care has been taken so that the three index stations are as close and as evenly distributed around the interpolation station as possible.

This particular set of four stations was selected because it exhibits many desired and convenient properties:

(1) the stations have an overlapping period of 55 years (1927-1981),

(2) for this 55 year period the record of the interpolation station (station A) is complete (no missing values),

(3) the three index stations have a small percent of missing values for the overlapping period (station 1: 2.7% missing, station 2: complete, and station 3: 1.2% missing values).

The 55 year length of the records is considered long enough to establish the historical statistics (e.g., monthly mean, standard deviation and skewness) and provides a monthly series of a satisfactory length (660 values) for fitting a univariate or multivariate ARMA model.

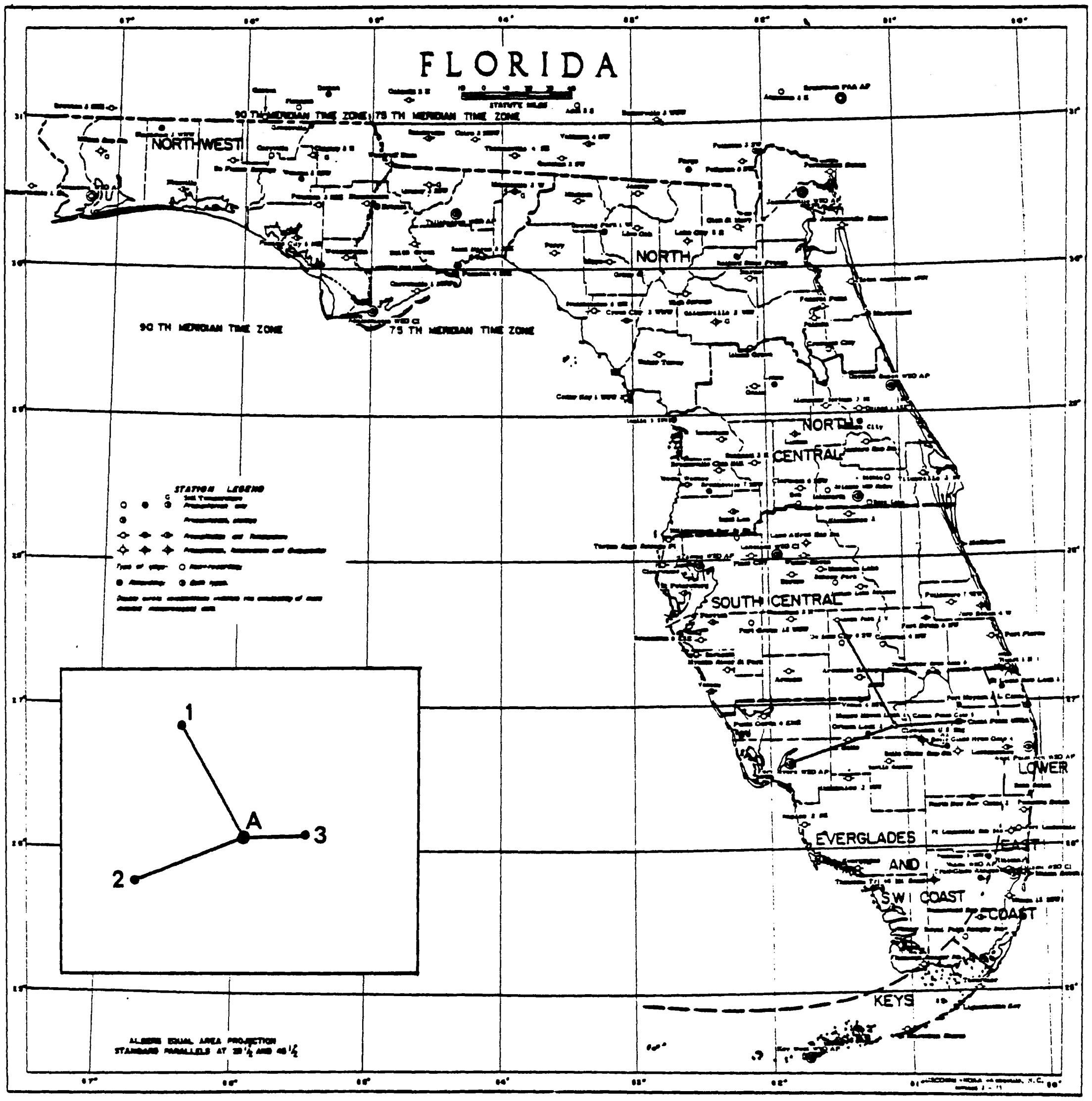


Fig. 5.1. The four south Florida rainfall stations used in the analysis.
A: 6038, Moore Haven Lock 1
1: 6013, Avon Park
2: 6093, Fort Myers WSO AP.
3: 6042, Canal Point USDA

The completeness of the series of the interpolation station permits the random generation of gaps in the series, corresponding to different percentages of missing values, with the method described in Chapter 1. After the missing values have been estimated by the applied models, the gaps are in-filled with the estimated values and the statistics of the new (estimated) series are compared with the statistics of the incomplete series and the statistics of the historical (actual) series. Also the statistical closeness of the in-filled (estimated) values to the hidden (actual) values provides a means for the evaluation and comparison of the methods.

When, for the estimation of a missing value of the interpolation station, the corresponding value of one or more index stations is also missing the latter is eliminated from the analysis, e.g., only the remaining one or two index stations are used for the estimation. Frequent occurrence of such concurrent gaps in both the interpolation and the index stations would alter the results of the applied method in a way that cannot be easily evaluated (e.g., another parameter such as the probability of having concurrent gaps should be included in the analysis). A small number of missing values in the selected index stations eliminates the possibility of such simultaneous gaps, and thus the effectiveness of the applied estimation procedures can be judged more efficiently.

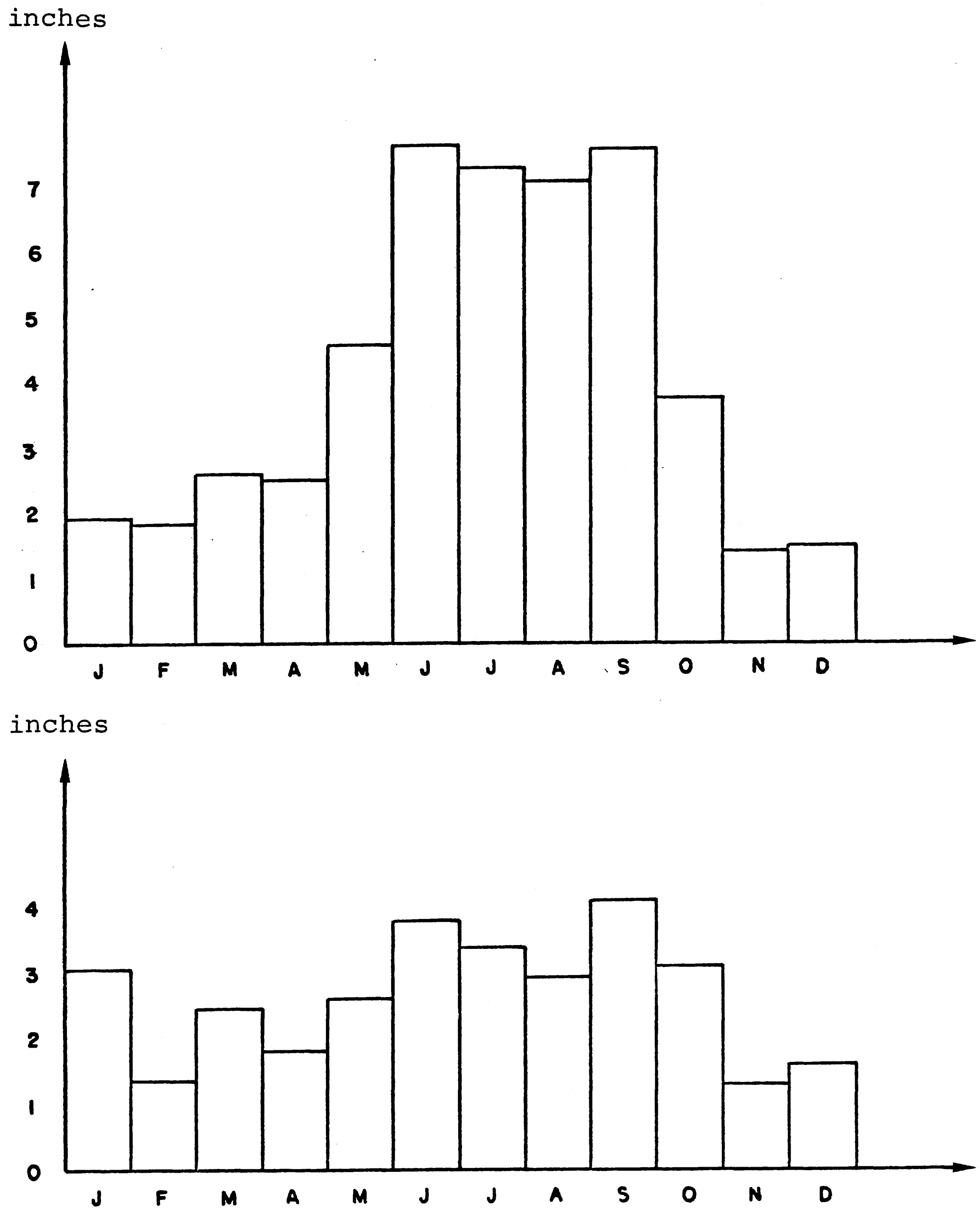
The statistical properties (e.g., monthly mean, standard deviation, skewness and coefficient of variation) of the truncated (to the 1927-1981 period) original monthly rainfall series for the four stations are shown on Tables C.1, C.2, C.3 and C.4 of Appendix C. Figure 5.2 shows the plot of the monthly means and standard deviations for station A. From these plots we observe that: (1) the plot of monthly means is in agreement with the typical plot for Florida shown in Fig. 1.1, and (2) months with a high mean usually have a high standard deviation. The only exception seems to be the month of January which in spite of its low mean exhibits a high standard deviation and therefore a very high coefficient of variation and an unusually high skewness. A closer look at the January rainfall values of station A shows that the unusual properties for that month are due to an extreme value of 21.4 inches of rainfall for January 1979, the other values being between 0.05 and 6.04 inches.

The three index stations 1, 2 and 3 are at distances 59 miles, 51 miles and 29 miles respectively from the interpolation station A.

Simplified Estimation Techniques

Techniques Utilized

From the simplified techniques presented in Chapter 2, the following four are applied for the estimation of missing



(b) monthly standard deviations

Fig. 5.2. Plot of the monthly means and standard deviations--station 6038 (1927 - 1981)

monthly rainfall values:

(1) the mean value method (MV)

(2) the reciprocal distances method (RD)

(3) the normal ratio method (NR), and

(4) the modified weighted average method (MWA).

These methods are all deterministic and are applied directly on the available data permitting thus a uniform and objective comparison of the results. The mean value plus random component method has not been included in this thesis.

The above four methods will be applied for five different percentages of missing values: 2%, 5%, 10%, 15% and 20%. These percentages cover almost 80% of all cases encountered in practice as has been shown in Table 1.1 (e.g., 80% of the stations have below 20% missing values). From the same table it can also be seen that almost 30% of the stations have below 5% missing values. Therefore, it would be of interest and practical use if we could generalize the results for the region of below 5% missing values since a large fraction of the cases in practice fall in this region.

The application of the first three methods (MV, RD, NR methods) is straightforward and no further comments need be made. However, some comments on the least squares (LS) method and the modified weighted average (MWA) method are necessary.

Least Squares Method (LS)

The least squares method although simple in principle involves an enormous amount of calculations, and for that reason it has been excluded from this study. For example, consider the case in which the interpolation station A is regressed on the three index stations 1, 2 and 3. The estimated values will be given by:

$$y' = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon \qquad (5.1)$$

where a, b_1, b_2, b_3 are the regression coefficients calculated from the available concurrent values of all the four variables. There are 12 such regression equations, one for each month. But if it happens that an index station (say, station 3) has a missing value simultaneously with the interpolation station, a new set of 12 regression equations is needed for the estimation, e.g.,

$$y' = a' + b_1' x_1 + b_2' x_2 + \varepsilon \qquad (5.2)$$

Unless this coincidence of simultaneously missing values is investigated manually so that only the needed least squares regressions are performed (Buck, 1960), all the possible combinations of regressions must otherwise be performed. This involves regressions among all the four variables (y; x_1, x_2, x_3), among the three of them (y; x_1, x_2), (y; x_1, x_3), (y; x_2, x_3) and between pairs of them (y; x_1), (y; x_2), (y; x_3), giving overall 7 sets of 12 regression equations. Because the regression coefficients are different for each percentage of missing values (since their calculation is based only on the existing concurrent values) the 84 (7 x 12) regressions must be repeated for each level of missing values (420 regressions overall for this study).

It could be argued that the same 12 regression equations (y; x_1, x_2, x_3) could be kept and a missing values x_i replaced by its mean $\bar{x}_i$ or by another estimate x_i'. In that case equation 5.1 would become

$$y' = a + b_1 x_1 + b_2 x_2 + b_3 x_3' + \varepsilon, \tag{5.3}$$

the coefficients of regression a, b_1, b_2, b_3 remaining unchanged. This in fact can be done, but then the method tested will not be the "pure" least squares method since the results will depend on the secondary method used for the estimation of the missing x_i values.

The coefficients a, b_1, b_2 and b_3 (equation 5.1) of the regression of the {y} series (of station A with 2% missing values) on the series {x_1}, {x_2} and {x_3} (of stations 1, 2 and 3 respectively) are shown in Table 5.1. In the same table the values of the squared multiple regression coefficient R^2 and the standard deviation of the {y} series are also shown. The numbers in parenthesis show the significance level α at which the parameters are significant (the percent probability of being nonzero is $(1-\alpha))100$. For

Table 5.1. Least Squares Regression Coefficients for Equation (5.1) and Their Significant Levels. The standard deviation, s, for each month is also given.

	a	b_1	b_2	b_3	R^2	s
	inches					inches
JAN	0.0059	0.1271	0.4994	0.3377	0.8046	3.076
	(0.9692)	(0.2790)	(0.0005)	(0.0017)	(0.0001)	
FEB	0.1355	0.2624	0.0086	0.5345	0.7033	1.365
	(0.5260)	(0.0025)	(0.9431)	(0.0001)	(0.0001)	
MAR	0.0052	0.1617	0.3457	0.4507	0.9142	2.464
	(0.9793)	(0.0138)	(0.0001)	(0.0001)	(0.0001)	
APR	0.7388	0.2405	0.2813	0.1919	0.4936	1.818
	(0.0273)	(0.0458)	(0.0156)	(0.1132)	(0.0001)	
MAY	2.1302	0.4046	-0.0591	0.2186	0.2752	2.583
	(0.0070)	(0.0115)	(0.7180)	(0.1308)	(0.0016)	
JUN	1.8765	0.2192	0.1108	0.3339	0.3351	3.812
	(0.1505)	(0.1576)	(0.4034)	(0.0133)	(0.0002)	
JUL	2.8601	-0.0345	0.3993	0.1885	0.2005	3.399
	(0.0750)	(0.7883)	(0.0131)	(0.1780)	(0.0154)	
AUG	2.0820	0.1771	0.2078	0.2660	0.1789	2.938
	(0.2065)	(0.1666)	(0.0787)	(0.0589)	(0.0248)	
SEP	0.0108	0.5102	0.2113	0.2450	0.5669	4.085
	(0.9916)	(0.0003)	(0.0893)	(0.0190)	(0.0001)	
OCT	-0.6985	0.3960	0.2287	0.4667	0.7749	3.073
	(0.0866)	(0.0020)	(0.0433)	(0.0001)	(0.0001)	
NOV	0.3167	0.3009	0.2473	0.1063	0.4575	1.228
	(0.1290)	(0.0030)	(0.0804)	(0.0069)	(0.0001)	
DEC	-0.2623	0.2332	0.3807	0.4381	0.7723	1.585
	(0.1987)	(0.1065)	(0.0084)	(0.0001)	(0.0001)	

example, for January the coefficient b_1 is not significant at the 5% significance level ($\alpha = 0.05$) since 0.279 is greater than 0.05, but the R^2 coefficient is significant even at 0.01% significance level ($\alpha = 0.0001$). The significance levels correspond to the "t-test" for the regression coefficients and to the "F-test" for the R^2 coefficients. The standard deviation, s, of the $\{y\}$ series is also listed since the random component is given by

$$\varepsilon = \sqrt{1-R^2}\ s \tag{5.4}$$

as has already been discussed in Chapter 2.

It is interesting to note, that although the multiple regression coefficient R^2 varies for each month from as low as 0.18 to as high as 0.91 it is always significant at the 5% significance level. The months of July and August exhibit the lowest (although significant) correlation coefficients as is expected for Florida. The physical reason for these low correlations is that in the summer most rainfall is convective, whereas in other months there is more cyclonic activity. Rainfall from scattered thunderstorms is simply not as correlated with that of nearby areas as is rainfall from broad cyclonic activity. Thus, on the basis of the regressions shown in Table 5.1, the least squares method would be expected to perform least well in the summer in Florida, but this point is not validated in this thesis.

Modified Weighted Average Method (MWA)

For the modified weighted average method the twelve (3x3) covariance matrices of the three index stations have been calculated for each month using equation (2.9) and (2.10), and are shown in Table C.11 (appendix C). Also the monthly standard deviations, s_y, have been estimated from the known {y} series, and the monthly standard deviations, s'_y have been calculated by equation (2.11) using the calculated covariance matrices. Notice that although the twelve s_y values (as calculated from the actual data and which we want to preserve) are different at different percentages of missing values, the twelve s'_y values (that depend only on the weights a_i and the covariance matrix of the index stations) are calculated only once. The correction coefficients f ($f = s_y/s'_y$) for each month and for each different percentage of missing values which must be applied on matrix A (equation 2.21) are shown in Table 5.2.

From this table it can be seen that if the simple weighted average scheme of equation (2.3) were used for the generation, the standard deviation of November would be overestimated (by a factor of approximately 2) and the standard deviation of all other months would be underestimated (e.g., by a factor of approximately 0.5 for the month of January). We also observe that due to small changes of s_y for different percentages of missing values, the correction factor f does not vary much either, but tends

Table 5.2. Correction Coefficient, f, for Each Month and for Each Different Percent of Missing Values ($f = s_y/s'_y$).

	2%	5%	10%	15%	20%
JAN	1.777	1.777	1.795	1.897	1.872
FEB	1.129	1.142	1.136	1.199	1.188
MAR	1.178	1.207	1.177	1.003	1.009
APR	1.089	0.980	1.061	1.051	1.054
MAY	1.269	1.197	1.212	1.222	1.360
JUN	1.214	1.173	1.192	1.228	1.242
JUL	1.338	1.345	1.386	1.390	1.491
AUG	1.424	1.414	1.425	1.432	1.369
SEP	1.313	1.328	1.325	1.210	1.331
OCT	1.258	1.273	1.218	1.229	1.314
NOV	0.533	0.537	0.509	0.583	0.572
DEC	1.161	1.140	1.169	1.172	1.248

to be slightly greater the greater the percent of missing values.

The modified weighted average scheme theoretically preserves the mean and variance of the series as has been shown in Chapter 2. But this is true for a series that has been generated by the model and not for a series that is a mix of existing values and values generated (estimated) by the model. This illustrates the difference between the two concepts: "generation of data by a model" and "estimation of missing values by a model." A method for generation of data which is considered "good" in the sense that it preserves first and second order statistics is not necessarily "good" for the estimation of missing values. In fact, it may give statistics comparable to the ones given from a simpler estimation technique which does not preserve the statistics, even as a generation scheme. Theoretically, for a "large" number of missing values, the estimation model operates as a generation model and thus preserves the "desired" statistics, but practically, for this large amount of missing values the "desired" statistics (calculated from the few existing values) are of questionable reliability. Only for augmentation of the time series (extension of the series before the first or after the last point) will the modified weighted average scheme or other schemes that preserve the "desired" statistics be expected to work better than the simple weighted average schemes.

One other disadvantage of the modified weighted average scheme as well as of the least squares scheme is that negative values may be generated by the model. Since all hydrological variables are positive, the negative generated values are set equal to zero, thus altering the statistics of the series. This is also true for all methods that involve a random component and is mainly due to "big" negative values taken on by the random deviate.

The number of negative values, estimated by the MWA method, which have been set equal to zero in the example that follows were 1, 1, 6, 4, and 9 values for the 2%, 5%, 10%, 15% and 20% levels of missing values, respectively.

The effect of the values arbitrarily set to zero cannot be evaluated exactly, but what can be intuitively understood is that a distortion in the distribution is introduced. A transformation that prevents the generation of negative values could be performed on the data before the application of the generation scheme. Such a transformation is, for example, the logarithmic transformation since its inverse applied on a negative value exists, and the mapping of the transformed to the original data and vice versa is one to one (this is not true for the square root transformation).

Comparison of the MV, RD, NR and MWA Methods

The performance of each method applied for the estimation of the missing values will be evaluated by comparing the estimated series (existing plus estimated values) to the incomplete series (really available in practice) and to the actual series (unknown in practice, but known in this artificial case). The criteria that will be used for the comparison of the method will be the following:

(1) the bias in the mean as measured (a) by the difference between the mean of the estimated series, $\bar{y}_e$, and the mean of the incomplete series, $\bar{y}_i$ (i = 1, 2, 3, 4, 5 for five different percentages of missing values), and (b) by the difference between the mean of the estimated series, $\bar{y}_e$ and the mean of the actual series, $\bar{y}_a$;

(2) the bias in the standard deviation as measured (a) by the ratio of the standard deviation of the estimated series, s_e, to the standard deviation of the incomplete series, s_i and (b) by the ratio of the standard deviation of the estimated series, s_e, to the standard deviation of the actual series, s_a;

(3) the bias in the lag-one and lag-two correlation coefficients as measured by the difference of the correlation coefficient of the estimated series, r_e, to the correlation coefficient of the actual series, r_a;

(4) the bias of the estimation model as given by the mean of the residuals, $\bar{y}_r$, i.e., the mean of the differences between the in-filled (estimated) and hidden (actual) values (this is also a check to

detect a consistent over- or under-estimation of the method);

(5) the accuracy as determined by the variance of the residuals (differences between estimated and actual values) of the whole series, s_r^2;

(6) the accuracy as determined by the variance of the residuals of only the estimated values, $s_{r,e}^2$; and

(7) the significance of the biases in the mean, standard deviation and correlation coefficients as determined by the appropriate test statistic for each (see appendix A).

Table 5.3 presents the statistics of the actual series (ACT), of the incomplete series (INC) and of the estimated series by the mean value method (MV), by the reciprocal distances method (RD), by the normal ratio method (NR) and by the modified weighted average method (MWA). The mean ($\bar{y}$), standard deviation (s), coefficient of variation (c_v) coefficient of skewness (c_s), lag-one and lag-two correlation coefficients (r_1, r_2) of the above series considered as a whole have then been calculated.

Regarding comparison of the means, the following can be concluded from Table 5.4:

(1) the bias in the mean in all cases is not significant at the 5% significance level as shown by the appropriate t-test;

Table 5.3. Statistics of the Actual (ACT), Incomplete (INC) and Estimated Series (MV, RD, NR, MWA).

	$\bar{y}$	s	c_v	c_s	r_1	r_2
ACT	4.126	3.673	89.040	1.332	0.366	0.134
			2% missing values			
INC	4.116	3.680	89.397	1.346	--	--
MV	4.125	3.663	88.808	1.335	0.371	0.130
RD	4.124	3.674	89.092	1.336	0.367	0.133
NR	4.114	3.666	89.104	1.339	0.368	0.131
MWA	4.113	3.674	89.331	1.342	0.363	0.131
			5% missing values			
INC	4.113	3.671	89.249	1.341	--	--
MV	4.101	3.610	88.040	1.352	0.372	0.139
RD	4.127	3.696	89.550	1.359	0.369	0.133
NR	4.105	3.674	89.501	1.349	0.367	0.131
NWA	4.116	3.720	90.386	1.388	0.364	0.126
			10% missing values			
INC	4.144	3.705	89.405	1.350	--	--
MV	4.134	3.603	87.152	1.346	0.379	0.159

continued

Table 5.3. Continued.

	$\bar{y}$	s	c_v	c_s	r_1	r_2
ACT	4.126	3.673	89.040	1.332	0.366	0.134
RD	4.150	3.689	88.884	1.301	0.380	0.166
NR	4.120	3.652	88.633	1.321	0.377	0.155
MWA	4.127	3.725	90.244	1.286	0.376	0.162
15% missing values						
INC	4.135	3.671	88.767	1.268	--	--
MV	4.106	3.513	85.567	1.270	0.399	0.133
RD	4.177	3.688	86.862	1.224	0.372	0.132
NR	4.135	3.691	86.854	1.236	0.379	0.133
MWA	4.134	3.650	88.291	1.248	0.357	0.123
20% missing values						
INC	4.082	3.701	90.673	1.404	--	--
MV	4.124	3.495	84.749	1.333	0.408	0.160
RD	4.231	3.723	87.993	1.865	0.370	0.156
NR	4.125	3.601	87.307	1.298	0.377	0.152
MWA	4.168	3.741	89.758	1.273	0.354	0.153

Table 5.4. Bias in the Mean

	INC	MV	RD	NR	MWA	
	$(\bar{y}_e - \bar{y}_i)$					$\bar{y}_i$
2%	0.	0.009	0.008	0.002	0.003	4.116
5%	0.	-0.012	0.014	-0.008	0.003	4.113
10%	0.	-0.010	0.006	-0.024	-0.017	4.144
15%	0.	-0.089	0.042	0.000	-0.001	4.135
20%	0.	0.042	0.149	0.043	0.086	4.082
	$(\bar{y}_e - \bar{y}_a)$					$\bar{y}_a$
2%	-0.010	-0.001	-0.002	-0.012	-0.013	4.126
5%	-0.013	-0.025	0.001	-0.021	-0.010	
10%	0.018	0.008	0.024	-0.006	0.001	
15%	0.009	-0.020	0.051	0.009	0.008	
20%	-0.044	-0.002	0.105	-0.001	0.042	

(2) the bias in the mean of the incomplete series is relatively small but becomes larger the higher the percent of missing values;

(3) at high percents of missing values the NR method gives the less biased mean;

(4) except for the RD method which consistently overestimates the mean (the bias being larger the higher the percent of missing values), the other methods do not show a consistent over or underestimation.

Regarding comparison of the variances the following can be concluded from Table 5.5:

(1) Although slight, the bias in the standard deviation is always significant, but this is so because the ratio of variances would have to equal 1.0 exactly to satisfy the F-test (i.e., be unbiased) with as large a number of degrees of freedom as in this study;

(2) the MV method always gives a reduced variance as compared to the variance of the incomplete series and of the actual series, the bias being larger the higher the percent of missing values;

(3) the bias in the standard deviation of the incomplete series is small;

(4) there is no consistent over or under-estimation of the variance by any of the methods (except the MV method);

Table 5.5. Bias in the Standard Deviation

	INC	MV	RD	NR	MWA	
			s_e/s_i			s_i
2%	1.	0.995	0.998	0.996	0.998	3.680
5%	1.	0.983	1.007	1.001	1.013	3.671
10%	1.	0.972	0.996	0.986	1.005	3.705
15%	1.	0.957	0.988	0.978	0.994	3.671
20%	1.	0.944	1.006	0.973	1.011	3.701
			s_e/s_a			s_a
2%	1.002	0.997	1.000	0.998	1.000	3.673
5%	0.999	0.983	1.006	1.000	1.013	
10%	1.009	0.981	1.004	0.994	1.014	
15%	0.999	0.956	0.988	0.978	0.994	
20%	1.008	0.952	1.014	0.980	1.019	

(5) the MWA method does not give less biased variance even at the higher percent of missing values tested, as compared to the RD and NR methods.

Regarding comparison of the correlation coefficients the following can be concluded from Table 5.6:

(1) the bias in the correlation coefficients is in all cases not significant at the 5% significance level as shown by the appropriate z-test;

(2) the MV method gives the largest bias in the correlation coefficients, the bias increasing the higher the percent of missing values, with a possible effect on the determination of the order of the model;

(3) all methods (except the MWA method) consistently overestimate the serial correlation coefficient of the incomplete series but not the serial correlation of the actual series and therefore is not considered a problem;

(4) the RD method seems to give a correlogram that closely follows the correlogram of the actual series.

Regarding accuracy of the methods the following can be concluded from Table 5.7:

(1) no method seems to consistently over or underestimate the missing values at all percent levels, but at high percent levels the missing values are overestimated by all methods;

Table 5.6. Bias in the Lag-One and Lag-Two Correlation Coefficients.

	INC	MV	RD	NR	MWA	
			$(r_{1,e} - r_{1,a})$			$r_{1,a}$
2%	--	0.005	0.001	0.002	-0.003	0.366
5%	--	0.006	0.003	0.001	-0.002	
10%	--	0.013	0.014	0.011	0.010	
15%	--	0.033	0.006	0.013	-0.009	
20%	--	0.042	0.004	0.011	-0.012	
			$(r_{2,e} - r_{2,a})$			$r_{2,a}$
2%	--	-0.004	-0.001	-0.003	-0.003	0.134
5%	--	0.005	-0.001	-0.003	-0.008	
10%	--	0.025	0.032	0.021	0.028	
15%	--	-0.001	-0.002	-0.001	-0.011	
20%	--	0.026	0.022	0.018	0.019	

Table 5.7. Accuracy--Mean and Variance of the Residuals
N_o = number of missing values
N^o = total number of values = 660.

	INC	MV	RD	NR	MWA	
		$\mu_r = \Sigma(y_e - y_a)/N_o$				N_o
2%	--	-0.043	-0.061	-0.570	-0.589	13
5%	--	-0.440	0.034	-0.380	-0.176	33
10%	--	0.007	0.156	-0.113	-0.046	62
15%	--	-0.175	0.338	0.074	0.105	98
20%	--	0.037	0.502	0.038	0.200	130
		$s^2_{r,e} = \Sigma(y_e - y_a)^2/(N_o-2)$				
2%	--	5.037	2.874	3.149	4.585	
5%	--	8.610	3.656	3.411	5.340	
10%	--	7.892	4.239	3.484	5.187	
15%	--	7.620	4.630	3.958	5.816	
20%	--	5.224	4.891	3.681	4.898	

Table 5.7. Continued.

	INC	MV	RD	NR	MWA
			$s_r^2 = \Sigma(y_e - y_a)^2/(N-2)$		
2%	--	0.084	0.048	0.053	0.077
5%	--	0.406	0.172	0.161	0.252
10%	--	0.720	0.387	0.318	0.473
15%	--	1.112	0.675	0.577	0.849
20%	--	1.016	0.951	0.716	0.953

(2) the NR method is the more accurate method especially at high percents of missing values (i.e., it gives the smaller mean and variance of the residuals).

Univariate Model

Model Fitting

Before considering the problem of missing values the problem of fitting an ARMA(p,q) model to the monthly rainfall series of the south Florida interpolation station will be considered.

The observed rainfall series has been normalized using the square root transformation and the periodicity has been removed by standardization. The reduced series, approximately normal and stationary, is then modeled by an ARMA(p,q) model. The ACF of the reduced series, as shown in Fig. 5.3, implies a white noise process since almost all the autocorrelation coefficients (except at lag-3 and lag-12) lie inside the 95 percent confidence limits.

Of course, it is unsatisfying to accept the white noise process as the "best" model for our series and an attempt is made to fit an ARMA(1,1) model to the series. The selection of an ARMA model and not an AR or MA model is based on the following reasons:

(1) The observed rainfall series contains important observational errors and so it is assumed to be the sum

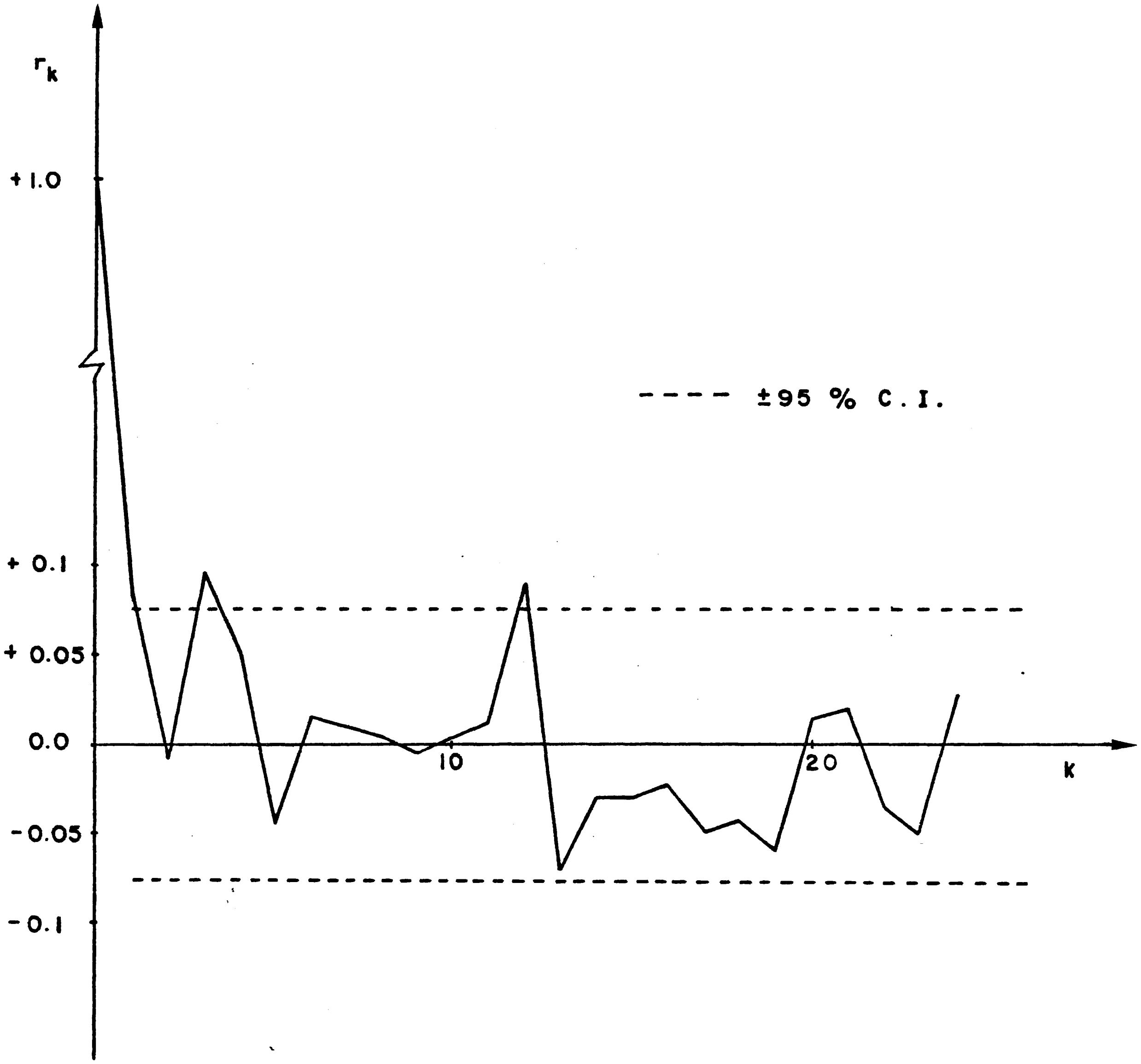


Fig. 5.3. Autocorrelation function of the normalized and standardized monthly rainfall series of Station A.

of two series: the "true" series and the observational error series (signal plus noise). Therefore, even if the "true" series obeys an AR process, the addition of the observational error series is likely to produce an ARMA model:

$$\begin{aligned} &AR(p) + \text{white noise} = ARMA(p,p) \\ &AR(p) + AR(q) = ARMA(p+q, \max(p,q)) \qquad (5.5) \\ &AR(p) + MA(q) = ARMA(p, p+q) \end{aligned}$$

The same can be said if the "true" series is an MA process and the observational error series an AR process but not if the latter is an MA process or a white noise process:

$$\begin{aligned} &MA(p) + AR(q) = ARMA(q,p+q) \\ &MA(p) + MA(q) = MA(\max(p,q)) \qquad (5.6) \\ &MA(p) + \text{white noise} = MA(p) \end{aligned}$$

(Granger and Morris, 1976; Box and Jenkins, 1976, Appendix A4.4).

It is understood, that the addition of any observational series to an ARMA process of the "true" series will give again an ARMA process. For example,

$$\begin{aligned} ARMA(p,q) + \text{white noise} &= ARMA(p,p) \text{ if } p > q \qquad (5.7) \\ &= ARMA(p,q) \text{ if } p \leq q \end{aligned}$$

from which it can also be seen that the addition of an observational error may not always change the order of the model of the "true" process.

(2) One other situation that leads exactly, or approximately, to ARMA models is the case of a variable which obeys a simple model such as AR(1) if it were recorded at an interval of K units of time but which is actually observed at an interval of M units (Granger and Morris, 1976, p. 251).

All these results suggest that a number of real data situations are all likely to give rise to ARMA models; therefore, an ARMA(1,1) model will be fitted to the observed monthly rainfall series of the south Florida interpolation station. The preliminary estimate of ϕ_1 (equation 3.23) is -0.08163, and the preliminary estimate of θ_1 (equations 3.21 for k = 0, 1, 2) is the solution of the quadratic equation

$$0.1656\ \theta_1^2 + 1.0204\ \theta_1 + 0.1656 = 0 \quad . \tag{5.8}$$

Only the one root $\theta_1 = -0.1667$ is acceptable, the second lying outside the unit circle. These preliminary estimates of ϕ_1 and θ_1 become now the initial values for the determination of the maximum likelihood estimates (MLE). In general, the choice of the starting values of ϕ and θ does not significantly affect the parameter estimates (Box and Jenkins, 1976, p. 236), but this was not the case for the

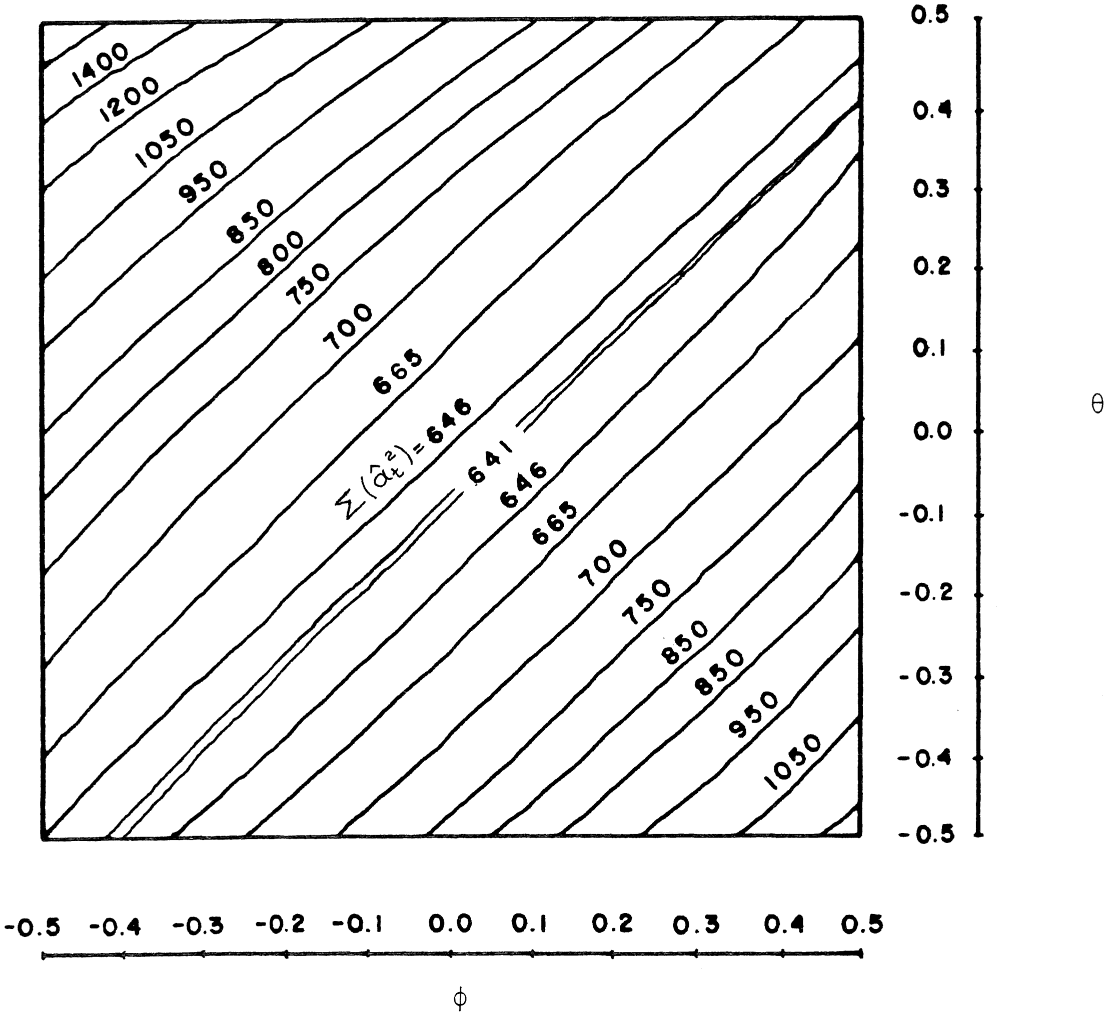


Fig. 5.4. Sum of squares of the residuals, $\Sigma(\hat{a}_t^2)$, of an ARMA (1,1) model fitted to the rainfall series of station A.

Table 5.8. Initial Estimates and MLE of the Parameters ϕ and θ of an ARMA(1,1) model fitted to the rainfall series of station A.

	Initial Estimates		Max. Likelihood Estimates	
Model	ϕ	θ	ϕ	θ
A	-0.0816	0.0	-0.0088	-0.0989
B	-0.0816	-0.1667	-0.3140	-0.4056
C	0.1	0.0	0.0537	-0.0278
D	-0.4	-0.5	-0.4064	-0.4939

south Florida rainfall series under study. In particular different initial estimates of ϕ_1 and θ_1 have been tested and the MLE of the parameters are compared in Table 5.8. The MLE have been calculated using the IMSL subroutine FTMXL which uses a modified steepest descent algorithm to find the values of ϕ and θ that minimize the sum of squares of the residuals (Box and Jenkins, 1976, p. 504).

The drastic changes in parameter values together with the idea that the process may be a white noise process suggest a plot of the sum of squares of the residuals for the visual detection of anomalies. The sum of squares grids and contours are shown in Fig. 5.4. We observe that there is not a well defined point where the sum of squares becomes a minimum but rather a line (contour of the value 641) on which the sum of squares has an almost constant value equal to the minimum. In such case combinations of parameter values give similar sum of squares of residuals and a change in the AR parameter can be nearly compensated by a suitable change in the MA parameter.

From the comparison of the parameters ϕ and θ (Table 5.8) of the four ARMA(1,1) models one cannot say that they all correspond to the same process. But this can in fact be illustrated by converting the four models to their "random shock form" (MA(∞) processes) or their "invertible form" (AR(∞) processes).

An ARMA(1,1) process

$$(1-\phi_1 B)\ z_t = (1-\theta_1 B)\ a_t \tag{5.9}$$

can be also written as

$$z_t = (1-\theta_1 B)\ (1-\phi_1 B)^{-1}\ a_t \tag{5.10}$$

which can be expanded in the convergent form

$$z_t = [1 + (\phi_1-\theta_1)B + \phi_1(\phi_1-\theta_1)B^2 + \phi_1^2(\phi_1-\theta_1)B^3 + \ldots]\ a_t \tag{5.11}$$

provided that the stationarity condition ($|\phi_1| < 1$) is satisfied. Then the four models of Table 5.8 become:

$$
\begin{aligned}
&(A) : \quad z_t = a_t + 0.090\, a_{t-1} - 0.001\, a_{t-2} + \dots \\
&(B) : \quad z_t = a_t + 0.092\, a_{t-1} - 0.029\, a_{t-2} + \dots \qquad (5.12) \\
&(C) : \quad z_t = a_t + 0.082\, a_{t-1} + 0.004\, a_{t-2} + \dots \\
&(D) : \quad z_t = a_t + 0.088\, a_{t-1} - 0.036\, a_{t-2} + \dots
\end{aligned}
$$

In the same way the ARMA(1,1) model may be written in the "invertible form"

$$(1-\phi_1 B)\ (1-\theta_1 B)^{-1}\ z_t = a_t \qquad (5.13)$$

which can be expanded as

$$[1 - (\phi_1-\theta_1)B - \theta_1(\phi_1-\theta_1)B^2 - \theta_1^2(\phi_1-\theta_1)B^3 - \dots]\ z_t = a_t \qquad (5.14)$$

given that the invertibility condition ($|\theta_1| < 1$) is satisfied. Then the four models become:

$$
\begin{aligned}
&(A) : \quad z_t = a_t + 0.090\, z_{t-1} - 0.009\, z_{t-2} + \dots \\
&(B) : \quad z_t = a_t + 0.092\, z_{t-1} - 0.037\, z_{t-2} + \dots \qquad (5.15) \\
&(C) : \quad z_t = a_t + 0.082\, z_{t-1} - 0.002\, z_{t-2} + \dots \\
&(D) : \quad z_t = a_t + 0.088\, z_{t-1} - 0.043\, z_{t-2} + \dots
\end{aligned}
$$

From the "random shock" form of the four models (equations 5.12) and from their "invertible form" (equations 5.15) the following remarks can be made:

(1) Although from the comparison of the ϕ and θ coefficients (Table 5.8) of the four ARMA(1,1) models one cannot say that they all correspond to the same process, the comparison of the MA coefficients (θ_1, θ_2, θ_3, ...) of equations (5.12) or the AR coefficients (ϕ_1, ϕ_2, ϕ_3, ...) of equations (5.15) imply that indeed all four models belong to the same process.

(2) Because the nonzero ϕ_2 (and θ_2) coefficients of z_{t-2} (and a_{t-2}) terms while small are of similar magnitude to the coefficients ϕ_1 (and θ_1), one cannot say that the "truncated" AR(1) or MA(1) model will fully describe the time series, but instead more terms are needed. On the other hand, we observe that the ϕ_1 coefficient so obtained (different for each model) is in the range of 0.082 to 0.090 and is greater than the coefficient ϕ_1 that would have been obtained by a direct fitting of an AR (1) model to the series (the latter would be $\phi_1 = r_1 = 0.0068$).

(3) It should also be noted that all the above models fitted to the series give residuals that pass the portemanteau goodness of fit test. As it can be seen from equation (5.12) the impulse response function (e.g., the weights ψ_j applied on the a_j's when the model is written in the "random shock form") dies off very quickly in all the models, and there is thus no doubt as to the application of the portemanteau test

(see Appendix A). The values of Q for each model (calculated from equation A.1 using K = 60) are: Q_A = 67.80, Q_B = 67.26, Q_C = 67.73 and Q_D = 67.39, all smaller than the χ^2 value with 58 degrees of freedom at a 5% significance level, $\chi^2_{58,5\%}$ = 79.1. It can also be seen that the values of Q for all models are almost equal, suggesting an equally good fit of the series by all the four models.

One other interesting question that could be asked is, given a specific ARMA(p,q) model whether or not this could have arisen from some simpler model. "Simplifications are not always possible as conditions on the coefficients of the ARMA model need to be specified for a simpler model to be realizable" (Granger and Morris, 1976, p. 252). At this stage with coefficients that are so instable it is meaningless to test the four ARMA models for simplification. However, this test will be made after a unique and stable model has been obtained through the following proposed algorithm.

Proposed Estimation Algorithm

The problem of estimation of missing values will be combined with the problem of stabilizing the coefficients of the ARMA(1,1) model in a recursive algorithm which will have solved both problems uniquely upon convergence.

The incomplete series ($\underline{S}_0$) is filled-in with some initial estimates of the missing values (these initial estimates can be simply the monthly means or even zeroes as will be shown). Denote by $\underline{S}_1$ this initial series. An ARMA (1,1) model is fitted to the series $\underline{S}_1$ and its coefficients ϕ_1 and θ_1 are used to update the first estimates of the missing values. For example, suppose that a gap of size k (k missing values) exists in the series $\underline{S}_0$:

$$\begin{array}{lllllllll} \text{Series } \underline{S}_0: & \ldots & z_{t-1} & z_t & \ldots & & z_{t+k+1} & z_{t+k+2} & \ldots \\ \text{Series } \underline{S}_1: & \ldots & z_{t-1} & z_t & z'_{t+1} & \ldots & z'_{t+k} & z_{t+k+1} & z_{t+k+2} \ \ldots \end{array} \tag{5.16}$$

where $z'_{t+1}, \ldots, z'_{t+k}$ are the initial estimates of the missing values. These values $z'_{t+1}, \ldots, z'_{t+k}$ are then replaced by the forecasted values $\hat{z}_t(1), \ldots, \hat{z}_t(k)$ by the model, made at origin t and for lead times $\ell = 1, \ldots, k$. These forecasts are the minimum mean square error forward forecasts as developed by Box and Jenkins (1976). For an ARMA(1,1) model with coefficients ϕ_1 and θ_1, the minimum mean square error forecasts $\hat{z}_t(\ell)$ of $z_{t+\ell}$, where ℓ is the lead time, are:

$$\begin{array}{ll} \hat{z}_t(\ell) = \phi_1 z_t - \theta_1 a_t & , \quad \ell = 1 \\ \hat{z}_t(\ell) = \phi_1 \hat{z}_t(\ell-1) & , \quad \ell = 2, \ldots, k \end{array} \tag{5.17}$$

from which it can be seen that only the one step ahead forecast depends directly on a_t, and the forecasts at longer lead times are influenced indirectly (Box and Jenkins, 1976, Ch. 5). The forecasting procedure in repeated for the estimation of all the gaps, and the newly estimated values are used in equations (5.17). These forecasts now become the new estimates of the missing values and they replace the old estimates giving the new series $\underline{S}_2$. An ARMA(1,1) model is then fitted to the new series $\underline{S}_2$ and the new coefficients ϕ_1 and θ_1 are found (different from the previous ones). Then the estimated values (forecasts from the previous model) are replaced by the forecasts by the new model, giving the new series $\underline{S}_3$, etc. The procedure is repeated until the model and the series stabilize in the sense that the parameters ϕ_1 and θ_1 of the model as well as the estimates of the missing values do not change between successive estimates within a specified tolerance.

Schematically the algorithm is presented in Fig. 5.5 where $\underline{S}_0$ denotes the incomplete series, $\underline{M}_0$ the method used for the initial estimation, $\underline{S}_i$ the estimated series at the ith iteration, and $\underline{M}_i$ the model (e.g., the set of parameters ϕ_1 and θ_1, $(\phi_1, \theta_1)_i$) fitted to the series $\underline{S}_i$. The notation $\underline{M}_i \xrightarrow{\sim} \underline{M}_{i+1}$ and $\underline{S}_i \xrightarrow{\sim} \underline{S}_{i+1}$ is introduced to denote the stabilization of the model and series respectively after i iterations. The above algorithm will be addressed as RAEMV-U (a recursive algorithm for the estimation of missing values--univariate model).

Application of the Algorithm on the Monthly Rainfall Series

The proposed recursive algorithm (RAEMV-U) has been applied for the estimation of missing monthly rainfall

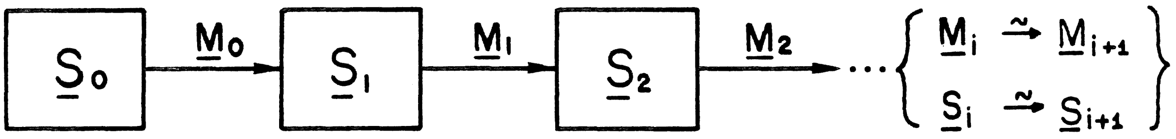

$\underline{S}_0$
$\underline{M}_0$
$\underline{S}_1$
$\underline{M}_1$
$\underline{S}_2$
$\underline{M}_2$
...
$\underline{M}_i \xrightarrow{\sim} \underline{M}_{i+1}$
$\underline{S}_i \xrightarrow{\sim} \underline{S}_{i+1}$

Fig. 5.5. Recursive algorithm for the estimation of missing values--univariate model (RAEMV-U). $\underline{S}_i$ denotes the series, and $\underline{M}_i$ the model, $(\underline{\phi},\theta)_i$, at the ith iteration.

values in the series of the south Florida interpolation station (station 6038). Different levels of percentage of missing values have been tested and the results for the 10% and 20% levels are presented herein. Tables 5.9 and 5.10 show the results for the 10% and 20% levels of missing values respectively. The starting series $\underline{S}_0$ is the incomplete series (with 10% or 20% the values missing). Four different methods $\underline{M}_0$ (MV, RD, NR, and zeros) have been applied to the incomplete series, $\underline{S}_0$, providing different starting series, $\underline{S}_1$, for the algorithm. Thus, its dependence on the initial conditions has also been tested.

Results of the Method

From Tables 5.9 and 5.10 the following can be concluded:

(1) The algorithm converges very rapidly and independently of the initial estimates, thus suggesting the convenient replacement of the missing values by zeros to start the algorithm.

(2) The greater the percent of missing values the slower the algorithm converges (6 iterations were needed for the 10% and 8 for the 20% to obtain accuracy to the third decimal place) as was expected since a larger part of the series is changing its values at each iteration and thus more iterations are needed to achieve equilibrium.

Table 5.9. Results of the RAEMV-U Applied at the 10% Level of Missing Values. Upper Value is ϕ_1, Lower Value is θ_1.

$\underline{M}_0$	MV	RD	NR	Zeroes
$\underline{M}_1$	0.0255 -0.0208	0.5092 0.5292	0.5018 0.4272	0.5059 0.2498
$\underline{M}_2$	0.4851 0.4323	0.5010 0.4166	0.5096 0.4336	0.4999 0.4313
$\underline{M}_3$	0.5149 0.4406	0.5110 0.4355	0.5094 0.4332	0.5088 0.4333
$\underline{M}_4$	0.5087 0.4322	0.5093 0.4329	0.5094 0.4333	0.5094 0.4333
$\underline{M}_5$	0.5096 0.4335	0.5095 0.4334	0.5096 0.4334	0.5095 0.4334
$\underline{M}_6$	0.5095 0.4333	0.5095 0.4333	0.5095 0.4333	0.5095 0.4334

Table 5.10. Results of the RAEMV-U Applied at the 20% Level of Missing Values. Upper Value is ϕ_1, Lower Value is θ_1.

$\underline{M}_0$	MV	RD	NR	Zeroes
$\underline{M}_1$	0.0954	0.5023	0.5021	0.5756
	-0.0069	0.4173	0.4159	0.2587
$\underline{M}_2$	0.0738	0.1167	0.1189	0.2926
	-0.0344	-0.0311	-0.0289	0.1187
$\underline{M}_3$	0.0789	0.0369	0.0377	0.0762
	-0.0276	-0.0693	-0.0688	-0.0458
$\underline{M}_4$	0.0774	0.0910	0.0908	0.0526
	-0.0296	-0.0125	-0.0128	-0.0503
$\underline{M}_5$	0.0778	0.0745	0.0746	0.0863
	-0.0291	-0.0334	0.0333	-0.0184
$\underline{M}_6$	0.0777	0.0786	0.0786	0.0756
	-0.0292	-0.0281	-0.0281	-0.0319
$\underline{M}_7$	0.0777	0.0775	0.0775	0.0783
	-0.0292	-0.0295	-0.0295	-0.0285
$\underline{M}_8$	0.0777	0.0778	0.0778	0.0776
	-0.0292	-0.0291	-0.0291	-0.0293

(3) For a specific percent of missing values the algorithm converges to the same point (e.g., same model and same series) independently of the initial estimates of the missing values.

(4) For a different percent of missing values the same series converges to a "different" point (e.g., "different" model and "different" series). This was expected since the constant information in the system (existing values) is different in each case, and thus a different model describes it better.

Diagnostic checking on the residuals from the two final models is performed using the portemanteau goodness of fit test. Denote the two models (at 10% and 20% levels) by $\underline{M}\text{-}U^{10}$ and $\underline{M}\text{-}U^{20}$ respectively, the U denoting that a univariate model has been fitted to the series. Then

$$\begin{aligned} \underline{M}\text{-}U^{10}: &\quad \phi = 0.5095 \quad \theta = 0.4333 \\ \underline{M}\text{-}U^{20}: &\quad \phi = 0.0777 \quad \theta = 0.0292. \end{aligned} \tag{5.18}$$

The values of Q for each model are $Q(M\text{-}U^{10}) = 26.54$ and $Q(M\text{-}U^{20}) = 30.22$ (calculated by equation A.1 using K=30) which are both smaller than the χ^2 value with 28 degrees of freedom at a 5% significance level: $\chi^2_{28,5\%} = 41.3$. Notice also that $Q(M\text{-}U^{10}) < Q(M\text{-}U^{20})$, indicating that the final model fitted to the series when 10% of the values were missing has a better fit than the model fitted to the series when 20% of the values were missing as expected.

Also, now that the final ARMA(1,1) model is stable we can ask the question "can it be simplified to an AR(1) plus white noise?". For an ARMA(1,1) process the simplification condition is

$$\frac{1}{1+\phi_1^2} \geq -\frac{\rho_1}{\phi_1} \geq 0 \qquad (5.19)$$

where

$$\rho_1 = \frac{\theta_1}{1+\theta_1^2} \qquad (5.20)$$

(Granger and Morris, 1976, p. 252). For the two models $\underline{M}$-U^{10} and $\underline{M}$-U^{20} of equations (5.18) the condition (5.19) gives

$$\begin{aligned} \underline{M}\text{-}U^{10} &: \quad 0.794 > 0.716 > 0 \\ \underline{M}\text{-}U^{20} &: \quad 0.994 > -0.375 \not> 0 \end{aligned} \qquad (5.21)$$

Although the first model barely satisfied the condition for simplification the second model does not, implying that an AR(1) process cannot describe the series as well as an ARMA (1,1) process. This result justified the selection of the ARMA(1,1) model for this rainfall series.

The statistical properties of the two final series (from the 10% and 20% missing values) have also been computed and are shown in Table 5.11 together with the ones of the actual series. The monthly statistics are also shown in Table C.13 (appendix C).

Table 5.11. Statistics of the Actual Series (ACT) and the Two Estimated Series (UN10, UN20).

	$\bar{y}$	s	c_v	c_s	r_1	r_2
ACT	4.126	3.673	89.04	1.332	0.366	0.134
UN10	4.105	3.609	87.920	1.354	0.384	0.157
UN20	4.043	3.492	86.381	1.373	0.410	0.160

Table 5.12 shows the bias in the mean, standard deviation and lag-one correlation coefficient so that the statistical closeness of the estimated series to the actual one can be evaluated. The bias in the mean and correlation coefficient is not significant at 5% significance level; however, the bias in the standard deviation does not pass the stringent F-test (requiring exact equality of standard deviations) and thus is significant.

Table 5.12. Bias in the Mean, Standard Deviation and Serial Correlation Coefficient-Univariate Model.

	$\bar{y}_e-\bar{y}_a$	s_e/s_a	$r_{1,e}-r_{1,a}$
UN10	-0.021	0.983	0.018
UN20	-0.083	0.951	0.044

Remarks

1. The forecasting procedure utilized for the estimation is the minimum mean square forward forecasting procedure of Box and Jenkins (1976). Damsleth (1980) introduced the method of optimal between-forecasts, combining the forward forecasts and backforecasts into between-forecasts with a minimum mean square error. He showed that the gain in forecast error by between-forecasting as compared to forward forecasting (or back-forecasting) an ARMA(1,1) model is proportional to $|\phi|^{k+1}$ where k is the size of the gap. Thus the gain rapidly becomes small, unless $|\phi|$ is very close to one and the size of the gap is very small. He also showed that the gain from between-forecasting can be substantial when θ is negative. Finally he concluded that "the reduction in forecast error variance by using this between-forecasting method is not very great for stationary series, but may be substantial when the

series is non-stationary" (Damsleth, 1980, p. 39). In our case, the use of the more complicated between-forecasting procedure does not seem to be justified. It has been shown that the simple Box-Jenkins forecasts work satisfactorily in the sense that rapid convergence to a "statistically acceptable" series occurs.

2. It is interesting to note that when the final estimates of the model (parameters of equations 5.18) are provided as initial estimates, the maximum likelihood estimates (calculated by a steepest descent algorithm) are equal to the initial estimates provided. This emphasizes the "uniqueness" of the stable model achieved by the proposed recursive algorithm.
3. It will also be interesting to check the threshold level of percent of missing values at which the algorithm starts to diverge. This is expected to happen at some level of percent of missing values (probably greater than 50%) when too much information in the system is changing at each iteration. At such high percents of missing values a more elaborate testing of the final model may also be needed.

Bivariate Model

Model Fitting

The lag-one multivariate autoregressive model of equation (4.3), suggested by Matalas (1967), preserves the lag-zero and lag-one auto- and cross-correlations. When applied to two stations the model is reduced to the bivariate Markov model:

$$\begin{bmatrix} z_{1,t} \\ z_{2,t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} z_{1,t-1} \\ z_{2,t-1} \end{bmatrix} + \begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} \eta_{1,t} \\ \eta_{2,t} \end{bmatrix} \qquad (5.22)$$

where the matrix B is a lower triangular matrix as suggested by Young (1968). The above model has been extensively used for the simultaneous generation of hydrologic series at two sites. An attempt will be made herein, to show how the above model can be used for the estimation of the missing values in one or both of the time series. A recursive algorithm analogous to the one proposed for the univariate case will be presented.

The special case that will be considered is the estimation of the missing values in the series of station 1, given the complete, concurrent, equal length series of station 2.

As has been extensively discussed in Chapter 4 incomplete data sets may result in inconsistent covariance matrices resulting in generated rainfall values that contain complex numbers. Therefore the incomplete series $\underline{S}_0$ of station 1 is first completed by the use of a simple estimation method $\underline{M}_0$ (e.g., MV, RD, NR or even replacement of missing values by zeroes) giving the complete series $\underline{S}_1$. Denote by $\underline{S}$ the complete and known series of station 2.

Then a bivariate AR(1) model is fitted to the series $\underline{S}_1$ and $\underline{S}$. Actually the model, as in the univariate case, is fitted to the residual series e.g., the normalized and standardized series. The following procedure is followed for the estimation of the parameters (matrices A and B) of the model: The lag-zero and lag-one correlation matrices, M_0 and M_1, of the residual series are computed

$$M_0 = \begin{bmatrix} 1 & r_{12}(0) \\ r_{12}(0) & 1 \end{bmatrix} , \quad M_1 = \begin{bmatrix} r_{11}(1) & r_{12}(1) \\ r_{21}(1) & r_{22}(1) \end{bmatrix} \qquad (5.23)$$

Then matrix A is given directly by the multiplication of the matrices M_1 and M_0^{-1} (equation B.8 of appendix B) and matrix C is computed from equation (B.13). Matrix B is given from the solution of equation $BB^T = C$, which in the case of B being a lower triangular matrix reduces to the direct calculation of the elements of B from equations (B.19).

Proposed Estimation Algorithm

An algorithm analogous to the one for the univariate case is also proposed for the bivariate case. After the incomplete series, $\underline{S}_0$ has been completed with a simple method $\underline{M}_0$, a bivariate AR(1) model is fitted to the complete series $\underline{S}_1$ and $\underline{S}$ as described earlier. The parameter matrices A and B of the fitted model $\underline{M}_1 = (A,B)_1$, are then used to construct new estimates for the "missing" values in the series $\underline{S}_1$. From equation (5.22) we can write that:

$$z_{1,t} = a_{11} z_{1,t-1} + a_{12} z_{2,t-1} + b_{11} \eta_{1,t} \tag{5.24}$$

$$z_{2,t} = a_{21} z_{1,t-1} + a_{22} z_{2,t-1} + b_{21} \eta_{1,t} + b_{22} \eta_{2,t} \, . \tag{5.25}$$

Since the second series is complete and known, equation (5.25) is ignored and only equation (5.24) is considered. Following the Box-Jenkins forecasting procedure, the mean square error forecasts $\hat{z}_{1,t}(\ell)$ of $z_{1,t+\ell}$, where ℓ is the lead time, are

$$\begin{aligned} \hat{z}_{1,t}(\ell) &= a_{11} z_{1,t} + a_{12} z_{2,t} \, , & \ell &= 1 \\ \hat{z}_{1,t}(\ell) &= a_{11} \hat{z}_{1,t}(\ell-1) + a_{12} \hat{z}_{2,t}(\ell-1) \, , & \ell &= 2, 3, \ldots, k \end{aligned} \tag{5.26}$$

where k is the number of values missing in each gap. The forecasting procedure is repeated for the estimation of all the gaps always using the newly estimated values in equations (5.26). These estimates then become the new estimates of the missing values, and they replace the old estimates in the series $\underline{S}_1$ giving the new series $\underline{S}_2$ and $\underline{S}$. Denote this new model by $\underline{M}_2 = (A,B)_2$, which is used in the same way as before to update the estimates. The procedure is repeated until convergence occurs in the sense that neither the model $\underline{M}_i$ nor the series $\underline{S}_i$ after the ith iteration change between iterations within a specified tolerance ($\underline{M}_i \xrightarrow{\sim} \underline{M}_{i+1}$ and $\underline{S}_i \xrightarrow{\sim} \underline{S}_{i+1}$).

Schematically the recursive algorithm for the estimation of missing values--bivariate model--1 station to be estimated (RAEMV-B1) is shown in Fig. 5.6.

The algorithm can be generalized to the case where a multivariate model of, say, K stations is used to estimate the missing values of L incomplete stations where $L \leq K$. Such a generalized algorithm can be economically written as RAEMV-MK.L. The algorithm for the case of a bivariate model with both records incomplete e.g., two series to be estimated (RAEMV-B2 or in the general form RAEMV-M2.2) is illustrated in Fig. 5.7. The notation is the same as before but two subscripts are used now for the series S, the first denoting the station (1 or 2) and the second denoting the iteration i (i=1, ...). In this case both equations (5.24) and (5.25) would be needed for the estimation of missing values existing in both series.

Application of the Algorithm on the Monthly Rainfall Series

The case study presented herein involves the estimation of the missing values of the rainfall series of station 6038 using a bivariate AR(1) model with the complete rainfall series of Station 6038. Thus the RAEMV-B1 illustrated in Fig. 5.6 has been used. Again, different levels of percentage of missing values have been tested, and the results for the 10% and 20% missing values are presented in Tables 5.13 and 5.14 respectively. The dependence of the algorithm on the starting values has been tested the same

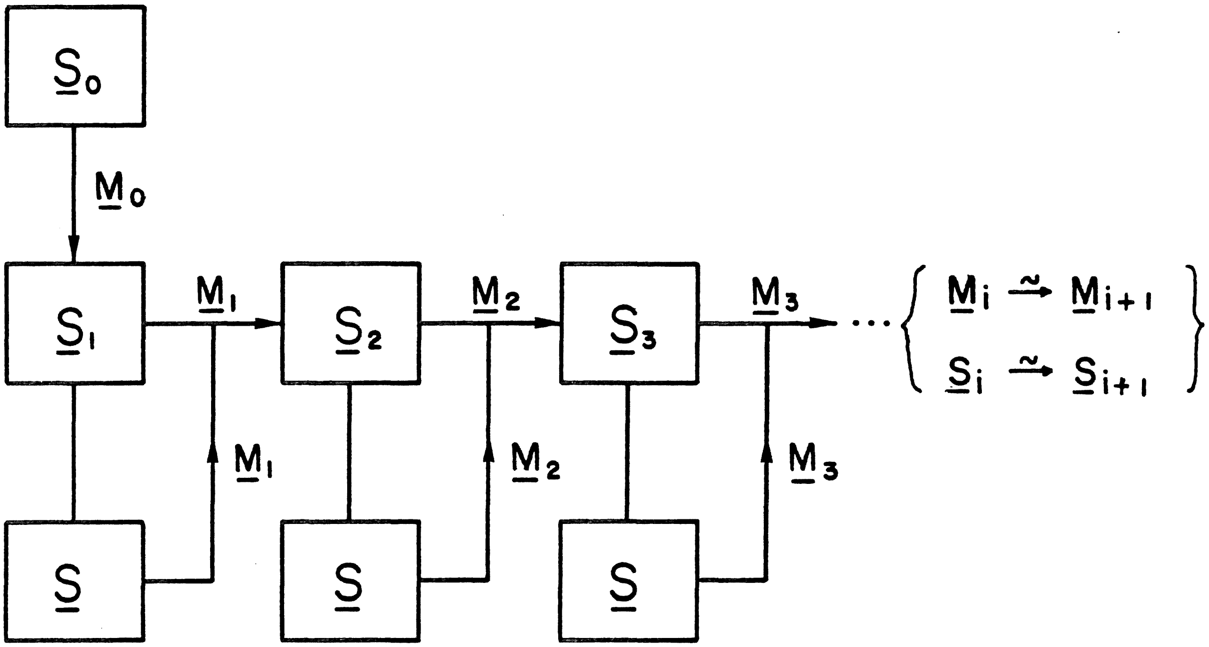


Fig. 5.6. Recursive algorithm for the estimation of missing values--bivariate model--1 station to be estimated. $\underline{S}_i$ denotes the series, and $\underline{M}_i$ the model, $\overline{(A,B)}_i$, at the ith iteration.

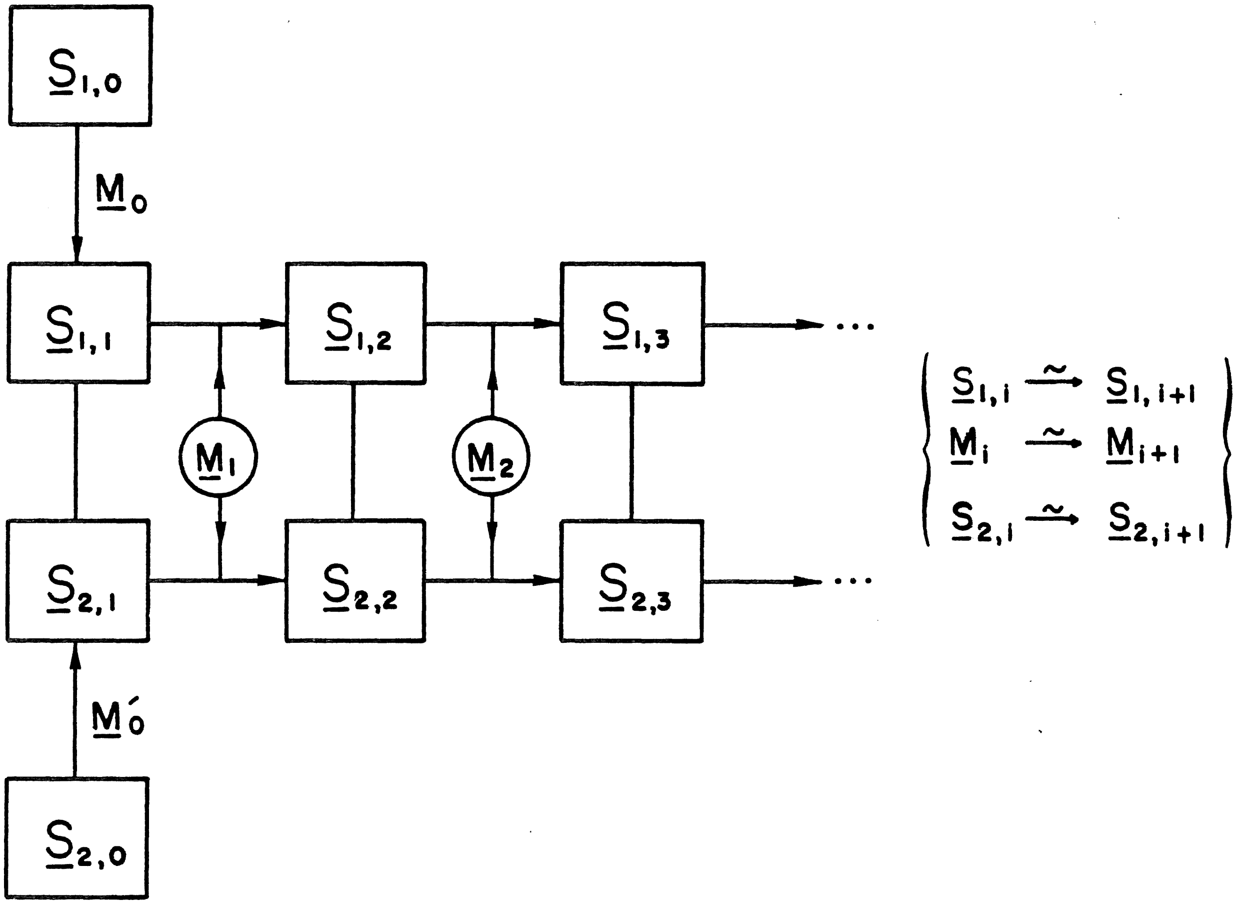


Fig. 5.7. Recursive algorithm for the estimation of missing values--bivariate model--2 stations to be estimated (RAEMV-B2). $\underline{S}_{1,i}$ $(\underline{S}_{2,i})$, denotes the series of station 1 (station 2), and $\underline{M}_i$ the model, $(A,B)_i$, at the ith iteration.

way as for the univariate case, e.g., by providing different initial series estimated by four different methods $\underline{M}_0$ (MV, RD, NR and zeroes).

Tables 5.13 and 5.14 show the cross-correlation matrices M_0 and M_1 at each iteration i and the model $\underline{M}_i = (A,B)_i$. It is interesting to follow the changes of the cross-correlation coefficients at each time step. Also notice that the autocorrelation coefficient (see equation 5.23) of the first series changes at each iteration (since new estimates of the missing values replace the old ones) but the autocorrelation coefficient of the second series remains unchanged (since the second series is complete and known).

From Tables 5.13 and 5.14 the following similar conclusions to the univariate case can be drawn:

(1) The algorithm converges rapidly, independently of the starting point (initial series). Thus, initial estimation of the missing values is not needed, and they may as well be replaced by zeroes.

(2) The convergence seems to be less sensitive to the percent of values missing, since in both the 10% and 20% levels convergence has been achieved in three to four iterations.

(3) For a specific percent of missing values the algorithm converges to the same point (e.g., same model, same series, and same correlation matrices) independently of the initial estimates of the missing values.

Table 5.13. Results of the RAEMV-B1 Applied at the 10% Level of Missing Values.

					$\underline{\underline{M}}_i$			
i	M_0		M_1		A		B	
				$\underline{\underline{M}}_0 = \mathrm{MV}$				
1	1.	0.330	0.004	0.137	-0.046	0.152	0.990	0.
	0.330	1.	0.042	0.315	-0.070	0.338	0.286	0.902
2	1.	-0.005	0.038	0.194	0.039	0.194	0.980	0.
	-0.005	1.	0.065	0.315	0.067	0.316	-0.071	0.944
3	1.	0.025	0.049	0.202	0.044	0.201	0.978	0.
	0.025	1.	0.069	0.315	0.061	0.314	-0.043	0.946
4	1.	0.025	0.049	0.201	0.044	0.200	0.979	0.
	0.025	1.	0.068	0.315	0.061	0.314	-0.042	0.946
				$\underline{\underline{M}}_0 = \mathrm{RD}$				
1	1.	0.554	0.124	0.249	-0.021	0.261	0.968	0.
	0.554	1.	0.201	0.315	0.038	0.294	0.492	0.811
2	1.	0.026	0.042	0.196	0.037	0.195	0.980	0.
	0.026	1.	0.070	0.315	0.062	0.314	-0.039	0.946
3	1.	0.025	0.048	0.201	0.043	0.200	0.979	0.
	0.025	1.	0.069	0.315	0.061	0.314	-0.042	0.946
4	1.	0.025	0.049	0.201	0.044	0.200	0.979	0.
	0.025	1.	0.068	0.315	0.061	0.314	-0.042	0.946

Continued

Table 5.13. Continued.

$\underline{\mathbf{M}}_0$ = NR								
1	1.	0.543	0.126	0.261	-0.022	0.273	0.965	0.
	0.543	1.	0.187	0.315	0.022	0.303	0.478	0.819
2	1.	-0.002	0.046	0.199	0.046	0.199	0.979	0.
	-0.002	1.	0.069	0.315	0.070	0.316	-0.070	0.944
3	1.	0.026	0.050	0.203	0.045	0.201	0.978	0.
	0.026	1.	0.069	0.315	0.061	0.314	-0.042	0.946
4	1.	0.025	0.049	0.202	0.044	0.200	0.978	0.
	0.025	1.	0.068	0.315	0.061	0.314	-0.042	0.946
$\underline{\mathbf{M}}_0$ = zeroes								
1	1.	0.258	0.463	0.172	0.448	0.057	0.885	0.
	0.258	1.	0.048	0.315	-0.036	10.385	0.247	0.915
2	1.	0.042	0.061	0.225	0.059	0.222	0.973	0.
	0.042	1.	0.081	0.315	0.068	0.313	-0.033	0.946
3	1.	0.029	0.048	0.203	0.042	0.201	0.978	0.
	0.029	1.	0.070	0.315	0.061	0.314	-0.038	0.946
4	1.	0.025	0.049	0.201	0.043	0.200	0.979	0.
	0.025	1.	0.068	0.315	0.061	0.314	-0.042	0.946

Table 5.14. Results of the RAEMV-B Applied at the 20% Level of Missing Values.

					$\underline{\underline{M}}_i$			
i	M_0		M_1		A		B	
				$\underline{\underline{M}}_0$ = MV				
1	1.	0.523	0.342	0.251	0.290	0.100	0.936	0.
	0.523	1.	0.257	0.315	0.126	0.249	0.446	0.831
2	1.	-0.025	0.369	0.307	0.377	0.316	0.874	0.
	-0.025	1.	0.256	0.315	0.264	0.322	-0.253	0.876
3	1.	-0.023	0.389	0.333	0.393	0.337	0.857	0.
	-0.012	1.	0.253	0.315	0.257	0.319	-0.255	0.877
4	1.	-0.012	0.389	0.332	0.393	0.337	0.858	0.
	-0.012	1.	0.253	0.315	0.257	0.319	-0.254	0.877
				$\underline{\underline{M}}_0$ = RD				
1	1.	0.588	0.320	0.290	0.228	0.156	0.939	0.
	0.588	1.	0.262	0.315	0.117	0.246	0.510	0.795
2	1.	-0.012	0.368	0.315	0.375	0.383	0.872	0.
	-0.023	1.	0.257	0.315	0.264	0.321	-0.254	0.875
3	1.	-0.012	0.388	0.334	0.392	0.338	0.857	0.
	-0.012	1.	0.253	0.315	0.257	0.319	-0.255	0.877
4	1.	-0.012	0.388	0.333	0.393	0.337	0.858	0.
	-0.012	1.	0.253	0.315	0.257	0.318	-0.254	0.877

Continued

Table 5.14. Continued.

				$\underline{\underline{M}}_0$ = NR				
1	1.	0.611	0.324	0.273	0.252	0.119	0.941	0.
	0.611	1.	0.279	0.315	0.137	0.232	0.534	0.777
2	1.	-0.022	0.372	0.311	0.379	0.320	0.872	0.
	-0.022	1.	0.258	0.315	0.265	0.321	-0.253	0.875
3	1.	-0.012	0.389	0.333	0.393	0.338	0.857	0.
	-0.012	1.	0.253	0.315	0.257	0.319	-0.255	0.877
4	1.	-0.012	0.389	0.332	0.393	0.337	0.857	0.
	-0.012	1.	0.253	0.315	0.257	0.318	-0.254	0.877
				$\underline{\underline{M}}_0$ = zeroes				
1	1.	0.321	0.601	0.201	0.599	0.009	0.799	0.
	0.321	1.	0.195	0.315	0.104	0.282	0.253	0.909
2	1.	0.006	0.423	0.340	0.421	0.337	0.841	0.
	0.006	1.	0.228	0.315	0.226	0.314	-0.233	0.892
3	1.	-0.012	0.392	0.332	0.397	0.337	0.856	0.
	-0.013	1.	0.249	0.315	0.253	0.319	-0.255	0.878
4	1.	-0.013	0.390	0.333	0.394	0.338	0.857	0.
	-0.013	1.	0.253	0.315	0.257	0.319	-0.255	0.877

(4) For a different percent of missing values the same series converges to a "different" point, but this is reasonable and expected since the constant information (existing values in the series) is different in each case, and a different model thus describes it better.

The statistical properties of the two final series (from the 10% and 20% missing values) are shown in Table 5.15 together with the ones of the actual series. The monthly statistics are also shown in Table C.14 (appendix C). Table 5.16 shows the statistical closeness of the two estimated series to the actual one. Again, the bias in the mean and correlation coefficient is not significant at the 5% significance level, but the bias in the standard deviation is.

Table 5.15. Statistics of the Actual Series (ACT) and the Two Estimated Series (B10 and B20).

	$\bar{y}$	s	c_r	c_s	r_1	r_2
ACT	4.126	3.673	89.04	1.332	0.366	0.134
B10	4.096	3.610	88.132	1.358	0.382	0.162
B20	4.077	3.523	86.421	1.341	0.416	0.165

Table 5.16. Bias in the Mean, Standard Deviation and Serial Correlation Coefficient-Bivariate Model.

	$\bar{y}_e-\bar{y}_a$	s_e/s_a	$r_{1,e}-r_{1,a}$
B10	-0.030	0.983	0.016
B20	-0.049	0.959	0.050

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

Summary and Conclusions

The objective of this study was to compare and evaluate different methods for the estimation of missing observations in monthly rainfall series. The estimation methods studied reflect three basic ideas:

(1) the use of regional information in four simple techniques:

- mean value method (MV),
- reciprocal distance method (RD),
- normal ratio method (NR),
- modified weighted average method (MWA);

(2) the use of a univariate stochastic (ARMA) model that describes the time correlation of the series;

(3) the use of a multivariate stochastic (ARMA) model that describes the time and space correlation of the series.

An algorithm for the recursive estimation of the missing values in a time series using the fitted univariate or multivariate ARMA model has been proposed and demonstrated. Apparently, the idea of the recursive estimation of missing values is known (Orchard and Woodbury, 1972; Beale and

Little, 1974), as well as the idea of using the fitted model to directly derive the estimates (Brubacher and Wilson, 1976; Damsleth, 1979). However it appears that a method which combines the above two ideas simultaneously in a recursive estimation of the missing values with parallel updating of the model has not been used before.

The proposed algorithm is general and can be used for the estimation of the missing values in any series that can be described by an ARMA model.

On the basis of the data from the four south Florida rainfall stations used in the analysis, the following conclusions can be drawn:

(1) All the simplified estimation techniques give unbiased (overall and monthly) means and correlation coefficients at the 5% significance level even for as high as 20% missing values.

(2) At high percentages of missing values (greater than 10%) the MV method gives the more biased (although not significantly so) correlation coefficients.

(3) All methods give a slightly biased overall variance but unbiased monthly variance at the 5% significance level, and the MV method gives the most biased variances for all percentages of missing values.

(4) The NR method gives the most and the MV the least accurate estimates, at almost all levels of percent missing values.

(5) The proposed recursive algorithm works satisfactorily in both the univariate and bivariate case. It converges rapidly and independently of the initial estimates and gives unbiased means and correlation coefficients at the 5% significance level.

(6) The use of a bivariate model as compared to a univariate one did not improve the estimates except for a slight improvement at 20% missing values. However, the use of a multivariate model based on three or four nearby stations is expected to give much better estimates. The use of three adjacent stations is the main reason for the better performance of the NR method over the more sophisticated univariate and bivariate ARMA models which use only zero and one additional stations.

If the purpose of estimation is to calculate the historical statistics of the series (e.g., mean, standard deviation, and autocorrelations) the selection of the method matters little, and the simplest one may be chosen. However, if it is desired to fit an ARMA model to the incomplete series, to be used, say, to construct forecasts, the estimation of the missing values and the parameters of the model by the proposed recursive algorithm is recommended. In this case the equilibrium state (i.e., final series and parameters of the model) achieved upon convergence is unique, depending only on the existing information in the system (available data) and not on any external information added to the system (by the replacement of the missing values with some derived estimates). The only assumption made is that the order of the ARMA model to be fitted to the series is known. In practical situations this is seldom a problem since the latter can be determined from the complete part of the series or from a series with similar characteristics. For example, if an ARMA(1,1) model is known to fit the monthly rainfall series well at a couple of nearby stations, there is little doubt that it will fit the incomplete monthly rainfall series equally well at the station of interest. Upon convergence, the recursive algorithm then gives the "best" estimates of the parameters of the model.

Further Research

Further research should include:

(1) application of the simple estimation techniques in short records where the biases may be significant for the methods with the poorer performance;

(2) test of the sensitivity of the recursive algorithm to the selection of the model (order of the model) when more than one model fits the data equally well;

(3) derivation of the threshold percent of missing values after which the algorithm diverges;

(4) application to the estimation of missing values in other hydrological series, e.g., runoff;

(5) trials of different forecasting procedures and determination of improvements obtained by the "between-forecasting procedure" in cases of a large number of single-value gaps, e.g., use of the average of a backwards and forwards ARMA model forecast;

(6) application of the concept of "missing values" for the estimation of erroneous values or outliers in a series to avoid errors when using the data, say, to construct forecasts; and

(7) estimation of values in a series that are affected by unusual circumstances, thereby permitting a measure of the magnitude of the unusual circumstance and the estimation of the effect of similar circumstances in the future (e.g., effect of a drought on water supply).

APPENDIX A

DEFINITIONS

1. Strict stationarity

A stochastic process is said to be strictly stationary if its statistics (e.g., mean, variance, serial correlation) are not affected by a shift in the time origin, that is, if the joint probability distribution associated with n observations $(z_1, z_2, \ldots, z_n)_t$ made at time origin t, is the same as that associated with n observations $(z_1, z_2, \ldots, z_n)_{t+k}$ made at time origin t+k. In other words, z(t) is a strictly stationary process when the two processes z(t) and z(t+k) have the same statistics for any k.

2. Weak stationarity

Weak stationarity of order f is when the moments of the process up to an order f depend only on time differences. Usually by weak stationarity we refer to second order stationarity, e.g., fixed mean and an autocovariance matrix that depends only on time differences (i.e., lags).

3. Gaussian process

If the probability distribution associated with any set of times is a multivariate normal distribution, the process is called a normal or Gaussian process. Since the multivariate normal distribution is fully described by its first and second order moments it follows that weak stationarity and an assumption of normality imply strict stationarity.

4. Non-stationarity

A stochastic process is said to be nonstationary if its statistical characteristics change with time. A homogeneous nonstationary process of order d is a process, for which the dth difference $\nabla^d z_t$ is a stationary process. For example a first order homogeneous nonstationary process is one that exhibits homogeneity apart from constant (e.g., a linear trend), and a second order nonstationary is the one that exhibits homogeneity apart from constant and slope (e.g., a parabolic trend).

5. Circular stationarity

A stochastic process is said to be circularly stationary with period τ, if the multivariate probability distribution of τ observations $(z_1, z_2, \ldots, z_\tau)_t$ made at time origin t, is the same as that associated with τ observations $(z_1, z_2, \ldots, z_\tau)_{t+\tau k}$ made at time origin $t + \tau k$, for $k = 1, 2, \ldots$. For example, a monthly hydrologic series has a period of 12 months, i.e., $\tau = 12$ and circular stationarity suggests that the probability distribution of a value of a particular month is the same for all the years.

6. Stationarity condition

A linear process can be always written in the random shock form:

$$\tilde{z}_t = \psi(B)\ a_t \tag{A.1}$$

where B is the backward shift operator defined by $Bz_t = z_{t-1}$; hence $B^m z_t = z_{t-m}$ and

$$\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots \tag{A.2}$$

is the so called transfer function of the linear system and is the generating function of the ψ weights. For the process to be stationary the ψ weights must satisfy the condition that $\psi(B)$ converges on or within the unit circle, e.g., for all $|B| \leq 1$.

7. Invertibility condition

The above model may also be written in the inverted form

$$\psi^{-1}(B)\ \tilde{z}_t = a_t \tag{A.3}$$

or

$$\pi(B)\ \tilde{z}_t = a_t \tag{A.4}$$

where

$$\pi(B) = 1 - \pi_1 B - \pi_2 B^2 - . . . \qquad (A.5)$$

is the generating function of the π weights. For the process to be invertible the π weights must satisfy the condition that $\pi(B)$ converges for all $|B| \leq 1$, that is on or within the unit circle. The invertibility condition is independent of the stationarity condition and is applicable also to the nonstationary linear models. The requirement of invertibility is needed in order to associate the present values of the process to the past values in a reasonable manner, as will be shown below.

8. Duality between AR and MA processes

In a stationary AR(p) process, a_t can be represented as a finite weighted sum of previous $\tilde{z}$'s,

$$\phi(B)\ \tilde{z}_t = a_t \qquad (A.6)$$

or $\tilde{z}_t$ as an infinite weighted sum of previous a's

$$\tilde{z}_t = \phi^{-1}(B)\ a_t \qquad (A.7)$$

Also, in an invertible MA(q) process, $\tilde{z}_t$ can be represented as a finite weighted sum of previous a's,

$$\tilde{z}_t = \theta(B)\ a_t \qquad (A.8)$$

or a_t as an infinite weighted sum of previous $\tilde{z}$'s

$$\theta^{-1}(B)\ \tilde{z}_t = a_t \qquad \text{(A.9)}$$

In other words, a finite AR process is equivalent to an infinite MA process, and a finite MA process to an infinite AR process. This principle of duality has further aspects, e.g., there is an inverse relationship between the autocorrelation and partial autocorrelation functions of AR and MA processes.

9. Physical interpretation of stationarity and invertibility

Consider an AR(1) process $(1 - \phi_1 B)\ z_t = a_t$. For this process to be stationary, the root of the polynomial $1 - \phi_1 B = 0$ must lie outside the unit circle, which implies that $B = \phi_1^{-1}$ must be greater than one, or $|\phi_1| < 1$. The process can be also written

$$\begin{aligned} z_t &= \phi_1 z_{t-1} + a_t \\ z_{t+1} &= \phi_1^2 z_{t-1} + \phi_1 a_t + a_{t+1} \\ z_{t+2} &= \phi_1^3 z_{t-1} + \phi_1^2 a_t + \phi_1 a_{t+1} + a_{t+2} \quad \text{etc.} \end{aligned} \qquad \text{(A.10)}$$

When $|\phi_1| > 1$ (or $|\phi_1| = 1$) the effect of the past on the present value of the time series increases (or stays the same) as the series moves into the future. Only when $|\phi_1| < 1$ (stationary process) does the effect of the past on the present decrease the further we move into the past, which is a reasonable and acceptable hydrologic fact (Delleur and Kavvas, 1978).

Consider now an MA(1) process $z_t = (1-\theta_1 B)a_t$. The invertibility condition implies that $|\theta_1| < 1$. The process can also be written in the form:

$$a_t = \frac{1}{1-\theta_1 B} z_t \tag{A.11}$$

where the polynomial $(1-\theta_1 B)^{-1}$ can be expanded in an infinite sum of convergent series only if $|\theta_1| < 1$. To illustrate the need for invertibility let us assume that $|\theta_1| > 1$. Then (A.11) can be written as

$$a_t = -\frac{1}{\theta_1 B} \frac{1}{(1 - \frac{1}{\theta_1 B})} z_t \tag{A.12}$$

and since $|\frac{1}{\theta_1 B}| < 1$, it can be expanded to the form

$$a_t = -(\frac{1}{\theta_1 B} + \frac{1}{\theta_1^2 B^2} + \frac{1}{\theta_1^3 B^3} + \ldots) z_t \tag{A.13}$$

or

$$a_t = - \frac{1}{\theta_1} z_{t+1} - \frac{1}{\theta_1^{\,2}} z_{t+2} - \frac{1}{\theta_1^{\,3}} z_{t+3} - \cdot \cdot \cdot \qquad (A.14)$$

which implies that future values are used to generate the present values. It becomes clear that the invertibility condition is required in order to assure hydrologic realizability.

10. The portemanteau lack of fit test

The portemanteau lack of fit test (Box and Jenkins, 1976, Ch. 8) considers the first K autocorrelations $r_k(\hat{a})$, $k = 1, 2, \ldots, K$, of the fitted residual series $\hat{a}$ of an ARIMA(p,d,q) process, to detect inadequacy of the model. It can be shown (Box and Pierce, 1970) that, if the fitted model is appropriate,

$$Q = (N-d) \sum_{k=1}^{K} r_k^2 \, (\hat{a}) \qquad (A.15)$$

is approximately distributed as $\chi^2(K-p-q)$ where K-p-q is the number of degrees of freedom, N is the total length of the series, and (N-d) is the number of observations used to fit the model. The adequacy of the model may be checked by comparing Q with the theoretical chi-square value $\chi^2(K-p-q)$ of a given significance level. If $Q < \chi^2(K-p-q)$, a_t is an independent series and so the model is adequate, otherwise the model is inadequate.

For the choice of K, Box and Jenkins suggest it to be "sufficiently large so that the weights ψ_j in the model, written in the form

$$\tilde{w}_t = \phi^{-1}(B)\ \theta(B)\ a_t = \psi(B)\ a_t \qquad (A.16)$$

will be negligibly small after j = K" (Box and Jenkins, 1976, p. 221). The IMSL subroutine FTCMP (IMSL - 0007, Ch. F) uses a value of K equal to N/10 + p + q to perform the portemanteau test.

Ozaki (1977) points out that "for the application of the portemanteau test, fast dying off of the impulse response function (weights ψ_j) of the model is a necessary condition" (Ozaki, 1977, p. 298). In cases where the impulse response function dies off rather slowly (possibly due to the near-nonstationarity of the model) when compared with the length of the series, the applicability of the portemanteau test is doubtful since the autocorrelations of the residuals may not be reliable at large lags.

11. Cumulative periodogram test

Another method used in the diagnostic checking stage of the Box-Jenkins procedure is the cumulative periodogram checking of the residuals. The normalized (area under the curve equal to one) cumulative periodogram for frequencies, f, between 0 and 0.5, of the fitted residuals $\hat{a}_t$, is compared with the theoretical cumulative periodogram of a white noise series which is a straight line joining the points (0, 0) and (0.5, 1). A periodicity in the residuals at frequency f_i is expected to show up as a deviation from the straight line at this frequency. Kolmogorov-Smirnov probability limits can be drawn on the cumulative periodogram plot to test the significance of such deviations. For a given level of significance α, the limit lines are drawn at distances $\pm K_\alpha / N'$ above and below the theoretical straight line, where $N' = (N-2)/2$ for N even and $N' = (N-1)/2$ for N odd. Approximate values of K_α for different levels of significance α, are:

α	0.01	0.05	0.10	0.20	0.25
K_α	1.63	1.36	1.22	1.07	1.02

(Box and Jenkins, 1976, p. 297). So, if more than αN of the plotted points fall outside the probability lines, the residual series may still have some periodicity; otherwise it may be concluded that the residuals are independent.

In practice, "because the a's are fitted values and not the true a's, we know that even when the model is correct they will not precisely follow a white noise process" and thus the cumulative periodogram test provides only a "rough guide" to the model inadequacy checking (Box and Jenkins, 1976, p. 297).

12. Akaike Information Criterion (AIC)

The AIC for an ARMA(p,q) model is given by

$$AIC(p,q) = N \log(\hat{\sigma}_a^2) + 2(p+q+2) + N\log 2\pi + N$$

where $\hat{\sigma}_a^2$ is the MLE of the residual variance given by

$$\hat{\sigma}_a^2 = \frac{1}{N-p-q} S(\underline{\phi},\underline{\theta}) \qquad (A.17)$$

and $\underline{\phi}$, $\underline{\theta}$ are the vectors of the parameters ϕ, θ which minimize the sum of squares of the residuals $\hat{a}_t$

$$S(\underline{\phi},\underline{\theta}) = \sum_{t=1}^{N} (\hat{a}_t)^2 \quad . \qquad (A.18)$$

For the purpose of comparison of models the definition of AIC can be replaced by

$$AIC(p,q) = N \log(\hat{\sigma}_a^2) + 2(p+q) \quad . \qquad (A.19)$$

Ozaki (1977) demonstrates that the inherent difficulties associated with the Box-Jenkins procedure (identification, estimation and diagnostic checking) for the selection of the model, when several models fit the data equally well, can be overcome by using the MAICE (minimum AIC estimation) procedure as the only objective criterion for the selection of the "best" approximating model among a set of possible models. He also points out that the AIC "measures both the fit of a model and the unreliability of a model" (Ozaki, 1977, p. 290).

13. Positive definite (semidefinite) matrix

A real symmetric matrix A is called positive definite (semidefinite) if and only if

$$X^T A X > 0 \qquad (\geq 0) \qquad \text{(A.20)}$$

for all vectors $X \neq 0$. The two following theorems hold:

Theorem 1: A matrix A is positive (semi-) definite if and only if all its characteristic values (i.e., eigenvalues) are (non-negative) positive.

Theorem 2: A matrix A is positive (semi-) definite if and only if all the successive principal minors of A are (non-negative) positive.

An obvious corollary of the above is that a positive semidefinite matrix is positive definite if and only if it is nonsingular i.e., none of its characteristic values are zero (Gantmacher, 1977, p. 305).

14. Test for differences in the means of two normal populations

Let μ_1, μ_2 denote the population means of two normal distributions and $\bar{x}_1$, $\bar{x}_2$ the sample means respectively. Let also assume that the variance of the two normal distributions are equal but unknown. The hypothesis H_o: $\mu_1 = \mu_2$ versus H_a: $\mu_1 \neq \mu_2$ is tested by calculating the statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{1/N_1 + 1/N_2}} \tag{A.21}$$

where

$$s^2 = \frac{(N_1-1)\ s_1^2 + (N_2-1)\ s_2^2}{N_1 + N_2 - 2} \tag{A.22}$$

which has a t distribution with $N_1 + N_2 - 2$ degrees of freedom. The H_o is rejected if

$$|t| > t_{1-\alpha/2\ ;\ N_1+ N_2- 2}\ . \tag{A.23}$$

Although the test is based on sample normality, for large samples, the Central Limit Theorem enables us to use the test as approximate test for nonnormal samples. If the two populations are of equal length, $N_1 = N_2 = N$, then equation (A.21) reduces to

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 + s_2^2)/N}} \qquad \text{(A.24)}$$

15. Test for equality of variances of two normal distributions

Let σ_1^2, σ_2^2 denote the population variances and s_1^2, s_2^2 the sample variances of two normal distributions. The hypothesis H_o: $\sigma_1^2 = \sigma_2^2$ versus H_a: $\sigma_1^2 \neq \sigma_2^2$ is tested by calculating the statistic

$$F_c = s_1^2/s_2^2 \qquad \text{(A.25)}$$

where s_1^2 is the larger sample variance. F_c is distributed as an F distribution with $N_1 - 1$ and $N_2 - 1$ degrees of freedom where N_1 is the length of the sample having the larger variance and N_2 is the length the sample with the smaller variance. H_o is rejected if

$$F_c > F_{N_2-1}^{N_1-1} \ ; \ 1 - \alpha \qquad \text{(A.25)}$$

16. Test for equality of correlation coefficients

Let ρ denote the population correlation coefficient and r the sample estimate of ρ. If the sample size is moderately large ($N \geq 25$) then the quantity W is approximately normally distributed with mean and variance 1/N-3 where

$$W = \frac{1}{2} \ln\left(\frac{1 + r}{1 - r}\right) \tag{A.27}$$

and

$$\omega = \frac{1}{2} \ln\left(\frac{1 + \rho}{1 - \rho}\right) \quad . \tag{A.28}$$

To test the hypothesis H_o: $\rho = r$ against the alternative H_a: $\rho \neq r$ the quantity

$$z = (W - \omega) \sqrt{N-3} \tag{A.29}$$

can be considered to be normally distributed with zero mean and unit variance. If $|z| > z_{1-\alpha/2}$ (z is the standard normal variable), H_o is rejected (see Haan, 1977, p. 223).

APPENDIX B

DETERMINATION OF MATRICES A AND B OF THE MULTIVARIATE AR(1) MODEL

Determination of matrix A

The multivariate lag-one autoregressive model is written as

$$Z_t = A\ Z_{t-1} + B\ N_t \tag{B.1}$$

Post-multiplying both sides of equation (B.1) by Z_{t-1}^T and taking expectations it becomes:

$$E[Z_t Z_{t-1}^T] = A\ E[Z_{t-1} Z_{t-1}^T] + B\ E[N_t Z_{t-1}^T] \tag{B.2}$$

By definition

$$M_0 = E[Z_t Z_t^T] \tag{B.3}$$

and

$$M_1 = E[Z_t Z_{t-1}^T] \tag{B.4}$$

and from the assumption of weak stationarity

$$M_0 = E[Z_{t-1} Z_{t-1}^T] \quad . \tag{B.5}$$

Also from the independent uncorrelated process N_t

$$E[N_t Z_{t-1}^T] = 0 \tag{B.6}$$

so that equation (B.2) becomes

$$M_1 = A \, M_0 \tag{B.7}$$

and solving for the parameter matrix A

$$A = M_1 \, M_0^{-1} \tag{B.8}$$

<u>Determination of matrix B</u>

Post-multiplying equation (B.1) by Z_t^T and taking expectations in both sides it becomes

$$\begin{aligned} E[Z_t Z_t^T] &= A \, E[Z_{t-1} Z_t^T] + B \, E[N_t Z_t^T] \\ &= A \, E[Z_t Z_{t-1}^T]^T + B \, E[N_t (AZ_{t-1} + BN_t)^T] \qquad (B.9) \\ &= A \, E[Z_t Z_{t-1}^T]^T + B \, E[N_t Z_{t-1}^T] \, A^T + B \, E[N_t N_t^T] \, B^T \end{aligned}$$

Because $E[N_t N_t^T] = I$, an identity matrix, and $E[N_t Z_{t-1}^T] = 0$ equation (B.9) can be written

$$M_0 = A\ M_1^T + B\ B^T \tag{B.10}$$

By substituting A from equation (B.8) and solving for $B\ B^T$

$$B\ B^T = M_0 - M_1\ M_0^{-1}\ M_1^T \tag{B.11}$$

Solution of equation $B\ B^T = C$

The right hand side of equation (B.11) involves the lag-zero and lag-one correlation matrices which can be estimated from the historical data and thus is a known quantity C. The problem that remains now, is to solve equation

$$B\ B^T = C \tag{B.12}$$

for B. A necessary and sufficient condition to have a real solution for B is that C must be a positive semidefinite matrix.

It can be proven (Valencia and Schaake, 1973) that if the correlation matrices M_0 and M_1 have been calculated using equal length records for all m sites, then the matrix

$$C = M_0 - M_1\ M_0^{-1}\ M_1^T \tag{B.13}$$

is always positive semidefinite and so a real solution for the matrix B exists. But this solution for B is not unique. An infinite number of matrices B exist that satisfy (B.12).

Proof: Let B denote a matrix solution of equation (B.12) and K denote an (mxm) matrix such that $K\ K^T = I$ where I is an (mxm) identity matrix. A matrix B_0 defined as

$$B_0 = B\ K \tag{B.14}$$

may be used in place of B in equation (B.12) since

$$B_0\ B_0^T = B\ K(B\ K)^T = B\ K\ K^T\ B^T = B\ B^T \tag{B.15}$$

There exists more than one matrix K such that $K\ K^T = I$, and therefore many solutions for matrix B exist, all valid since the elements of B have no physical significance as far as synthetic hydrology is concerned (Matalas, 1967).

Several techniques have been proposed for the solution of equation (B.12). Fiering (1964) and Matalas (1967) suggested the use of principal component analysis and Moran (1970) used canonical correlation analysis. Young (1968) assumed that B is a lower triangular matrix, based on the fact that $C = B\ B^T$ is a symmetric matrix, and gave a unique recursive solution for the elements of B. Let us examine this case closely:

(1) $C = B\,B^T$ is symmetric for any B. The (i,j)th element of matrix C is

$$c_{ij} = \sum_{k=1}^{m} b_{ik}b'_{kj} \tag{B.16}$$

and the across the diagonal element is

$$c_{ji} = \sum_{k=1}^{m} b_{jk}b'_{ki} \tag{B.17}$$

where the prime denotes a transposed element. Thus, $b'_{kj} = b_{jk}$ and $b'_{ki} = b_{ik}$, which implies that $c_{ij} = c_{ji}$ and therefore C is symmetric for any B.

(2) That C is symmetric implies that m(m+1)/2 equations are required to specify it, and so m(m+1)/2 non zero elements of matric B are needed. Thus, since the (mxm) matrix B has m^2 elements there are m(m-1)/2 elements that can be set to zero. So the assumption of a lower triangular matrix B is valid.

(3) The assumption of a lower triangular matrix B allows a recursive solution for the coefficients of B. This will be illustrated in the (2x2) case, and the reader is referenced to Young and Pisano (1968) for the general case.

$$\begin{bmatrix} b_{11} & 0 \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} b_{11} & b_{21} \\ 0 & b_{22} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

or (B.18)

$$\begin{bmatrix} b_{11}^2 & b_{11}\, b_{21} \\ b_{21}\, b_{11} & (b_{21}^2 - b_{22}^2) \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

from which

$$\begin{aligned} b_{11} &= \quad c_{11} \\ b_{21} &= c_{21}/b_{11} \\ b_{22} &= \sqrt{c_{11} - b_{21}^2} \ , \end{aligned} \tag{B.19}$$

with the constraints

$$c_{11} \geq 0 \quad \text{and} \quad c_{22} \geq c_{12}^2/c_{11} \ . \tag{B.20}$$

APPENDIX C
DATA USED AND STATISTICS

Table C.1. 55 years of monthly rainfall data for the South Florida Station 6038.

***** STATION 6038, MOORE HAVEN LOCK 1 *****

1927	0.11	2.09	1.70	2.02	1.94	10.79	5.79	8.61	6.99	4.12	0.38	0.39
1928	0.42	2.31	2.46	1.52	4.19	8.12	5.43	11.82	14.60	0.47	0.97	0.31
1929	0.82	0.14	0.52	1.55	2.73	9.35	8.44	4.93	13.45	1.71	1.27	1.39
1930	0.49	3.23	4.76	4.12	11.33	17.85	4.72	11.61	11.26	6.33	0.45	2.33
1931	2.58	0.76	5.90	3.44	1.59	1.20	2.68	10.34	5.06	1.94	0.08	0.35
1932	1.97	3.13	2.87	1.76	6.05	4.96	6.25	15.71	5.99	2.93	3.28	0.07
1933	1.65	0.19	3.88	6.92	3.89	4.66	5.36	5.77	2.75	5.18	0.92	0.28
1934	1.33	2.89	2.73	2.22	6.43	4.36	8.48	6.20	4.18	5.54	3.58	0.26
1935	0.52	1.00	0.03	5.18	3.57	5.84	5.09	5.50	9.53	1.42	1.71	1.48
1936	2.23	4.97	1.95	2.55	5.41	14.59	2.99	5.79	11.51	3.55	0.58	1.18
1937	2.07	1.70	4.83	4.89	4.94	4.29	13.79	4.71	4.48	8.72	5.47	0.44
1938	0.61	0.57	0.34	0.21	6.28	7.40	8.20	2.39	2.23	3.92	1.52	0.11
1939	0.18	0.35	0.79	3.08	4.48	3.61	16.13	10.42	4.20	3.60	1.45	1.45
1940	2.37	3.07	5.55	2.06	3.36	4.96	7.92	10.43	14.13	0.32	0.42	3.91
1941	5.73	3.86	3.68	5.62	3.30	4.87	13.23	6.71	8.54	2.92	1.66	1.52
1942	2.80	3.51	4.55	5.64	1.99	9.51	4.81	5.66	4.16	0.03	0.46	1.62
1943	0.35	0.37	2.72	3.91	3.43	5.02	8.04	8.07	3.07	2.67	1.69	0.20
1944	0.98	0.12	2.35	5.41	1.52	5.50	8.36	5.42	9.23	3.47	0.07	0.27
1945	1.82	0.27	0.17	3.20	2.22	7.07	9.47	6.86	8.38	4.92	0.53	0.57
1946	0.68	0.76	2.53	0.27	7.52	5.74	6.90	4.49	7.77	1.16	2.16	0.90
1947	0.70	1.64	8.73	0.55	4.80	15.02	6.43	10.74	10.57	6.18	4.33	1.51
1948	4.16	0.38	0.62	3.15	2.24	4.67	6.00	3.94	21.55	2.42	0.57	0.57
1949	0.05	0.03	0.46	1.64	3.13	6.56	9.40	12.51	10.22	0.73	0.96	2.74
1950	0.06	0.72	1.40	2.88	3.29	4.55	7.53	8.86	2.77	5.54	1.57	1.45
1951	0.15	1.99	0.82	3.31	4.47	5.02	11.63	5.03	6.20	7.74	1.36	0.11
1952	0.92	5.02	1.50	2.25	10.74	7.56	7.05	8.09	6.35	11.11	0.19	0.46
1953	1.45	2.57	0.76	4.03	2.78	6.52	9.13	5.65	14.16	9.67	0.55	1.25
1954	0.38	1.72	2.24	3.52	11.96	12.53	10.58	5.96	6.48	2.63	1.19	1.89
1955	2.78	1.27	1.26	1.72	3.91	13.17	5.80	3.59	7.07	2.55	0.28	1.18
1956	0.86	1.04	0.40	1.58	1.13	5.43	3.53	4.67	5.18	6.47	0.13	0.52
1957	1.74	3.73	6.09	4.06	5.58	4.35	6.59	7.59	9.50	1.20	0.24	7.58
1958	6.04	0.84	7.03	5.84	4.91	5.93	8.32	4.12	3.09	4.59	0.47	5.77
1959	1.09	1.08	5.82	1.99	6.07	10.16	5.60	6.12	12.00	12.36	1.29	1.02
1960	0.31	4.43	1.37	6.55	2.77	11.35	11.11	6.37	11.30	5.99	1.21	0.69
1961	2.71	2.16	3.56	2.44	6.12	7.17	3.74	4.73	2.64	0.66	1.41	0.33
1962	0.88	0.47	3.57	2.60	2.33	11.46	5.46	7.71	8.78	1.20	4.03	0.22
1963	0.86	3.64	0.49	0.80	8.82	6.92	1.08	6.06	3.52	0.65	2.68	4.20
1964	2.55	4.80	0.61	0.67	2.34	5.20	4.78	8.89	3.46	2.74	0.65	0.72
1965	0.42	3.59	3.16	1.76	1.11	10.16	5.57	2.78	4.71	9.06	0.34	1.89
1966	5.47	3.67	0.42	3.01	5.97	9.26	10.93	11.19	6.76	2.62	0.11	0.40
1967	0.84	1.69	0.24	0.14	2.58	11.27	7.02	3.74	8.53	3.37	0.08	1.95
1968	0.58	1.72	1.03	0.85	8.64	10.73	7.13	4.23	6.81	3.21	2.25	0.21
1969	1.76	2.28	6.19	0.69	4.10	10.09	3.68	10.04	8.49	11.75	1.46	3.82
1970	3.55	2.40	12.63	0.02	2.98	8.74	5.91	7.35	3.46	4.70	0.13	0.28
1971	0.25	0.51	0.37	0.14	1.50	13.86	7.28	8.29	7.18	6.35	0.90	1.20
1972	0.30	1.55	2.24	2.34	7.52	10.50	2.77	6.40	0.93	0.40	2.21	1.39
1973	2.72	2.73	3.34	1.02	5.88	10.48	8.01	5.58	8.43	1.38	0.03	1.52
1974	0.14	1.36	0.08	0.97	3.00	14.91	18.56	7.99	5.91	1.35	1.64	1.71
1975	0.20	1.95	0.74	1.22	4.89	5.29	7.00	3.13	11.11	4.88	0.27	0.38
1976	0.65	1.41	1.59	1.81	4.43	3.10	9.98	12.31	5.74	0.80	1.88	2.31
1977	4.87	1.38	1.12	0.20	5.17	3.74	6.19	5.51	6.29	1.01	5.33	4.74
1978	1.78	1.39	2.64	2.06	8.38	5.43	9.32	2.67	6.40	2.23	2.13	4.39
1979	21.40	0.23	2.30	0.84	7.64	1.09	1.45	5.66	17.69	1.90	1.83	1.96
1980	2.76	1.08	2.32	5.29	2.23	3.10	7.58	7.61	6.88	1.47	2.20	0.62
1981	0.87	1.52	1.28	0.38	2.06	3.33	3.70	10.29	4.54	0.24	1.27	0.15

Table C.2. 55 years of monthly rainfall data for the South Florida Station 6013.

***** STATION 6013, AVON PARK *****

1927	0.10	1.87	2.29	1.52	0.31	8.59	5.39	5.93	3.98	3.80	0.40	1.71
1928	0.26	1.14	3.12	3.66	3.51	6.90	13.01	9.66	10.64	2.05	1.03	0.35
1929	1.70	-1.00	1.35	2.78	5.62	8.42	5.61	10.55	11.59	2.40	0.56	2.29
1930	4.00	4.17	6.59	3.95	7.55	11.37	4.49	7.06	18.22	2.42	1.25	4.13
1931	3.92	2.36	3.75	5.25	6.10	3.74	8.15	6.37	7.84	2.98	0.18	1.47
1932	0.63	0.14	1.99	2.08	5.95	9.29	4.68	2.80	4.06	4.50	2.48	0.07
1933	1.97	2.35	1.70	5.90	3.66	4.77	13.78	-1.00	11.71	1.94	3.47	0.27
1934	1.22	2.80	3.58	4.32	7.15	10.94	4.13	-1.00	3.17	0.11	0.93	1.00
1935	0.41	1.15	0.81	6.03	2.87	6.87	-1.00	9.93	11.35	2.99	1.05	2.39
1936	4.83	8.35	5.52	1.67	2.59	10.87	-1.00	7.99	9.99	3.87	1.07	2.14
1937	2.63	5.13	3.31	4.06	1.65	-1.00	5.29	6.27	6.47	6.47	5.44	0.87
1938	1.44	1.43	1.45	0.42	3.43	4.64	8.13	4.24	2.81	6.44	2.50	0.19
1939	1.52	1.20	1.34	4.66	5.85	7.91	8.22	19.85	6.22	4.63	0.50	0.61
1940	3.83	3.66	3.58	1.54	5.30	8.43	11.76	4.02	9.94	0.68	0.10	4.43
1941	4.01	3.02	2.92	4.73	1.04	9.52	15.20	3.11	4.89	2.62	2.49	1.98
1942	4.48	4.72	3.86	2.67	6.43	8.52	8.76	5.19	5.37	0.13	0.0	3.54
1943	1.21	0.46	4.94	1.69	8.83	5.76	7.86	10.02	3.98	4.35	1.32	0.59
1944	-1.00	-1.00	-1.00	5.73	2.07	7.39	11.17	6.42	3.39	4.45	0.26	0.51
1945	1.95	0.03	0.40	1.61	2.45	14.09	14.48	2.79	8.43	5.94	0.49	2.00
1946	1.14	2.11	1.08	0.20	6.03	8.02	9.88	6.04	8.09	4.74	2.06	1.31
1947	1.92	3.82	6.18	4.65	3.57	12.77	10.50	9.30	14.31	2.97	2.65	1.65
1948	4.03	0.51	0.83	6.00	2.34	4.39	18.99	6.72	16.10	6.99	1.99	1.50
1949	0.13	0.08	0.92	3.30	2.66	6.74	6.48	16.12	8.18	0.70	1.79	0.41
1950	0.0	0.66	1.46	3.15	2.42	2.08	3.38	5.90	7.83	7.56	0.32	1.79
1951	0.22	2.57	0.64	10.35	0.33	6.88	5.30	8.72	3.99	5.94	-1.00	0.90
1952	1.30	4.61	5.49	0.97	5.48	7.38	7.23	8.46	5.42	6.80	1.60	1.15
1953	3.27	2.58	6.90	7.45	0.83	13.16	5.52	11.00	12.71	6.82	7.44	2.40
1954	1.78	1.96	1.62	4.71	3.12	18.95	4.73	6.31	6.20	1.60	1.60	1.97
1955	2.73	1.06	1.67	1.31	1.62	5.27	6.65	1.86	8.93	2.46	0.56	0.74
1956	0.26	0.94	1.54	2.23	1.95	9.13	4.76	10.95	6.76	7.78	0.22	0.22
1957	2.14	5.10	4.77	6.07	10.91	9.37	12.74	6.99	7.08	1.45	1.30	2.12
1958	8.33	3.50	5.55	3.43	4.16	6.77	4.45	6.31	4.97	2.75	0.91	3.96
1959	1.23	3.60	7.35	3.06	6.47	15.17	7.03	8.20	12.06	11.26	1.73	2.47
1960	0.55	6.54	5.52	3.00	2.28	7.06	13.67	8.07	14.82	3.06	0.28	1.02
1961	2.30	3.22	3.02	2.06	4.18	9.56	4.09	4.77	2.86	2.11	0.58	0.78
1962	1.62	1.53	3.38	3.30	1.21	10.80	2.96	8.42	7.07	1.23	2.68	1.42
1963	2.35	6.13	1.22	0.81	13.06	7.28	7.24	6.29	10.10	0.45	5.28	3.59
1964	2.97	4.58	3.81	2.28	3.24	6.08	9.44	5.28	7.31	0.61	0.77	1.08
1965	1.08	4.37	6.85	2.91	1.44	9.53	13.66	4.75	7.67	4.26	1.19	2.39
1966	5.95	6.05	0.77	2.98	5.08	9.68	8.27	8.98	7.85	2.02	0.15	1.36
1967	0.65	2.81	0.51	0.0	-1.00	-1.00	9.74	9.94	7.15	0.86	0.36	2.42
1968	0.58	1.91	1.29	0.43	8.73	16.73	8.19	6.32	4.40	3.94	2.73	0.35
1969	1.89	1.80	6.89	0.97	1.86	11.92	5.34	8.88	7.84	7.91	1.64	4.35
1970	2.99	2.03	5.23	0.22	3.92	4.51	14.93	5.33	5.84	2.25	0.54	1.06
1971	0.22	2.52	0.95	0.49	2.34	6.22	5.59	8.29	6.17	7.11	0.63	1.92
1972	0.93	3.47	3.74	2.24	4.75	8.30	9.67	7.23	0.36	1.98	4.95	2.80
1973	-1.00	1.57	3.06	5.61	2.06	3.64	8.50	10.71	7.59	4.43	0.80	-1.00
1974	-1.00	1.26	-1.00	-1.00	-1.00	20.14	9.64	3.53	3.22	0.36	0.23	2.20
1975	0.50	1.93	1.98	0.23	5.30	5.45	5.90	8.52	9.14	6.23	0.49	0.28
1976	0.51	0.54	2.46	1.59	6.20	7.66	8.84	7.80	6.29	2.08	1.81	1.91
1977	2.69	1.66	0.46	0.26	3.99	4.95	8.27	4.38	4.03	1.62	4.39	2.61
1978	2.96	4.32	2.29	0.13	5.17	10.05	13.36	4.13	2.02	1.42	0.49	3.23
1979	6.53	1.12	2.44	1.87	7.76	10.17	4.05	4.92	13.37	1.18	1.23	1.58
1980	2.42	3.46	1.80	5.41	3.15	5.09	4.60	6.55	3.88	4.19	2.68	1.09
1981	0.57	4.16	2.13	0.17	2.21	7.56	6.57	6.49	8.01	0.61	1.03	0.55

Note: -1 indicates missing value.

Table C.3. 55 years on monthly rainfall data for the South Florida Station 6093.

***** STATION 6093, FORT MYERS WSO AP *****

1927	0.30	0.76	1.42	0.80	1.23	8.04	8.78	3.14	5.59	1.78	0.30	0.71
1928	0.23	2.05	0.51	1.44	2.61	9.25	12.26	13.95	11.78	3.22	0.71	0.30
1929	1.09	0.08	1.03	0.88	7.82	8.30	6.68	5.56	15.44	3.42	0.30	1.31
1930	1.09	2.88	5.08	5.89	6.80	14.61	4.65	5.97	13.73	1.88	0.13	2.45
1931	3.53	3.76	6.64	2.92	2.58	3.96	6.33	7.27	6.44	0.86	0.09	1.83
1932	0.70	0.53	1.93	1.06	7.03	3.59	7.91	17.64	6.08	5.37	0.71	0.30
1933	0.25	2.60	3.93	6.06	6.86	5.02	9.20	4.51	4.63	2.08	1.09	0.13
1934	0.76	5.93	0.75	0.92	5.78	11.56	6.09	3.55	8.30	1.59	0.66	0.31
1935	0.24	1.81	0.0	3.50	2.36	6.42	9.30	9.38	14.49	0.30	0.83	1.58
1936	3.33	5.50	1.69	1.14	6.11	20.25	8.54	7.50	3.56	5.39	2.78	1.34
1937	0.52	3.68	3.74	1.38	0.94	10.75	5.13	7.00	3.04	5.88	1.44	0.72
1938	2.20	0.34	0.70	0.33	2.91	8.24	12.71	5.28	5.12	3.57	0.39	0.21
1939	0.45	0.87	0.04	8.42	3.01	16.43	7.69	6.97	12.83	5.81	1.80	1.01
1940	3.79	4.00	4.41	1.73	0.73	10.52	3.50	8.69	13.02	0.61	0.13	5.42
1941	3.02	3.82	6.88	7.66	1.16	7.12	15.28	7.46	6.09	0.96	2.48	0.99
1942	1.60	3.35	2.31	4.54	3.38	11.15	10.66	9.18	5.37	0.50	0.08	1.80
1943	0.74	0.71	1.61	4.45	5.96	16.06	12.24	8.59	5.68	3.56	2.37	0.48
1944	1.20	0.0	3.76	0.85	4.00	3.73	5.09	5.89	3.56	5.77	0.0	0.32
1945	2.19	0.68	0.10	0.21	1.58	11.97	12.41	11.06	5.71	5.19	0.03	1.45
1946	0.35	2.24	0.19	0.01	6.71	10.19	5.78	6.47	5.21	1.34	3.39	0.57
1947	0.83	2.92	8.94	2.82	6.47	12.84	11.17	9.40	16.32	4.97	2.05	1.44
1948	4.16	0.06	0.83	1.57	2.19	5.06	10.08	4.98	14.05	3.90	0.45	0.63
1949	0.01	0.07	0.13	5.50	4.03	7.53	13.32	7.60	12.70	3.60	1.27	1.62
1950	0.0	0.08	0.49	0.08	4.14	4.84	6.83	5.93	8.32	3.26	0.02	2.20
1951	0.38	1.96	1.13	2.71	2.14	9.19	11.44	10.30	3.48	11.91	1.14	0.14
1952	1.28	4.34	2.05	0.78	1.75	7.95	5.74	8.39	12.35	8.34	0.75	0.71
1953	1.71	2.01	0.68	2.28	0.41	12.81	9.34	4.32	15.58	6.68	1.07	1.18
1954	0.30	2.53	2.13	3.49	4.08	4.78	9.19	6.84	10.31	1.82	2.33	1.93
1955	2.68	1.16	0.32	0.97	3.23	8.53	8.76	4.29	10.50	2.15	0.52	0.85
1956	0.57	1.06	0.05	3.50	4.76	4.67	5.34	8.03	6.00	4.42	1.35	0.10
1957	0.78	3.68	4.73	2.69	7.97	4.85	12.52	9.39	8.77	3.19	1.52	3.55
1958	6.04	1.26	10.31	2.18	6.22	7.37	10.92	4.12	8.89	4.57	1.43	3.36
1959	1.48	1.72	6.33	1.75	4.74	16.10	6.17	5.75	6.89	12.04	1.92	1.79
1960	0.46	3.66	1.87	3.83	2.20	5.20	13.76	5.66	11.93	3.01	2.02	0.73
1961	3.31	1.88	3.58	0.46	4.92	9.75	9.82	13.41	2.80	3.16	1.12	0.53
1962	0.43	0.54	2.65	1.37	0.34	12.08	6.01	10.89	14.54	5.44	3.01	0.85
1963	0.81	4.65	0.59	0.27	7.58	7.70	4.06	3.98	7.49	0.05	3.45	2.27
1964	2.88	3.30	2.12	0.80	0.50	4.58	2.28	4.26	9.45	1.38	0.22	1.06
1965	1.24	2.99	2.91	2.39	4.70	7.78	12.05	6.57	4.35	4.42	0.58	0.85
1966	3.39	1.06	0.37	3.03	1.61	12.42	8.22	8.10	4.18	2.14	0.18	0.29
1967	1.15	2.15	0.72	0.0	1.46	7.41	6.69	15.86	7.04	3.08	0.92	2.91
1968	0.40	2.08	0.65	0.57	10.32	15.03	9.85	11.44	8.92	7.99	2.88	0.16
1969	1.44	2.87	4.74	0.15	4.71	10.63	7.11	8.49	16.60	11.03	0.22	3.95
1970	4.36	2.20	18.58	0.0	6.36	7.47	4.74	4.82	8.29	1.19	0.46	0.37
1971	0.85	1.55	0.55	0.70	3.77	6.18	9.50	8.06	9.21	6.49	0.16	0.30
1972	0.77	2.14	4.72	0.27	5.20	7.86	9.72	16.22	2.33	2.20	3.85	1.43
1973	3.14	2.23	3.89	1.71	0.78	3.99	9.57	8.66	8.38	0.16	0.10	1.72
1974	0.36	0.81	0.03	0.11	2.40	20.10	14.47	7.70	4.31	0.19	1.46	0.89
1975	0.26	0.27	1.47	0.80	2.78	10.55	10.81	7.74	12.59	3.05	0.49	0.69
1976	0.21	1.20	0.91	0.90	5.22	10.59	6.14	8.95	8.81	1.96	2.10	1.68
1977	3.53	0.15	0.09	0.76	6.51	8.96	9.60	10.58	9.21	0.43	1.50	2.74
1978	2.48	3.36	3.43	2.35	2.52	6.75	10.29	10.90	5.18	1.45	0.04	4.35
1979	7.45	1.92	0.43	3.12	5.32	8.31	5.96	14.79	13.65	0.39	0.48	5.16
1980	2.44	1.04	3.59	1.52	8.73	1.99	7.02	8.79	4.64	1.54	3.15	0.55
1981	0.80	1.65	1.29	0.08	3.07	11.79	8.24	16.73	6.70	0.40	0.71	0.73

Table C.4. 55 years of monthly rainfall data for the South Florida Station 6042.

***** STATION 6042, CANAL POINT USDA *****

1927	0.33	1.80	2.37	1.08	1.54	6.31	7.32	8.14	3.31	3.35	0.49	0.40
1928	0.19	1.38	3.48	1.72	3.10	5.42	14.57	14.13	16.45	0.77	1.24	0.20
1929	1.34	0.07	0.60	2.32	5.43	11.74	11.26	6.31	10.70	3.08	0.69	1.08
1930	2.54	3.03	4.32	9.25	6.10	16.96	4.08	3.07	5.36	5.14	0.67	2.77
1931	2.05	0.91	4.27	5.71	3.05	0.49	3.33	4.67	5.64	4.43	0.70	4.62
1932	0.26	2.38	0.87	2.67	3.49	11.26	4.91	9.91	2.40	4.51	25.09	0.16
1933	1.54	0.35	4.73	6.42	1.31	7.62	14.02	8.51	8.16	4.36	1.84	0.09
1934	0.25	5.36	2.77	7.64	6.27	7.96	5.20	8.14	11.69	-1.00	-1.00	-1.00
1935	0.16	2.81	0.17	5.45	0.76	6.11	3.98	3.62	11.90	4.44	0.57	1.22
1936	2.40	5.69	3.27	0.39	6.10	14.29	5.44	8.59	4.08	2.84	5.08	1.65
1937	4.30	1.81	4.88	3.36	1.92	4.44	14.62	9.37	5.88	6.50	2.23	0.26
1938	0.12	0.84	1.08	0.45	3.13	6.67	7.28	5.52	8.45	3.69	0.97	0.10
1939	0.38	0.08	1.26	2.82	4.29	8.87	6.40	12.26	8.86	5.55	0.42	2.32
1940	-1.00	-1.00	-1.00	0.38	5.61	8.63	8.79	8.22	6.09	1.20	0.57	4.76
1941	5.72	4.63	3.74	6.68	2.23	3.90	14.73	4.78	6.40	4.92	1.72	1.50
1942	1.34	2.77	6.36	2.36	4.92	14.11	3.62	4.42	4.93	2.06	2.15	2.47
1943	0.31	0.45	2.08	1.33	1.86	8.83	11.73	6.56	5.10	2.81	2.08	0.38
1944	0.98	0.04	4.17	2.71	3.98	3.40	5.66	5.81	4.73	8.35	0.30	0.43
1945	0.47	0.88	0.03	0.70	3.11	10.93	10.83	7.24	13.71	4.10	0.49	0.53
1946	1.13	0.84	4.31	0.0	10.60	11.20	8.59	6.98	12.28	1.54	5.08	2.13
1947	0.42	2.66	8.52	5.16	4.46	10.90	11.56	10.66	17.61	9.72	5.28	1.16
1948	3.70	0.48	0.78	5.18	1.30	2.17	7.62	8.41	16.14	2.74	0.38	0.34
1949	0.40	0.80	0.52	1.94	1.64	15.69	6.28	12.16	7.36	1.94	1.09	6.47
1950	0.30	0.79	3.04	0.87	2.14	2.15	6.71	4.20	3.20	11.17	1.07	1.25
1951	0.04	2.06	1.01	5.41	5.68	6.34	9.16	8.68	5.38	10.58	0.98	0.90
1952	1.68	5.20	0.92	2.99	3.27	3.46	8.13	8.74	4.90	13.72	0.18	0.07
1953	1.83	1.89	2.69	4.20	0.84	7.85	14.00	12.24	11.02	7.65	2.10	1.82
1954	0.35	1.96	2.71	7.57	6.77	12.78	8.08	8.27	5.45	2.95	0.56	1.60
1955	1.31	2.20	2.08	2.67	1.55	12.93	8.45	7.27	4.46	1.70	0.27	2.03
1956	0.72	1.11	0.03	1.92	3.04	3.70	7.34	3.08	14.09	6.16	0.38	0.50
1957	3.88	2.57	2.97	5.73	11.35	5.20	10.89	4.97	12.68	3.15	0.77	5.75
1958	8.73	0.61	5.10	4.35	6.33	4.86	7.79	6.60	6.26	6.07	0.62	6.35
1959	2.20	0.01	5.73	3.90	10.03	9.19	12.52	5.29	7.72	9.66	2.18	1.72
1960	0.05	4.59	0.99	4.33	3.20	6.80	7.83	6.16	12.89	4.00	2.01	0.70
1961	3.67	0.43	4.17	2.03	8.82	3.21	9.25	10.79	1.19	4.55	0.97	0.20
1962	1.22	0.57	3.05	4.08	2.12	7.01	10.45	5.14	9.88	1.70	2.19	0.31
1963	0.99	4.18	0.71	0.09	6.41	7.68	1.60	5.54	3.61	2.58	1.62	6.09
1964	3.32	2.06	0.93	3.67	2.05	13.52	9.02	8.59	5.65	6.63	0.45	4.37
1965	0.97	4.54	2.20	2.04	4.50	10.25	8.10	7.22	7.32	13.24	0.32	1.13
1966	4.09	2.27	1.01	3.02	5.46	9.81	12.03	5.66	5.77	6.60	0.31	0.84
1967	0.66	2.55	1.00	0.0	1.36	6.33	7.73	3.48	4.37	3.45	0.13	1.40
1968	0.29	2.27	0.80	0.33	7.26	19.18	10.35	4.21	10.55	7.36	1.77	0.02
1969	1.66	1.76	4.74	1.87	7.17	9.93	3.36	8.09	5.82	8.44	2.09	2.14
1970	3.13	2.89	14.55	0.0	6.92	3.10	9.45	13.07	2.19	3.79	0.17	0.10
1971	0.40	1.12	0.40	0.16	6.74	8.43	5.07	5.40	6.47	8.09	1.80	1.97
1972	2.33	1.99	2.09	4.03	-1.00	9.99	-1.00	2.50	1.77	1.72	4.15	2.42
1973	2.66	1.99	2.00	0.84	5.03	4.62	6.03	4.30	5.74	3.38	0.98	1.77
1974	2.12	0.58	0.22	1.37	6.01	10.43	6.87	5.89	7.14	2.06	1.60	0.95
1975	0.46	4.15	1.00	1.09	10.13	7.34	7.72	4.52	8.95	4.36	0.82	0.21
1976	0.43	2.11	0.30	1.79	8.74	7.85	2.07	7.49	2.96	0.26	2.26	2.41
1977	3.62	0.46	0.55	1.11	3.01	5.83	2.06	6.84	13.28	1.39	6.17	6.59
1978	2.34	1.42	3.73	2.02	5.69	15.47	6.22	10.41	8.03	4.57	2.37	4.55
1979	-1.00	-1.00	-1.00	-1.00	4.65	2.34	2.85	4.09	11.96	3.52	2.52	2.10
1980	3.06	1.89	1.94	5.08	4.15	5.10	7.52	5.96	16.08	1.42	1.59	0.62
1981	0.54	1.62	2.27	0.16	3.18	7.16	4.05	13.50	5.12	0.35	1.97	0.27

Note: -1 indicates missing value.

Table C.5. Monthly statistics of stations 6038, 6013, 6093, 6042.

***** STATION 6038 *****

VARIABLE	N	MEAN	STANDARD DEVIATION	SKEWNESS	C. V.
JAN	55	1.927	3.063	5.016	159.002
FEB	55	1.878	1.368	0.664	72.856
MAR	55	2.595	2.456	1.762	94.631
APR	55	2.507	1.818	0.674	72.498
MAY	55	4.575	2.584	1.032	56.482
JUN	55	7.606	3.776	0.646	49.646
JUL	55	7.235	3.358	1.008	46.420
AUG	55	7.033	2.897	0.724	41.193
SEP	55	7.567	4.085	1.081	53 983
OCT	55	3.747	3.073	1.138	82.017
NOV	55	1.379	1.283	1.532	93.042
DEC	55	1.457	1.555	1.975	106.686

***** STATION 6013 *****

VARIABLE	N	MEAN	STANDARD DEVIATION	SKEWNESS	C. V.
JAN	52	2.093	1.780	1.344	85.043
FEB	53	2.718	1.828	0.798	67.259
MAR	53	2.987	2.006	0.676	67.151
APR	54	2.928	2.209	0.864	75.460
MAY	53	4.192	2.655	1.073	63.326
JUN	53	8.613	3.694	1.129	42.892
JUL	53	8.307	3.664	0.793	44.108
AUG	53	7.258	3.148	1.487	43.370
SEP	55	7.521	3.732	0.716	49.620
OCT	55	3.500	2.488	0.798	71.082
NOV	54	1.567	1.551	1.833	98.982
DEC	54	1.687	1.135	0.727	67.280

***** STATION 6093 *****

VARIABLE	N	MEAN	STANDARD DEVIATION	SKEWNESS	C. V.
JAN	55	1.636	1.587	1.531	97.018
FEB	55	2.039	1.450	0.588	71.102
MAR	55	2.619	3.206	2.779	122.418
APR	55	1.995	1.953	1.474	97.916
MAY	55	4.049	2.414	0.381	59.615
JUN	55	9.105	4.082	0.777	44.835
JUL	55	8.672	2.976	0.123	34.313
AUG	55	8.309	3.490	0.974	41.997
SEP	55	8.553	3.988	0.407	46.624
OCT	55	3.474	2.877	1.295	82.817
NOV	55	1.175	1.048	0.881	89.186
DEC	55	1.399	1.260	1.555	90.105

***** STATION 6042 *****

VARIABLE	N	MEAN	STANDARD DEVIATION	SKEWNESS	C. V.
JAN	53	1.686	1.688	1.812	100.077
FEB	53	1.960	1.475	0.881	75.242
MAR	53	2.632	2.501	2.388	95.010
APR	54	2.860	2.278	0.758	79.661
MAY	54	4.626	2.659	0.666	57.482
JUN	55	8.141	4.107	0.530	50.446
JUL	54	7.861	3.408	0.256	43 353
AUG	55	7.194	2.853	0.636	39.655
SEP	55	7.802	4.126	0.660	52.880
OCT	54	4.709	3.163	1.056	67.163
NOV	54	1.972	3.489	5.743	176.912
DEC	54	1.818	1.863	1.362	102.457

Table C.6. Station 6038--monthly statistics of the incomplete and estimated series--2% missing values.

***** STATION 6038 (2% MIS.) *****

VARIABLE	N	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	54	1.885	3.076	163.191	5.083
FEB	54	1.906	1.365	71.646	0.643
MAR	54	2.632	2.464	93.616	1.743
APR	55	2.507	1.818	72.498	0.674
MAY	54	4.624	2.583	55.855	1.016
JUN	53	7.510	3.812	50.762	0.717
JUL	53	7.308	3.399	46.513	0.953
AUG	53	7.030	2.938	41.798	0.725
SEP	55	7.567	4.085	53.983	1.081
OCT	55	3.747	3.073	82.017	1.138
NOV	54	1.324	1.228	92.781	1.645
DEC	53	1.457	1.585	108.781	1.943

***** USING THE MEAN VALUE (2% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	1.885	3.048	161.666	5.127
FEB	1.906	1.353	70.976	0.649
MAR	2.632	2.441	92.746	1.759
APR	2.507	1.818	72.498	0.674
MAY	4.624	2.559	55.336	1.025
JUN	7.510	3.741	49.813	0.730
JUL	7.308	3.336	45.643	0.969
AUG	7.030	2.883	41.016	0.738
SEP	7.567	4.085	53.983	1.081
OCT	3.747	3.073	82.017	1.138
NOV	1.324	1.217	91.911	1.659
DEC	1.457	1.555	106.738	1.977

***** RECIPROCAL DISTANCES METHOD (2% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	1.921	3.059	159.265	5.039
FEB	1.878	1.368	72.821	0.664
MAR	2.598	2.453	94.413	1.766
APR	2.507	1.818	72.498	0.674
MAY	4.563	2.598	56.929	1.011
JUN	7.595	3.805	50.105	0.672
JUL	7.282	3.341	45.883	0.987
AUG	6.997	2.891	41.310	0.764
SEP	7.567	4.085	53.983	1.081
OCT	3.747	3.073	82.017	1.138
NOV	1.376	1.278	92.825	1.527
DEC	1.460	1.556	106.582	1.968

***** NORMAL RATIO METHOD (2% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	1.927	3.064	158.978	5.014
FEB	1.876	1.370	73.020	0.660
MAR	2.598	2.453	94.437	1.766
APR	2.507	1.818	72.498	0.674
MAY	4.560	2.603	57.080	1.002
JUN	7.538	3.757	49.839	0.703
JUL	7.279	3.339	45.877	0.993
AUG	6.977	2.896	41.516	0.779
SEP	7.567	4.085	53.983	1.081
OCT	3.747	3.073	82.017	1.138
NOV	1.349	1.231	91.259	1.566
DEC	1.448	1.556	107.430	1.992

***** MODIFIED WEIGHTED AVERAGE (2% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	1.953	3.089	158.165	4.889
FEB	1.871	1.377	73.587	0.643
MAR	2.591	2.459	94.897	1.757
APR	2.507	1.818	72.498	0.674
MAY	4.551	2.615	57.459	0.977
JUN	7.573	3.812	50.339	0.684
JUL	7.235	3.362	46.473	1.000
AUG	6.963	2.909	41.779	0.773
SEP	7.567	4.085	53.983	1.081
OCT	3.747	3.073	82.017	1.138
NOV	1.348	1.230	91.241	1.571
DEC	1.449	1.556	107.386	1.986

Table C.7. Station 6038--monthly statistics of the incomplete and estimated series--5% missing values.

***** STATION 6038 (5% MIS.) *****

VARIABLE	N	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	54	1.885	3.076	163.191	5.083
FEB	52	1.925	1.381	71.771	0.616
MAR	48	2.619	2.526	96.460	1.804
APR	51	2.368	1.731	73.093	0.666
MAY	51	4.521	2.437	53.900	0.853
JUN	52	7.411	3.683	49.699	0.719
JUL	53	7.355	3.356	45.626	0.984
AUG	54	7.062	2.917	41.303	0.696
SEP	53	7.474	4.132	55.287	1.147
OCT	52	3.790	3.109	82.021	1.131
NOV	52	1.344	1.238	92.137	1.643
DEC	55	1.457	1.555	106.686	1.975

***** USING THE MEAN VALUE (5% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.885	3.048	161.666	5.127
FEB	1.925	1.343	69.758	0.633
MAR	2.619	2.357	89.985	1.922
APR	2.368	1.666	70.330	0.689
MAY	4.521	2.345	51.866	0.884
JUN	7.411	3.580	48.299	0.738
JUL	7.356	3.293	44.772	1.001
AUG	7.062	2.890	40.919	0.702
SEP	7.474	4.055	54.255	1.167
OCT	3.790	3.021	79.711	1.161
NOV	1.344	1.203	89.555	1.688
DEC	1.457	1.555	106.686	1.975

***** RECIPROCAL DISTANCES METHOD (5% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.921	3.059	159.265	5.039
FEB	1.867	1.370	73.336	0.683
MAR	2.580	2.426	94.058	1.812
APR	2.429	1.786	73.539	0.667
MAY	4.417	2.414	54.646	0.885
JUN	7.613	3.776	49.603	0.653
JUL	7.259	3.332	45.904	1.036
AUG	7.039	2.895	41.128	0.722
SEP	7.733	4.294	55.526	1.022
OCT	3.837	3.070	80.022	1.098
NOV	1.375	1.270	92.366	1.578
DEC	1.457	1.555	106.686	1.975

***** NORMAL RATIO METHOD (5% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.927	3.064	158.978	5.014
FEB	1.856	1.377	74.160	0.683
MAR	2.557	2.423	94.741	1.847
APR	2.403	1.752	72.939	0.660
MAY	4.434	2.398	54.081	0.878
JUN	7.536	3.690	48.961	0.643
JUL	7.223	3.365	46.590	1.012
AUG	7.062	2.890	40.919	0.702
SEP	7.691	4.221	54.883	1.008
OCT	3.773	3.041	80.588	1.157
NOV	1.335	1.231	92.198	1.608
DEC	1.457	1.555	106.686	1.975

***** MODIFIED WEIGHTED AVERAGE (5% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.953	3.089	158.165	4.889
FEB	1.849	1.387	74.995	0.657
MAR	2.561	2.459	96.012	1.756
APR	2.405	1.773	73.744	0.644
MAY	4.403	2.438	55.374	0.846
JUN	7.584	3.787	49.934	0.666
JUL	7.197	3.396	47.191	0.973
AUG	7.019	2.907	41.409	0.725
SEP	7.796	4.425	56.753	1.085
OCT	3.816	3.092	81.032	1.106
NOV	1.349	1.221	90.547	1.624
DEC	1.457	1.555	106.686	1.975

Table C.8. Station 6038--monthly statistics of the incomplete and estimated series--10% missing value.

***** STATION 6038 (10% MIS.) *****

VARIABLE	N	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	50	1.848	3.108	168.167	5.328
FEB	47	1.924	1.374	71.410	0.705
MAR	52	2.509	2.463	98.180	1.904
APR	49	2.565	1.874	73.066	0.631
MAY	48	4.488	2.468	54.985	1.051
JUN	51	7.807	3.742	47.931	0.675
JUL	50	7.223	3.308	45.803	1.069
AUG	49	7.160	2.940	41.064	0.697
SEP	51	7.582	4.124	54.393	1.078
OCT	50	3.706	2.976	80.318	1.261
NOV	50	1.315	1.173	89.225	1.498
DEC	51	1.486	1.595	107.372	1.943

***** USING THE MEAN VALUE (10% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.848	2.961	160.178	5.572
FEB	1.923	1.268	65.926	0.760
MAR	2.509	2.394	95.411	1.955
APR	2.566	1.767	68.873	0.665
MAY	4.488	2.302	51.295	1.120
JUN	7.807	3.601	46.120	0.700
JUL	7.223	3.151	43.632	1.118
AUG	7.160	2.772	38.715	0.736
SEP	7.582	3.969	52.341	1.117
OCT	3.706	2.835	76.501	1.318
NOV	1.314	1.117	85.021	1.568

***** RECIPROCAL DISTANCES METHOD (10% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.872	3.025	161.608	5.225
FEB	1.876	1.366	72.829	0.736
MAR	2.610	2.462	94.323	1.745
APR	2.608	1.941	74.405	0.762
MAY	4.430	2.462	55.572	0.999
JUN	7.621	3.808	49.967	0.629
JUL	7.435	3.313	44.566	0.861
AUG	7.158	2.832	39.562	0.715
SEP	7.681	4.042	52.631	1.014
OCT	3.746	3.032	80.926	1.200
NOV	1.323	1.156	87.370	1.421
DEC	1.445	1.553	107.435	2.007

***** NORMAL RATIO METHOD (10% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.883	3.025	160.648	5.217
FEB	1.817	1.341	73.800	0.803
MAR	2.590	2.448	94.535	1.775
APR	2.556	1.870	73.141	0.678
MAY	4.498	2.432	54.064	0.960
JUN	7.632	3.814	49.972	0.647
JUL	7.263	3.188	43.889	1.046
AUG	7.121	2.800	39.326	0.753
SEP	7.624	4.019	52.721	1.064
OCT	3.660	2.963	80.954	1.192
NOV	1.347	1.175	87.209	1.382
DEC	1.451	1.553	107.011	2.000

***** MODIFIED WEIGHTED AVERAGE (10% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.908	3.165	165.872	4.645
FEB	1.841	1.402	76.132	0.696
MAR	2.616	2.478	94.722	1.728
APR	2.596	1.959	75.448	0.780
MAY	4.425	2.532	57.217	0.947
JUN	7.533	3.935	52.237	0.489
JUL	7.399	3.359	45.398	0.826
AUG	7.115	2.895	40.689	0.723
SEP	7.681	4.086	53.193	1.003
OCT	3.693	3.119	84.458	1.155
NOV	1.299	1.129	86.871	1.557
DEC	1.419	1.568	110.497	1.971

Table C.9. Station 6038--monthly statistics of the incomplete and estimated series--15% missing values.

***** STATION 6038 (15% MIS.) *****

VARIABLE	N	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	46	1.927	3.284	170.428	4.910
FEB	45	1.750	1.449	82.791	0.895
MAR	43	2.448	2.099	85.738	1.171
APR	47	2.484	1.857	74.749	0.749
MAY	47	4.674	2.490	53.272	0.887
JUN	48	7.727	3.854	49.876	0.592
JUL	50	7.111	3.467	48.762	1.101
AUG	45	7.245	2.954	40.777	0.664
SEP	47	7.293	3.766	51.632	0.673
OCT	49	3.656	3.003	82.138	1.182
NOV	46	1.488	1.343	90.292	1.417
DEC	49	1.470	1.598	108.686	1.975

***** USING THE MEAN VALUE (15% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	1.927	2.998	155.539	5.338
FEB	1.750	1.308	74.733	0.984
MAR	2.449	1.851	75.601	1.314
APR	2.484	1.714	69.007	0.808
MAY	4.673	2.298	49.174	0.956
JUN	7.728	3.596	46.529	0.631
JUL	7.111	3.303	46.450	1.152
AUG	7.244	2.667	36.813	0.730
SEP	7.293	3.475	47.657	0.725
OCT	3.656	2.831	77.431	1.247
NOV	1.488	1.226	82.403	1.540
DEC	1.470	1.506	102.471	2.085

***** RECIPROCAL DISTANCES METHOD (15% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	2.015	3.019	149.847	5.144
FEB	1.898	1.443	76.039	0.757
MAR	2.672	2.557	95.699	2.011
APR	2.659	1.993	74.944	0.696
MAY	4.711	2.574	54.652	0.775
JUN	7.676	3.765	49.050	0.578
JUL	7.159	3.338	46.633	1.079
AUG	7.174	2.745	38.261	0.734
SEP	7.401	3.749	50.660	0.714
OCT	3.893	3.118	80.103	1.068
NOV	1.402	1.272	90.778	1.534
DEC	1.463	1.537	105.049	1.979

***** NORMAL RATIO METHOD (15% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	2.017	3.029	150.183	5.095
FEB	1.808	1.370	75.774	0.812
MAR	2.621	2.377	90.718	1.575
APR	2.571	1.878	73.030	0.625
MAY	4.750	2.599	54.711	0.794
JUN	7.585	3.732	49.208	0.639
JUL	7.129	3.362	47.161	1.086
AUG	7.130	2.739	38.419	0.757
SEP	7.314	3.718	50.843	0.697
OCT	3.815	2.989	78 337	1.009
NOV	1.412	1.278	90.467	1.507
DEC	1.463	1.538	105.085	1.994

***** MODIFIED WEIGHTED AVERAGE (15% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C.V.	SKEWNESS
JAN	2.152	3.093	143.705	4.701
FEB	1.853	1.467	79 163	0.844
MAR	2.621	2.547	97.169	1.981
APR	2.490	1.969	79.056	0.564
MAY	4.760	2.707	56.857	0.797
JUN	7.557	3.871	51.220	0.529
JUL	7.053	3.369	47.772	1.113
AUG	7.094	2.835	39.965	0.700
SEP	7.315	3.857	52.732	0.766
OCT	3.865	3.203	82.859	1.120
NOV	1.413	1.248	88.354	1.623
DEC	1.431	1.553	108.479	1.956

Table C.10. Station 6038--monthly statistics of the incomplete and estimated series--20% missing values.

***** STATION 6038 (20% MIS.) *****

VARIABLE	N	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	47	1.856	3.240	174.615	5.048
FEB	45	1.909	1.436	75.213	0.656
MAR	45	2.412	2.112	87.561	1.144
APR	48	2.573	1.862	72.393	0.701
MAY	43	4.797	2.769	57.730	0.912
JUN	43	7.306	3.900	53.376	0.818
JUL	44	7.306	3.720	50.916	0.887
AUG	42	7.023	2.826	40.233	0.383
SEP	43	7.528	4.142	55.024	1.291
OCT	45	3.841	3.210	83.576	1.180
NOV	42	1.364	1.317	96.582	1.603
DEC	43	1.573	1.703	108.218	1.765

***** USING THE MEAN VALUE (20% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.856	2.991	161.109	5.434
FEB	1.909	1.296	67.886	0.720
MAR	2.411	1.906	79.048	1.257
APR	2.572	1.738	67.547	0.748
MAY	4.797	2.442	50.905	1 023
JUN	7.307	3.439	47.067	0 917
JUL	7.307	3.319	45.430	0 983
AUG	7.022	2.462	35.061	0 435
SEP	7.528	3.653	48.524	1 448
OCT	3.841	2.898	75.446	1 296
NOV	1.363	1.148	84.213	1.820
DEC	1.573	1.501	95.482	1 981

***** RECIPROCAL DISTANCES METHOD (20% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.930	3.032	157.096	5.151
FEB	1.870	1.343	71.811	0 686
MAR	2.638	2.532	95.978	2.121
APR	2.520	1.868	74.132	0.665
MAY	4.651	2.636	56.673	0.916
JUN	7.529	3.629	48.200	0 662
JUL	7.643	3.501	45.811	0.669
AUG	7.060	2.531	35.852	0.400
SEP	7.839	4.065	51.860	0.950
OCT	3.840	3.065	79.805	1.128
NOV	1.710	2.417	141.338	4.585
DEC	1.540	1.594	103.496	1.811

***** NORMAL RATIO METHOD (20% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	1.952	3.037	155.594	5.110
FEB	1.828	1.340	73.293	0.780
MAF	2.580	2.344	90.853	1.626
APR	2.530	1.870	73.914	0.665
MAY	4.671	2.658	56.899	0.859
JUN	7.181	3.591	50.001	0.894
JUL	7.383	3.404	46.102	0.873
AUG	6.950	2.570	36.975	0.428
SEP	7.684	3.930	51.143	1.087
OCT	3.779	3.001	79.420	1.232
NOV	1.459	1.418	97.228	1.850
DEC	1.499	1.543	102.983	1.948

***** MODIFIED WEIGHTED AVERAGE (20% MIS.) *****

VARIABLE	MEAN	STANDARD DEVIATION	C. V.	SKEWNESS
JAN	2 027	3.143	155.040	4.605
FEB	1.823	1.371	75.175	0.674
MAR	2.591	2.525	97.451	2.107
APR	2.501	1.886	75.377	0.644
MAY	4.705	2.770	58.864	0.779
JUN	7.283	3.690	50.666	0.732
JUL	7.572	3.610	47.677	0.736
AUG	6.941	2.593	37.356	0.517
SEP	7.829	4.323	55.215	0 814
OCT	3.738	3.188	85.288	1.043
NOV	1 510	1.655	109.608	2.927
DEC	1.494	1.648	110.308	1.751

Table C.11. Lag-zero covariance matrices of the monthly rainfall series of stations 6013, 6093, 6042. All matrices are symmetric.

	3.169					3.342		
JAN:	2.393	2.518		,	FEB:	1.813	2.101	
	2.343	1.723	2.848			1.496	1.505	2.175
	4.022					4.881		
MAR:	3.501	10.275		,	APR:	1.959	3.814	
	2.677	6.921	6.254			3.545	2.378	5.190
	7.047					13.647		
MAY:	3.610	5.826		,	JUN:	8.235	16.666	
	3.539	2.871	7.071			7.373	7.138	16.865
	13.425					9.907		
JUL:	3.655	8.854		,	AUG:	-0.628	12.177	
	3.378	2.546	11.615			1.630	0.260	8.138
	13.928					6.189		
SEP:	9.080	15.902		,	OCT:	5.032	8.278	
	5.913	6.443	17.022			4.516	5.896	10.004
	2.406					1.289		
NOV:	0.822	1.098		,	DEC:	1.045	1.588	
	1.395	0.547	12.174			1.145	1.510	3.471

Table C.12. Normality transformations applied on the monthly rainfall data of Station 6038.

***** STATION 6038 (NO TRANSFORMATION) *****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.927	3.063	5.016	159.002
FEB	1.878	1.368	0.664	72.856
MAR	2.595	2.456	1.762	94.631
APR	2.507	1.818	0.674	72.498
MAY	4.575	2.584	1.032	56.482
JUN	7.606	3.776	0.646	49.646
JUL	7.235	3.358	1.008	46.420
AUG	7.033	2.897	0.724	41.193
SEP	7.567	4.085	1.081	53.983
OCT	3.747	3.073	1.138	82.017
NOV	1.379	1.283	1.532	93.042
DEC	1.457	1.555	1.975	106.686

***** STATION 6038 (LOGARITHMIC TRANSFORMATION) *****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	-0.009	0.532	-0.266	-5598.555
FEB	0.103	0.466	-1.224	452.424
MAR	0.188	0.521	-0.936	276.239
APR	0.218	0.502	-1.549	230.355
MAY	0.593	0.251	-0.164	42.276
JUN	0.822	0.246	-0.875	29.918
JUL	0.809	0.227	-1.034	28.124
AUG	0.810	0.184	-0.226	22.643
SEP	0.814	0.253	-0.645	31.143
OCT	0.385	0.488	-1.330	126.596
NOV	-0.088	0.519	-0.755	-588.139
DEC	-0.068	0.479	-0.173	-706.420

***** STATION 6038 (POWER=0.25) *****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.040	0.314	0.738	30.141
FEB	1.096	0.256	-0.540	23.382
MAR	1.161	0.312	-0.108	26.870
APR	1.175	0.283	-0.698	24.109
MAY	1.421	0.203	0.143	14.281
JUN	1.620	0.218	-0.355	13.449
JUL	1.606	0.199	-0.456	12.387
AUG	1.603	0.168	0.024	10.457
SEP	1.614	0.227	-0.140	14.051
OCT	1.292	0.316	-0.334	24.436
NOV	0.990	0.268	-0.130	27.100
DEC	0.998	0.270	0.364	27.061

***** STATION 6038 (POWER=0.35) *****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.083	0.461	1.219	42.593
FEB	1.154	0.361	-0.327	31.285
MAR	1.257	0.458	0.163	36.456
APR	1.274	0.407	-0.443	31.909
MAY	1.644	0.328	0.264	19.933
JUN	1.975	0.366	-0.180	18.530
JUL	1.949	0.332	-0.238	17.051
AUG	1.942	0.283	0.121	14.592
SEP	1.965	0.382	0.040	19.458
OCT	1.456	0.480	-0.060	32.927
NOV	1.006	0.370	0.110	36.724
DEC	1.017	0.383	0.581	37.667

***** STATION 6038 (SQUARE ROOT) *****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.178	0.740	2.043	62.830
FEB	1.265	0.533	-0.048	42.115
MAR	1.442	0.724	0.540	50.192
APR	1.459	0.620	-0.117	42.466
MAY	2.059	0.584	0.445	28.373
JUN	2.671	0.693	0.053	25.944
JUL	2.618	0.625	0.073	23.873
AUG	2.598	0.539	0.265	20.760
SEP	2.654	0.729	0.295	27.446
OCT	1.768	0.795	0.282	44.953
NOV	1.051	0.529	0.461	50.341
DEC	1.067	0.570	0.908	53.473

Table C.13. Statistics of the estimated series--univariate model

**** UNIVARIATE MODEL (10% MIS.) ****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.854	2.980	5.462	160.709
FEB	1.891	1.278	0.815	67.608
MAR	2.503	2.402	1.943	95.956
APR	2.499	1.805	0.701	72.242
MAY	4.504	2.305	1.095	51.181
JUN	7.809	3.601	0.697	46.120
JUL	7.185	3.161	1.141	43.995
AUG	7.067	2.809	0.780	39.757
SEP	7.563	3.975	1.127	52.553
OCT	3.594	2.871	1.369	79.865
NOV	1.314	1.132	1.515	86.162
DEC	1.475	1.539	2.019	104.304

**** UNIVARIATE MODEL (20% MIS.) ****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.777	2.997	5.480	168.623
FEB	1.846	1.303	0.854	70.585
MAR	2.334	1.914	1.364	81 989
APR	2.523	1.743	0.828	69.053
MAY	4.713	2.449	1.119	51.959
JUN	7.199	3.446	1.009	47.865
JUL	7.216	3.325	1.062	46.072
AUG	6.961	2.465	0.510	35.406
SEP	7.420	3.659	1.531	49 316
OCT	3.719	2.910	1.408	78.235
NOV	1.302	1.153	1.954	88.580
DEC	1.498	1.509	2.101	100.727

Table C.14. Statistics of the estimated series--bivariate model

**** BIVARIATE MODEL (10% MIS.) ****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C V.
JAN	1.825	2.972	5.532	162.862
FEB	1.869	1.287	0.841	68.866
MAR	2.483	2.398	1.976	96.608
APR	2.534	1.781	0.698	70 275
MAY	4.491	2.327	1.086	51 825
JUN	7.749	3.612	0.740	46.611
JUL	7.220	3.166	1.104	43 856
AUG	7.110	2 784	0.779	39.159
SEP	7.523	3.979	1.154	52 888
OCT	3.592	2.863	1.392	79.709
NOV	1.293	1.123	1.599	86.839
DEC	1.459	1.539	2.052	105.450

**** BIVARIATE MODEL (20% MIS.) ****

VARIABLE	MEAN	STANDARD DEVIATION	SKEWNESS	C.V.
JAN	1.798	3.013	5.374	167 566
FEB	1.835	1.337	0.790	72.847
MAR	2.353	1.936	1.283	82.300
APR	2.551	1.775	0.763	69 591
MAY	4.759	2.490	1.005	52.327
JUN	7.280	3.490	0.897	47 932
JUL	7.263	3.353	0.994	46.161
AUG	6.941	2.515	0.506	36 239
SEP	7.482	3.729	1.404	49.839
OCT	3.786	2 941	1.294	77.684
NOV	1.322	1.159	1.872	87.644
DEC	1.551	1.527	1.926	98.456

APPENDIX D

COMPUTER PROGRAMS

RAEMV-U

(Recursive Algorithm for the Estimation of Missing Values - Univariate Model)

Input

The program inputs the time series; the parameters of the normality transformation to be performed (power transformation); the number of gaps (not necessarily the number of missing values unless all the gaps are singles); and for each gap the starting and ending point (counting starts from the first value in the series). For the first iteration the missing values in the original series (usually indicated by a code or by a negative value) are initialized to zeroes or to some other desired initial estimates.

Program Description

The main program reads the input data and then subsequently calls subroutine ARMA (each call corresponds to one iteration). Subroutine ARMA performs the following calculations each time it is called:

(1) The input series is transformed to normal (using the selected transformation) and stationary (by subtracting the monthly means and dividing by the standard deviations).

(2) The mean, variance, autocovariance function (ACVF), autocorrelation function (AGF) and partial autocorrelation function (PACF) of the transformed series are computed by calling the IMSL subroutine FTAUTO.

(3) Preliminary estimates of the p AR parameters, and q MA parameters are computed by calling the IMSL subroutines FTARPS and FTMPS subsequently.

(4) Maximum likelihood estimates (MLE) of the AR and MA parameters are computed and the residual series is calculated by calling the IMSL subroutine FTMXL.

(5) The mean, variance, ACVF, ACF and PACF of the residual series are computed by calling the IMSL subroutine FTAUTO.

(6) The parameters of the fitted model (MLE) are used to estimate the missing values in all the gaps by the Box-Jenkins minimum mean square error forecasting procedure.

(7) The inverse normality and stationarity transformations are performed on the series and the estimated complete series is output.

The estimated series (output from the first call) now becomes the input series for the second call and the above seven steps are repeated. The subroutine ARMA is called as many times as needed until stabilization of the parameter estimates and of the missing values estimates occur. The program is initialized to five calls (more can be easily added as needed), and a stabilization check for the parameters is provided so that the iterations stop when the two parameters remain constant to the second decimal place.

The computation and printing of the ACVF, ACF and PACF of the transformed and residual series (steps 2 and 5) are not necessary and can be eliminated from the program without any problem. However, their inclusion permits the checking of the goodness of the fitted model at each iteration by diagnostic checking applied on the residuals. A listing of the program in FORTRAN follows.

RAEMV-B

(Recursive Algorithm for the Estimation of Missing Values - Bivariate Model)

The special case of having only the one series incomplete and the other complete will be considered here. However, the program can be easily modified to include the case of having both the series incomplete.

Input

The program inputs the two time series, the parameters of the normality transformation to be performed on each series, the number of gaps and the position of each gap for the incomplete series. The missing values in the incomplete series are initialized to zeros or to some other values.

Program description

The main program reads the input data and then subsequently calls subroutine BIVAR (each call corresponds to one iteration). Subroutine BIVAR performs the following each time is called:

(1) The two input series are transformed to normal and stationary by calling subroutine STAT.

(2) The lag-zero and lag-one autocovariances and cross-covariances of the two series are computed by calling the IMSL subroutine FTCRXY.

(3) The parameter matrices A and B are calculated. Inversion and multiplication of matrices are performed by the IMSL subroutines LINV2F, VMULFF and VMULFP.

(4) The parameter matrices A and B are used to estimate the missing values of the incomplete series.

(5) The inverse normality and stationarity transformations are performed on the two series, and the estimated complete series is output.

The estimated series (output from the first call) now becomes the input series for the second call and the above five steps are repeated until stabilization of the matrices A and B occurs. No check for stabilization is provided by the program (eight values must be checked simultaneously) but instead the subroutine is called for a prefixed number of times. A listing of the computer program in FORTRAN follows.

```
C
C-----------------------------------------------------------
C
C                PROGRAM RAEMV-U
C
C          RECURSIVE ALGORITHM FOR THE ESTIMATION OF
C          MISSING VALUES - UNIVARIATE MODEL
C
C-----------------------------------------------------------
C
      DIMENSION RAIN(60,12),NYEAR(60)
      DIMENSION VRAIN(800),E1RAIN(60,12),VRAIN1(800),
     .   E2RAIN(60,12),VRAIN2(800),
     .   E3RAIN(60,12),VRAIN3(800),
     .   E4RAIN(60,12),VRAIN4(800),
     .   E5RAIN(60,12),VRAIN5(800),
     .   E6RAIN(60,12),VRAIN6(800)
      DIMENSION LI(200),LG(200),ISI(200),IEI(200)
      COMMON/A/ ID,NYEAR
      COMMON/B/ N,C,P
      COMMON/C/ NG,ISG(200),IEG(200)
C
C     READ INPUT PARAMETERS
C     HEADER..TITLE
C     N.......NUMBER OF YEARS
C     NG......NUMBER OF GAPS
C     LI......LENGTH OF INTEREVENT
C     LG......LENGTH OF GAP
C     C,P......PARAMETERS OF THE TRANSFORMATION
C              TRANSFORMED SERIES  Y=(X+C)**P
C
      READ(5,10) HEADER
 10   FORMAT(20A4)
C
      READ(5,20) C,P
 20   FORMAT(2F5.2)
C
      READ(5,30) N,NG
 30   FORMAT(2(I4/))
C
      DO 5 I=1,NG
 5    READ(5,40) LI(I),LG(I)
 40   FORMAT(2I4)
C
      READ(10,110) (ID,(NYEAR(I),(RAIN(I,J),J=1,12)),I=1,N)
 110  FORMAT(A4,I3,1X,12F6.2)
C
C     FROM THE INPUT VARIABLES LI AND LG TWO ARRAYS OF
C     LENGTH NG ARE COMPUTED. THEN THE STARTING POINT
C     OF THE KTH GAP IS ISG(K) AND THE ENDING POINT IS
C     IEG(K).
C
      ISI(1)=1
      IEI(1)=LI(1)
      ISG(1)=IEI(1)+1
      IEG(1)=ISG(1)+LG(1)-1
      DO 11 I=2,NG
      ISI(I)=IEG(I-1)+1
      IEI(I)=ISI(I)+LI(I)-1
      ISG(I)=IEI(I)+1
      IEG(I)=ISG(I)+LG(I)-1
 11   CONTINUE
C
      WRITE(6,15) HEADER
 15   FORMAT(20A4,///)
C
C     PRINT THE POSITION OF THE GAPS FOR CHECKING
C
      WRITE(6,100) (I,ISG(I),IEG(I),I=1,NG)
 100  FORMAT(3I6)
C
C     INITIALIZE THE MISSING VALUES TO ZERO
C
      DO 50 I=1,N
      DO 50 J=1,12
      IF(RAIN(I,J).EQ.-1) RAIN(I,J)=0.
 50   CONTINUE
C
```

```
C      SUBROUTINE ARMA IS CALLED TO FIT AN ARMA(P,Q) MODEL
C      TO THE INPUT SERIES. THE PARAMETERS OF THE MODEL
C      ARE USED FOR THE ESTIMATION OF THE MISSING VALUES.
C
       CALL ARMA(RAIN,VRAIN,E1RAIN,VRAIN1,PHI1,THETA1)
       CALL ARMA(E1RAIN,VRAIN1,E2RAIN,VRAIN2,PHI2,THETA2)
       CALL ARMA(E2RAIN,VRAIN2,E3RAIN,VRAIN3,PHI3,THETA3)
       IF((PHI3-PHI2).LE.0.001.AND.(THETA3-THETA2).LE.0.001)
      .  GO TO 999
       CALL ARMA(E3RAIN,VRAIN3,E4RAIN,VRAIN4,PHI4,THETA4)
       IF((PHI4-PHI3).LE.0.001.AND.(THETA4-THETA3).LE.0.001)
      .  GO TO 999
       CALL ARMA(E4RAIN,VRAIN4,E5RAIN,VRAIN5,PHI5,THETA5)
       IF((PHI5-PHI4).LE.0.001.AND.(THETA5-THETA4).LE.0.001)
      .  GO TO 999
       CALL ARMA(E5RAIN,VRAIN5,E6RAIN,VRAIN6,PHI6,THETA6)
C
C
 999   STOP
       END
C
C--------------------- SUBROUTINE ARMA ---------------------------
C
C      SUBROUTINE ARMA FITTS AN ARMA(P,Q) MODEL TO THE INPUT
C      SERIES EACH TIME IS CALLED. THE MISSING VALUES ARE
C      ESTIMATED BY THE BOX-JENKINS FORECASTING PROCEDURE
C      AND THE ESTIMATED SERIES IS SAVED TO BE THE INPUT
C      SERIES TO THE NEXT CALL.
C
       SUBROUTINE ARMA(RAIN,VRAIN,ERAIN,EVRAIN,PHI1,THETA1)
C
       REAL MEAN(13),MP,LP
       DIMENSION ERAIN(60,12),EVRAIN(800),Z(800)
       DIMENSION RAIN(60,12),NYEAR(60),IND(8),PHI(11),
      .  THETA(11),SUMSQ1(11,11),ACV(300),AC(300),PACV(300),
      .  VRAIN(800),TEMP(800),WKAREA(1600),A(1600),GR(1600)
       DIMENSION YTOTAL(61),STD(13)
       COMMON/A/ ID,NYEAR
       COMMON/B/ N,C,P
       COMMON/C/ NG,ISG(200),IEG(200)
C
C      INPUT THE VALUES OF PHI AND THETA FOR WHICH YOU WANT
C      THE SUM OF SQUARES SURFACE TO BE CALCULATED.
C
       DATA PHI/-.5,-.4,-.3,-.2,-.1,0.,.1,.2,.3,.4,.5/
       DATA THETA/0.,.1,.2,.3,.4,.5,.6,.7,.8,.9,1./
C
C      PRINT OUT THE ORIGINAL SERIES. THIS IS THE SERIES
C      THAT EACH ITERATION STARTS WITH.
C
       WRITE(6,112)
 112   FORMAT(1H1,//,50X,'TABLE 1',//40X,
      $    ' MONTHLY RAINFALL SERIES',///)
       WRITE(6,111) (ID,(NYEAR(I),(RAIN(I,J),J=1,12)),I=1,N)
 111   FORMAT(1X,A4,I3,1X,12F6.2)
C
C      COMPUTE POWER  TRANSFORMATION OF THE SERIES
C
       DO 10 I=1,N
       DO 10 J=1,12
       RAIN(I,J)=(RAIN(I,J)+C)**P
 10    CONTINUE
C
C      COMPUTE YEARLY TOTALS
C
 25    DO 22 I=1,N
       YTOTAL(I)=0.
       DO 23 J=1,12
       YTOTAL(I)=YTOTAL(I)+RAIN(I,J)
 23    CONTINUE
 22    CONTINUE
C
C      COMPUTE MONTHLY MEANS AND STANDARD DEVIATIONS
C
       DO 32 J=1,12
       MEAN(J)=0.
       STD(J)=0.
       DO 34 I=1,N
       MEAN(J)=MEAN(J)+RAIN(I,J)/FLOAT(N)
       STD(J)=STD(J)+RAIN(I,J)**2
 34    CONTINUE
 32    CONTINUE
C
       DO 36 I=1,12
      STD(I)=((STD(I)-MEAN(I)**2*FLOAT(N))/(FLOAT(N)-1.))**0.5
 36   CONTINUE
C
      MEAN(13)=0.
      STD(13)=0.
      DO 38 I=1,N
      MEAN(13)=MEAN(13)+YTOTAL(I)/FLOAT(N)
      STD(13)=STD(13)+YTOTAL(I)**2
 38   CONTINUE
      STD(13)=((STD(13)-MEAN(13)**2*FLOAT(N))/(FLOAT(N)-1.))**0.5
C
C     NOW STANDARDIZE THE MONTHLY SERIES
C
      DO 42 I=1,N
      DO 42 J=1,12
      RAIN(I,J)=(RAIN(I,J)-MEAN(J))/STD(J)
 42   CONTINUE
C
C     STORE THE MATRIX SERIES IN A VECTOR SERIES
C
      DO 30 I=1,N
      DO 30 J=1,12
      K=J+(I-1)*12
      VRAIN(K)=RAIN(I,J)
 30   CONTINUE
      NN=N*12
C
C     COMPUTE AC, PAC, AND ACV OF THE SERIES USING
C     SUBROUTINE FTAUTO
C
      L=30
      CALL FTAUTO(VRAIN,NN,L,L,7,AMEAN,ACV(1),ACV(2),AC(2),
     $            PACV(2),WKAREA)
C
C     SET AC AND PACV OF LAG ZERO TO ONE
C
      AC(1)=1.
      PACV(1)=1.
      WRITE(6,130) AMEAN,ACV(1)
 130  FORMAT(1H1,///,15X,'STANDARDIZED TRANSFORMED SERIES',//,
     1  5X,'MEAN..........',F15.7,//,
     2  5X,'VARIANCE.....',F15.7,///)
      WRITE(6,135)
 135  FORMAT(5X,'LAG',12X,'AC',12X,'PACV',/,2X,45('-')/)
C
      SSQ=0.
      LP1=L+1
      DO 40 I=1,LP1
      IM1=I-1
      WRITE(6,140) IM1,AC(I),PACV(I)
      SSQ=SSQ+AC(I)**2
 40   CONTINUE
      SSQ=SSQ-1
 140  FORMAT(3X,I5,2F15.7)
      WRITE(6,145) SSQ
 145  FORMAT(///,2X,'SUM OF AC**2 (NOT INCLUDING LAG 0)',F15.6)
C
C     PRELIMINARY ESTIMATE OF AR PARAMETER AND OVERALL MA
C     CONSTANT USING SUBROUTINE FTARPS
C
      CALL FTARPS(ACV,AMEAN,1,1,ARPS,PMAC,WKAREA)
      WRITE(6,150) ARPS,PMAC
 150  FORMAT(///,2X,'-----SUBROUTINE FTARPS-----',//,
     1    5X,'ESTIMATE OF AR PARAMETER (ARPS).......',F15.6,//,
     2    5X,'OVERALL MA CONSTANT (PMAC).............',F15.6,//)
C
C     ESTIMATE MA PARAMETER AND WHITE NOISE VARIANCE USING
C     SUBROUTINE FTMPS
C
      CALL FTMPS(ACV,ARPS,1,1,PM^S,WNV,WKAREA,IER)
      WRITE(6,155) PMAS,WNV
 155  FORMAT(///,2X,'-----SUBROUTINE FTMPS------',//,
     1    5X,'ESTIMATE OF MA PARAMETER (PMAS)......',F15.6,//,
     2    5X,'WHITE NOISE VARIANCE (WNV)...........',F15.6)
C
C     STORE VECTOR VRAIN BECAUSE FTMXL WILL DESTROY IT
C
      DO 45 I=1,NN
      TEMP(I)=VRAIN(I)
 45   CONTINUE
C
C     GENERATE SUM OF SQUARES SURFACE OF THE RESIDUALS
C     LET ITHETA AND IPHI BE THE ROW AND COLUMN NUMBER
C     CORRESPONDING TO THE MINIMUM SUM OF SQUARES OF
```

```
C      THE RESIDUALS
C
       ITHETA=1
       IPHI=1
       DO 50 I1=1,11
       DO 55 I2=1,11
       ETA=TEMP(2)-PHI(I2)*TEMP(1)
       SUMSQ1(I1,I2)=ETA**2
       DO 60 I3=3,NN
       ETA1=TEMP(I3)-PHI(I2)*TEMP(I3-1)+THETA(I1)*ETA
       ETA=ETA1
       SUMSQ1(I1,I2)=SUMSQ1(I1,I2)+ETA1**2
 60    CONTINUE
       IF(SUMSQ1(I1,I2).GT.SUMSQ1(ITHETA,IPHI)) GO TO 55
       ITHETA=I1
       IPHI=I2
 55    CONTINUE
 50    CONTINUE
C
C      WRITE OUT THE SUM OF SQUARES SURFACE OF THE RESIDUALS
C
       WRITE(6,160)
 160   FORMAT(1H1,///,50X,'TABLE 2',//,15X,'SUM OF SQUARES OF THE',
      $  ' RESIDUALS OF THE STANDARDIZED TRANSFORMED SERIES',
      $  ///,52X,'PHI')
       WRITE(6,165) (PHI(I),I=1,11)
 165   FORMAT(5X,'THETA',2X,11(3X,F5.2,1X)/)
       WRITE(6,170) (THETA(I),(SUMSQ1(I,J),J=1,11),I=1,11)
 170   FORMAT((5X,F5.2,3X,11(F8.2,1X)))
C
C      LET THE VALUES OF ARPS AND PMAS THAT LED TO MINIMUM SUM
C      OF SQUARES OF RESIDUALS BE INPUT TO SUBROUTINE FTMXL.
C      COMPUTE IMPROVED ESTIMATES OF ARPS, PMAC, PMAS AND WNV
C      USING SUBROUTINE FTMXL
C
       DATA IND/0,1,1,0,75,4,1,3/
       IND(1)=NN
       PMAS=THETA(ITHETA)
       ARPS=PHI(IPHI)
C
C      IF PHI=1 SET PHI=0.99 SINCE PHI=1 IS NOT DESIRABLE
C
       IF(PMAS.EQ.1.) PMAS=0.99
       WRITE(6,180) ARPS,PMAS
 180   FORMAT(///,2X,'THE SUM OF SQUARES OF RESIDUALS IS MINIMUM',
      $ ' FOR:',   //,4X,'ARPS= ',F15.6,/,4X,'PMAS= ',F15.6,/)
C
       CALL FTMXL(TEMP,IND,ARPS,PMAS,PMAC,WNV,GR,A,IER)
       WRITE(6,190) ARPS,PMAS,PMAC,WNV
 190   FORMAT(///,2X,'-----SUBROUIINE FTMXL-----',//,
      1  5X,'ESTIMATE OF AR PARAMETER (ARPS)........',F15.6,//,
      2  5X,'ESTIMATE OF MA PARAMFTER (PMAS)........',F15.6,//,
      3  5X,'OVERALL MA CONSTANT (PMAC).............',F15.6,//,
      4  5X,'WHITE NOISE VARIANCE (WNV).............',F15.6)
       PHI1=ARPS
       THETA1=PMAS
C
C      FIND AC AND PACV OF RESIDUALS THAT ARE STORED IN THE VECTOR
C      <A> AS OUTPUT FROM SUBROUTINE FTMXL
C
       CALL FTAUTO(A,NN,L,L,7,AMEAN,ACV(1),ACV(2),AC(2),
      $            PACV(2),WKAREA)
       AC(1)=1.
       PACV(1)=1.
C
C      WRITE OUT AC AND PACV .COMPUTE SUM OF AC**2
C
       SSQR=0.
       WRITE(6,200)
 200   FORMAT(1H1,///,15X,'RESIDUAL SERIES ',///)
       WRITE(6,135)
       DO 70 I=1,LP1
       IM1=I-1
       WRITE(6,140) IM1,AC(I),PACV(I)
       SSQR=SSQR+AC(I)**2
 70    CONTINUE
       SSQR=SSQR-1.
       WRITE(6,145) SSQR
C
       DO 15 I=1,NN
 15    EVRAIN(I)=VRAIN(I)
C
C      GENERATE RANDOM NUMBERS N(0,1) TO BE USED FOR THE
C      FORECASTING
```

```
C
      DSEED=123457.DO
      CALL GGNML(DSEED,NN,Z)
C
C
      DO 20 I=1,NG
      I1=ISG(I)
      I2=IEG(I)
      K=I2-I1+1
      IF(K.GT.1) GO TO 51
      EVRAIN(I1)=PHI1*VRAIN(I1-1)-THETA1*Z(I1-1)
      GO TO 20
 51   EVRAIN(I1)=PHI1*VRAIN(I1-1)-THETA1*Z(I1-1)
      DO 31 L=2,K
 31   EVRAIN(I1+L-1)=PHI1*EVRAIN(I1+L-2)
 20   CONTINUE
C
C     APPLY THE INVERSE TRANSFORMATIONS ON THE SERIES.
C
      PP=1/P
      DO 61 I=1,N
      K1=(I-1)*12+1
      K2=I*12
      DO 71 L=K1,K2
      J=L-(I-1)*12
      ERAIN(I,J)=(EVRAIN(L)*STD(J)+MEAN(J))**PP
 71   CONTINUE
 61   CONTINUE
C
C
      RETURN
      END
//GO.SYSIN DD *
 ***** STATION 6038 - UNIVARIATE MODEL *****
   0.    .5
  55   25
   1    4
  27    1
   5    3
  11    5
   3    2
  63    3
   1    3
  10    4
   2    1
  19    2
  83    1
  11    1
  14    2
  36    2
  31    1
  33    7
  49    2
  19    7
  21    2
  39    1
  11    2
   2    1
  25    1
   2    2
  30    2
//GO.FT10F001 DD DSN=UF.B0063401.S7.C60381,DISP=(OLD,KEEP)
```

```
C
C-----------------------------------------------------------------
C
C                        PROGRAM RAEMV-B
C
C           RECURSIVE ALGORITHM FOR THE ESTIMATION OF
C           MISSING VALUES - BIVARIATE MODEL
C
C-----------------------------------------------------------------
C
      DIMENSION RAIN1(60,12),VR1(800),RAIN2(60,12),VR2(800),
     1   E1R1(60,12),V1R1(800),E1R2(60,12),V1R2(800),A1(2,2),
     2   B1(2,2),M01(2,2),M11(2,2)
      DIMENSION E2R1(60,12),V2R1(800),E2R2(60,12),V2R2(800),
     1   A2(2,2),B2(2,2),M02(2,2),M12(2,2)
      DIMENSION E3R1(60,12),V3R1(800),E3R2(60,12),V3R2(800),
     1   A3(2,2),B3(2,2),M03(2,2),M13(2,2)
      DIMENSION E4R1(60,12),V4R1(800),E4R2(60,12),V4R2(800),
     1   A4(2,2),B4(2,2),M04(2,2),M14(2,2)
      DIMENSION E5R1(60,12),V5R1(800),E5R2(60,12),V5R2(800),
     1   A5(2,2),B5(2,2),M05(2,2),M15(2,2)
      DIMENSION LI(200),LG(200),ISI(200),IEI(200)
      COMMON/A/ ID1,ID2,NYEAR(60)
      COMMON/B/ N,C,P
      COMMON/C/ NG,ISG(200),IEG(200)
C
C     READ INPUT PARAMETERS
C     HEADER..TITLE
C     N.......NUMBER OF YEARS
C     NG......NUMBER OF GAPS
C     LI......LENGTH OF INTEREVENT
C     LG......LENGTH OF GAP
C     C,P.....PARAMETERS OF THE TRANSFORMATION
C             TRANSFORMED SERIES   Y=(X+C)**P
C
      READ(5,10) HEADER
 10   FORMAT(20A4)
C
      READ(5,20) C,P
 20   FORMAT(2F5.2)
C
      READ(5,30) N,NG
 30   FORMAT(2(I4/))
C
      DO 5 I=1,NG
 5    READ(5,40) LI(I),LG(I)
 40   FORMAT(2I4)
C
      READ(10,100) (ID1,(NYEAR(I),(RAIN1(I,J),J=1,12)),I=1,N)
      READ(11,100) (ID2,(NYEAR(I),(RAIN2(I,J),J=1,12)),I=1,N)
 100  FORMAT(A4,I3,1X,12F6.2)
C
C     FROM THE INPUT VARIABLES LI AND LG TWO ARRAYS OF
C     LENGTH NG ARE COMPUTED. THEN THE STARTING POINT
C     OF THE KTH GAP IS ISG(K) AND THE ENDING POINT IS
C     IEG(K).
C
      ISI(1)=1
      IEI(1)=LI(1)
      ISG(1)=IEI(1)+1
      IEG(1)=ISG(1)+LG(1)-1
      DO 11 I=2,NG
      ISI(I)=IEG(I-1)+1
      IEI(I)=ISI(I)+LI(I)-1
      ISG(I)=IEI(I)+1
      IEG(I)=ISG(I)+LG(I)-1
 11   CONTINUE
C
      WRITE(6,15) HEADER
 15   FORMAT(20A4,///)
C
C     PRINT THE POSITIONS OF THE GAPS FOR A CHECK
C
      WRITE(6,101) (I,ISG(I),IEG(I),I=1,NG)
 101  FORMAT(3I6)
C
C     INITIALIZE THE MISSING VALUES OF THE INCOMPLETE
C     SERIES
```

```
C
      DO 60 I=1,N
      DO 60 J=1,12
      IF(RAIN1(I,J).EQ.-1) RAIN1(I,J)=0.
 60   CONTINUE
C
C     PRINT OUT THE SERIES WHICH IS TO BE ESTIMATED
C
      WRITE(6,66)
      WRITE(6,102) (ID1,(NYEAR(I),(RAIN1(I,J),J=1,12)),I=1,N)
 102  FORMAT(1X,A4,I3,1X,12F6.2)
 66   FORMAT(1H1)
C
C     SUBROUTINE BIVAR IS CALLED TO FIT A BIVARIATE AR(1)
C     MODEL TO THE TWO INPUT SERIES. IT ESTIMATES ALSO
C     THE MISSING VALUES OF THE ONE SERIES AND SAVES IT
C     TO BE INPUT TO THE NEXT CALL
C
      CALL BIVAR(RAIN1,VR1,RAIN2,VR2,E1R1,V1R1,E1R2,V1R2,
     .           A1,B1,MO1,M11)
      CALL BIVAR(E1R1,V1R1,E1R2,V1R2,E2R1,V2R1,E2R2,V2R2,
     .           A2,B2,MO2,M12)
      CALL BIVAR(E2R1,V2R1,E2R2,V2R2,E3R1,V3R1,E3R2,V3R2,
     .           A3,B3,MO3,M13)
      CALL BIVAR(E3R1,V3R1,E3R2,V3R2,E4R1,V4R1,E4R2,V4R2,
     .           A4,B4,MO4,M14)
      CALL BIVAR(E4R1,V4R1,E4R2,V4R2,E5R1,V5R1,E5R2,V5R2,
     .           A5,B5,MO5,M15)
C
      STOP
      END
C
C------------------- SUBROUTINE BIVAR ---------------------------
C
C     SUBROUTINE BIVAR FITS A BIVARIATE AR(1) MODEL TO
C     THE TWO INPUT SERIES EACH TIME IS CALLED. IT
C     ESTIMATES ALSO THE MISSING VALUES OF THE ONE
C     SERIES AND THE ESTIMATED SERIES IS SAVED TO
C     BE INPUT TO THE NEXT CALL
C
C
      SUBROUTINE BIVAR(RAIN1,VR1,RAIN2,VR2,ERAIN1,EVR1,ERAIN2,
     .                 EVR2,A,B,MO,M1)
C
      DIMENSION WKAREA(200)
      DIMENSION RAIN1(60,12),VR1(800),RAIN2(60,12),VR2(800)
      DIMENSION ERAIN1(60,12),EVR1(800),ERAIN2(60,12),EVR2(800)
      DIMENSION XM1(12),XM2(12),STD1(12),STD2(12)
      DIMENSION A(2,2),B(2,2),C(2,2),D(2,2)
      REAL MO(2,2),M1(2,2),MOINV(2,2)
      COMMON/A/ ID1,ID2,NYEAR(60)
      COMMON/B/ N,C,P
      COMMON/C/ NG,ISG(200),IEG(200)
      COMMON/D/ XM1,XM2,STD1,STD2,X1,X2,ST1,ST2
C
C     CALL SUBROUTINE STAT TO NORMALIZE AND STANDARDIZE
C     THE SERIES AND COMPUTE THE STATISTICS
C
      CALL STAT(RAIN1,XM1,STD1,VR1,X1,ST1)
      CALL STAT(RAIN2,XM2,STD2,VR2,X2,ST2)
C
C     CALL THE IMSL SUBROUTINE FTCRXY TO COMPUTE AUTO-
C     AND CROSS-COVARIANCES OF THE SERIES
C
      CALL FTCRXY(VR1,VR2,N,X1,X2,0,N,C120,IER)
      CALL FTCRXY(VR1,VR1,N,X1,X1,-1,N,C111,IER)
      CALL FTCRXY(VR2,VR2,N,X2,X2,-1,N,C221,IER)
      CALL FTCRXY(VR1,VR2,N,X1,X2,-1,N,C121,IER)
      CALL FTCRXY(VR2,VR1,N,X2,X1,-1,N,C211,IER)
C
      MO(1,1)=1.
      MO(2,2)=1.
      MO(1,2)=C120/(ST1*ST2)
      MO(2,1)=MO(1,2)
C
      M1(1,1)=C111/(ST1*ST1)
      M1(2,2)=C221/(ST2*ST2)
      M1(1,2)=C121/(ST1*ST2)
      M1(2,1)=C211/(ST1*ST2)
C
      WRITE(6,66)
 66   FORMAT(1H1)
C
C     PRINT OUT THE CORRELATION MATRICES MO AND M1
```

```
C
      WRITE(6,110) ((MO(I,J),J=1,2),I=1,2)
      WRITE(6,111) ((M1(I,J),J=1,2),I=1,2)
 110  FORMAT(5X,'CORRELATION MATRIX:  MO',//,((5X,2F10.3)/))
 111  FORMAT(5X,'CORRELATION MATRIX:  M1',//,((5X,2F10.3)/))
C
C     CALCULATE THE PARAMETER MATRICES A AND B
C
      CALL LINV2F(MO,2,2,MOINV,0,WKAREA,IER)
      CALL VMULFF(M1,MOINV,2,2,2,2,2,A,2,IER)
      CALL VMULFP(A,M1,2,2,2,2,2,D,2,IER)
C
      DO 10 I=1,2
      DO 10 J=1,2
 10   C(I,J)=MO(I,J)-D(I,J)
C
      B(1,1)=C(1,1)**0.5
      B(2,1)=C(1,2)/B(1,1)
      B(2,2)=(C(2,2)-C(1,2)**2/C(1,1))**0.5
      B(1,2)=0.
C
C     PRINT OUT THE MATRICES A AND B
C
      WRITE(6,140) ((A(I,J),J=1,2),I=1,2)
      WRITE(6,141) ((B(I,J),J=1,2),I=1,2)
 140  FORMAT(5X,'COEFFICIENT MATRIX:  A',//,((5X,2F10.3)/))
 141  FORMAT(5X,'COEFFICIENT MATRIX:  B',//,((5X,2F10.3)/))
C
      NN=N*12
      DO 15 I=1,NN
      EVR2(I)=VR2(I)
 15   EVR1(I)=VR1(I)
C
C     ESTIMATE THE GAPS OF THE INCOMPLETE SERIES
C
      DO 20 I=1,NG
      I1=ISG(I)
      I2=IEG(I)
      K=I2-I1+1
      DO 40 L=1,K
 40   EVR1(I1+L-1)=A(2,1)*EVR1(I1+L-2)+A(2,2)*EVR2(I1+L-2)
 20   CONTINUE
C
C     PERFORM INVERSE TRANSFORMATIONS
C
      PP=1/P
      DO 50 I=1,N
      DO 50 J=1,12
      L=J+(I-1)*12
      ERAIN1(I,J)=(EVR1(L)*STD1(J)+XM1(J))**PP
      ERAIN2(I,J)=(EVR2(L)*STD2(J)+XM2(J))**PP
 50   CONTINUE
C
C     PRINT OUT THE ESTIMATED SERIES
C
      WRITE(6,66)
      WRITE(6,101) (ID1,(NYEAR(I),(ERAIN1(I,J),J=1,12)),I=1,N)
 101  FORMAT(1X,A4,I3,1X,12F6.2)
C
      RETURN
      END
C
C---------------- SUBROUTINE STAT ----------------------------
C
C     SUBROUTINE STAT TRANSFORMS THE ORIGINAL SERIES TO
C     NORMAL AND STATIONARY AND COMPUTES THE STATISTICS
C     OF THE TRANSFORMED SERIES.
C
      SUBROUTINE STAT(RAIN,XM,STD,VRAIN,X,ST)
      DIMENSION RAIN(60,12),VRAIN(800),XM(12),STD(12)
      COMMON/A/ ID1,ID2,NYEAR(60)
      COMMON/B/ N,C,P
      COMMON/C/ NG,ISG(200),IEG(200)
C
      DO 10 I=1,N
      DO 10 J=1,12
      RAIN(I,J)=(RAIN(I,J)+C)**P
 10   CONTINUE
C
C     COMPUTE MONTHLY MEANS AND STANDARD DEVIATIONS OF
C     THE NORMALIZED SERIES
C
      DO 20 J=1,12
      XM(J)=0.
      STD(J)=0.
      DO 25 I=1,N
      XM(J)=XM(J)+RAIN(I,J)/FLOAT(N)
      STD(J)=STD(J)+RAIN(I,J)**2
 25   CONTINUE
 20   CONTINUE
C
      DO 30 I=1,12
      STD(I)=((STD(I)-XM(I)**2*FLOAT(N))/(FLOAT(N)-1.))**0.5
 30   CONTINUE
C
C     NOW,STANDARDIZE THE SERIES
C
      DO 40 I=1,N
      DO 40 J=1,12
      RAIN(I,J)=(RAIN(I,J)-XM(J))/STD(J)
 40   CONTINUE
C
C     COMPUTE MEAN AND STD OF THE WHOLE SERIES
C
      NN=N*12
      IC=0
      DO 50 I=1,N
      DO 50 J=1,12
      IC=IC+1
      VRAIN(IC)=RAIN(I,J)
 50   CONTINUE
C
      X=0.
      ST=0.
      DO 60 I=1,NN
      X=X+VRAIN(I)
      ST=ST+VRAIN(I)**2
 60   CONTINUE
      X=X/FLOAT(NN)
      ST=((ST-X**2*FLOAT(NN))/(FLOAT(NN)-1.))**0.5
C
      RETURN
      END
//GO.SYSIN DD *
 ***** STATION 6038 - BIVARIATE MODEL *****
   .0   .5
  55  25
   1   4
  27   1
   5   3
  11   5
   3   2
  63   3
   1   3
  10   4
   2   1
  19   2
  83   1
  11   1
  14   2
  36   2
  31   1
  33   7
  49   2
  19   7
  21   2
  39   1
  11   2
   2   1
  25   1
   2   2
  30   2
//GO.FT10F001 DD DSN=UF.B0063401.S7.C603820,DISP=(OLD,KEEP)
//GO.FT11F001 DD DSN=UF.B0063401.S7.B6093,DISP=(OLD,KEEP)
/*EOJ
/*EOJ
```

REFERENCES

Afifi, A.A., and Elashoff, R.M., 1966, "Missing observations in multivariate statistics I: Review of the literature," J. Am. Stat. Assoc., 61:595-604.

Anderson, D.G., 1979. "Satelite versus conventional methods in hydrology" in Satellite Hydrology, American Water Resources Association, Minneapolis.

Anderson, T. W., 1957, "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing," J. Am. Stat. Assoc., 52:200-203.

Ansley, G.F., Spivey, W.A., and Worblski, W.J., 1977, "A class of transformations for Box-Jenkins's seasonal modelling," Appl. Stat., 26:173-178.

Beale, E.M.L., and Little, R.J.M., 1975, "Missing values in multivariate analysis," J. R. Stat. Soc., B37:129-145.

Beard, L.R., 1973, "Hydrologic data fill-in and network design," in Design of Water Resources Projects with Inadequate Data, Proc. of the Madrid Symposium, June, 1973.

Bendat, J.S., and Piersol, A.G., 1967, Measurement and Analysis of Random Data, John Wiley & Sons, New York, 3rd. printing.

Bloomfield, P., 1970, "Spectral analysis with randomly missing observations," J. R. Stat. Soc., B32:369-380.

Box, G.E. P., and Cox, D.R., 1964, "An analysis of transformation (with discussion)," J. R. Stat. Soc., B26:211-252.

Box G.E.P., and Jenkins, G.M., 1973, "Some comments on a paper by Chatfield and Prothero and on a review by Kendall (with discussion)," J. R. Stat. Soc., A135:337-345.

Box, G.E.P., and Jenkins, G.M., 1976, Time Series Analysis Forecasting and Control, Holden-Day, San Francisco, Revised ed.

Box, G.E.P., and Pierce, D.A., 1970, "Distribution of residual autocorrelations in autoregressive-integrated moving average time series models," J. Am. Stat. Assoc., 64:1509-1526.

Brubacher, S.R., and Tunnicliffe Wilson, G., 1976, "Interpolating time series with application to the estimation of holiday effects on electricity demand," Appl. Stat., 25:107-116.

Buck, S.F., 1960, "A method of estimation of missing values in multivariate data suitable for use with an electronic computer," J. R. Stat. Soc., B22:302-307.

Chatfield, C., 1980, The Analysis of Time Series: An Introduction, Chapman and Hall, London, 2nd ed.

Chatfield, C., and Prothero, D.L., 1973a, "Box-Jenkins seasonal forecasting: Problems in a case study (with discussion)," J. R. Stat. Soc., A136:295-336.

Chatfield, C., and Prothero, D.L., 1973b, "Reply by Dr. Chatfield and Dr. Prothero on the paper 'Some comments on a paper by Chatfield and Prothero and on a review by Kendall' by Box, G.E.P., and Jenkins, G.M.," J. R. Stat. Soc., A136:347-352.

Crosby, D.S., and Maddoc, T., 1970, "Estimating coefficients of a flow generator for monotone samples of data," Water Resour. Res., 6(4):1079-1086.

Damsleth, E., 1980, "Interpolating missing values in a time series," Scand. J. Stat., 7:33-39.

Dean, J.D., and Snyder, W.M., 1977, "Temporally and areally distributed rainfall," J. of the Irrigation and Drainage Div., ASCE, 103(IR2):221-229.

Delleur, J.W., and Kavvas, M.L., 1978, "Stochastic models for monthly rainfall forecasting and synthetic generation," J. Appl. Meteor., 17(10):1528-1536.

Draper, N.R., and Cox, D.R., 1969, "On distributions and their transformation to normality," J. R. Stat. Soc., B31:472-476.

Draper, N.R., and Smith, H., 1966, Applied Regression Analysis, John Wiley & Sons, New York.

Durbin, J., 1960, "The fitting of time series models," Rev. Int. Inst. Stat., 28:233.

Fiering, M.B., 1964, "Multivariate technique for synthetic hydrology," J. Hydraul. Div., ASCE, 90(HY5):43-60.

Fiering, M.B., 1968, "Schemes for handling inconsistent matrices," Water Resour. Res., 4(2):291-297.

Fiering, M.B., and Jackson, B.B., 1971, "Synthetic Hydrology," Monograph No. 1, American Geophysical Union, Washington, D.C.

Finzi, G., Todini, E., and Wallis, J.R., 1977, "SPUMA: Simulation package using Matalas algorithm," in Mathematical Models for Surface Water Hydrology, Ed. by Ciriani, T.A., Maione, U., and Wallis, J.R., John Wiley & Sons, London.

Finzi, G., Todini, E., and Wallis, J.R., 1975, "Comment upon multivariate synthetic hydrology," Water Resour. Res., 11(6):844-850.

Gantmacher, F.R., 1977, The Theory of Matrices, Vol. I, Chelsea Publ. Company, New York.

Granger, C.W.J., and Morris, M.J., 1976, "Time series modelling and interpretation," J. R. Stat. Soc., A139:246-257.

Haan, C.T., 1977, Statistical Methods in Hydrology, Iowa State Univ. Press, Ames.

Hamrick, R.L., 1972, "South Florida's 'unmanaged' resource," In Depth Report, Central and South Florida Flood Control District 1:1-12.

Hannan, E.J., 1960, Time Series Analysis, Chapman and Hall, London.

Hashino, M., 1977, "A similar storm method on filling data voids," in Modeling Hydrologic Processes, Ed. by Morel-Seytoux, H., Salas, J.D., Sanders, T.G., and Smith, R.E., Water Resour. Res. Publications, Fort Collins, Colorado.

Hinkley, D., 1977, "On quick choice of power transformation," Appl. Stat., 26(1):67-70.

IMSL LIB-0007, 1979, Reference Manual, Edition 7, Revised.

Jenkins, G.M., and Watts, D.G., 1969, Spectral Analysis and its Applications, Holden-Day, San Francisco, 2nd printing.

John, J.A., and Draper, N.R., 1980, "An alternative family of transformations," Appl. Stat., 29(2):190-197.

Jones, R.H., 1962, "Spectral analysis with regularly missed observations," Ann. Math. Stat., 32:455-61.

Kahan, J.P., 1974, "A method for maintaining cross and serial correlations and the coefficient of skewness under generation in a linear bivariate regression model," Water Resour. Res., 10(6):1245-1248.

Kavvas, M., and Delleur, J., 1975, "Removal of Periodicities by differencing and monthly mean substraction," J. Hydrol., 26:335-353.

Kottegoda, N.T., and Elgy, J., 1977, "Infilling missing flow data," in Modeling Hydrologic Processes, Ed. by Morel-Seytoux, H., Salas, J.D., Sanders, T.G., and Smith, R.E., Water Resour. Res. Publications, Fort Collins, Colorado.

Linsley, R.K., Jr., Kohler, M.A., and Paulhus, J.L.H., 1978, Hydrology for Engineers, McGraw-Hill Book Co., New York, 2nd ed.

Marshall, R.J., 1980, "Autocorrelation estimation of time series with randomly missing observations," Biometrika, 67(3):567-570.

Matalas, N.C., 1967, "Mathematical assessment of synthetic hydrology," Water Resour. Res., 3(4):937-945.

Matalas, N.C., 1978, "Generation of multivariate synthetic flows," in Mathematical Models for Surface Water Hydrology, Ed. by Ciriani, T.A., Maione, U., and Wallis, J.R., John Wiley & Sons, London.

Mejia, J.M., Rodriguez-Iturbe, I., and Cordova, J.R., 1974, "Multivate generation of mixtures of normal and log-normal variables," Water Resour. Res., 10(4):691-693.

Moran, P.A.P., 1970, "Simulation and evaluation of complex water systems operations," Water Resour. Res., 6(6):1737-1742.

Neave, H.R., 1970, "Spectral analysis with initially scarce data," Biometrika, 57:111-122.

O'Connell, P.E., 1973, "Multivariate synthetic hydrology: a correction," J. Hydr. Div., ASCE, Tech. notes, 9(HY12): 2391-2396.

O'Connell, P.E., 1974, "Stochastic modelling of long-term persistence in streamflow sequences,", Ph.D. thesis, University of London, London, England.

Orchard, T., and Woodbury, M.A., 1972, "A missing information principle: Theory and applications," in Proc. 6th Berkeley Symp. Math. Statist. Prob., Vol I:697-715.

Ozaki, T., 1977, "On the order determination of ARIMA models," Appl. Stat., 26:290-301.

Parzen, E., 1963, "On spectral analysis with missing observations and amplitude modulation," Sankhya, A25:383-392.

Paulhus, J.L.H., and Kohler, M.A., 1952, "Interpolation of missing precipitation records," Mon. Weather Review, 80:129-133.

Pegram, G.G.S., and James, W., 1972, "Multilag multivariate autoregressive model for the generation of operational hydrology," Water Resour. Res., 8(4):1074-1076.

Roesner, L.A., and Yevjevich, V., 1966, "Mathematical models for time series of monthly precipitation and monthly runoff," Hydrology paper No. 15, Colorado State University, Fort Collins, Colorado.

Salas, J.D., Delleur, J.W., Yevjevich, V., and Lane, W.L., 1980, Applied Modeling of Hydrologic Time Series, Water Resour. Res. Publ., Fort Collins, Colorado.

Salas, J.D., and Pegram, G.G.S., 1977, "A seasonal multivariate multilag autoregressive model in hydrology," in Modeling hydrologic processes, Ed. by Morel-Seytoux, H., Salas, J.D., Sanders, T.G., and Smith, R.E., Water Resour. Publications, Fort Collins, Colorado.

Slack, J.R., 1973, "I would if I could (self-denial by conditional models)," Water Resour. Res., 9(1):247-249.

Scheinok, P.A., 1965, "Spectral analysis with randomly missed observations: The binomial case," Ann. Math. Stat., 36:971-977.

Schlesselman, J., 1971, "Power families: A note on the Box and Cox transformation," J. R. Stat. Soc., B33:307-311.

Shearman, R.J., and Salter, P.M., 1975, "An objective rainfall interpolation and mapping technique," Hydrological Sciences Bulletin, 20(3):353-363.

Stidd, C.K., 1953, "Cube-root-normal precipitation distributions," Trans. Amer. Geophys. Union, 34:31-35.

Stidd, C.J., 1968, "A three parameter distribution for precipitation data with a straight-line plotting method," Proc. 1st Statist. Meteorol. Conf., Amer. Meteor. Soc., Hartford, Connecticut, pp. 158-162.

Stidd, C.K., 1970, "The nth root normal distribution of precipitation," Water Resour. Res., 6(4):1095-1103.

Tukey, J.W., 1957, "On the comparative anatomy of transformation," Ann. of Math. Stat., 28:602-632.

Valencia, D.R., and Schaake, J.C., Jr., 1973, "Disaggregation processes in stochastic hydrology," Water Resour. Res., 9(3):580-585.

Wastler, T.A., 1969, Spectral Analysis, Applications in Water Pollution Control, U.S. Dept of the Interior, Federal Water Pol. Control Adm., Washington, D.C.

Wei, T.C., and McGuiness, J.L., 1973, "Reciprocal distance squared method, a computer technique for estimating areal precipitation," ARS NC-8, U.S., Dept. of Agriculture, Washington, D.C.

Wilson, G.T., 1973, "Contribution to discussion of 'Box-Jenkins seasonal forecasting: Problems in a case study," by C. Chatfield and D.L. Prothero, J. R. Stat. Soc., A136:315-319.

Wold, H.O., 1938, A Study of the Analysis of Stationary Time Series, Almquist and Wicksell, Uppsala, 2nd ed., 1954.

Yevjevich, V.M., 1972, "Structural analysis of hydrologic time series," Hydrol. paper No. 56, Colorado State University, Fort Collins, Colorado.

Young, G.K., 1968, "Discussion of 'Mathematical assessment of synthetic hydrology' by N. G. Matalas," Water Resour. Res., 4(3):681-682.

Young, G.K., and Pisano, W.C., 1968, "Operational hydrology using residuals," J. Hydr. Div., ASCE, 94(HY4):909-923.

Yule, G.U., 1927, "On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers," in Statistical Papers of George Undy Yule, selected by Stuart, A., and Kendall, M., Hafner Publ. Co., New York, 1971.