WATER RESOURCES research center

Publication No. 85

DISCRETE-TIME POINT PROCESS MODELS FOR
DAILY RAINFALL

by

Efi Foufoula-Georgiou

University of Florida

Gainesville

UNIVERSITY OF FLORIDA

DISCRETE-TIME POINT PROCESS MODELS FOR DAILY RAINFALL

By

EFI FOUFOULA-GEORGIOU

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1985

ACKNOWLEDGMENTS

Wayne Huber (Department of Environmental Engineering Sciences, University of Florida) was my dissertation advisor. His guidance and moral support during my four years of graduate studies have contributed much to my overall research performance. I am especially indebted to him for having granted me complete freedom to pursue my own interests which, in April, 1984, carried me all the way to the University of Washington.

Dennis Lettenmaier (Department of Civil Engineering, University of Washington) created the opportunity for me to be a visiting graduate student for one year at the University of Washington. Working with him has been most challenging, inspiring, and certainly rewarding. His stamina, enthusiasm, and clear-cut outlook on things have deeply influenced my approach not only to research problems, but also to other aspects of life.

While at the University of Washington, I had the good fortune to meet and benefit from the expertise and warm friendship of Stephen Burges (Department of Civil Engineering). Long discussions on point processes with Peter Guttorp (Department of Statistics) have always managed to return my enthusiasm when I was discouraged. The sincere help of Antonio Possolo (Department of Statistics) during the last months of this work is also gratefully acknowledged.

Jim Smith (Potomac River Basin, Maryland) has been of invaluable help. He willingly shared his experience on point processes and daily rainfall modeling through correspondence and telephone calls. This help is gratefully acknowledged and has had a considerable impact on this work.

My good friend and husband Tryphon Georgiou (Department of Electrical Engineering, Florida Atlantic University) has triggered my early interest in stochastic processes. I have benefited the greatest from him through many "back-of-the-envelope" technical discussions. He also deserves special thanks for so unselfishly encouraging me to pursue my research interests 4,000 miles away from home.

I wish to thank the members of my supervisory committee at the University of Florida, James Heaney and Warren Viessman (Department of Environmental Engineering), Barry Benedict (Department of Civil Engineering), and Mark Yang (Department of Statistics). Their guidance throughout my graduate work has had a positive impact on this work.

Daily rainfall data were kindly supplied to me by Malcolm Leytham (University of Washington), Bob Dickinson (University of Florida), and Shaw-Pin Miaou (University of Texas). Without the computer resources provided to me by the College of Engineering, University of Washington, this work would not have been possible.

The efficient and accurate typing and the competent organizing abilities of Stephanie Boyd (University of Washington) made the management of this manuscript an easy task. The help of Lori Nagle, Cristal Weber, and Shirley Kook, in the final production is also greatefully acknowledged.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

DISCRETE-TIME POINT PROCESS MODELS FOR
DAILY RAINFALL

By

Efi Foufoula-Georgiou

May, 1985

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

Several authors have recently had apparent success in applying continuous-time point process models to daily rainfall observation sequences. In this work it is shown that major problems arise when the observation sequence represents cumulative rainfall amounts over a period (e.g., one day) which is on the order of the process interarrival time. In particular, the use of continuous-time point process models for daily rainfall occurrences may result in incorrect inferences about the underlying rainfall generating mechanisms. This was confirmed by the statistical analysis of six daily rainfall records from diverse climatologic regimes throughout the U.S. (Snoqualmie Falls, Washington; Roosevelt, Arizona; Austin, Texas; Miami, Florida; Philadelphia, Pennsylvania; and Denver, Colorado). In addition, the use of continuous-time point process models for generation of daily rainfall sequences leads to serious upward biases

in the event-interarrival times and in dependence structures that may be much different than those of the apparent rainfall generating process.

In this work, a discrete-time point process model has been developed and its structural properties derived. In the proposed process the sequence of times between events is formed through sampling from two geometric distributions, according to transition probabilities specified by a Markov chain. This process belongs to the class of semi-Markov (or Markov renewal) processes and is a non-renewal, clustered (relative to the Bernoulli) process which reduces to a renewal process with a mixture distribution for the interarrival times as a special case. Several methods for fitting the proposed model have been studied and an approximate maximum likelihood estimation method has been found to perform adequately, especially for daily rainfall structures with not very significant autocorrelation structures.

The semi-Markov model was fitted to the daily rainfall occurrences of the Snoqualmie Falls and Roosevelt stations, both on a monthly and seasonal basis. The fit of the model was assessed by the preservation of selected statistical properties of the series which were not used directly in the estimation. It was shown that the fitted model gave a theoretical spectrum of counts surprisingly close to the empirical one. Also, the model was quite successful in preserving the distributional properties of the cumulative rainfall amounts over longer periods of time, particularly for the Snoqualmie Falls station.

CHAPTER 1
INTRODUCTION

> It was six men of Indostan
> To learning much inclined,
> Who went to see the Elephant
> (Though all of them were blind),
> That each by observation
> Might satisfy his mind.

Rainfall is the result of a complex atmospheric process evolving continuously over space and time. At any time, rainfall fields are characterized by their areal extent and their spatially variable intensity. Austin and Houze (1972) and Hobbs and Locatelli (1978) have classified rainfall fields according to their areal extent and lifetimes as synoptic, large mesoscale, small mesoscale and rain cells. Synoptic rainfall fields cover areas on the order of 10^4 km^2 and have a lifetime of one to several days; large mesoscale fields cover areas of 10^3 - 10^4 km^2 and have a lifetime of several hours; small mesoscale fields have areal extent of 10 - 10^2 km^2 and a lifetime of approximately one hour; rain cells have areal extent of 1 - 10 km^2 and lifetimes of a few minutes to 1/2 hour. The system of rainfall fields is hierarchical in the sense that larger rainfall fields usually contain one or more of the smaller ones. The continuous movement, build-up, and dissipation of rainfall fields determines the rainfall intensity variations in space and time.

Space-time modeling of an observed rainfall sequence at a point based on a mathematical description of the underlying atmospheric processes would be an extremely complicated, if not impossible, task. The need for mathematically tractable descriptions of rainfall for operational purposes, i.e., forecasting for day-to-day operation of hydrologic systems, has motivated treatment of rainfall as a stochastic process. Approaches to the space-time stochastic modeling of rainfall have recently been suggested by Waymire et al. (1984) and Kavvas and Herd (1984). In the present work, only the time variability of rainfall, i.e., the characterization of a precipitation observation sequence at a single station, is considered.

Point rainfall is the precipitation intercepting a small area such as the opening of a rain guage; it may be treated as a continuous-time intermittent process, say with intensity $\xi(t)$. Precipitation measurements are recorded for cumulative amounts over discrete time intervals such as minutes, hours, or days. Let $\{Y_i\}_T$, i = 1,2,3,... denote the discrete sequence of rainfall observations over an arbitrary time interval T. The continuous process $\xi(t)$ is related to the discrete process $\{Y_i\}_T$ by

$$Y_{i(T)} = \int_{t_{i-1}}^{t_i} \xi(\tau)\, d \qquad (1.1)$$

where $t_i - t_{i-1} = T$ is the time of measurement. Figure 1.1 illustrates the relationship between $\xi(t)$ and $Y_{i(T)}$: the continuous process $\xi(t)$ is integrated over, say, daily time intervals to give the sequence of daily data $\{Y_i\}_T$, where T = 1 day.

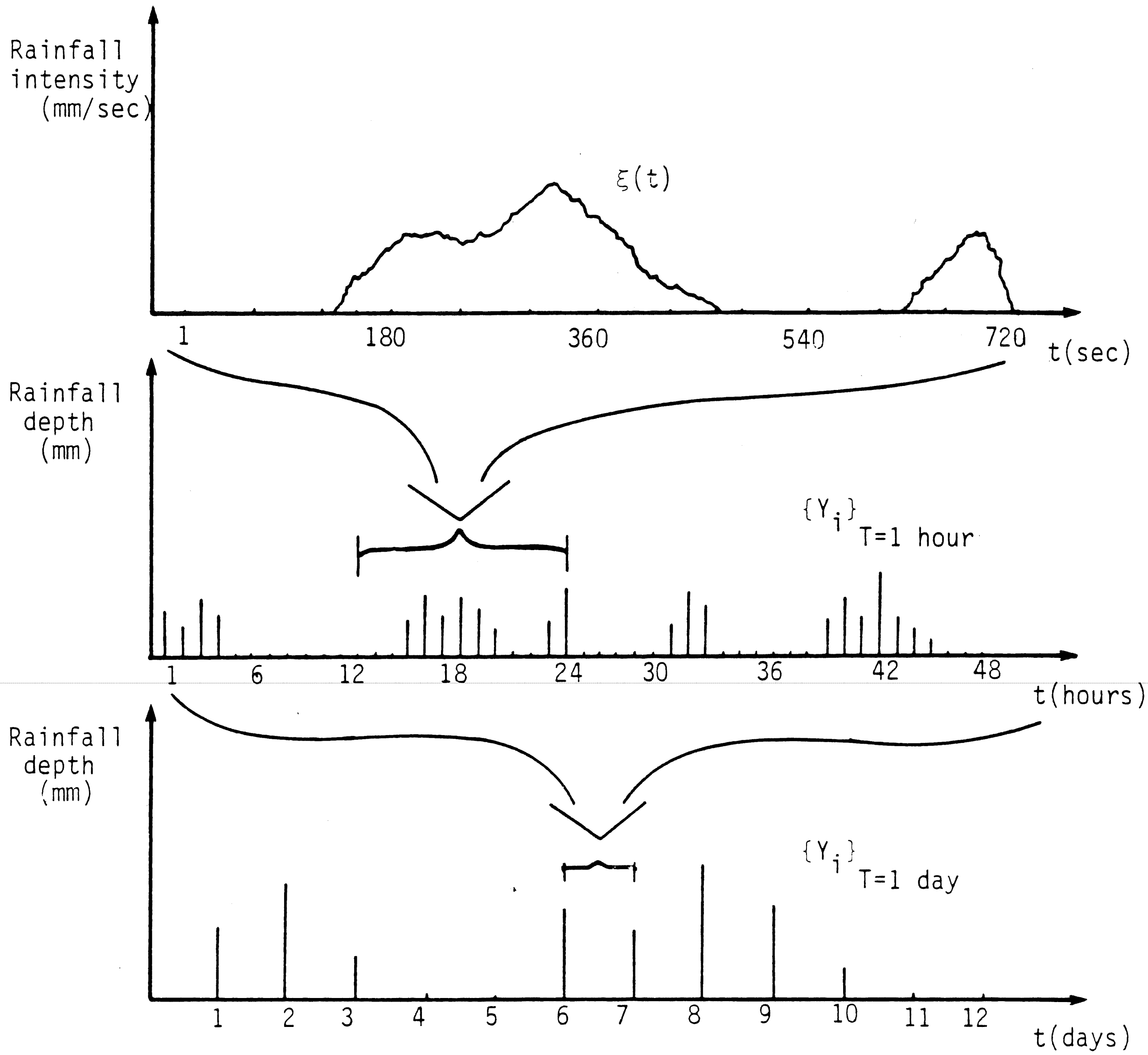


Figure 1.1 Continuous rainfall process $\xi(t)$ and discrete hourly and daily rainfall sequences $\{Y_i\}_T$.

Rodriguez-Iturbe et al. (1984) view $\xi(t)$ as "a generalized stochastic process" representing the instantaneous rate of rainfall. By postulating several continuous-time models for $\xi(t)$, they have derived distributional properties of the discrete accumulated amounts $\{Y_i\}_T$ for an arbitrary time scale T. This approach raises some fundamental questions, as Diggle (1984) has recently pointed out, specifically what is a suitable class of models for $\xi(t)$, and how can inferences about $\xi(t)$ be made, given data in the form of daily or hourly accumulated amounts? Rodriguez-Iturbe et al. (1984) assessed the validity of possible models for $\xi(t)$ by comparing parameters estimated from hourly and daily data with the theoretical parameters of $\{Y_i\}_T$, T = 1 hour and 1 day, derived for several candidate models for $\xi(t)$. This approach, however, is primarily of theoretical interest because it does not suggest a model for the observed discrete rainfall sequences $\{Y_i\}_T$ but rather for the unobserved continuous process $\xi(t)$, and the derived distributional properties of $\{Y_i\}_T$ do not lead to a parsimonious representation of the discrete process.

In this work a somewhat different approach is suggested. The appropriateness of several model structures for the discrete process $\{Y_i\}_T$ has been examined, with emphasis on a one day interval. The ultimate goal is to derive a realistic parameter-parsimonious model suitable for the analysis and synthesis of daily rainfall sequences.

Although daily rainfall observation sequences are only one possible discrete aggregation of the time-continuous rainfall process, the daily scale is of special interest for several reasons. Many water resource systems are operated on a daily time step. For example, operational decisions for small reservoirs for water supply and irrigation scheduling (e.g., Ramirez-Rodriguez and Bras, 1982) are often made on a daily time scale, and therefore an adequate mathematical description of the daily rainfall input is necessary. More generally, though, one day may be considered as the upper limit of event scale for precipitation; larger scale precipitation sequences no longer reflect individual precipitation events, and take on a fundamentally different statistical structure. Another reason for modeling daily rainfall is that most U.S. rainfall stations are cooperative, that is, the data are not collected by the National Weather Service. Most cooperative stations report daily precipitation totals. Manned or remote National Weather Service stations are much fewer in number; generally, it is these stations that collect hourly or shorter increment data. It should be emphasized that the model developed herein is not restricted to the daily time scale; it is expected that much of the work will also be applicable to smaller time scales such as hourly. Nonetheless, the emphasis in this work is on the daily time scale.

The stochastic structure of daily rainfall occurrences has been extensively studied over the past two decades. A classification of the modeling approaches and a literature review of the models proposed are given in Chapter 2. This work concentrates on only one approach, namely, the point process approach. A point process is defined as a sequence of events completely characterized by the location (in time or space) of the events. The daily rainfall occurrence process may be viewed as a point process in which an event takes place any time the cumulative rainfall amount over a one-day period exceeds a specified threshold value, as for example 0.01 inches. Under the above definition of event, and given that a day can be either dry (no rain) or wet (rain exceeding a threshold value), a point process model for daily rainfall occurrences will identify only the "state" of each day i.e., dry or wet. In that sense, the point process modeling approach for daily rainfall is conceptually equivalent to the discrete binary series approach in which a sequence of zeroes and ones (for no rain and rain) is formed and subsequently modeled. The important difference, however, between the two approaches is that the theory of point processes permits construction of models with much more flexible dependence structures, as compared with the structures that can be formulated under the theory of binary time series.

An example of a discrete binary series model is a Markov chain, which has been used by a number of authors to model daily rainfall occurrences (e.g., Chin, 1977). Newer developments in discrete binary models include the work of Chang et al. (1984) who introduced the discrete autoregressive moving average (DARMA) models for daily rainfall and applied them to the daily rainfall sequences of nine stations in Indiana. While newer approaches in discrete binary models, such as the DARMA models, may meet some of the inadequacies of Markov chains for daily rainfall occurrences, these models lead to complicated recursive formulas for the distributional properties of the interarrival times (e.g., Chang et al., 1982). It is the author's feeling that the point process modeling methodology provides much more elegant mathematical formulations of a stochastic process, and for this reason discrete binary models are not considered further.

Point process models for the areal distribution of rainfall were first introduced by LeCam (1961). Later Kavvas and Delleur (1975) applied the point process methodology to daily rainfall occurrences. The powerful theory of point processes was illustrated in the hydrological literature by Waymire and Gupta (1981a, b, and c) in a series of three papers. The suitability of the flexible point process model structures for small-time-increment rainfall sequences encouraged further study and Smith (1981) proposed a different model for daily rainfall occurrences. Details of these models will be given in Chapter 2. The important point to be made here is that previous studies have applied time-continuous point process models to the daily rainfall occurrences. However, the point process of daily rainfall occurrences is discrete in time. The discreteness stems from the definition of the event as a day with rainfall above a threshold value. It should be noted here that throughout the discussion that follows we have used the short term continuous (or discrete) point process instead of the accurate term continuous-time (or discrete-time) point process, primarily for convenience.

In this work it will be demonstrated that continuous point process models are not operationally useful for daily rainfall and that discrete point process models are needed instead. In addition, it will be shown that the theory of continuous point processes is not appropriate for modeling daily rainfall occurrences and that inferences made about the underlying rainfall generating mechanisms by comparing sample properties of daily rainfall occurrences to the independent Poisson process may be misleading. In view of the above, a discrete point process methodology will be suggested and a discrete point process model with demonstrated flexibility introduced and applied to representative daily rainfall occurrence structures. Methods for fitting this model will also be studied. Finally, the model will be coupled with a model for the non-zero daily rainfall amounts to give an operational, parsimonious model for analysis and synthesis of daily rainfall sequences. Further, it will be shown that such a model may be able to preserve the distributional properties of the cumulative rainfall amounts over periods of specified length, e.g., a week or month. This is an important property of a daily rainfall model especially when it is used for rainfall-runoff studies, where mass balance over long periods of time is desired.

In summary, this dissertation is structured as follows. In Chapter 2, a classification and brief review of available daily rainfall models are given. The inappropriateness of the continuous point process models for the discrete daily rainfall occurrences is demonstrated in Chapter 3. In Chapter 4, the statistical analysis of six daily rainfall records with respect to the rainfall occurrences and amounts is presented. In Chapter 5, a discrete point process model is developed and its statistical properties are derived. Methods for fitting the developed model are suggested and compared in Chapter 6. In Chapter 7, the discrete point process is fitted to the daily rainfall occurrences of two stations and coupled with a model for the non-zero daily rainfall amounts. The satisfactory performance of the model is assessed by checking the extent to which several rainfall statistics are preserved. The summary, conclusions, and recommendations for further research are given in Chapter 8.

CHAPTER 2
LITERATURE REVIEW

> The First approached the Elephant,
> And happening to fall
> Against his broad and sturdy side,
> At once began to bawl:
> "God bless! but the Elephant
> Is very like a wall!"

Due to the intermittency of small-time-increment (i.e., hourly or daily) rainfall processes, standard time-series analysis methods are not applicable. Instead, the most commonly used approach to modeling daily rainfall is to model the rainfall occurrences separately from the non-zero rainfall amounts, and then superimpose the two models. This chapter classifies the existing daily rainfall occurrence and amounts models and, under each category, a review of selected work is given. For supplementary review papers, the reader is referred to Roldan and Woolhiser (1982), Woolhiser and Roldan (1982), Court (1979), and Waymire and Gupta (1981a).

2.1 Models for the Daily Rainfall Occurrences

2.1.1 The "Wet-Dry Spell" Approach

In this approach any uninterrupted sequence of wet days (i.e., days with total rainfall above a specified threshold value) defines an event (see Fig. 2.1a). Such an occurrence model is completely specified by

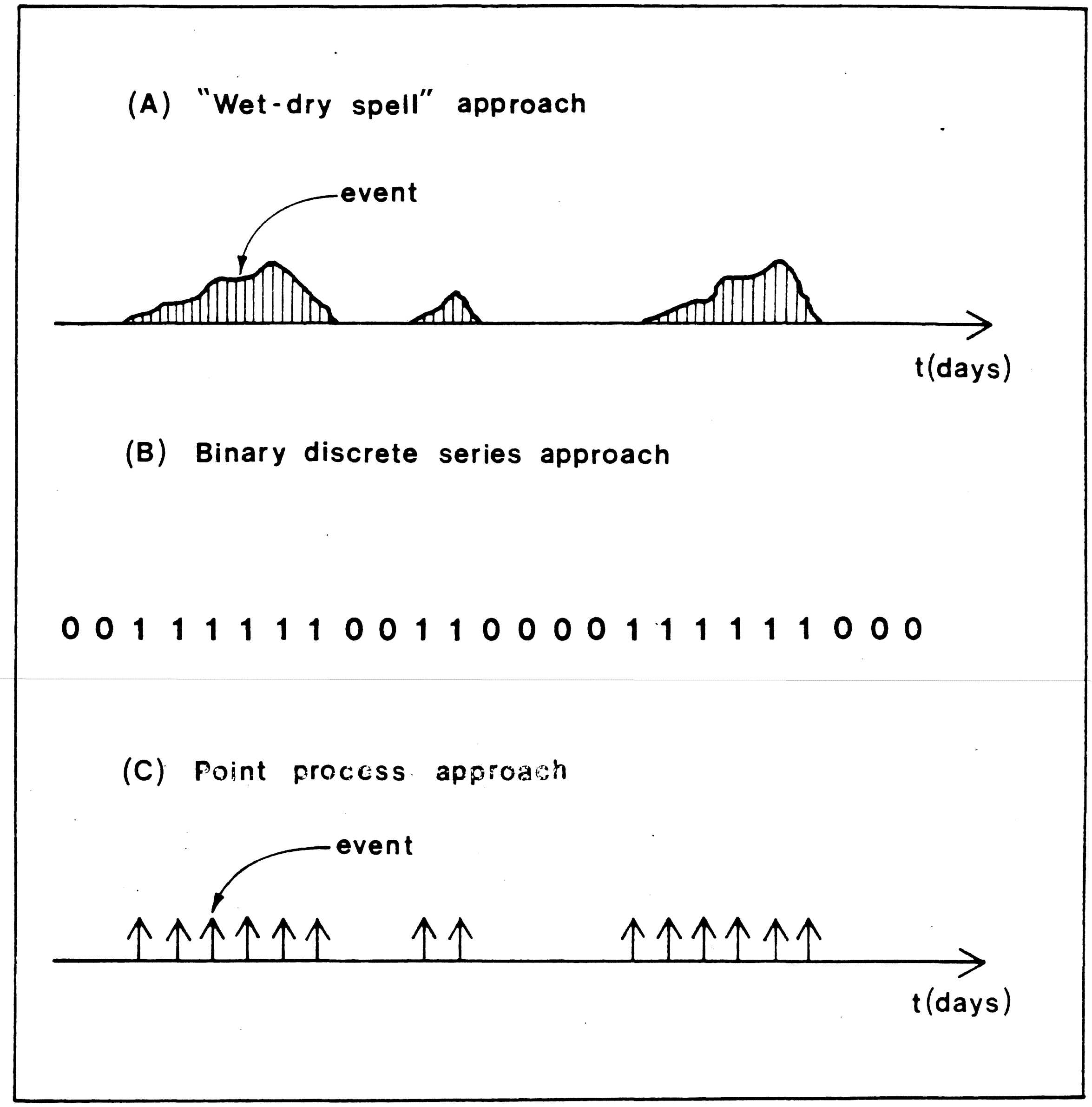


Figure 2.1 Different approaches to modeling daily rainfall occurrences.

the probability laws of the length of the wet periods (storm duration) and the length of the dry periods (time between storms).

This model structure, with exponential distributions for the lengths of the dry and wet periods, was used by Thom (1958) and Green (1964), among others. Grace and Eagleson (1966) used a Weibull distribution for the wet-period lengths and applied the model to short time increment (on the order of minutes and hours) rainfall occurrences. Todorovic and Yevjevich (1969) and Eagleson (1978) conducted later studies using this modeling approach.

In probabilistic terminology the above model is an alternating renewal model. The term renewal stems from the implied independence between the dry and wet period lengths, and the term alternating is used to indicate that a wet (dry) period is always followed by a dry (wet) period, i.e., no transition to the same state is possible. In many early studies, such a model with exponential distributions for the dry period lengths was referred to as a Poisson model. This is an inaccurate terminology resulting from the assumption that an event, which in this case corresponds to a wet period, occurs instantaneously at the middle or end of the wet period.

A recent study of an alternating renewal model for daily rainfall was reported by Galloy et al. (1981). They used discrete negative binomial distributions for the wet and dry period lengths and implemented the theory of point processes to derive the statistical properties of intervals (times between events) and counts (number of events in a time interval).

The main problem with the wet-dry spell approach to modeling daily or other small-time-increment rainfall lies in the modeling of the rainfall amounts. Due to the definition of an event as an uninterrupted sequence of wet days, the cumulative rainfall amounts corresponding to each event are conditional on the wet-period length. Therefore, conditional probability distributions have to be fitted to the amounts. This can pose problems, especially for events of extreme duration, where not many points are available for identification and fitting of a probability density function.

2.1.2 The Binary Discrete Time-Series Approach

The daily rainfall series consists of either rainy or dry days, and therefore can be viewed as a binary series of zeroes and ones, with zero corresponding to a dry day, and one to a wet day (see Fig. 2.1b.). A probabilistic model is then sought which can adequately describe this sequence of zeroes and ones. Alternatively, this binary process can be thought of as being formed by a sequence of Bernoulli trials i.e., repetitive trials without replacement (see, for example, Feller, 1968), with two possible outcomes, zero and one. The outcomes can be either independent (giving rise to a Bernoulli process), or dependent (giving rise, for example, to a Markov chain). The independent Bernoulli process is not adequate to describe the dependence present in the daily rainfall occurrences (see, for example, Smith and Schreiber, 1973). Markov chain models are the simplest models with a dependence structure and have been extensively used for modeling daily rainfall.

Markov chains. A Markov chain is a sequence of discrete random variables, X_n, and is said to be of order k if k is the smallest positive integer such that the following equation of conditional probabilities is satisfied for all n:

$$P\{X_n | X_{n-i},\ i = 1, \ldots\} = P\{X_n | X_{n-1}, X_{n-2}, \ldots, X_{n-k}\} \quad (2.1)$$

A complete treatment of the theory of Markov chains can be found in Cox and Miller (1965), Parzen (1962), and Çinlar (1975), among others. A two-state Markov chain (appropriate for the zero-one rainfall occurrence process) is completely specified by the transition probability matrix:

$$\underline{P} = \begin{bmatrix} p_0 & 1-p_0 \\ 1-p_1 & p_1 \end{bmatrix}$$

where p_0 is the probability of a dry day following a dry day, and p_1 is the probability of a wet day following a wet day.

Markov chains have been extensively used for modeling daily rainfall occurrences. Gabriel and Neumann (1957,1962) used a first-order homogeneous (i.e., constant parameters) Markov chain for the winter daily rainfall occurrences at Tel-Aviv, while Caskey (1963) and Weiss (1964) used a non-homogeneous (i.e., time varying parameters) Markov chain for several stations in the northern U.S. Hopkins and Robillard (1964) used a first order Markov chain for the daily rainfall occurrences in Canada and found that it was not adequate to describe the months with few rainy days. Feyerherm and Bark (1967) showed the inadequacy of a first-order Markov chain in describing the higher-order dependence structure present in daily rainfall, and they proposed a second-order Markov chain for the daily rainfall occurrences at Indiana, Iowa, and Kansas. Wiser (1965) and Green (1965) also concluded that the geometric memory of the first-order Markov chain is not adequate to describe long droughts or long wet spells. Smith and Schreiber (1973) found a non-homogeneous first-order Markov chain superior to an independent Bernoulli model for the seasonal thunderstorm rainfall in the southwest U.S. Woolhiser and Pegram (1979) studied Markov chain models with seasonally varying parameters using Fourier series.

In deciding the order of a Markov chain, Tong (1975) used the Akaike Information Criterion (AIC), while Hoel (1954) presented a likelihood ratio goodness of fit test. Chin (1977) identified the Markov chain orders of 25-year daily rainfall records in the United States, using the AIC, and illustrated their dependence on the season and geographical location.

For the simultaneous modeling of daily rainfall occurrences and amounts, multiple-state Markov chains have also been considered. Khanal and Hamrick (1974) used a 14-state Markov chain model for each month of the year, for daily rainfall sequences from Florida. Haan et al. (1976) separated the year into four seasons and used a seven-state first-order Markov chain for the daily rainfall in Kentucky. They assumed uniform distributions for the rainfall amounts in all but the last state, in which a shifted exponential was found more appropriate due to the larger variability in the amounts. Carey and Haan (1978) used a three-state first-order Markov chain with two different Gamma distributions for the amounts in the two wet states, which were further combined to the same pooled distribution due to the large number of parameters in their twelve-season model.

In conclusion, Markov chain models provide a simple mathematical representation of the daily rainfall occurrence process which may be adequate for some specific sites and seasons. However, their Markovian structure cannot describe the long term persistencies (i.e., long wet or dry spells) and the effect of clustering (i.e., higher likelihood of having an event due to an event at a previous time) present in the short-time-increment rainfall occurrences. Long term persistence in the daily rainfall may be caused, for example, by cyclonic activity persisting during certain seasons (Petterssen, 1969), and clustering may be the result of frontal thunderstorms with a relatively long life cycle (Kavvas and Delleur, 1975).

Discrete Autoregressive Moving Average (DARMA) models. A DARMA(p,q) model, where p is the order of the autoregressive and q the order of the moving average component, is a sequence $\{X_n\}$ formed by a probabilistic combination of elements of a sequence $\{Y_n\}$ which is independent and identically distributed (i.i.d.). For the binary DARMA models $\{Y_n\}$ is assumed to be i.i.d. with a Bernoulli distribution, i.e., $P(Y_n = 0) = \pi_0$, $P(Y_n = 1) = \pi_1$, and $\pi_0 + \pi_1 = 1$. For illustration purposes, the DARMA(1,0) and DARMA (0,1) models are defined by a sequence $\{X_n\}$ such that

$$\text{DARMA}(1,0):\ X_n = \begin{cases} Y_n & \text{with probability } \theta \\ Y_{n-1} & \text{with probability } 1-\theta \end{cases} \tag{2.3}$$

$$\text{DARMA}(0,1):\ X_n = \begin{cases} X_{n-1} & \text{with probability } \phi \\ Y_n & \text{with probability } 1-\phi\ . \end{cases} \tag{2.4}$$

For further details on these models and derivation of their statistical properties, see Jacobs and Lewis (1978a,b) and Chang et al. (1982).

DARMA models for daily rainfall were first used by Buishand (1978) to analyze wet-dry spells in the Netherlands. Subsequently, Chang et al. (1982, 1984) applied DARMA models to daily rainfall sequences in

Indiana. They derived the probability density functions of the wet and dry period lengths as functions of the DARMA model parameters ϕ, θ, and the marginal distribution function $\pi = (\pi_0, \pi_1)$.

Although DARMA models may be an improvement over Markov chains, in the sense that they can accommodate longer term persistence in the series in a more parsimonious way than a high-order Markov chain, their linear structure is still not able to describe the clustered short-term dependence known to be present in the daily rainfall occurrences (Kavvas and Delleur, 1975). Also, their mathematical framework seems to permit derivation only of interval properties (i.e., probability distributions of run lengths) and not of counting properties (i.e., distributions of number of events in a time interval). Given that analysis of the second order properties of intervals and counts are not, in general, equivalent (see, for example, Cox and Lewis, 1978), it is advantageous to be able to use both for model identification and fitting, and this can be effectively done in the point process mathematical framework.

2.1.3 The Point-Process Approach

By defining an event as the occurrence of a day with a total rainfall amount exceeding a specified threshold (i.e., the occurrence of a wet day), the sequence of daily rainfall forms a point process. With the above definition of events, a wet period of several days is treated as a group of instantaneous rainfall events occurring at one-day intervals and, therefore, the interarrival times are positive integer values (1, 2, 3, ... days). Such a point process is discrete. A major issue, which is deferred to Chapter 3, is how to accommodate this feature within the framework of continuous point processes. The present discussion is limited to continuous point processes.

The theory of continuous point processes has been studied by Cox and Lewis (1978), Cox and Isham (1980), Çinlar (1975), Parzen (1962), Lewis (1972), Vere-Jones (1970), and others. Waymire and Gupta (1981b, c) give an excellent review of the theory of point processes and relate it to the stochastic modeling of hydrologic series.

The simplest continuous point process is the Poisson process, whose formal definition and properties can be found in many probability theory texts (see for example Çinlar, 1975, Ch.4). In a Poisson process, the times between events are independent and identically distributed (i.i.d.) random variables having an exponential distribution, and the number of events in a time interval t is an i.i.d. random variable having a Poisson distribution. The non-homogeneous (time-varying parameters) Poisson process has been applied by Todorovic and Yevjevich (1969) and Gupta and Duckstein (1975) to the modeling of rainfall occurrences.

Kavvas and Delleur (1975) observed that the daily rainfall occurrences in Indiana exhibit a clustering which might be satisfactorily modeled by the class of Poisson cluster models (e.g., Cox and Isham, 1980, Ch.3) and in particular by the Neyman-Scott (N-S) models. A N-S process is a two-level process. At the primary level, the rainfall generating mechanisms (RGM) occur according to a Poisson model with rate of occurrence h_0 (i.e., mean interarrival time $1/h_0$). Each RGM gives rise to a group of rainfall events, and each of these groups is called a cluster. Within each cluster, the occurrence of events is completely specified by a distribution for the number of events and a distribution for their positions relative to the cluster centers. Kavvas and Delleur (1975) assumed a geometric distribution with parameter p for the number of rainfall events in a cluster and an exponential distribution with parameter θ for the distances of events from their cluster centers. For these distributions, the final observed process has a rate of occurrence $m = h_0/p$. Applications of the N-S model include modeling of the areal clustering of rainfall (LeCam, 1961) and modeling of earthquake occurrences (Vere-Jones, 1970).

Smith (1981) introduced another point process model, namely a doubly stochastic Poisson model, to describe the clustering observed in the daily rainfall occurrences of the summer season (July to October) rainfall in the Potomac river basin. In a doubly stochastic Poisson model (also known as a Cox model) the rate of occurrence of the process alternates between two states, one zero and the other positive. During periods when the intensity is zero, no events can occur. Smith and Karr (1983) assumed that during periods with positive intensity, events occur according to a Poisson process with rate of occurrence λ, and that the sequence of states visited forms a Markov chain. This model is a renewal model (i.e., interarrival times are independent) and was termed the RCM model (Renewal Cox model with Markovian intensity).

In summary, two main classes of continuous point process models (namely the N-S model and the RCM model) have been studied for the daily rainfall occurrence process. Both of these models are overdispersed relative to Poisson, that is the variance of the number of events in an arbitrary time interval is greater than the mean number of events in that interval, as compared to the Poisson model in which the variance is equal to the mean. Our analysis has pointed out that structures of daily rainfall occurrences that are underdispersed relative to Poisson are possible (more regular occurrence of events than that of a Poisson process), a feature that cannot be reproduced by either the N-S or the RCM models. More significant, however, is the question of whether continuous models are appropriate for modeling discrete daily rainfall occurrences. This is an especially significant issue when the time scale is daily, since the observation sequence represents cumulative rainfall amounts over a period (i.e., one day) which is on the order of the process interarrival time. In Chapter 3, it will be shown that the more natural way to proceed is to model the daily rainfall occurrence process as a discrete point process. First, however, a review of recent work on modeling rainfall amounts is given.

2.2 Models for the Non-Zero Daily Rainfall Amounts

If the non-zero daily rainfall amounts process is independent, then it is completely characterized by its marginal probability density function (pdf). The marginal probability distributions most commonly used are the following:

(1) The exponential distribution (Todorovic and Woolhiser, 1971; Richardson, 1981) which is a one-parameter distribution. Skees and Shenton (1974) and Mielke and Johnson (1974) suggested that the exponential distribution has a thinner tail than that observed in daily rainfall amounts.

(2) The mixed exponential distribution (Smith and Schreiber, 1973; Woolhiser and Pegram, 1979; Woolhiser and Roldan, 1982) whose coefficient of variation is always greater than unity, as is usually the case in the daily rainfall amounts. This distribution has the appealing interpretation of being the superposition of two or more exponential distributions produced by, say, different mechanisms. The mixed exponential distribution was found to be the best of four candidates for

daily rainfall amounts in a comparison study by Woolhiser and Roldan (1982). Everitt and Hand (1981, Ch.3) discuss methods of identifying and fitting mixed distributions.

(3) The gamma distribution which has been used extensively (see, for example, Ison et al., 1971; Buishand, 1978; Carey and Haan, 1978).

(4) The Kappa or generalized beta distribution introduced by Mielke (1973) and Mielke and Johnson (1974).

(5) Empirical distributions as, for example, that used by Cole and Sheriff (1972) or other special distributions. For example, Katz (1977) used a chain-dependent distribution assuming that the rainfall amounts are independent but that the distribution function depends on whether the previous day was wet or dry. Buishand (1978) distinguished between three different types of wet days (DWD, DWW, WWD, and WWW where D stands for dry and W for wet) and fitted different Gamma distributions to each of the three rainfall amounts. All these distributions, however, have the disadvantage of too many parameters.

Woolhiser and Roldan (1982) present a comparison of several distributions (chain-dependent and independent exponential, gamma, and mixed exponential distributions) for five U.S. stations in Kansas, Missouri, Florida, Wyoming, and Indiana. Using the Akaike Information Criterion these distributions ranked from best to worst as mixed exponential, independent gamma, chain-dependent gamma, and exponential. It should be noted that in the above study the degree of dependence of rainfall amounts in consecutive days was not tested and independence was assumed.

If a dependence structure is present, then more complicated models have to be used. Commonly used time-series models, such as those described by Box and Jenkins (1976), are not appropriate because rainfall amounts are bounded from below by zero, and are therefore positively skewed. The class of Exponential-ARMA (EARMA), Gamma-AR (GAR), or standard ARMA models together with normalization transformations might be considered, however. ARMA models with skewed marginal pdf's (especially exponential and Gamma) have been extensively studied by Lawrance and Lewis (1980), Lawrance (1980), Gaver and Lewis (1980), Jacobs and Lewis (1977), Lawrance and Lewis (1977), and Lewis (1978) and have been applied to hydrology by Obeysekera and Salas (1983) for streamflow modeling. Raudkivi and Lawgun (1972) have proposed a scheme for modeling serially correlated data with skewness described by a Pearson type 3 distribution. Although they have applied their technique to rainfall durations (defined as the length of non-zero 10-minute interval rainfall depths), the potential use in modeling daily rainfall amounts seems straightforward.

CHAPTER 3
CONTINUOUS VERSUS DISCRETE POINT PROCESS MODELS FOR DAILY RAINFALL OCCURRENCES

The Second, feeling of the tusk,
    Cried, "Ho! what have we here
So very round and smooth and sharp?
    To me 'tis mighty clear
This wonder of an Elephant
    Is very like a spear!"

When daily rainfall occurrences are modeled as a continuous point process, it is implied that events can occur anywhere on the time axis, i.e., that multiple occurrences during a day are possible. The only information, however, that is contained in the daily rainfall occurrence data is whether a day is dry or wet, i.e., whether or not at least one event has occurred during a day, and not the number of events. With the continuous-point-process interpretation of the daily rainfall occurrences, one faces the problem of inferring the properties of the (unobservable) continuous counting process from the discrete sampled data. Brillinger (1978) and Guttorp and Thompson (1983) have studied this problem and have proposed methods of estimating the parameters of a continuous point process, as well as approximately reconstructing the locations of events, from the discrete sampled data. Such methods, however, are applicable only when the sampled counting process provides at least the information of the number of events during the sampling intervals, as for example in the process of daily traffic fatalities where the number of fatalities during a day are recorded but not the exact times of occurrence.

Daily rainfall occurrence data, on the other hand, do not contain information about the number of "events" in a day, and therefore their interpretation as a continuous point process is complicated. For example, if a continuous point process model is used for generation of (synthetic) daily rainfall occurrences, the most natural approach to discretizing the continuous sequence is to lump all the occurrences during a one day interval to only one point, say, at the end of that day. The result of such an operation is a discretized point process with a lower rate of occurrence and altered statistical properties. The greater the rate of occurrence, i.e., the more frequent the events, the more serious the discretization effect will be. For rates corresponding to daily rainfall occurrences (λ = 0.5 to 0.2 days^{-1}, for mean interarrival times of 2 to 5 days), these effects are fairly significant, in contrast with rates corresponding to, say, hourly rainfall events or occurrence of wet periods, i.e., interrupted sequences of rainy hours or days. In general, the use of continuous point process models for discrete observation sequences will present major problems when the observation sequences represent cumulative rainfall amounts over a period which is on the order of the process interarrival time.

How much such a discretization scheme affects a Poisson process with rate of occurrence λ can easily be shown analytically. For example, it can be shown that in order to obtain a discrete point process with rate of occurrence λ, a continuous Poisson process with rate of occurrence equal to $-\ln(1 - \lambda) > \lambda$ is required. Similar results for non-Poisson processes such as Poisson cluster processes are not easily obtained in closed form and are better studied via simulation.

Modeling daily rainfall occurrences by using continuous models with adjusted parameters to compensate for the effects of discretization are awkward at best and generally impractical. The natural approach to modeling the daily rainfall occurrences is to use only the information provided by the data, i.e., to view the rainy days as constituting all of the events of the process, and to generate rainy or dry days rather than continuous events. Some of the implications of this viewpoint are considered in this chapter. First, however, some definitions and general properties of a continuous point process, necessary for the development of the rest of this work, are given.

3.1 Statistical Background on Stationary Point Processes

Let an event $\mathcal{E}$ occur at times t_1, t_2, $t_3, \ldots$, and let $X_r = t_r - t_{r-1}$ $(r = 1,2,3,\ldots)$ be continuous random variables identically distributed with common pdf $f(x)$. The variable X is called the interarrival time, or time between events, or simply interval. A point process is stationary if the joint distribution of the number of events in a set of k fixed intervals, for all k = 1,2,3,..., is invariant under translation (Cox and Lewis, 1978). Two immediate consequences of stationarity are that (1) the distibution of the number of events in an interval depends only on the length of the interval, and (2) there is no trend in the mean rate of occurrence of events, i.e., the expected number of events in an interval is proportional to the length of that interval.

Depending on whether or not an event occurs at time t_0, the process starts with an arbitrary event or at an arbitrary time (synchronous and asynchronous sampling, respectively, in the terminology of Lawrance (1972)). For the rainfall occurrences, we have considered the process as starting at an arbitrary event but not including it. This implies that the pdf of the time to the first event is the same as the pdf of all the other subsequent interarrival times.

Continuous point processes have been extensively studied in the statistical literature (see for example, Cox and Lewis, 1978; Cox and Isham, 1980; Çinlar, 1975; Lewis, 1972; Srinivasan, 1974). They have found extensive applications: in queueing theory (Khintchine, 1960); modeling times to computer failure (Lewis, 1964); earthquake occurrences (Vere-Jones, 1970); traffic data (Bartlett, 1963); spatial distibution of galaxies (Neyman and Scott, 1958); and rainfall occurrences (LeCam, 1961; Kavvas and Delleur, 1975; Waymire and Gupta, 1981; and Smith and Karr, 1983). Additional applications can be found in a series of papers edited by Lewis (1972).

In studying a series of events (point process), two properties are of interest: the interval properties dealing with the times between events, and the counting properties dealing with the number of events in time periods of specified length. The second order properties of intervals and counts which will be used in this work are introduced below.

Interval properties. Let $\{X_i\}$ be the series of interarrival times. We denote the mean, variance, and coefficient of variation of X by E(X), Var(X), and c_v, respectively. Standard time series analysis, using the autocorrelation function and the power spectrum of the series, can be applied to test the presence of autocorrelation. Independent interarrival times (inferred by an autocorrelation function not significantly different from zero, or a constant power spectrum) indicate that the process is a renewal point process. Special care with the properties of the autocorrelation coefficients is needed, however, due to the non-normality (high skewness) of interarrival times (Lewis, 1972). For example, Moran (1970) and Cox (1966) have shown that the variance of the first autocorrelation coefficient tends to be smaller for random variables with long tails than for variables with a normal distribution.

The departure of the coefficient of variation, c_v, from the value of one for the exponential distribution, is used as a rough measure of the departure of the process from the Poisson process (Cox and Lewis, 1978). A value of $c_v > 1$ indicates overdispersion relative to the Poisson process ("random" clustering), and a value of $c_v < 1$ indicates underdispersion ("regular" clustering).

Let $F(x) = P(X \leq x)$ be the cumulative probability distribution of the interarrival times. Then, the probability of exceedence

$$R(x) = P(X > x) = 1 - F(x) \tag{3.1}$$

is called the survivor function, and its logarithm is the log-survivor function. It can be easily checked from (3.1) and the pdf of an exponential distribution that the log-survivor function of a Poisson process with rate of occurrence λ is a straight line with slope equal to $-\lambda$. In analyzing a series of events, deviations of the empirical log-survivor function from a straight line indicate deviation of the process from the Poisson. In particular, for a renewal process, a convex log-survivor function implies a coefficient of variation less than one, while the opposite holds for a concave log-survivor function (Watson and Wells, 1961). Such relationships are helpful in determining the marginal distribution of the interarrival times and in identifying possible models for the process.

Counts properties. Let $\{N_t\}$ denote the counting process of an asynchronous point process (i.e., a process which starts at an arbitrary time), and $\{N_t'\}$ denote the counting process of a synchronous point process (i.e., a process which starts with an arbitrary event). Notice that $\{N_t'\}$, the number of events in (0,t], is the counting process of a series of events that starts with an event but does not include it. The obvious relationship between the sequence of intervals $\{X_i\}$ and the counting process $\{N_t'\}$ is

$$P(N_t' < r) = P(x_1 + x_2 + \dots + x_r > t),\ r = 1,2,\dots \quad (3.2)$$

(Cox and Lewis, 1978).

The following counting properties are of interest:

(1) The mean value function, M(t), defined as

$$M(t) = E(N_t). \quad (3.3)$$

For any stationary process, $M(t) = t/E(X) = mt$, where $m = 1/E(X)$ is the intensity or rate of occurrence of the process.

(2) The renewal function, H(t), defined as

$$H(t) = E(N_t'). \tag{3.4}$$

For large t, $H(t) \rightarrow M(t)$. For a Poisson process $H(t) = mt$.

(3) The renewal density, or conditional intensity function, h(t), defined as

$$h(t) = \lim_{\Delta t \to 0} \frac{E(N_{t}',\ _{t+\Delta t})}{\Delta t} = \frac{dH(t)}{dt} \tag{3.5}$$

(Cox and Lewis, 1978). Notice that h(t) is not a pdf, but instead $h(t)\Delta t$ is the probability of having an event in a small interval Δt near t. Since multiple events are not permitted (this is the so-called orderliness requirement; see Daley and Vere-Jones, 1972), the probability of more than one event in an interval of length Δt is $O(\Delta t^2)$, and therefore:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(\text{event in } (t_0+t,\ t_0+t+\Delta t)\ /\ \text{event at } t_0)}{\Delta t}, \tag{3.6}$$

where the event at t_0 is an arbitrary event in the stationary process. The renewal density of a Poisson process is constant and equal to the intensity of the process, m.

(4) The variance time curve, V(t), defined as

$$V(t) = Var(N_t). \tag{3.7}$$

For a Poisson process, $\{N_t\}$ has a Poisson distribution for which the variance is equal to the mean, and therefore $V(t) = mt$. Procedures for estimating the empirical variance time curve are given in Cox and Lewis (1978).

(5) The index of dispersion function, $I(t)$, defined as

$$I(t) = \frac{V(t)}{M(t)} = \frac{V(t)\ E(X)}{t}, \qquad (3.8)$$

which has the constant value of one for the Poisson process. An empirical $I(t) < 1$ for all t implies underdispersion relative to Poisson, and an $I(t) > 1$ for all t implies overdispersion (analogously to the coefficient of variation of the interarrival times relative to one, the value for the exponential distribution).

(6) The covariance density, $\gamma_+(\tau)$, defined as

$$\gamma_+(\tau) = \lim_{\Delta t \to 0} \frac{\mathrm{cov}(N_{t+\tau+\Delta t,\ t+\tau},\ N_{t,\ t+\Delta t})}{(\Delta t)^2} \qquad (3.9)$$

$$= \lim_{\Delta t \to 0} \frac{\mathrm{cov}(\Delta N_{t+\tau},\ \Delta N_t)}{(\Delta t)^2},$$

which can be interpreted as the autocovariance function of the differential process $\Delta N_t = \lim_{\Delta t \to 0} N_{t,t+\Delta t} = \lim_{\Delta t \to 0}(N_{t+\Delta t} - N_t)$. The differential process, $\{\Delta N_t\}$, can be thought of as an instantaneous process having zeroes at all points except for spikes (delta functions) at the points of occurrence of events. The covariance density, $\gamma_+(\tau)$, is a measure of the likelihood of two events occurring τ units apart (Cox and Lewis, 1978, ch.4). For a Poisson process with intensity m,

$$\gamma_+(\tau) = \begin{cases} m, & \text{for } \tau = 0 \\ 0, & \text{otherwise.} \end{cases} \qquad (3.10)$$

(7) The spectrum of counts, $g_+(\omega)$, which is the Fourier transform of the covariance density

$$g_+(\omega) = \int_0^{\infty} \gamma_+(\tau) e^{-i\omega\tau} d\tau. \qquad (3.11)$$

The spectrum of counts is a useful tool in the statistical analysis of series of events and is preferable to other functions due to its superior sampling properties (Bartlett, 1963). For a Poisson process the spectrum of counts has a constant value equal to m/π.

3.2. Poisson Versus Bernoulli Processes

In this section, the properties of the Bernoulli process, which is the discrete analogue of the Poisson process, are studied and compared with those of the Poisson process. This comparison reveals that if indeed the discrete daily rainfall occurrences were an independent process, i.e., a Bernoulli process, if modeled as a continuous point process they would be interpreted as underdispersed relative to the (continuous) Poisson process. On the other hand, selected daily rainfall structures underdispersed relative to the

Poisson process are, in fact, all shown to be overdispersed relative to Bernoulli.

3.2.1. Statistical Properties of the Bernoulli Process

Consider a sequence of independent repeated trials with two possible outcomes, success and failure. Let p denote the probability of success at each trial and N_r denote the number of successes in r trials. Then, N_r has a binomial probability distribution

$$P(N_r = k) = \binom{r}{k} p^k (1-p)^{r-k} \quad , k = 0,1,2,..., \tag{3.12}$$

and the number of trials between the n'th and (n+1)st success, X_n, has a geometric distribution

$$P(X_n = k) = p(1-p)^{k-1} \quad , k = 1,2,..., \tag{3.13}$$

for all n. In the discrete-time point process terminology, a success corresponds to the occurrence of an event (i.e., a rainy day); N_r to the counting process, that is the number of events in (0,r]; and X_n to the time between events.

The Bernoulli process is the discrete analogue of the Poisson process in the sense that it is characterized by independent intervals and independent counting increments and is discrete in time. This lack of memory property is the result of the geometric distribution for the times between events, analogously to the exponential for the Poisson (see Feller, 1968, p.329 for a proof).

The statistical properties (i.e., mean, variance, and higher moments) of the geometric and binomial distributions are well known (see for example, Parzen, 1962). For this work, some additional properties of the Bernoulli counting process are of interest, such as the spectrum of counts, and this derivation is given below. Since it is more natural to discuss the daily rainfall occurrences with respect to time instead of trials, the familiar terminology of continuous point processes has been retained, with the understanding that time t in a Bernoulli process, or in a general discrete point process, corresponds to t discrete time units (i.e., t days).

Let f(x) be a probability density function (pdf) defined as the continuous representation of the geometric probability mass function (pmf) of (3.13). Then,

$$f(x) = \sum_{k=1}^{\infty} p(1-p)^{k-1} \delta(x-k), \tag{3.14}$$

where $\delta(\cdot)$ is the Dirac delta function. Let $*f(s)$ denote the Laplace transform of f(x) defined as

$$*f(s) = \int_0^{\infty} e^{-sx} f(x)dx.$$

The symbol $*f(s)$ is used to indicate the Laplace transform of a generalized function of the form (3.14) from the Laplace transform $f^*(s)$ of a standard continuous function f(x). Notice that $*f(s)$ is an exponential function of s, since the Laplace transform of $\delta(x-k)$ is $\mathcal{L}(\delta(x-k)) = e^{-sk}$. It is easily shown that the Laplace transform of f(x) of (3.14) is

$$*f(s) = \frac{pe^{-s}}{1 - (1-p)e^{-s}} . \tag{3.15}$$

Making use of a standard result of the renewal theory (see for example, Cox and Lewis, 1978, Ch.4), the spectrum of counts of a stationary renewal point process is given as

$$g_+(\omega) = \frac{p}{\pi}\left[1 + \frac{*f(i\omega)}{1-*f(i\omega)} + \frac{*f(-i\omega)}{1-*f(-i\omega)}\right]. \tag{3.16}$$

Substitution of (3.15) into (3.16) yields

$$g_+(\omega) = \frac{p(1-p)}{\pi}. \tag{3.17}$$

The statistical properties of interest for a Bernoulli process, with a probability of success p, are given in Table 3.1. In the same table, the corresponding properties of a Poisson process with rate of occurrence λ are also given.

It is convenient to notice here that $*f(s) = \psi(-s)$, where $\psi(\cdot)$ is the moment generating function of the probability law of x (eq. 3.14), and is defined as $\psi(z) = E[e^{zx}]$, i.e., as the expectation of the exponential function e^{zx} (see, for example, Parzen, 1960, p.215). The equivalence of these terms will be used in Chapter 5.

3.2.2 Comparison of a Poisson and a Bernoulli Process

Consider a sequence of daily rainfall occurrences with mean interarrival time $\bar{x}$. If a Bernoulli process were fit to the series, the estimate of its probability of success, p, would be $\hat{p} = 1/\bar{x}$.

Table 3.1 Comparison of Poisson and Bernoulli Processes

	Poisson	Bernoulli
	λ = rate of occurrence	p = prob. of success
Interarrival times: $\{X_i\}$	$f(x) = \lambda e^{-\lambda x}$, $\lambda > 0$	$p(x) = p(1-p)^{x-1}$, $0 \leq p \leq 1$
	$E(X) = 1/\lambda$	$E(X) = 1/p$
	$Var(X) = 1/\lambda^2$	$Var(X) = (1-p)/p^2$
	$c_v = 1$[a]	$c_v = \sqrt{1-p} < 1$
	$c_s = 2$[b]	$c_s = \frac{2-p}{\sqrt{1-p}} > 2$
Number of events: $\{N_t\}$	$p(N_t=k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$	$p(N_t=k)=\binom{k}{t}p^k(1-p)^{t-k}$ t = discrete time
	$E(N_t) = \lambda t$	$E(N_t) = pt$
	$Var(N_t) = \lambda t$	$Var(N_t) = p(1-p)t$
Conditional intensity function	$h(t) = \lambda$	$h(t) = p$
Log survivor function	$\ell n[R(x)] = -\lambda x$	$\ell n[R(x)] = \ell n(1-p)\ x$
Variance time curve	$V(t) = \lambda t$	$V(t) = p(1-p)t$
Index of dispersion function	$I(t) = 1, \forall t$	$I(t) = 1-p < 1, \forall t$
Spectrum of counts	$g_+(\omega) = \lambda/\pi$, $\omega \geq 0$	$g_+(\omega) = p(1-p)/\pi$, $\omega \geq 0$
Normalized spectrum of counts	$g_+'(\omega) = 1$, $\omega \geq 0$	$g_+'(\omega) = 1-p < 1$, $\omega \geq 0$

[a] c_v = coefficient of variation
[b] c_s = skewness coefficient

Similarly, if a Poisson process were fit, the estimate of the rate of occurrence, λ, would be $\hat{\lambda} = 1/\bar{x}$. Thus $\hat{p} = \hat{\lambda}$. Notice, however, from Table 3.1 how different the other properties of the two processes are. In particular, the Bernoulli process has a coefficient of variation of intervals and an index of dispersion function of the counts always less than one, which imply underdispersion relative to Poisson. This means that inferences about over- and under-dispersion of the daily rainfall occurrences would be different depending on whether the empirical functions of the process were compared to those of a Poisson or to those of a Bernoulli process. It seems only natural that a discrete point process model, such as daily rainfall occurrences, should be compared with the discrete independent Bernoulli process and not with the continuous Poisson process. This is an important observation and has immediate consequences in the interpretation of the statistical functions of the daily rainfall occurrence process. In the next section, the effects of using a continuous point process model for the generation of a discrete sequence will be studied, analytically for a Poisson model and via simulation for a Neyman-Scott model.

3.3 Effects of Discretization on a Continuous Point Process

When a continuous point process is used for generation of (synthetic) daily rainfall occurrences, the most natural approach to discretizing a continuous synthetic sequence is to lump all the occurrences during a day at one point, such as, the end of that day. The resulting discrete point process has different statistical properties than the continuous one. How much these two structures differ will be illustrated below, first for a Poisson process and then for a Neyman-Scott process.

Let F(x) denote the cumulative distribution function of the exponential pdf of the intervals of a Poisson process. The discretization scheme suggested above is equivalent to replacing the continuous exponential distribution of the intervals with a discretized one, so that

$$
\begin{aligned}
P_d(x = 1) &= P_c(x \leq 1) = F(1) = 1 - e^{-\lambda}, \\
P_d(x = 2) &= P_c(1 < x \leq 2) = F(2) - F(1) = e^{-\lambda}(1 - e^{-\lambda}), \\
&\cdot \\
&\cdot \\
&\cdot \\
P_d(x = k) &= P_c(k-1 < x \leq k) = F(k) - F(k-1) = e^{-(k-1)\lambda}(1 - e^{-\lambda}),
\end{aligned}
\tag{3.18}
$$

where P_d and P_c denote probabilities of a discrete variable and a continuous variable, respectively. Notice that the resulting discrete distribution, $P_d(x)$, is geometric with parameter:

$$
\gamma = 1 - e^{-\lambda}, \tag{3.19}
$$

implying that the discretized process is a Bernoulli process with a probability of occurrence (or rate of occurrence) equal to γ, a value always less than λ. All the other properties of the discretized process can be obtained by substituting the value of γ for p in the right-hand column of Table 3.1.

For data generated from a Poisson process, figures 3.1 and 3.2 show the effects of discretization on some commonly used counting properties, i.e., the spectrum of counts, log-survivor function, variance time curve and index of dispersion. A period of observation of 1000 time units (for example, days) was used, since this is approximately the length of series available in a month by month analysis of thirty years of daily data. Notice the agreement of the analytical and simulation results; the empirical functions of the discretized process differ from those of the Poisson and the differences are larger for the higher rates of occurrence. Also, notice that the discretized process is always underdispersed relative to Poisson.

Another important issue raised from Figures 3.1 and 3.2 is the data requirements to obtain reliable estimates of the empirical functions. It can be seen from Figure 3.1 that, although the length of observation is the same for all cases, the empirical functions of the continuous Poisson are closer to the theoretical ones the larger the number of events is. This implies that fewer years of daily rainfall data during rainy seasons contain the same information as more years during dry seasons, and therefore caution must be applied when interpreting the statistical properties of the rainfall occurrences during seasons with few rainy days.

Figure 3.3 illustrates the effect of discretization on a clustered Neyman-Scott process. (This process and the meaning of its parameters have been discussed in Chapter 2.) Although the effects of discretization cannot be directly associated with the parameter values

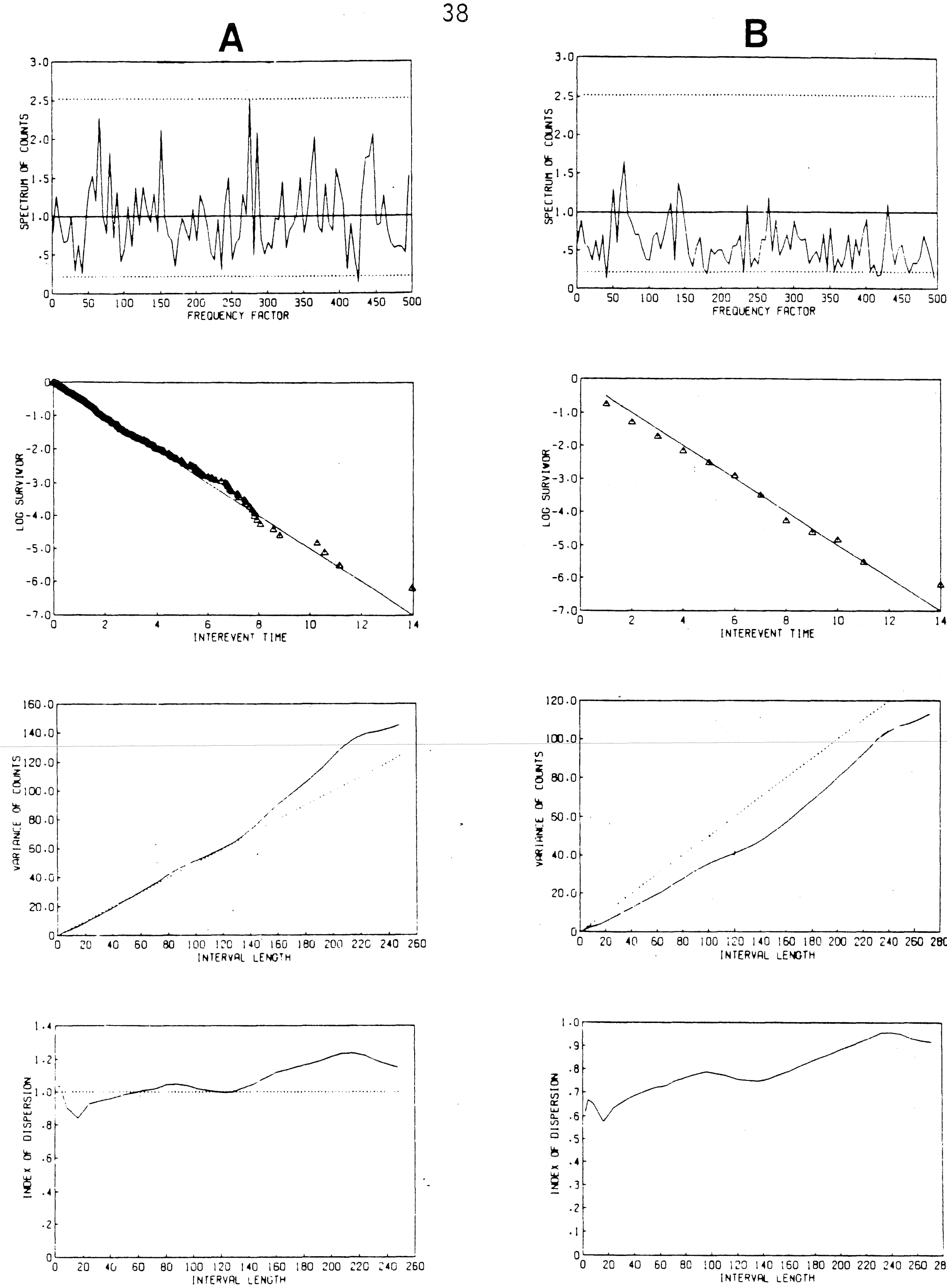


Figure 3.1 Effects of discretization on a Poisson process with rate of occurrence $\lambda = 0.50$.
(A) Continuous process, (B) Discretized process.

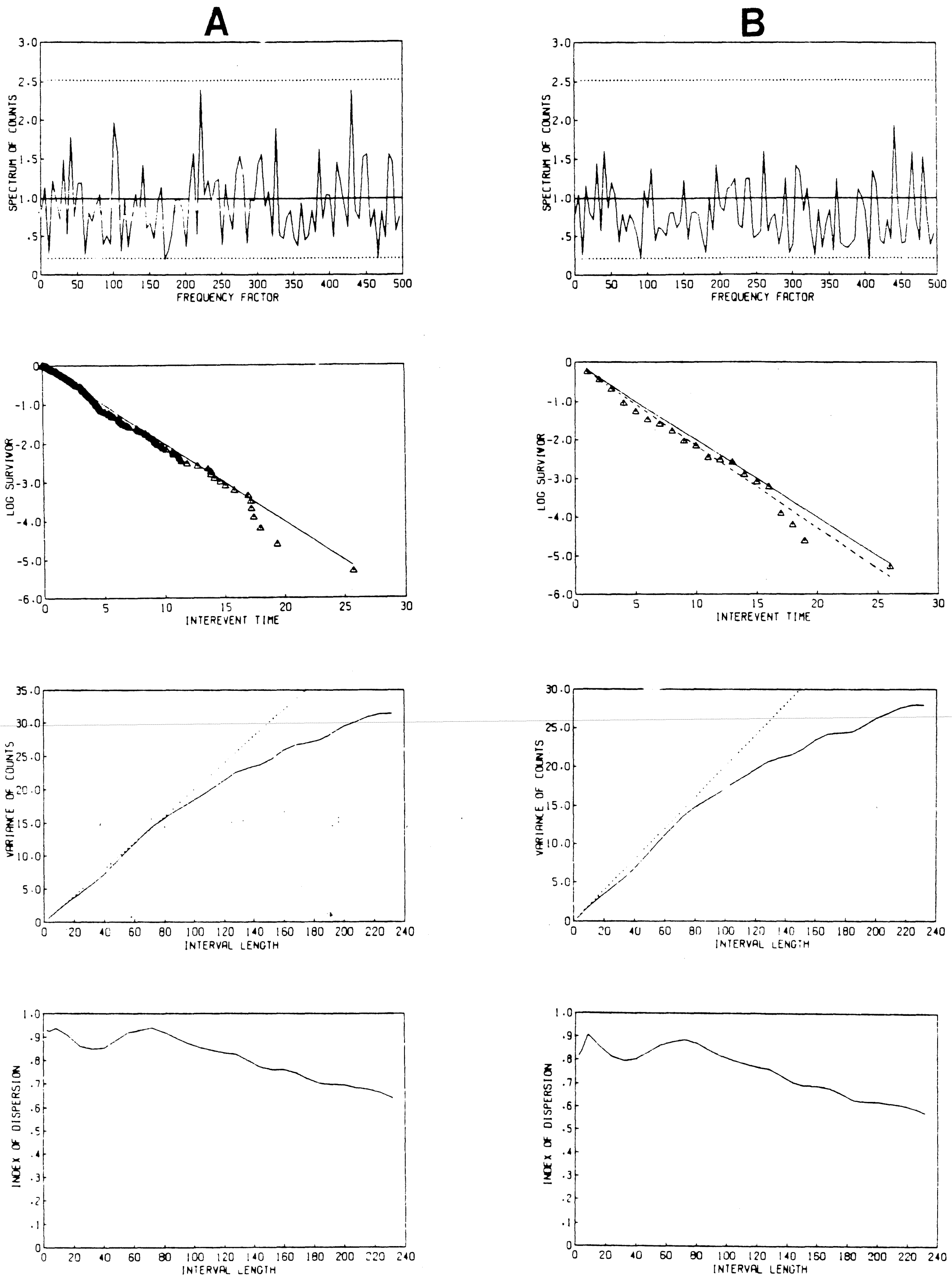


Figure 3.2 Effects of discretization on a Poisson process with rate of occurrence $\lambda = 0.20$.
(A) Continuous process, (B) Discretized process.

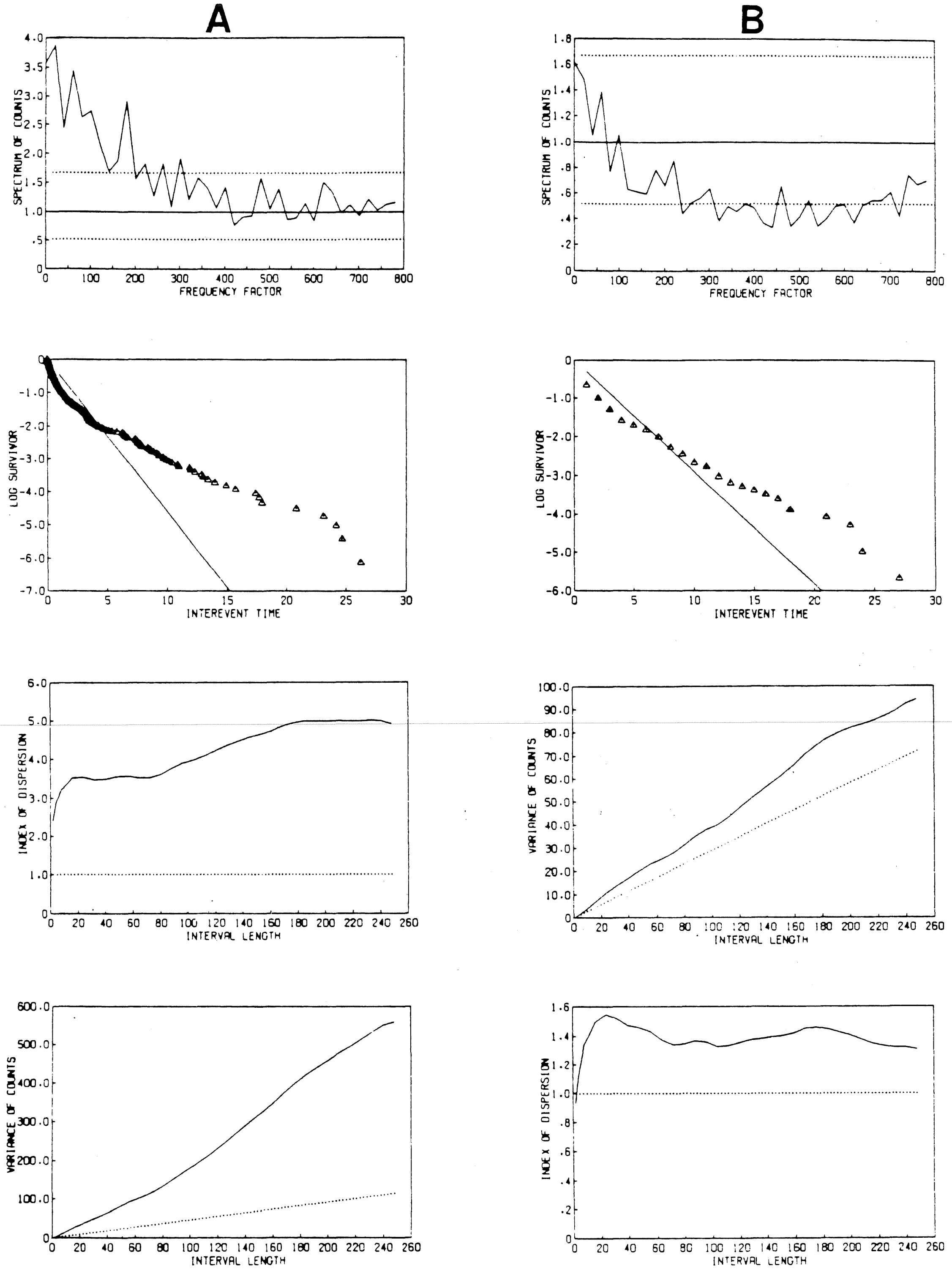


Figure 3.3 Effects of discretization on a Neyman-Scott process with parameters $h_0 = 0.20$, $p = 0.40$, and $\theta = 0.70$. (A) Continuous process, (B) Discretized process.

as for a discretized Poisson process, it is still apparent that the effects are similar and are greater the more clustered the process is.

3.4 Implications on Modeling Daily Rainfall

Recently, several authors have had apparent success with the application of time-continuous point process models to daily rainfall observation sequences. In this chapter we have shown that the practice of using continuous point process models for discrete observation sequences can give misleading results regarding inferences about over- and under-dispersion of the process and, therefore, incorrect conclusions about the underlying rainfall generating mechanism. Moreover, continuous point process models cannot be used for generation of daily rainfall sequences, which in many cases may be the purpose of modeling rainfall in the first place.

A discrete point process modeling approach which uses the Bernoulli process (the discrete analogue of the Poisson process) as its basis for comparison has been suggested. Inferences about clustering (over- and under-dispersion) in daily rainfall should, therefore, be made by comparing the empirical properties of the process to those of the Bernoulli and not to those of the Poisson, as has usually been the practice.

In the next chapter, six daily rainfall time series from stations throughout the U.S. are analyzed to give further insight into the structure of daily rainfall occurrence processes. On the basis of the preliminary theoretical analysis given in this chapter and the results of the data analysis, the inappropriateness of the continuous point process models for daily rainfall is conclusively demonstrated.

CHAPTER 4
AN EXPLORATION OF DAILY RAINFALL STRUCTURES

The Third approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a Snake!"

Previous studies on point-process modeling of daily rainfall occurrences have been confined to the analysis of a single season within which the process has been assumed homogeneous (i.e., stationary), and/or to the analysis of stations with similar probabilistic structures, i.e., stations from particular geographic regions. For example, Kavvas and Delleur (1975) analyzed seventeen daily rainfall records, all from Indiana, and applied a homogenization scheme to cope with trends and seasonality over the seven-year period studied. A time varying function, $\lambda(t)$, was fit to the mean rate of occurrence:

$$\lambda(t) = \exp \{\alpha_1 + \alpha_2 t + \alpha_3 t^2 + \sum_{i=1}^{r} R_i \sin(\omega_i t + \theta_i)\} . \qquad (4.1)$$

Under the Poisson hypothesis, the original time increments, Δt, were rescaled to $\Delta\tau = \lambda(t)\Delta t$, where $\Delta\tau$ is referred to as the intrinsic time scale. In eq. (4.1), α_1, α_2, and α_3 are parameters to model the long term trends; R_i, ω_i, and θ_i are respectively the amplitude, frequency and phase angle of the i'th significant periodicity (Kavvas and Delleur, 1975). It should be noted, however, that this homogenization scheme removes the nonstationarity only from the first moment for a non-Poisson process, such as daily rainfall, and not from higher moments. In addition, long term trends and periodicities identified from seven years of data cannot reasonably be extrapolated, and therefore the model is limited to analysis, rather than generation of synthetic sequences. Smith and Karr (1983) analyzed the summer season (July to October) rainfall occurrences for seven stations in the Potomac river basin. Twenty-seven years of daily rainfall occurrences for Denver, Colorado, were analyzed by Ramirez-Rodriguez and Bras (1982) for the period May 15 to September 11, and by Rodriguez-Iturbe et. al. (1984) for the period May 15 to June 16.

All of these studies found that the daily rainfall occurrence process is overdispersed relative to the Poisson process (i.e., the clustering of events is more random than in a Poisson process) and these results have formed the basis for applications of continuous cluster models, such as the Neyman-Scott model, discussed in Chapter 2. In this chapter, it will be shown, using six records of daily precipitation from sites throughout the continental U.S., that: (1) the daily rainfall occurrence process during many seasons of several sites is actually underdispersed relative to the Poisson, and (2) more importantly, as shown in Chapter 3, the proper basis for comparison is the (discrete) Bernoulli process, with respect to which the rainfall occurrence process is overdispersed. Moreover, the analysis presented in this chapter shows that the structure of the daily precipitation has strong seasonal variations; in many cases the season-to-season variation in model structure (as opposed to model parameters) is as significant as site-to-site climatic effects.

4.1 Selection, Description, and History of the Stations Analyzed

Six U.S. stations were selected for the analysis. These stations are located in regions of different climatologic regimes and exhibit widely different rainfall structures. Figure 4.1 shows the distributions of the regional monthly depths and the location of the stations. The station locations are sufficiently diverse to represent the major climatic types within the continental U.S.

Additional information on the stations is given in Table 4.1. The effect of the time of observation on the daily rainfall structures is not thought to be a serious problem. A different time of observation might have resulted in a different daily rainfall sequence, but the rainfall statistics (i.e., sequence of wet and dry days) are not likely to be much different as long as there is no significant diurnal periodicity in rainfall. Of course, the division of a storm into two when the observation time falls within its duration is a problem, but this is inherent to any discretized sequence of a continuous process.

Major changes in the location of the recording gages, measurement equipment, time of observation etc. could introduce artificial trends in the recorded sequences. Therefore, an inspection of the history of the analyzed stations was performed. All the changes reported in the National Oceanic and Atmospheric Administration (NOAA) Climatological Data Publications for the six stations of interest are shown in Table 4.2. Apart from a major change of 225 ft. in elevation for the station at Roosevelt, Arizona, the other changes do not seem

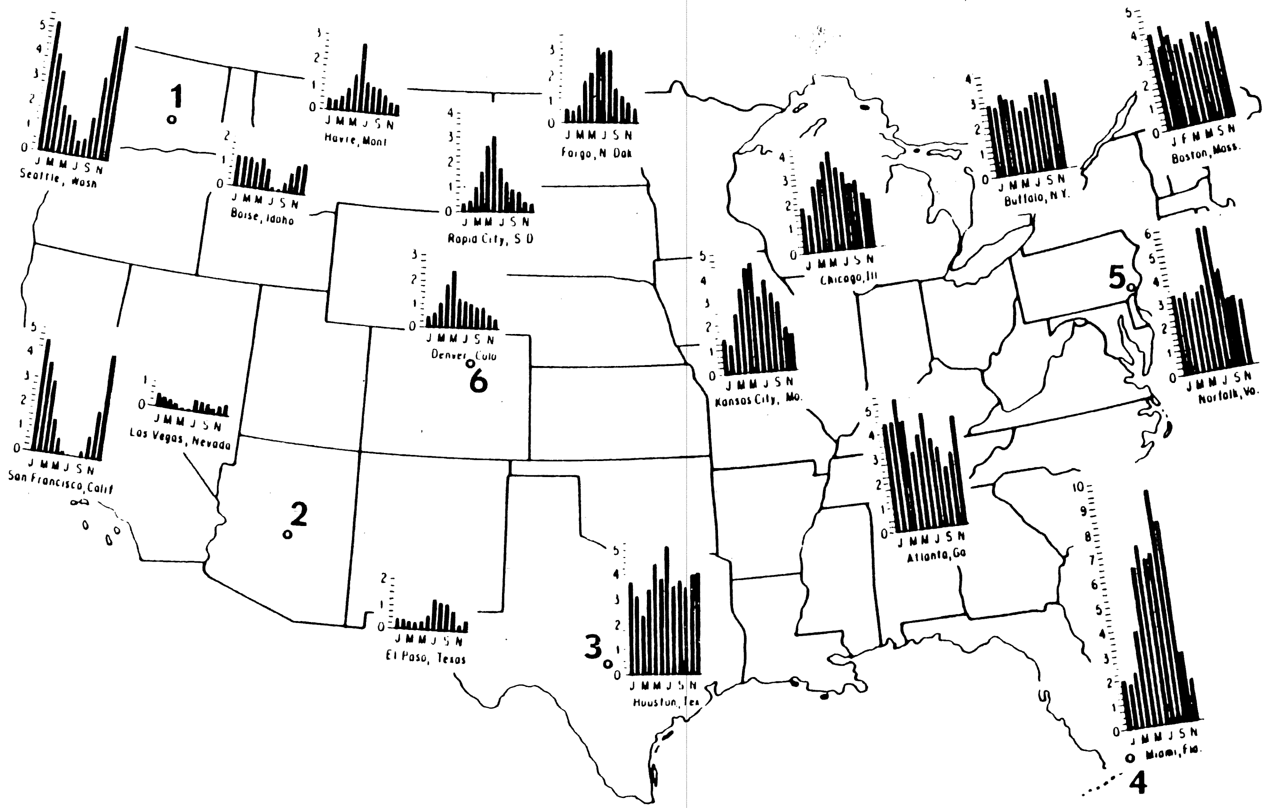


Figure 4.1 The six stations selected for the statistical analysis of their daily rainfall structures.

1. Snoqualmie Falls, Washington
2. Roosevelt, Arizona
3. Austin, Texas
4. Miami, Florida
5. Philadelphia, Pennsylvania
6. Denver, Colorado

Table 4.1 Information on the Six Daily Rainfall Stations Analyzed

Station Name	Station ID	Years Analyzed	Latitude	Longitude	Elevation (ft)	Observation Time
Snoqualmie Falls	45-7773	1948-1977	47 33	121 51	440	5 pm
Roosevelt	02-7281	1948-1977	33 40	111 09	2005	7 am
Austin, Ap	41-0428	1948-1977	30 18	97 42	597	midnight
Miami, Ap	08-5663	1949-1978	25 48	80 16	12	midnight
Philadelphia, Ap	36-6889	1948-1977	39 53	75 15	10	midnight
Denver, Ap	05-2220	1949-1978	39 46	104 52	5286	midnight

Table 4.2 History of the Six Rainfall Stations Analyzed

1. Snoqualmie Falls, Washington (ID: 45-7773)

February 1953:	Latitude from 47° 31' to 47° 33'
February 1958:	Elevation from 430 ft. to 440 ft.
April 1967:	Observation time from 5pm to midnight

2. Roosevelt, Arizona (ID: 02-7281)

July 1954:	Observation time from 7am to 8am
October 1961:	Elevation from 2230 ft. to 2005 ft.
November 1979:	Observation time from 8am to 7am

3. Austin WSO Ap., Texas (ID: 41-0428)

July 1961:	Elevation from 615 ft. to 597 ft.
January 1970:	Equipment from weighing to recording

4. Miami WSO Ap., Florida (ID: 08-5663)

June 1958:	Latitude from 25° 49' to 25° 48' Longitude from 80° 17' to 80° 16' Elevation from 8 ft. to 7 ft.
May 1977:	Longitude from 80° 16' to 80° 18' Elevation from 7 ft. to 12 ft.

5. Philadelphia WSO Ap., Pennsylvania (ID: 36-6889)

October 1953:	Elevation from 20 ft. to 13 ft.
January 1958:	Longitude from 75 14 to 75 15 Elevation from 13 ft. to 7 ft.
May 1965:	Elevation from 7 ft. to 5 ft.
April 1976:	Elevation from 5 ft. to 10 ft.

6. Denver WSO Ap., Colorado (ID: 05-2220)

March 1981:	Elevation from 5298 ft. to 5292 ft.
June 1963:	Elevation from 5292 ft. to 5283 ft.
January 1970:	Latitude from 39° 45' to 39° 46' Longitude from 104° 53' to 104° 52'

significant enough to have had major effects in the measured rainfall. It should be noted that none of the six stations is located in a major downtown metropolitan area so the direct effects of urbanization should not have been significant.

To verify the stationarity of the records, a graphical trend analysis on both the occurrence and amounts processes was conducted. Figures A.7-A.12 of Appendix A show plots of the total number of rainy days over a year and the total annual rainfall amounts as functions of the year for all the six stations. No significant trends in either the occurrence or the amounts process are apparent from this graphical analysis. Formal statistical tests, such as Cramer's statistic (Cramer, 1946) for a trend in the rate of occurrence of events, were not applied since these tests require a Poisson hypothesis and their performance is unknown when the true process is clustered.

4.2 Statistical Analysis of Daily Rainfall Sequences

Thirty years of daily rainfall during the period 1948 to 1977 (1949-1978 for Miami) were analyzed for the six stations shown in Table 4.1. The statistical properties of the occurrence processes (i.e., dependence structure and first- and second-order properties of the non-zero precipitation sequences) were estimated from the daily rainfall data. In addition, the cross-correlation functions of the amounts with the preceding and following interarrival times were estimated.

4.2.1 Seasonality of Daily Rainfall Sequences

The daily rainfall process is a non-stationary (periodic) process for both the rate of occurrence of events and the daily amounts. Therefore, a time-varying model is needed to accommodate this non-stationarity. Apart from the case of a simple Poisson model with time-varying rate of occurrence, the generalization of other models, such as Poisson cluster models and doubly stochastic Poisson models, is in most cases mathematically intractable (Cox and Lewis, 1978; Srinivasan, 1974; and others). Use of a homogenization scheme, such as that of Kavvas and Delleur (1975), to transform the data prior to the data analysis is rejected for two reasons: (1) homogenization schemes are based on the Poisson hypothesis and therefore remove the non-stationarity only from the first moment, and (2) the inverse transformation is not valid for a non-Poisson process and therefore the model cannot be used for generation purposes. Hence, it seems that the best approach is to model the daily rainfall process by seasons within which the process is assumed homogeneous. This approach has been followed herein.

The transient effects caused by crossing from one season to the next are neglected in this formulation. For the formation of the daily rainfall occurrence series, a dry period (i.e., an uninterrupted sequence of dry days) was assigned to the month or season in which it started, regardless of the ending month or season. In other words, if the last rainy day in July was on July 25, and the next rainy day was on August 10, a dry period of 16 days was assigned to the month of July. This is believed to be the most natural approach to handle the transient effects from season to season. Other workers (e.g., Chang et al., 1984) have used an abrupt transition between seasons; for the above example, their approach would have assigned a dry sequence of 6 days to July and a dry sequence of 10 days to August.

An important issue in modeling the non-homogeneous daily rainfall structure is the selection of seasons. One approach would be to consider each month as a separate season. However, grouping the data into seasons of more than one month each is desirable for several reasons: (1) the sample size of data for the estimation of the model parameters is increased; (2) the representation of the process is more parsimonious; (3) transient effects from season to season are reduced; and (4) computational effort is reduced. An alternate approach is to separate the year into seasons of equal length based on preliminary statistical analysis of the average number of events per month, average storm depths, and other summary statistics. However, the proper selection of seasons depends not only on the number of events in a given period, but also on the distribution of events within that period. Second-order properties of counts, which provide information about the distribution of events, should therefore be used in season identification procedures. A season discrimination methodology, based on all the statistical properties of intervals and counts, will be discussed and implemented in Chapter 7. The first step, however, is the selection of a small time period (i.e., a few days or one month) over which the process can be safely assumed homogeneous. The statistical properties of the process within each of these small periods are then analyzed and compared so that those periods with similar statistical structures can be grouped together. Unless predictable climatic changes are known to occur within a month, a monthly period can usually be assumed homogeneous. In this work, a month by month statistical analysis has been carried out as a first step for all six stations. Based on the results of this analysis, longer homogeneous periods (seasons) are identified in Chapter 7.

4.2.2 Second-Order Properties of Intervals and Counts

The sequence of interarrival times (times between events) is a discrete positive valued sequence, whose dependence structure and marginal pdf are to be identified. Table A.1 gives the first five autocorrelation coefficients of the interarrival time sequences for the six stations. An approximate test for their significance results from assuming that the autocorrelation coefficient, ρ_j, has a $N(\mu,\sigma^2) = N(0, 1/(n-j))$ distribution. Lewis et al. (1969) comment that this test is applicable "provided that the marginal distribution of intervals is not too highly skewed and that the number of events is greater than 100." Using this test, the significance (at the 5 percent and 1 percent levels) of the autocorrelation coefficients has been tested and the results are shown in Table A.1. Only a few autocorrelation coefficients were significant. However, this test is weak for skewed data and not directly appropriate for discrete time series. Non-parametric tests, such as exponential-score product moment statistics (Cox and Lewis, 1978), are particularly useful for short and highly skewed series, but due to the problem of ties in the series of interarrival times, they are difficult to apply.

Another way of testing independence of discrete data could be to use tests for independence in Markov chains (Billingsley, 1961). For example, Cox and Lewis (1978, p.177) present a case of a discrete point process, where the standard test on the autocorrelation function failed to indicate significant dependence in the series of intervals, whereas significant dependencies were identified from a contingency table of conditional transition probabilities. For the daily rainfall occurrences, informal tests have indicated significant autocorrelation structures. These tests consist of comparing conditional probabilities of transition to intervals of lengths ℓ_i, i=1,2,..., from intervals of a particular length ℓ, $p(\ell_i/\ell)$, i.e., p(1/1), p(2/1), p(3/1), etc. These conditional probabilities should not be significantly different for an independent process; however, this was not the case for the daily rainfall sequences. Therefore, in general it was concluded that the daily rainfall occurrences at the stations analyzed are not generated from an independent Bernoulli process.

Table A.2 shows the mean, variance, coefficient of variation, and skewness coefficient of the interarrival times for all six stations. The coefficient of variation is not always greater than one (recall that values less than one imply a process underdispersed relative to the Poisson). In particular, the winter months (October - February) for Snoqualmie Falls, the summer months (May, June) for Roosevelt, the summer months (June - September) for Miami, and most of the months (January - April, June, July, November, and December) for Philadelphia have a coefficient of variation less than unity. Therefore, for all these months, the Poisson cluster models and the renewal Cox models are precluded since both have a coefficient of variation of intervals greater than one.

Figures A.7-A.12 of Appendix A, show the empirical normalized spectrum of counts, log-survivor, variance time curve and index of dispersion functions on a monthly basis for the six stations studied. On the same plots, the corresponding functions for a Poisson process have been plotted. Recall from Table 3.1 of Chapter 3 that the corresponding functions for a Bernoulli process are as follows:

Normalized spectrum of counts:	Constant line of 1 for Poisson (1-m) for Bernoulli
Log survivor function:	Straight line with slope -m for Poisson ln(1-m) for Bernoulli
Variance time curve:	Straight line with slope m for Poisson m(1-m) for Bernoulli
Index of dispersion:	Constant line of 1 for Poisson (1-m) for Bernoulli

where m is the estimated rate of occurrences of the process. In view of the above and the discussion in Chapter 3, interpretations of these functions, i.e., clustering (over- or under-dispersion) relative to the Poisson and Bernoulli processes is possible. Consider, for example, the month of January for Snoqualmie Falls. The theory of continuous point processes would infer that this process is underdispersed relative to the Poisson, implying that events occur more regularly than in a Poisson process. However, these functions show that the process is overdispersed relative to the Bernoulli process, that is, rainfall events occur more randomly than in an independent discrete point process. The above example illustrates the inappropriateness of the continuous point process theory for modeling the discrete daily rainfall occurrences.

4.2.3 Second-Order Properties of the Rainfall Amounts

The sequence of non-zero rainfall amounts is a continuous positive time series whose autocorrelation structure and marginal pdf are to be identified. Table A.3 gives the first five autocorrelation coefficients of these sequences for all twelve months for the six stations analyzed. Only a few were significantly different from zero, for example, the first autocorrelation coefficient of the winter months (December through April) for Snoqualmie Falls. It should be noted that due to the non-normality of these sequences, the standard ARMA-type models cannot be used. Depending on the marginal pdf's, either a normalization transformation may be applied on the data and standard ARMA models be used or the exponential ARMA (EARMA) or Gamma AR (GAR) models of Lawrance and Lewis (1977) may be used directly. More references on the EARMA and GAR models have been given in Chapter 2.

Table A.4 gives the statistics of the storm depth sequences. The coefficient of variation is always greater than one, and varies from 1.07 to 1.30 for Snoqualmie Falls, 1.12 to 1.58 for Roosevelt, 1.36 to 1.78 for Austin, 1.30 to 2.23 for Miami, and 1.12 to 1.55 for Philadelphia, and 1.07 to 1.88 for Denver.

4.2.4 Cross-Correlational Properties of Intervals and Amounts

The cross-correlation coefficients of the event rainfall amounts with the interarrival times preceding and following that event are given in Table A.4 for all six stations. Only Snoqualmie Falls has a significant cross correlation between the daily rainfall amounts and the immediately following interarrival time for the months of July through January. Except for occasional exceptions, the other stations do not show significant cross dependence structure.

4.3 Discussion on the Second-Order Properties of Intervals and Counts

In this section a more detailed discussion is given of the statistical properties shown in Figures A.7-A.12 (i.e., spectrum of counts, log-survivor function, variance time curve and index of dispersion). The last two functions are straightforward. The only point to be made is that their values at high lags (i.e., long interval lengths) are of interest, since these values will better depict the deviations of the process from an independent (Poisson or Bernoulli) process. For small time intervals (i.e., a few days) many processes appear to be independent (local independence). However, the estimation of these functions should not extend to more than about 20-25 percent of the length, T_0, of the observed series to avoid excessive sampling variability. Cox and Lewis (1978, p. 116) discuss several estimators for the variance time curve, as well as sampling properties of these estimators.

The log survivor function has been plotted in discrete time to illustrate the discreteness of the interarrival times. For example, for an interarrival time, $x = x_0$, multiple points (triangles) are shown on the plot to illustrate the number of ties, i.e., number of intervals of length x_0. To interpret the log-survivor function, i.e., concavity or convexity and slope, only the lowermost points (triangles) at each entry are needed. Also, the full length of interarrival times has been retained to illustrate extreme situations. These extreme points, however, are less reliable and should be given less weight when the log-survivor function is used for model fitting.

The spectrum of counts needs special attention because of the discreteness of the data. For a continuous point process where, theoretically at least, events can occur arbitrarily close to each other, the spectrum of counts extends to infinite frequencies $\omega > 0$. For the daily rainfall occurrences, however, events cannot occur closer than one day apart and this introduces a cutoff frequency (Nyquist frequency), $\omega_N = \pi$, or equivalently $f_N = 1/2$ days^{-1}. The value plotted in the abscissa of the spectrum of counts plots is called the frequency factor and is defined as $j = \omega T/2\pi$ where T is the total length of observation. Therefore, the frequency factor corresponding to the Nyquist frequency is $j_N = T/2$. This is the maximum value over which the spectrum of counts should be computed. Guttorp and Thompson (1983) discuss aliasing of the spectrum of counts estimated from discrete sampled counting processes. They show that this can be severe, especially when the spectrum of counts does not decrease rapidly with respect to the sampling interval. For the daily rainfall occurrences, the estimated spectrum of counts began rising at high frequencies, apparently due to aliasing, but the effects of aliasing introduced into lower frequencies cannot be easily assessed.

Lewis (1970) gives a useful discussion of the theory, computation and application of the spectrum of counts. For a discrete point process he proposes a different estimator for the spectrum of counts. This estimator is based on the Fourier transform of the autocorrelation function of the binary series of zeros and ones. The reader is also referred to Bartlett (1963) for a lengthy discussion of the spectral analysis of point processes. Sampling properties of the spectral estimates are also given in the above papers and in Cox and Lewis (1978, p.126).

Notice that the normalized spectra of counts for most of the months decrease with increasing frequency to a value less than one and approximately equal to $1-\lambda$, where λ is the estimated rate of occurrence. For the months that have coefficients of variation less than one, the spectrum of counts is either approximately constant (indicating an independent Bernoulli process) or increases slightly for low frequencies and then decreases. Such spectra of counts are usually consistent with variance time curves below that of the Poisson process. This indicates underdispersion relative to Poisson. However, most of these structures are overdispersed relative to the Bernoulli, since the variance time curve of the Bernoulli process has a slope equal to $\lambda(1-\lambda) < \lambda$.

It should be noted that inferences about clustering require only the shape of the spectrum (decreasing or constant, etc.) and not the absolute values. In that sense, the clustering found in daily rainfall occurrences is still valid, the only change is that clustering should be assessed relative to a discrete independent Bernoulli point process rather than the continuous Poisson.

4.4 Need for a Discrete Clustered Point Process Model for Daily Rainfall

This chapter, together with Chapter 3, has demonstrated that continuous point process models are not appropriate for modeling daily rainfall sequences. In addition, using the proposed discrete-time point process methodology, it has been shown that, indeed, the daily rainfall process is a clustered overdispersed process i.e., the rainfall events tend to occur more randomly than in an independent arrival process. Therefore, the need for discrete clustered point process models for daily rainfall sequences has become apparent. The development of such a model is the subject of the next chapter.

CHAPTER 5
DEVELOPMENT OF DISCRETE POINT PROCESS MODELS FOR THE DAILY RAINFALL OCCURRENCES

> The Fourth reached out an eager hand,
> And felt about the knee.
> "What most this wondrous beast is like
> Is mighty plain," quoth he:
> "'Tis clear enough the Elephant
> Is very like a tree!"

In this chapter, a discrete clustered point process model is defined and developed. The model belongs to the general class of Markov renewal models which were introduced by Smith (1955), and later studied by Pyke (1961 a,b) and Cox (1963). An extensive bibliography of theoretical developments and applications of the Markov renewal models is given by Teugels (1976). In the words of Çinlar (1975), a Markov renewal process can be pictured as follows: "Think of a particle which moves from one state to another with random sojourn times in between; the successive states visited form a Markov chain, and a sojourn time has a distribution which depends on the state being visited as well as the next state to be entered" (Çinlar, 1975, p. 313). In the most general Markov renewal process with k states, it is assumed that there are k^2 different type of intervals (sojourn times), independently distributed with probability distributions $f_{ij}(x)$ $(i,j = 1, \ldots, k)$, which are sampled in accordance with a Markov chain with transition probability matrix $\underline{P}$. Thus, if the Markov chain has made a transition to state i and the next transition is to state j, an event of probability p_{ij}, then the time between these transitions has probability distribution $f_{ij}(x)$.

Markov renewal theory combines elements of Markov chain theory and renewal theory to give more general non-Markovian, non-renewal processes. It will be seen later that Markov chains, Markov processes, renewal processes and alternating renewal processes are all special cases of a general Markov renewal process. It should be made clear that the times between events are not independent as the term renewal implies, but instead are conditionally independent. This conditional independence gives a limited Markov property to the process, in the sense that the future of the process is independent of the past given the present state, provided that the present is an occurrence of an event. The above statement is an informal definition of the Markov renewal process. A formal definition follows:

DEFINITION: For each $n \varepsilon N$, let a random variable S_n take values in a countable set of states $E = \{1,2, \ldots\}$, and a random variable T_n take values in $R_+ = [0,+\infty)$ such that $0 = T_0 \leq T_1 \leq T_2 \leq \ldots$. The stochastic process $(S,T) = \{S_n, T_n ; n \quad N\}$ is said to be a Markov renewal process with state space E provided that

$$P\{S_{n+1} = j , T_{n+1} - T_n < t \mid S_0, \ldots, S_n ; T_0, \ldots, T_n\}$$
$$= P\{S_{n+1} = j , T_{n+1} - T_n < t \mid S_n\}$$

for all $n \varepsilon N$, $j \varepsilon E$, and $t \varepsilon R_+$ (Çinlar, 1975, p. 313).

Many authors (for example, Cox and Lewis, 1978; Cox and Isham, 1980) refer to the Markov renewal processes as semi-Markov processes, while others make a distinction between the two terms and reserve the term semi-Markov for the state of the process as a function of time (see for example Çinlar, 1975, p. 316). In this work, the terms Markov renewal and semi-Markov refer to the same process (defined above) and are used interchangeably, with preference on the term semi-Markov.

We will consider here only the case where there are two types of intervals i.e., a two-state semi-Markov process. In the next section the definition and statistical properties of a general two-state semi-Markov process are presented. In section 5.2, a specific discrete semi-Markov model for the daily rainfall occurrences is defined and its statistical properties derived.

5.1 Statistical Properties of a General Two-State Semi-Markov Model

In a two-state semi-Markov model it is assumed that there are two types of intervals sampled from two different probability distributions, $f_1(x)$ and $f_2(x)$, according to a probability transition matrix:

$$\underline{P} = \begin{pmatrix} a_1 & 1-a_1 \\ 1-a_2 & a_2 \end{pmatrix} , \quad 0 \leq a_1, a_2 \leq 1. \tag{5.1}$$

In equation (5.1),

$$a_1 = \text{prob(type 1 interval} \rightarrow \text{type 1 interval)},$$
$$a_2 = \text{prob(type 2 interval} \rightarrow \text{type 2 interval)},$$

or alternatively, given that the interval x_{i-1} has the pdf $f_1(x)$, the probability that x_i has the pdf $f_2(x)$ is $1-a_1$, etc. Notice that if all the intervals take the same constant value with probability one, then a one-state semi-Markov model reduces to a Markov chain. In other words, the semi-Markov process can be viewed as a generalization of a Markov chain process in which the time spent in a particular state between transitions is no longer geometrically distributed.

Let the row vector $\underline{R}^{(n)} = (p_1^{(n)}, p_2^{(n)})$ denote the probability that the n'th interval will be of type 1 or type 2 when the initial probability of the first interval being of type 1 or type 2 is given by $\underline{R}^{(0)} = (p_1^{(0)}, p_2^{(0)})$. It can be easily shown (Cox and Miller, 1965) that

$$\underline{R}^{(n)} = \underline{R}^{(n-1)}\ \underline{P} = \underline{R}^{(n-2)}\ \underline{P}^2 = \ldots = \underline{R}^{(0)}\underline{P}^n. \tag{5.2}$$

Thus, given the initial probabilities $\underline{R}^{(0)}$ and the transition probability matrix $\underline{P}$, the probability that the n'th interval will be of type 1 or type 2 can be found. The matrix $\underline{P}^n$ is called the n-step transition probability matrix, and the probabilities $p_1^{(i)}$, $p_2^{(i)}$ are called interval-type probabilities (in contrast to the state occupation probabilities of a Markov chain process).

After a sufficiently long period of time, the system settles down to a condition of statistical equilibrium in which the interval-type probabilities are independent of the initial conditions. Then, there is an equilibrium probability distribution $\underline{e} = (e_1, e_2)$, which, letting $n \to \infty$ in (5.2), satisfies

$$\underline{e} = \underline{e}\ \underline{P}. \tag{5.3}$$

The solution of (5.3) with respect to the row vector $\underline{e}$, subject to the constraint $e_1+e_2 = 1$, e_1, $e_2 \geq 0$, gives the equilibrium interval-type probabilities associated with the transition probability matrix $\underline{P}$ of (5.1) as

$$e_1 = \frac{1-a_2}{2-a_1-a_2}, \quad e_2 = \frac{1-a_1}{2-a_1-a_2}, \tag{5.4}$$

(e.g., Cox and Miller, 1965). For instance, the probability e_1 is the unconditional probability that an arbitrary interval will be of type 1. Note that

$$e_1 + e_2 = 1. \tag{5.5}$$

From the theory of Markov chains we know that

$$\underline{P}^{(n)} = \begin{bmatrix} e_1 & e_2 \\ e_1 & e_2 \end{bmatrix} + (a_1+a_2-1)^n \begin{bmatrix} e_2 & -e_2 \\ -e_1 & e_1 \end{bmatrix} \longrightarrow \begin{bmatrix} e_1 & e_2 \\ e_1 & e_2 \end{bmatrix},$$

and therefore from (5.2)

$$\underline{R}^{(n)} \dashrightarrow \underline{R}^{(0)} \begin{bmatrix} e_1 & e_2 \\ e_1 & e_2 \end{bmatrix} = (e_1, e_2) = \underline{e},$$

so that the system tends to a statistical equilibrium with a rate depending on the value of $(a_1 + a_2 - 1)^n$ which tends to zero as n increases (Cox and Miller, 1965). The value of $(a_1 + a_2 - 1)$ is less than unity in modulus, except in the trivial cases (i) $a_1 + a_2 = 0$, i.e., $a_1 = 0$, $a_2 = 0$ in which the system alternates deterministically between the two states and (if the initial state is given the behavior of the system is non-random), and (ii) $a_1 + a_2 = 2$, i.e., $a_1 = 1$, $a_2 = 1$ in which the system remains forever in its initial state. For $a_1 + a_2 = 1$ the process is a renewal process, and the transition probabilities of the Markov chain are equal to the equilibrium probabilities, i.e., $a_1 = e_1$ and $a_2 = e_2$.

5.1.1 Interval Properties

The pdf of the intervals of the process is given as

$$f(x) = e_1 f_1(x) + e_2 f_2(x), \tag{5.6}$$

where e_1 and e_2 are the equilibrium probabilities given in (5.4). It is easy to show that the mean, variance, and survivor function of the interarrival times x are given as

$$E(X) = e_1\mu_1 + e_2\mu_2, \tag{5.7a}$$

$$\mathrm{var}(X) = e_1\sigma_1^2 + e_2\sigma_2^2 + e_1e_2(\mu_1-\mu_2)^2, \tag{5.7b}$$

$$R(x) = e_1R_1(x) + e_2R_2(x), \tag{5.7c}$$

where μ_i, σ_i^2, i = 1,2, are the means and variances, respectively, and the subscripts indicate the types of the intervals. The autocovariance function of a two-state semi-Markov process can be shown to be

$$\mathrm{Cov}(X_i, X_{i+k}) = (\mu_1-\mu_2)^2 e_1 e_2 \beta^k \ , \ k = 1,2,\ldots, \tag{5.8}$$

where

$$\beta = a_1 + a_2 - 1 \tag{5.8a}$$

and therefore the autocorrelation function can be written as

$$\rho_k = \frac{(\mu_1-\mu_2)^2 \ e_1 e_2}{e_1\sigma_1^2 + e_2\sigma_2^2 + e_1 e_2(\mu_1-\mu_2)^2}\beta^k \tag{5.9}$$

$$= c\beta^k,$$

in which

$$c \equiv \frac{(\mu_1-\mu_2)^2 \ e_1 e_2}{e_1\sigma_1^2 + e_2\sigma_2^2 + e_1 e_2(\mu_1-\mu_2)^2}. \tag{5.9a}$$

Consequently, the spectral density function of the intervals, given in terms of ρ_k is

$$f_+(\omega) = \frac{1}{\pi} \{1 + 2 \sum_{k=1}^{\infty} \rho_k \cos(k\omega)\} \ , \ 0 \leq \omega \leq \pi, \tag{5.10}$$

which takes the form

$$f_+(\omega) = \frac{1}{\pi} \{1 + 2c \frac{\beta\cos\omega - \beta^2}{1 + \beta^2 - 2\beta\cos\omega} \}, \tag{5.11}$$

where c is defined in (5.9a) and β in (5.8a).

Notice that the autocovariance function of the intervals of a two-state semi-Markov model and therefore the power spectrum depend on the pdf's $f_1(x)$ and $f_2(x)$ only through their means and variances. This can provide a helpful first check for the appropriateness of a semi-Markov structure for a series of events, since no assumption about the pdf's of the intervals is required.

5.1.2 Counts Properties

Cox (1963) first showed that the Laplace transform of the conditional intensity function of a two-state semi-Markov model is given as

$$h^*(s) = \frac{e_1 f_1^*(s) + e_2 f_2^*(s) + (1 - a_1 - a_2)\, f_1^*(s) f_2^*(s)}{1 - a_1 f_1^*(s) - a_2 f_2^*(s) - (1 - a_1 - a_2)\, f_1^*(s) f_2^*(s)}, \qquad (5.12)$$

where $f_1^*(s)$ and $f_2^*(s)$ are the Laplace transforms of the pdf's $f_1(x)$ and $f_2(x)$, respectively. Explicit formulae for h(t) exist whenever the inversion of (5.12) is possible. Given h(t), all the other properties of counts can be obtained from the following general relationships:

$$\gamma_+(\tau) = m[h(\tau) - m], \qquad (5.13)$$

$$H(t) = \int_0^t h(\tau)\, d\tau, \qquad (5.14)$$

$$V^*(s) = \frac{m}{s^2} + \frac{2h^*(s)m}{s^2} - \frac{2m^2}{s^3}, \qquad (5.15)$$

$$g_+(\omega) = \frac{m}{\pi} [1 + h^*(i\omega) + h^*(-i\omega)], \qquad (5.16)$$

(see Cox and Lewis, 1978, for proofs). In the above equations, the rate of occurrence, m, is given in terms of the transition probabilities as

$$m = \frac{1}{e_1\mu_1 + e_2\mu_2} = \frac{2 - a_1 - a_2}{(1 - a_2)\mu_1 + (1 - a_1)\mu_2}. \qquad (5.17)$$

5.2 A Discrete Semi-Markov Model for the Daily Rainfall Occurrences

Daily rainfall occurrences are the result of the interaction of several rainfall generating mechanisms. For example, the first rainy day in a wet period may be the result of a frontal storm passing over a region, whereas subsequent rainy days in the same wet period may be just aftereffects (secondary events). In that sense, times between events may come from different probability distributions, for instance, one with a smaller coefficient of variation for the secondary events, and one with a large coefficient of variation for the primary events. The sequence of event types is governed by transition probabilities, with higher probabilities of having secondary events after a primary event or after a small number of secondary events.

In view of this, a two-state semi-Markov model is proposed for the daily rainfall occurrences, in which the times between events are sampled from two different geometric distributions with parameters p_1 and p_2, according to a Markov chain with the transition probability matrix $\underline{P}$ of (5.1). The notation SMGG will be used to denote a two-state semi-Markov model (SM) with both type 1 and type 2 interarrival times having geometric distributions (GG). The statistical properties of intervals and counts for a SMGG process are derived below.

Let $f_1(x)$ and $f_2(x)$, defined as

$$f_i(x) = \sum_{k=1}^{\infty} p_i(1 - p_i)^{k-1} \delta(x - k) \text{ , for } i = 1,2 \quad , \tag{5.18}$$

be the continuous representations of the geometric probability mass functions (pmf) of the interarrival times. Notice that $f_1(x)$ and $f_2(x)$ are probability density functions (pdf). The statistical properties of intervals and counts of a SMGG process are given in the following propositions.

PROPOSITION 1: The moment generating function of the interarrival times of a SMGG process is given as

$$\psi(s) = \frac{1}{(2-a_1-a_2)}\left[(1-a_2)\frac{p_1e^s}{1-(1-p_1)e^s} + (1-a_1)\frac{p_2e^s}{1-(1-p_2)e^s}\right]. \tag{5.19}$$

Moments of the interarrival times are obtained from

$$E(X^k) = (-1)^k \left.\frac{d^k \psi(s)}{ds^k}\right|_{s=o}. \tag{5.20}$$

Proof: The proof follows immediately from (5.6) by noting that the moment generating function of a geometric distribution with parameter p, is given as

$$\psi(s) = \frac{pe^s}{1-(1-p)e^s}.$$

Corollaries: The mean interarrival time of a SMGG process is

$$E(X) = \frac{1}{(2-a_1-a_2)}\left[\frac{1-a_2}{p_1} + \frac{1-a_1}{p_2}\right]. \qquad (5.21)$$

The variance of the interarrival times of a SMGG process is

$$var(X)=\frac{1}{(2-a_1-a_2)}\left[(1-a_2)\frac{1-p_1}{p_1^2}+(1-a_1)\frac{1-p_2}{p_2^2}+\frac{(1-a_1)(1-a_2)}{(2-a_1-a_2)}\left(\frac{1}{p_1}-\frac{1}{p_2}\right)^2\right]. \qquad (5.22)$$

The survivor function of the interarrival times of a SMGG process is

$$R(x) = \frac{1}{(2-a_1-a_2)}\left[(1-a_2)(1-p_1)^x + (1-a_1)(1-p_2)^x\right], \quad x=1,2,\ldots \qquad (5.23)$$

Equation (5.23) follows immediately from (5.7c) by noting that the survivor function of a geometric distribution is given as $(1-p)^x$.

PROPOSITION 2: The autocorrelation function, ρ_k, of the interarrival times of a SMGG process is given as

$$\rho_k = c(a_1+a_2-1)^k, \tag{5.24}$$

where

$$c = \frac{(1-a_1)(1-a_2)}{2-a_1-a_2} \frac{(p_1-p_2)^2}{(1-a_2)(1-p_1)p_2^2+(1-a_1)(1-p_2)p_1^2+(1-a_1)(1-a_2)(p_1-p_2)^2}. \tag{5.25}$$

Proof: Eq. (5.25) follows from (5.9) after substituting the means and variances of the two geometric distributions as functions of the parameters p_1 and p_2.

Note on terminology: For a discrete point process, we introduce the term conditional probabilities of occurrence for the discrete sequence of conditional probabilities $\{h_k\}$, $k = 1,2,\ldots$, analogously to the conditional intensity function $h(t)$ of a continuous point process. The relationship between $h(t)$ and h_k is simply

$$h(t) = \sum_{k=1}^{\infty} h_k \ \delta(t-k), \tag{5.25}$$

where $\delta(\cdot)$ is the Dirac delta function. The interpretation of $\{h_k\}$ remains the same as in the continuous case; values of h_k greater than the constant (unconditional) probability of occurrence m imply a greater likelihood of having an event at time $(t+k)$ due to an event at time t. The conditional probability of occurrence sequence $\{h_k\}$, from which all the other statistical properties of the counting process can be derived, is given in the following proposition.

PROPOSITION 3: The conditional probability of occurrence sequence $\{h_k\}$ of a SMGG process is given as

$$h_k = m + AW^k \qquad , k = 1,2,\ldots, \tag{5.26}$$

where

$$A = e_1 p_1 + e_2 p_2 - m \tag{5.27}$$

and

$$W = 1-p_1(1-a_1)-p_2(1-a_2). \tag{5.28}$$

The equilibrium probabilities appearing in (5.27) are given in (5.4) and the mean intensity of the process, m, (i.e., the unconditional probability of occurrence of an event), can be given in terms of the transition probabilities and the parameters of the geometric distributions as

$$m = \frac{p_1 p_2 (2-a_1-a_2)}{p_1(1-a_1)+p_2(1-a_2)}. \tag{5.29}$$

Proof: The Laplace transform, $\overset{*}{h}(s)$, of the intensity function, h(t), is obtained by substituting in (5.12) the expressions for $f_i^*(s)$, i = 1,2 , which are the Laplace transforms of $f_i(x)$, i = 1,2, defined in (5.18), i.e.,

$$f_i^*(s) = \frac{p_i e^{-s}}{1-(1-p_i)e^{-s}}. \tag{5.30}$$

After lengthly algebraic manipulations, $\overset{*}{h}(s)$ is inverted to give h(t) whose discrete analogue is h_k of (5.26). More details on this derivation are given in Appendix B.

REMARK 1: The conditional intensity function h(t) of a SMGG process tends monotonically to the mean rate of occurrence m, as t becomes large. Specifically, h(t) decreases geometrically to the constant intensity m, since A can be shown to be positive and $0 < W < 1$. This implies that the semi-Markov process exhibits clustering. Although the shape of the conditional intensity function is only indicative of the presence, but not the type, of clustering, the fact that the coefficient of variation of the interarrival times can take values greater or less than one (see equations 5.21 and 5.22) suggests that the semi-Markov process can accommodate rainfall occurrence structures which the Neyman-Scott process and doubly stochastic Poisson processes cannot.

COROLLARY 1: The expected number of events in an interval of length t, given that the interval starts with an event, is given as

$$H(t) = mt + A \sum_{k=1}^{t} W^k$$

$$= mt + AW \frac{W^t - 1}{W-1} \quad , \; t = 1,2,\ldots, \qquad (5.32)$$

where the parameters m, A and W have been defined in (5.29), (5.27), and (5.28), respectively, and t is referring to discrete time units.

Proof of Corollary 1: The mean function H(t) of a continuous point process is defined as the integral of the conditional intensity function h(t) in (5.14). For discrete point process we can write by analogy,

$$H(t) = \sum_{k=1}^{t} h_k = \sum_{k=1}^{t} (m + AW^k), \qquad (5.33)$$

from which (5.32) follows immediately.

COROLLARY 2: The variance of the number of events in an interval of length t, V(t), where the interval starts with an event, is given as

$$V(t) = mt - m^2t^2 + 2m \sum_{k=1}^{t-1} (t-k)h_k, \qquad (5.34)$$

where h_k is given by (5.26).

Proof of Corollary 2: The Laplace transform $V^*(s)$ of the variance time curve is given in terms of the Laplace transform of the conditional intensity function by (5.15). By taking inverse Laplace transforms:

$$V(t) = m\mathcal{L}^{-1}(\frac{1}{s^2}) - 2m^2\mathcal{L}^{-1}(\frac{1}{s^3}) + 2m\mathcal{L}^{-1}[\frac{h^*(s)}{s^2}],$$

and therefore

$$V(t) = mt - m^2t^2 + 2m \int_0^t \int_0^\sigma h(\tau)\,d\tau\,d\sigma \quad ,$$

which in the discrete domain can be written as

$$V(t) = mt - m^2t^2 + 2m \sum_{i=1}^{t} \sum_{k=1}^{i} h_k$$

$$= mt - m^2t^2 + 2m \sum_{k=1}^{t-1} (t-k)h_k .$$

This completes the proof of Corollary 2.

PROPOSITION 4: The spectrum of counts, $g_+(\omega)$, of a SMGG process is given as

$$g_+(\omega) = \frac{m}{\pi} [1 - m - 2A\frac{W - \cos\omega}{1 - 2W\cos\omega + W^2}], \qquad (5.35)$$

where m, A, and W have been defined previously.

Proof: Eq. (5.35) is derived by substituting into (5.16) the expression of $h^*(s)$ from (5.12) and performing lengthy algebraic manipulations. Expression of $f_1^*(s)$ and $f_2^*(s)$ needed in (5.12) are obtained from (5.30).

5.3 Discussion

In this chapter a discrete semi-Markov model was introduced and its statistical properties derived. It was seen that the model has considerable flexibility (see Remark 1) in the sense that it can model structures with different types of clustering. It remains to explore parameter estimation methods, and to apply the model to observed daily precipitation series. In the next chapter, methods for fitting the model are studied.

CHAPTER 6
FITTING THE DISCRETE SEMI-MARKOV MODEL

The Fifth who chanced to touch the ear,
    Said: "E'en the blindest man
Can tell what this resembles most;
    Deny the fact who can,
This marvel of an Elephant
    Is very like a fan!"

The discrete semi-Markov model developed in Chapter 5 has four parameters: a_1, a_2, p_1, and p_2. These parameters are: a_1, the transition probability from type 1 to type 1 interval; a_2, the transition probability from type 2 to type 2 interval; p_1, the parameter of the geometric distribution of the type 1 intervals; and p_2 the parameter of the geometric distribution of the type 2 intervals. Note that the type 1 and type 2 intervals are in general indistinguishable from each other by direct observation of the series of daily rainfall events. Thus, the transition probabilities a_1 and a_2 cannot be estimated directly from the data, but instead have to be estimated together with the parameters of the two geometric distributions, p_1 and p_2. The estimation methods studied are the method of moments (MOM) and two approximate maximum likelihood (ML) estimation methods.

6.1 Method of Moments

The first three moments and the lag-one covariance of the interarrival times of the semi-Markov model SMGG are given as functions of the parameters a_1, a_2, p_1 and p_2 as follows:

$$E(X) = \frac{1}{2-a_1-a_2}\left[\frac{1-a_2}{p_1} + \frac{1-a_1}{p_2}\right], \tag{6.1}$$

$$E(X^2) = \frac{1}{2-a_1-a_2}\left[\frac{(1-a_2)(2-p_1)}{p_1^2} + \frac{(1-a_1)(2-p_2)}{p_2^2}\right], \tag{6.2}$$

$$E(X^3) = \frac{1}{2-a_1-a_2}\left[\frac{(1-a_2)(p_1^2- p_1+6)}{p_1^3} + \frac{(1-a_1)(p_2^2- p_2+6)}{p_2^3}\right], \tag{6.3}$$

$$c_1 = \frac{1}{(2-a_1-a_2)^2}\ (1-a_1)(1-a_2)\left(\frac{1}{p_1} - \frac{1}{p_2}\right)^2 (a_1+a_2-1). \tag{6.4}$$

The above four equations can be numerically solved for a_1, a_2, p_1, and p_2 using, for instance, the Newton-Raphson method. Since all four parameters are probabilities, they must lie inside the interval [0,1]. Therefore, a transformation was applied to these parameters to unconstrain them, and the search carried out in the unconstrained space. Denoting by y the real parameter, $y \varepsilon (\ell,u)$, and by y' the unconstrained parameter, $y' \varepsilon (-\infty,+\infty)$, the following transformation was used:

$$y = \ell + (u-\ell)\ \sin^2(y') \tag{6.5}$$

where ℓ and u are the lower and upper bounds on the parameters. The values of $\ell = 0.01$ and $u = 0.99$ were used to avoid numerical problems at the bounds.

Due to the long tail of the probability distribution of the interarrival times, the use of the third moment in the estimation is not desirable. A modified method of moments estimation which involves the median instead of the third moment, was tested. The median x_m of the pdf of the interarrival times of the semi-Markov model SMGG, is given by the following equation:

$$\frac{1}{(2-a_1-a_2)}\left[(1-a_2)(1-p_1)^{x_m} + (1-a_1)(1-p_2)^{x_m}\right] = 0.5. \qquad (6.6)$$

Modified method of moments estimates were then obtained by simply substituting (6.6) for (6.3).

6.2 Approximate Maximum Likelihood Estimates (MLE)

Let I_i, $i = 1,2,\dots n$, denote the type of the ith interval in one realization of length n of the point process. Then, $I_i \varepsilon \{I, II\}$, where I stands for type 1 interval and II for type 2 interval. Let $\underline{I}$ also denote the vector $(I_1, I_2, \dots, I_n)^T$, that is, the vector of the types of intervals of all n interarrival times of the given realization. The general form of the likelihood function of a two-state semi-Markov model can be expressed as:

$$\begin{aligned} L(x) &= \sum_{\underline{I}} p(x_n, \dots, x_1 | I_n, \dots I_1)\, p(I_n, \dots, I_1) \\ &= \sum_{\underline{I}} \underbrace{\prod_{i=1}^{n} p(x_i|I_i)}_{(A)} \; \underbrace{\prod_{i=1}^{n-1} p(I_{i+1}|I_i)\; p(I_1)}_{(B)} \end{aligned} \qquad (6.7)$$

where the summation is over all possible vectors $\underline{I}$, i.e., over all possible vectors of length n formed by the two elements I and II. Observe in the above expression that the term (A) depends only on the vector of interarrival times $\underline{x} = (x_1, x_2, \ldots, x_n)^T$ and the parameters of the two geometric distributions p_1, and p_2, while the second term (B) depends only on the transition probabilities, a_1, and a_2, of the Markov chain of intervals, i.e.,

$$L(x) = \sum_{\underline{I}} f(\underline{x}, p_1, p_2)\, f(a_1, a_2) \tag{6.8}$$

where f(·) denotes function of (·). It becomes apparent from (6.7) that the likelihood function of the semi-Markov model cannot be expressed in a tractable closed form as function of the parameters a_1, a_2, p_1, p_2 and the vector of observations $\underline{x}$. Although numerical evaluation of the likelihood function is possible, it is infeasible for typical sample sizes of several hundred values, since it requires double summations over all possible vectors $\underline{I}$.

In view of the above, an approximate maximum likelihood estimation procedure has been developed. This procedure consists of two steps. The first step involves the maximum likelihood estimation of e_1, p_1, and p_2, where e_1 is the equilibrium probability of the Markov chain of intervals. Given the equilibrium probability e_1, the transition probabilities, a_1 and a_2, of the Markov chain of intervals are subsequently obtained. It is understood that although the parameters p_1 and p_2 are exact maximum likelihood estimates, the parameters a_1 and a_2 are not and thus the method is termed approximate maximum likelihood. Details of this method follow.

The probability mass function (pmf) of the interarrival times of the semi-Markov model is given as:

$$p(x) = e_1 p_1 (1-p_1)^{x-1} + (1-e_1) p_2 (1-p_2)^{x-1} \tag{6.9}$$

where e_1 is the equilibrium probability of the Markov chain of intervals, i.e., the unconditional probability of any interval being of type 1. The log-likelihood function $L'(x)$ is

$$\begin{aligned} L'(x) &= \ln[L(x)] = \ln[\prod_{i=1}^{n} (p(x_i))] = \sum_{i=1}^{n} \ln[p(x_i)] \\ &= \sum_{i=1}^{n} \ln[e_1 p_1 (1-p_1)^{x_i - 1} + (1-e_1) p_2 (1-p_2)^{x_i - 1}]. \end{aligned} \tag{6.10}$$

Parameter estimates for e_1, p_1 and p_2 can be obtained by maximizing $L'(x)$, for instance, using the simplex method of Nelder and Mead (1965). The optimization is carried out in the unconstrained space using again the transformation (6.5). Estimates of the parameters a_1 and a_2 can subsequently be obtained by one of the two methods described below.

6.2.1 Estimation of the Transition Probabilities Using the First Autocorrelation Coefficient

The first autocorrelation coefficient of the semi-Markov model is given as

$$r_1 = c(a_1 + a_2 - 1), \tag{6.11}$$

where

$$c = \frac{e_1(1-e_1)(1/p_1 - 1/p_2)^2}{e_1\frac{1-p_1}{p_1^2} + (1-e_1)\frac{1-p_2}{p_2^2} + e_1(1-e_1)(\frac{1}{p_1} - \frac{1}{p_2})^2}, \qquad (6.12)$$

and e_1 is the equilibrium probability of the Markov chain, given in terms of the transition probabilities as

$$e_1 = (1-a_2)/(2-a_1-a_2). \qquad (6.13)$$

Equations (6.11) and (6.13) can be solved for a_1 and a_2, giving

$$a_1 = (1-e_1)(r_1/c + 1) + 2e_1 - 1$$

and (6.14)

$$a_2 = e_1(r_1/c + 1) - 2e_1 + 1.$$

From the above two equations it can be shown that for acceptable parameter estimates, that is, $0 < a_1, a_2 < 1$, the following inequality must hold:

$$-\min(\frac{e_1}{1-e_1}, \frac{1-e_1}{e_1}) < \frac{r_1}{c} < 1. \qquad (6.15)$$

Note that the value $\min(e_1/(1-e_1)\ ,\ (1-e_1)/e_1) = \min\ (e_1/e_2,\ e_2/e_1)$ corresponds to the ratio of the smallest to the largest equilibrium probability, a value always less than 1. The inequality (6.15), therefore, is consistent with the requirement that the autocorrelation function of the process, given as

$$r_k = c(a_1 + a_2 - 1)^k \tag{6.16}$$

is less than one in absolute value. Note also that the equal signs in (6.15) are not permitted since from (6.11) they can be shown to correspond to the following trivial cases. The right hand side equal sign implies $a_1 + a_2 = 2$, and therefore $a_1 = 1$ and $a_2 = 1$, in which case the system remains forever in the initial state. The left hand side equal sign implies $a_1 = 0$ and $a_2 = 0$, in which case the system alternates deterministically between the two states and given the initial state, the behavior of the system is non-random.

Therefore, estimates for a_1 and a_2 cannot be obtained by the above method, unless the ratio of the estimated first autocorrelation coefficient r_1 to the value c, satisfies (6.15). The value of c is obtained from (6.12) using the values of e_1, p_1, p_2 estimated from the approximate maximum likelihood function. It was found that (6.15) was not satisfied in general, and therefore the transition probabilities cannot always be estimated using this method.

6.2.2. Bayesian Approach to Estimation of the Transition Probabilities

Let $\xi_{1,i}$, $\xi_{2,i}$ denote the conditional probabilities of an interval having length x_i given that it is of type 1 (I) or type 2

(II), respectively. In view of the geometric distributions for the two types of intervals, these probabilities can be written as

$$\xi_{1,i} = p(x=x_i|I) = p_1(1-p_1)^{x_i-1}$$

and (6.17)

$$\xi_{2;i} = p(x=x_i|II) = p_2(1-p_2)^{x_i-1}.$$

The conditional probabilities of an interval being of type 1 or type 2 given that it has length $x=x_i$, i.e., $p(I|x=x_i)$, $p(II|x=x_i)$ can now be determined. Using Bayes theorem

$$p(I|x=x_i) = \frac{p(I,\ x=x_i)}{p(x=x_i)} = \frac{p(I)\ p(x=x_i|I)}{p(x=x_i)} = \frac{e_1\xi_{1,i}}{p(x=x_i)} \qquad (6.18)$$

and analogously,

$$p(II|x=x_i) = \frac{e_2\xi_{2,i}}{p(x=x_i)} = \frac{(1-e_1)\xi_{2,i}}{p(x=x_i)} \qquad (6.19)$$

where $\xi_{1,i}$ and $\xi_{2,i}$ are given in (6.17) as functions of the parameters p_1 and p_2, and $p(x=x_i)$ is given from (6.9) for $x=x_i$.

The transition probabilities a_1 and a_2 can be estimated as

$$a_1 = p(I|I) = \sum_{i=1}^{n-1} p(I|x_i \wedge I|x_{i+1}) \Big/ \sum_{i=1}^{n} p(I|x_i)$$

$$= \sum_{i=1}^{n-1} p(I|x_i)\ p(I|x_{i+1}) \Big/ \sum_{i=1}^{n} p(I|x_i) \qquad (6.20)$$

where $p(I|x_i)$, $p(I|x_{i+1})$ are given in (6.18), and $p(x=x_i)$, $p(x=x_{i+1})$ in (6.9).

6.3 Monte Carlo Tests of Estimators

The fitting methods considered are basically variations of method of moments (MOM) and Maximum Likelihood (ML) estimation methods. In particular, the following five methods were tested for efficiency and consistency in estimation:

M1: MOM on $E(X)$, $E(X^2)$, $E(X^3)$ and Cov(1) $\rightarrow a_1, a_2, p_1, p_2$

M2: MOM on $E(X)$, $E(X_2)$, median and Cov(1) $\rightarrow a_1, a_2, p_1, p_2$

M3: MOM on $E(X)$, $E(X_2)$, $E(X^3)$ $\rightarrow e_1, p_1, p_2$; coupled with $r_1 \rightarrow a_1, a_2$

M4: ML $\rightarrow e_1, p_1, p_2$; coupled with $r_1 \rightarrow a_1, a_2$

M5: ML $\rightarrow e_1, p_1, p_2$; Bayesian approach $\rightarrow a_1, a_2$.

Recall from Chapter 5 that depending on whether $a_1 + a_2 > 1$ (or < 1), the first and all the odd-lagged autocorrelation coefficients of the interarrival times become positive (or negative). Most of the daily rainfall structures analyzed exhibited a positive autocorrelation structure of intervals, although few significant lag-one autocorrelation coefficients were present as, for example, for the station of Denver. In view of the above, Monte Carlo tests of estimators were performed for sets of parameter values for a_1 and a_2

such as both of the above model structures are covered.

The first set of parameters tested was { $a_1 = 0.4$, $a_2 = 0.3$, $p_1 = 0.8$, $p_2 = 0.2$ } . These parameter values correspond to an occurrence process with a mean interarrival time of 2.98 days, a standard deviation of 3.59 days (coefficient of variation $c_v = 1.2$), a skewness coefficient equal to 3.01, and a first autocorrelation coefficient $r_1 = -0.08$. The conditional intensity function for these parameters takes the form $h_k = 0.335 + 0.186(0.38)^k$, which indicates a clustering of counts. Such statistics are representative of daily rainfall occurrence processes, as can be seen from Table A.2 of Appendix A. Five hundred synthetic sequences of 50, 100, 200, 500 and 800 events were generated from a semi-Markov model with the above parameters. The implied rate of occurrence of the process is $m = 1/2.98 = 0.34$ days^{-1}, and therefore these sequences correspond to approximately 150, 300, 600, 1500, and 2400 days of observation, respectively.

The five methods (M1, M2, M3, M4, and M5) discussed previously were fitted to all synthetic sequences. The bias and standard deviation of the estimated parameters are given in Table 6.1. As was expected, the consistency (bias) and efficiency (variability) of the estimators improve with the number of events available for the estimation. The best estimators in terms of root mean square error (RMSE = $((\text{bias})^2 + \text{variance})^{1/2}$) were methods M4 and M5 which are the two approximate maximum likelihood estimation methods using the first autocorrelation coefficient and a Bayesian approach, respectively. Method M4 has a low bias but a large variance as compared to method M5 which has a larger bias but a much smaller variance for the parameters

a_1 and a_2. It is also observed from Table 6.1 that in terms of RMSE method M4 is the best for large sample sizes (larger than 500 events) whereas method M5 is the best for small sample sizes. This was expected given that method M4 involves an estimate of the first autocorrelation coefficient of the intervals. In addition, method M5 always gives a fit, whereas method M4 failed in a number of cases.

The second set of parameters tested was $\{a_1 = 0.9,\ a_2 = 0.6,\ p_1 = 0.8,\ p_2 = 0.4\}$. These parameters correspond to an occurrence process with mean interarrival time 1.5 days (mean rate of occurrence, $m = 0.667$ days^{-1}), a standard deviation of 1.11 days (coefficient of variation $c_v = 0.74$), and skewness coefficient $c_s = 4.02$. The autocorrelation function of the process is $r_k = r_1(a_1 + a_2 - 1)^{k-1}$, where $r_1 = 0.1$, and $(a_1 + a_2 - 1) = 0.5$, and the conditional intensity function is $h_k = 0.667 + 0.05(0.76)^k$. These functions indicate a strong dependence stucture in the intervals but a relatively small clustering in the counts. The results of the estimation methods are shown in Table 6.2. Method M5 performed poorly, whereas method M4 gave satisfactory parameter estimates. These results suggest that method M4 may perform better when a strong autocorrelation in the interarrival times is present, and method M5 when the clustering of counts is the more significant element of dependence.

The effect of the first autocorrelation coefficient of the process on the consistency and efficiency of the estimators a_1, a_2, p_1 and p_2 was also tested. For the discussion that follows, the convention is made that e_1 corresponds to the geometric distribution

Table 6.1 Monte Carlo Results on Estimators for a Semi-Markov Model with Parameters $a_1 = 0.4$, $a_2 = 0.3$, $p_1 = 0.8$, and $p_2 = 0.2$.

				Bias				Standard Deviation			
N	m	m'	Method	a_1	a_2	p_1	p_2	a_1	a_2	p_1	p_2
50	500	68	M1	0.0674	-0.0332	0.0172	0.0018	0.2162	0.1926	0.1007	0.0450
		129	M2	0.0790	0.0204	-0.0674	0.0066	0.2492	0.2287	0.1060	0.0485
		159	M3	0.1134	0.0176	-0.0044	-0.0144	0.2729	0.2433	0.2062	0.0677
		318	M4	0.0261	0.0255	0.0207	0.0031	0.2346	0.1237	0.1237	0.0518
		500	M5	0.1029	0.1258	0.0122	0.0080	0.1546	0.1517	0.1330	0.0582
100	500	122	M1	0.0091	-0.0292	0.0252	-0.0020	0.1009	0.1813	0.1177	0.0317
		215	M2	0.0424	-0.0214	-0.0673	0.0017	0.2061	0.1766	0.0878	0.0360
		238	M3	0.0433	0.0385	0.0182	-0.0049	0.2506	0.2369	0.1838	0.0509
		412	M4	0.0004	-0.0047	0.0124	0.0016	0.1988	0.1745	0.1012	0.0355
		500	M5	0.1023	0.1315	0.0138	0.0055	0.1204	0.1154	0.1072	0.0412
200	500	188	M1	0.0128	-0.0092	0.0173	-0.0011	0.1634	0.1227	0.1118	0.0259
		375	M2	0.0658	-0.0614	-0.0813	-0.0059	0.1684	0.1510	0.0706	0.0273
		291	M3	0.0317	0.0053	-0.0091	-0.0090	0.2152	0.1776	0.1845	0.0402
		482	M4	-0.0029	-0.0152	0.0016	0.0004	0.1523	0.1405	0.0770	0.0255
		500	M5	0.1109	0.1255	0.0039	0.0014	0.0838	0.0824	0.0793	0.0267
500	400	229	M1	0.0061	-0.0040	0.0213	0.007	0.1200	0.1125	0.1103	0.0208
		347	M2	0.0506	-0.0602	-0.0766	-0.0052	0.1320	0.1132	0.0574	0.0210
		322	M3	0.0197	-0.0085	0.0012	-0.0057	0.1743	0.1311	0.1544	0.0309
		397	M4	-0.0043	-0.0005	0.0030	0.0017	0.1082	0.0966	0.0510	0.0179
		400	M5	0.1069	0.1293	0.0039	0.0021	0.0576	0.0560	0.0518	0.0185
800	250	163	M1	0.0105	-0.0029	0.0169	-0.0004	0.1223	0.1040	0.1144	0.0196
		233	M2	0.0504	-0.0524	-0.0743	-0.0055	0.1098	0.0926	0.0511	0.0182
		212	M3	0.0090	-0.0075	0.0028	-0.0056	0.1609	0.1305	0.1500	0.0286
		250	M4	-0.0035	0.0001	0.0009	0.0016	0.0854	0.0732	0.0399	0.0147
		250	M5	0.1080	0.1285	0.0011	0.0017	0.0451	0.0443	0.0401	0.0148

N = number of events in each sequence
m = number of sequences
m'= number of sequences a method succeeded

Table 6.2 Monte Carlo Results on Estimators for a Semi-Markov Model with Parameters $a_1 = 0.9$, $a_2 = 0.6$, $p_1 = 0.8$, and $p_2 = 0.4$.

N	m	m'	Method	Bias				Standard Deviation			
				a_1	a_2	p_1	p_2	a_1	a_2	p_1	p_2
50	500	61	M1	-0.1587	-0.1560	0.0737	0.0129	0.2078	0.2712	0.0826	0.1027
		0	M2	-	-	-	-	-	-	-	-
		180	M3	-0.1158	-0.1837	0.0448	-0.1292	0.2303	0.2623	0.1400	0.1995
		191	M4	-0.1623	-0.1412	0.0775	0.0085	0.2279	0.2736	0.0984	0.1388
		500	M5	-0.2396	-0.2484	0.0259	0.1003	0.2439	0.2435	0.1105	0.1756
100	500	149	M1	-0.1298	-0.1452	0.0589	0.0189	0.1820	0.2573	0.0744	0.0953
		0	M2	-	-	-	-	-	-	-	-
		258	M3	-0.0974	-0.1357	0.0437	-0.0650	0.2059	0.2682	0.1271	0.1759
		287	M4	-0.1030	-0.1105	0.0489	-0.0030	0.2003	0.2813	0.0928	0.1249
		500	M5	-0.1897	-0.2939	0.0297	0.0510	0.2015	0.2000	0.0981	0.1501
200	500	254	M1	-0.0994	-0.0896	0.0506	0.0185	0.1514	0.2293	0.0706	0.0839
		0	M2	-	-	-	-	-	-	-	-
		321	M3	-0.0785	-0.0913	0.0453	-0.0247	0.1570	0.2533	0.0956	0.1438
		353	M4	-0.0623	-0.0891	0.0263	-0.0031	0.1522	0.2530	0.0783	0.1052
		500	M5	-0.1635	-0.3170	0.0234	0.0197	0.1820	0.1819	0.0826	0.1170

with the higher parameter i.e., the distribution with the shorter tail. A value of $e_1 > 0.5$, therefore, implies that shorter interarrival times have greater probability of occurrence. The parameter set considered is $\{e_1 = 0.6, p_1 = 0.9, p_2 = 0.1\}$. Depending on selected values for the first autocorrelation coefficient r_1, several sets of transition probabilities (a_1, a_2) were obtained and the corresponding processes tested. Table 6.3 and Figure 6.1 show the results of this experiment. It is observed that the bias in the parameters a_1 and a_2 increases as a function of r_1, and that the bias in a_2 (where a_2 corresponds to the transition probability of the longer tail geometric distribution) is always greater than the bias in a_1. In terms of RMSE it can be concluded that method M4 performs better when a strong dependence srtucture is present and method M5 when a strong clustering of counts is present. Method M5 seems to give parameter estimates which are more variable but less biased than those of method M4.

From the above preliminary analysis, our conclusions regarding the estimators tested can be summarized as follows. MOM estimates have a very high likelihood of failure and should be avoided given the long tail distribution of the interarrival times. Methods M4 and M5 seem to perform the best and it is suggested that method M4 should be used whenever feasible, especially for large sample sizes (on the order of 500 events). For smaller sample sizes it is suggested that method M4 is used when the dependence of intervals is stronger than the clustering of counts, and method M5 when the clustering is stronger than the dependence in intervals. In view of the above

Table 6.3 Monte Carlo Results on Estimators for Semi-Markov Models with Various Sets of Parameters (a_1, a_2) Consistent with the Fixed Parameters e_1 = 0.6, p_1 = 0.9, and p_2 = 0.1.

N	m		m'	Method	Bias a_1	Bias a_2	Bias p_1	Bias p_2	Standard Deviation a_1	Standard Deviation a_2	Standard Deviation p_1	Standard Deviation p_2
100	500		37	M1	0.0668	0.0269	-0.0621	0.0017	0.1266	0.1324	0.1028	0.0109
		$a_1+a_2=0.5$	223	M2	-0.0174	0.0415	-0.1176	0.0053	0.1800	0.1220	0.0702	0.0200
		$a_1=0.4$	126	M3	0.0006	0.0762	-0.0538	0.0004	0.2117	0.1593	0.1686	0.0278
		$a_2=0.1$	363	M4	0.0023	0.0276	0.0085	0.0036	0.1303	0.1143	0.0521	0.0186
		$r_1=-0.17$	500	M5	0.0883	0.1402	0.0027	0.0030	0.0747	0.0633	0.0536	0.0180
100	500		65	M1	0.0604	-0.0565	0.0872	-0.0027	0.1187	0.1251	0.1399	0.0197
		$a_1+a_2=0.8$	391	M2	-0.0246	-0.0219	-0.1272	0.0061	0.1663	0.1553	0.0847	0.0198
		$a_1=0.52$	202	M3	0.0263	0.0419	-0.1104	-.0053	0.2488	0.2174	0.2490	0.0464
		$a_2=0.28$	483	M4	-0.0152	-0.0212	0.0046	-0.0022	0.1293	0.1484	0.0543	0.0182
		$r_1=-0.068$	500	M5	0.0334	0.0519	0.0046	0.0031	0.0706	0.0729	0.0544	0.0182
100	500		59	M1	0.0665	-0.0905	-0.4215	-0.0054	0.1160	0.1906	0.1319	0.0191
		$a_1+a_2=1$	411	M2	-0.0053	-0.0276	-0.1373	0.0053	0.1525	0.1852	0.0863	0.0201
		$a_1=0.6$	238	M3	0.0665	0.0015	-0.1295	0.0065	0.2236	0.2467	0.2774	0.0649
		$a_2=0.4$	497	M4	-0.0153	-0.0284	0.0020	0.0026	0.1342	0.1709	0.0541	0.0169
		$r_1=0$	500	M5	-0.0041	-0.0071	0.0020	0.0026	0.0701	0.0793	0.0539	0.0169
100	500		58	M1	0.0019	-0.1123	-0.1245	-0.0055	0.1014	0.2055	0.1502	0.0225
		$a_1+a_2=1.5$	310	M2	0.0039	-0.0523	-0.1629	-0.0032	0.1018	0.1819	0.0859	0.0197
		$a_1=0.8$	243	M3	0.0291	-0.1113	-0.1233	0.0005	0.1465	0.2257	0.2784	0.0592
		$a_2=0.7$	464	M4	-0.0167	-0.0455	0.0025	0.0029	0.1180	0.1763	0.0605	0.0189
		$r_1=+0.17$	500	M5	-0.0944	-0.1526	0.0039	0.0033	0.0748	0.0896	0.0596	0.0187

Note: N, m and m' have been defined in Table 6.1.

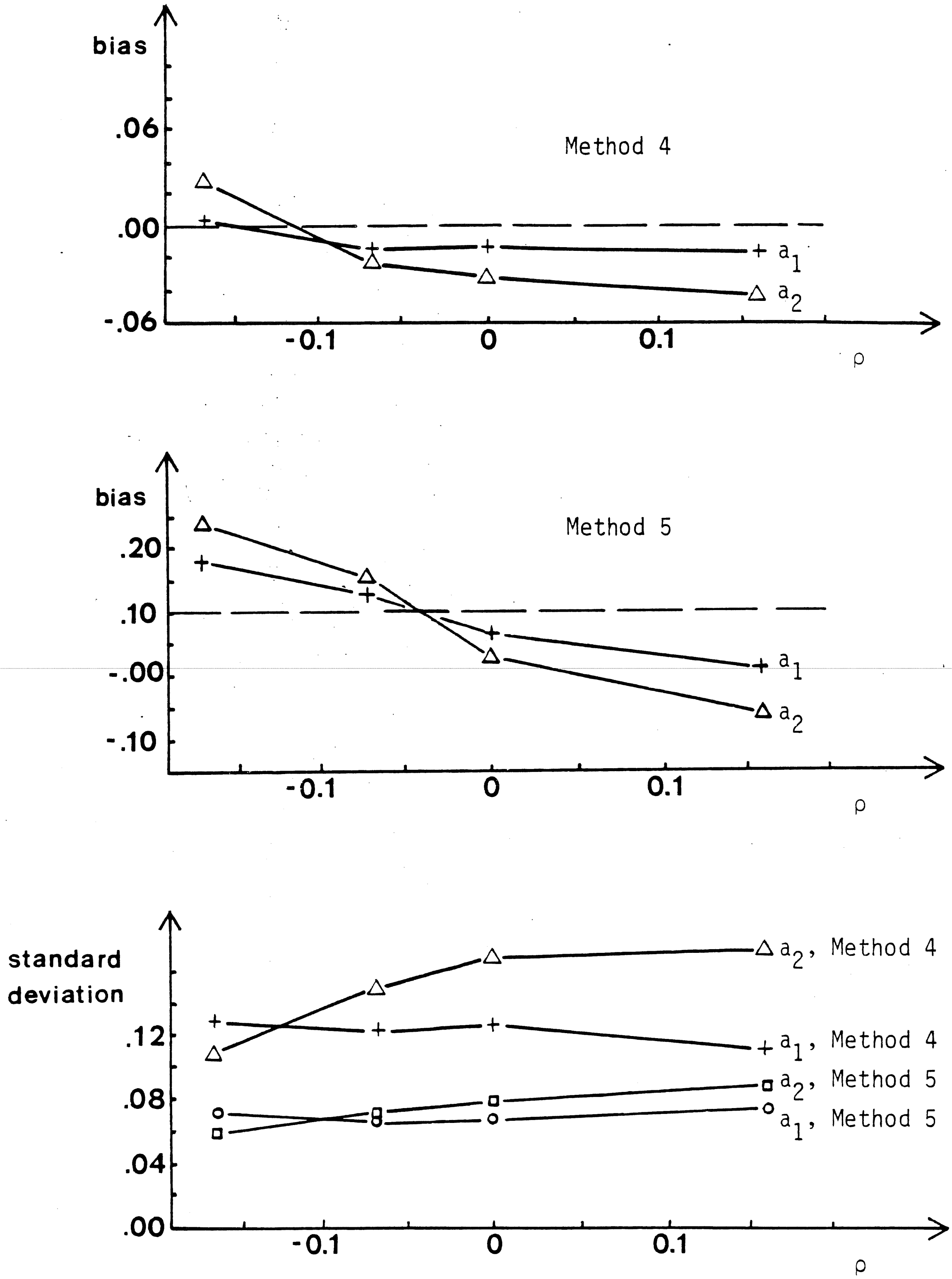


Figure 6.1 Bias and variability of the approximate maximum likelihood estimators as functions of the first autocorrelation coefficient. Model is that of Table 6.3.

results, only methods M4 and M5 were used for fitting the semi-Markov model to the daily rainfall occurrences. These results are given in the next chapter.

CHAPTER 7
APPLICATION OF THE SEMI-MARKOV MODEL TO DAILY RAINFALL OCCURRENCES

> The Sixth no sooner had begun
>   About the beast to grope,
> Than, seizing on the swinging tail
>   That fell within his scope,
> "I see," quoth he, "the Elephant
>   Is very like a rope!"

The analysis of the six daily rainfall records in Chapter 4 has revealed that the Snoqualmie Falls and Roosevelt stations may be of special interest. This is due to the tremendous variability they exhibit within a year, as well as to the non-Poissonian clustered rainfall occurrence structures within each of the seasons. In addition, Snoqualmie Falls lies in a significantly different climatologic regime than Roosevelt, and therefore these two stations have different underlying rainfall generating mechanisms. For example, the mean interarrival time for Snoqualmie Falls ranges from 1.34 days for December to 4.26 days for July, while for Roosevelt the corresponding figures are 4.67 for July to 22.52 for May. For these reasons, the Snoqualmie Falls and Roosevelt daily rainfall sequences were selected to demonstrate the fitting of the semi-Markov model.

7.1 Selection of Seasons and Seasonal Statistical Analysis

The objective of the season discrimination methodology is to identify periods (seasons) within the year in which the statistical structure of the process remains constant, i.e., does not vary over time. The statistical structure of the daily rainfall process is completely characterized by the probability laws of two properties: the number of events (rainy days) within a season, and the daily rainfall amounts. For the number of events within a season, the relevant properties to be examined are the probability distribution function of the interarrival times and the second order properties of counts, specifically the spectrum of counts, variance time curve and index of dispersion. For the non-zero daily rainfall amounts, the property of interest is the probability distribution function. All these properties must be examined in parallel for a successful selection of seasons. It is also understood that physical considerations (including an understanding of the climatic conditions of the region) and subjective judgment play an important role in this process.

For the Snoqualmie Falls and Roosevelt stations, the following homogeneous seasons were identified after careful examination:

	Snoqualmie Falls	Roosevelt
Season 1:	Jan, Feb, Mar	Jan, Feb, Mar, Apr
Season 2:	Apr, May, Jun	May, Jun
Season 3:	Jul, Aug	Jul
Season 4:	Sept, Oct	Aug
Season 5:	Nov, Dec	Sept, Oct
Season 6:	-	Nov, Dec

A statistical analysis, similar to that of Chapter 4, was performed on a seasonal basis and the results are given in Tables

7.1 - 7.5. For the Roosevelt station, 30 years of data (1948-1977) were analyzed, whereas for Snoqualmie Falls the analysis was performed only on the last 15 years of the record (1963-1977). The reason for the different record lengths is that Snoqualmie Falls has a fairly high rate of occurrence, i.e., large number of events in each season, which permits a reliable analysis for a shorter recording length, at a considerable savings in computer resources.

Figures 7.1 and 7.2 show the properties of intervals (log-survivor function) and counts (spectrum of counts, variance time curve and index of dispersion) for the two stations. Comparison of the seasonal empirical curves with the corresponding curves for the months constituting each season (Figures A.7-A.12 of Appendix A) revealed no significant differences confirming the selection of seasons.

7.2 Fitting the Semi-Markov Model to the Daily Rainfall Occurrences

From the Monte Carlo analysis in Chapter 6, it was concluded that the two most consistent and efficient estimation methods are the approximate maximum likelihood (ML) methods M4 and M5 (see Chapter 6 for details). These methods will be referred to in this chapter as ML1 and ML2, respectively. It is also recalled that although the Bayesian estimation method (ML2) always gives parameter estimates, method ML1, based on the first autocorrelation coefficient, gives parameter estimates if and only if the constraint of (6.15) is satisfied.

The results of fitting the semi-Markov model to the seasons for Snoqualmie Falls and Roosevelt are shown in Table 7.6. The interpretation of the estimated parameters for Snoqualmie Falls suggest that the interarrival times of the process are sampled from two geometric distributions, one with mean at the order of 1 day ($p_1 \simeq 0.9$) and the other at the order of 2.5 to 7 days ($p_1 \simeq 0.4$ to 0.15). For this station, it is also observed that the transition probability a_1 is always greater than 0.5 which suggests that small interarrival times are most likely to be followed by small interarrival times, an indication of clustering. For the Roosevelt station, method ML1 did not give feasible parameter estimates for one season (month of July), and both methods gave a value of p_1 at the bound ($p_1 = 0.99$) for three out of six seasons. Problems with fitting the model to this station were expected given the small number of events available for estimation.

The assessment of the goodness of fit of the semi-Markov model was performed by comparing empirical functions of the data which were not used in the estimation with their theoretical counterparts. Figures 7.3 and 7.4 show these comparisons for some selected seasons and stations. It is observed that the theoretical spectra of counts are surprisingly close to the empirical ones, especially for Snoqualmie Falls. This is a sign of a good fit, given that this function was not explicitly used in the estimation. The agreements for the variance time curves generally is not as good. This is not surprising since the estimated variance time curve is much more variable than the estimated spectrum of counts. It should be noted that the model does a good job in preserving the probability distribution of the interarrival times as expected, since this information is used explicitly in the estimation.

Table 7.1 Autocorrelation Coefficients of Interarrival Times--Seasonal Analysis (S_i corresponds to the ith season)

	S_1	S_2	S_3	S_4	S_5	S_6
			(a) Snoqualmie Falls			
r_1	0.047	-0.054	-0.010	0.025	0.036	
r_2	-0.026	0.074	0.057	-0.026	0.034	
r_3	0.032	0.016	-0.022	0.032	-0.040	
r_4	-0.022	-0.009	0.051	-0.053	-0.040	
r_5	0.003	0.042	-0.050	-0.042	-0.031	
			(b) Roosevelt			
r_1	0.009	-0.058	-0.112	-0.013	-0.051	-0.020
r_2	0.067	-0.249*	0.073	-0.056	0.072	-0.050
r_3	0.088	0.048	-0.008	0.056	0.081	-0.074
r_4	-0.036	-0.044	0.006	0.073	0.008	0.019
r_5	-0.006	-0.050	-0.098	-0.073	0.032	-0.035

Table 7.2 Statistics of the Interarrival Times--Seasonal Analysis

Season	$\bar{x}$	s_x	c_v	c_s	Number of Events
		(a) Snoqualmie Falls (15 years)			
1	1.496	1.377	0.920	4.217	896
2	2.101	2.603	1.239	4.085	672
3	3.715	5.235	1.409	2.924	246
4	2.271	2.776	1.222	3.781	391
5	1.393	1.125	0.808	4.212	657
		(b) Roosevelt (30 years)			
1	7.821	13.601	1.739	3.437	502
2	18.973	19.088	1.006	0.837	75
3	4.671	4.759	1.019	1.739	152
4	6.052	10.880	1.798	4.345	191
5	8.153	11.631	1.427	2.278	222
6	8.165	14.565	1.784	5.579	231

Table 7.3 Autocorrelation Coefficients of Non-Zero Daily Rainfall Amounts--Seasonal Analysis (S_i corresponds to the ith season)

	S_1	S_2	S_3	S_4	S_5	S_6
			(a) Snoqualmie Falls			
r_1	0.238**	0.058	0.094	0.054	0.130**	
r_2	0.008	0.016	0.011	0.016	0.023	
r_3	-0.014	-0.07	-0.046	0.057	0.014	
r_4	0.011	-0.026	-0.023	0.025	-0.003	
r_5	0.051	-0.010	0.012	0.014	-0.009	
			(b) Roosevelt			
r_1	0.069	0.117	-0.072	0.049	0.109	0.082
r_2	0.008	-0.115	-0.092	-0.030	-0.035	0.080
r_3	0.029	0.185	0.042	-0.030	0.012	0.212**
r_4	-0.043	-0.114	0.023	0.015	0.057	0.011
r_5	-0.034	-0.087	-0.177*	0.074	-0.025	0.059

Table 7.4 Statistics of the Non-Zero Daily Rainfall Amounts--Seasonal Analysis

Season	$\bar{x}$	s_x	c_v	c_s
	(a) Snoqualmie Falls (15 years)			
1	0.373	0.456	1.222	3.019
2	0.240	0.281	1.170	2.559
3	0.216	0.274	1.270	2.116
4	0.311	0.342	1.099	1.963
5	0.407	0.474	1.164	2.165
	(b) Roosevelt (30 years)			
1	0.268	0.312	1.166	2.263
2	0.217	0.269	1.241	2.099
3	0.262	0.352	1.343	2.757
4	0.278	0.367	1.322	2.801
5	0.345	0.518	1.501	3.240
6	0.350	0.437	1.248	2.224

Table 7.5 Cross Correlation Coefficients of the Non-Zero Daily Rainfall Amounts with Preceding and Following Interarrival Times (X_i = interarrival time following the event P_i)

Season	(X_{i-2}, P_i)	(X_{i-1}, P_i)	(X_i, P_i)	(X_{i+1}, P_i)
		(a) Snoqualmie Falls		
1	-0.030	-0.093**	-0.147**	-0.089**
2	-0.069	-0.087*	0.068	0.041
3	0.086	0.079	-0.206**	-0.197**
4	0.036	-0.123*	-0.162**	-0.062
5	0.011	-0.111*	-0.091	-0.065
		(b) Roosevelt		
1	0.003	-0.043	-0.093*	-0.045
2	0.013	0.050	-0.154	0.014
3	-0.835	0.019	-0.014	-0.051
4	-0.035	-0.019	-0.007	0.008
5	-0.103	-0.13	-0.001	-0.018
6	0.018	-0.036	-0.096	-0.105

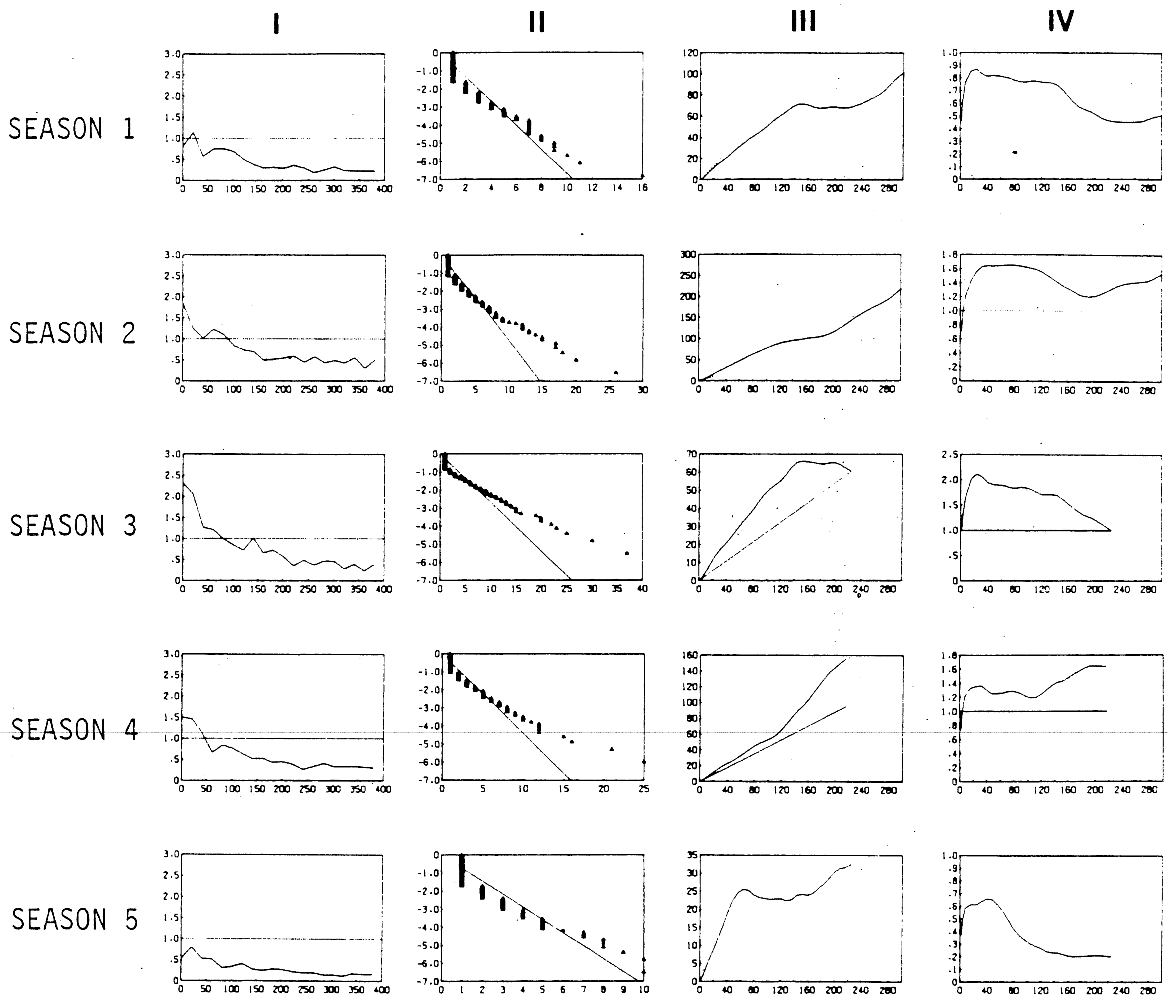


Figure 7.1 Statistical properties of intervals and counts for Snoqualmie Falls--seasonal analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

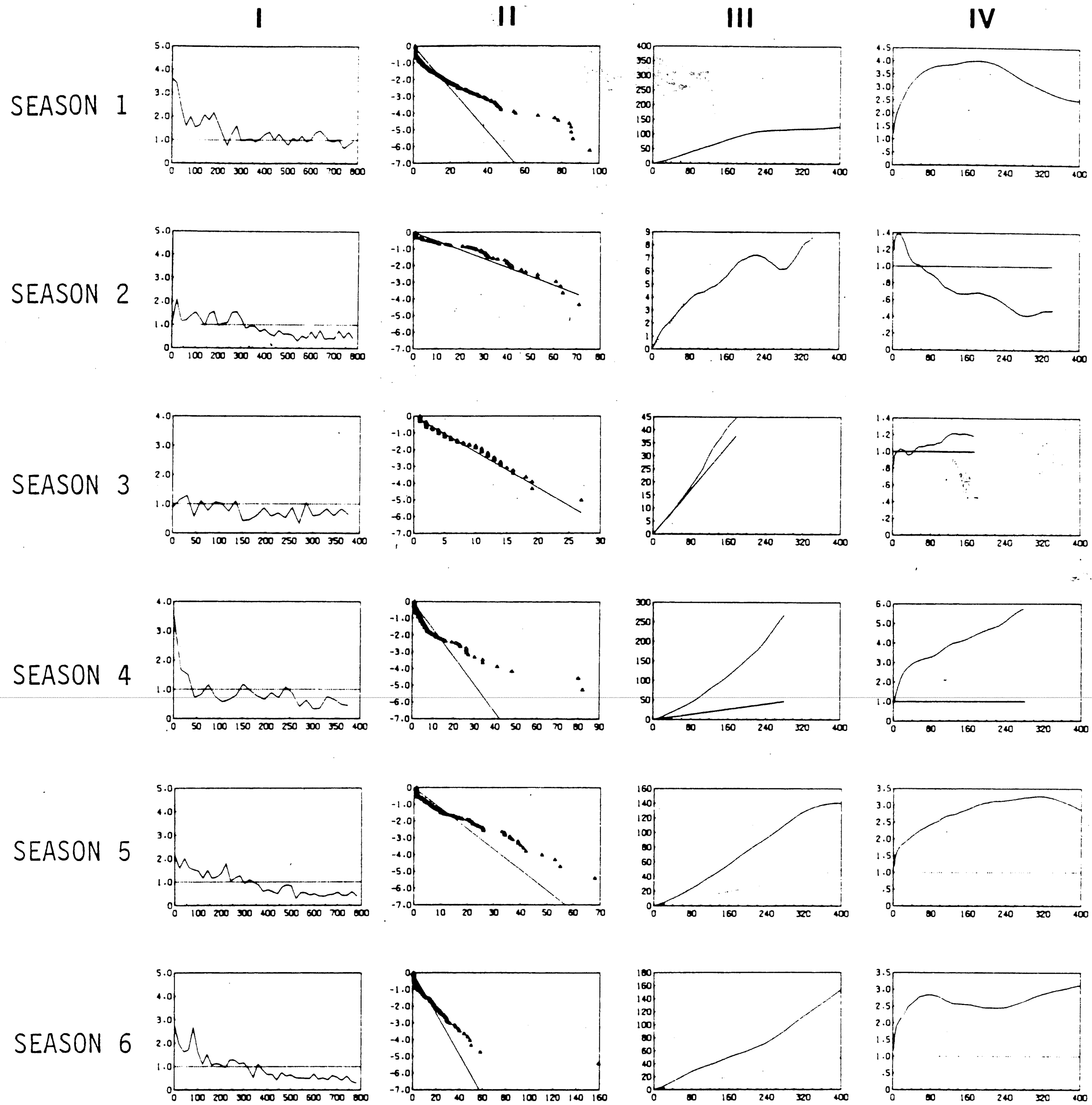


Figure 7.2 Statistical properties of intervals and counts for Roosevelt--seasonal analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

Table 7.6. Results of Fitting the Semi-Markov Model to the Daily Rainfall Occurrences

Season	Method	a_1	a_2	p_1	p_2	e_1
		(a) Snoqualmie Falls				
1	ML1	0.776	0.380	0.958	0.364	0.735
	ML2	0.740	0.279			
2	ML1	0.599	0.227	0.905	0.248	0.659
	ML2	0.651	0.326			
3	ML1	0.534	0.434	0.929	0.144	0.549
	ML2	0.539	0.430			
4	ML1	0.631	0.454	0.916	0.248	0.597
	ML2	0.601	0.407			
5	ML1	0.759	0.369	0.971	0.425	0.723
	ML2	0.407	0.273			
		(b) Roosevelt				
1	ML1	0.463	0.568	0.917	0.075	0.446
	ML2	0.411	0.523			
2	ML1	0.049	0.661	0.990	0.039	0.263
	ML2	0.123	0.680			
3	ML1	-	-	0.624	0.165	0.688
	ML2	0.271	0.668			
4	ML1	0.782	0.183	0.376	0.053	0.789
	ML2	0.790	0.210			
5	ML1	0.245	0.556	0.990	0.080	0.370
	ML2	0.386	0.635			
6	ML1	0.376	0.551	0.990	0.075	0.419
	ML2	0.374	0.548			

ML1 = Approximate maximum likelihood estimates (MLE) coupled with the first autocorrelation coefficient

ML2 = Approximate MLE with a Bayesian approach

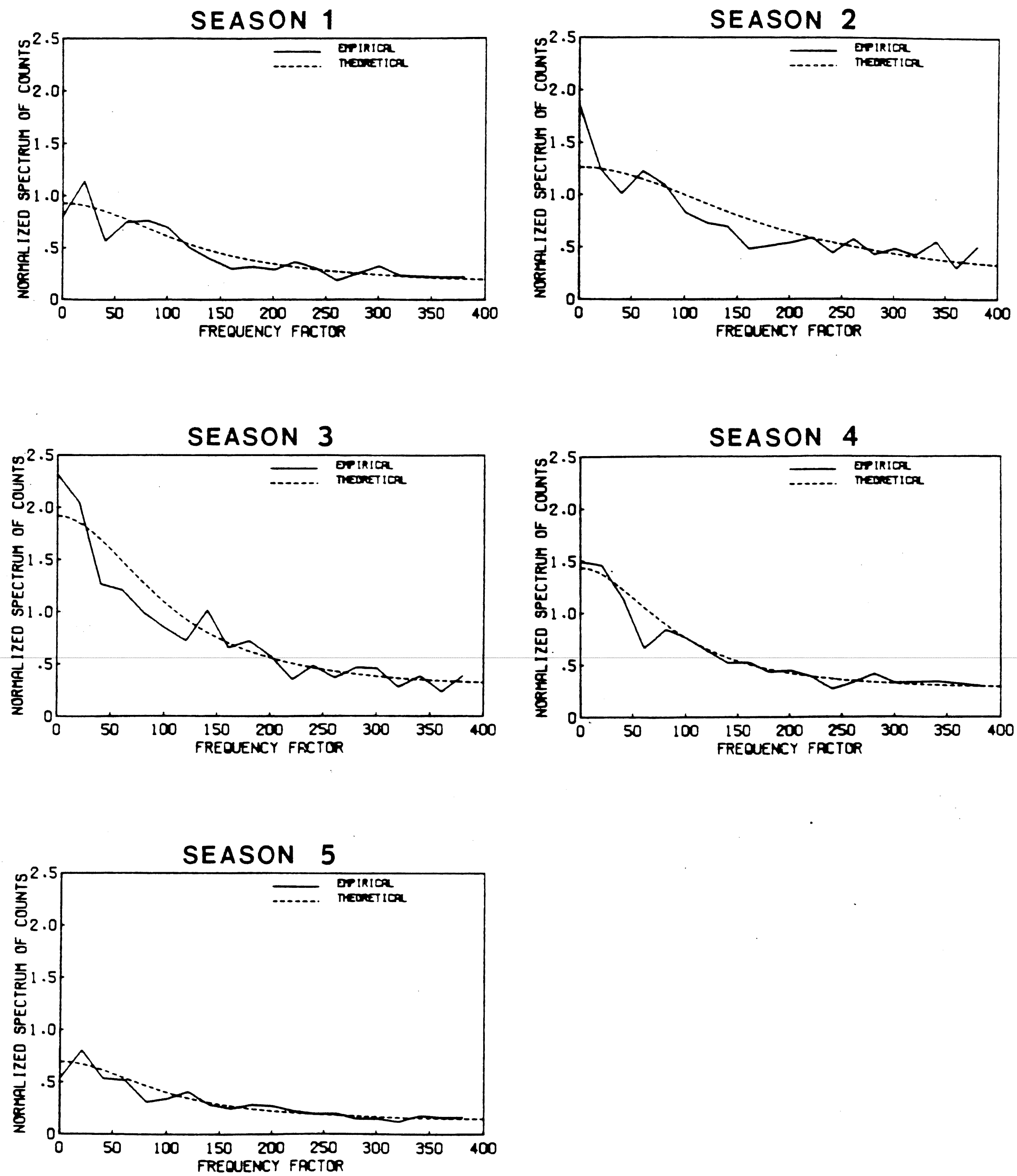


Figure 7.3 Comparison of empirical and theoretical spectra of counts for Snoqualmie Falls--seasonal analysis.

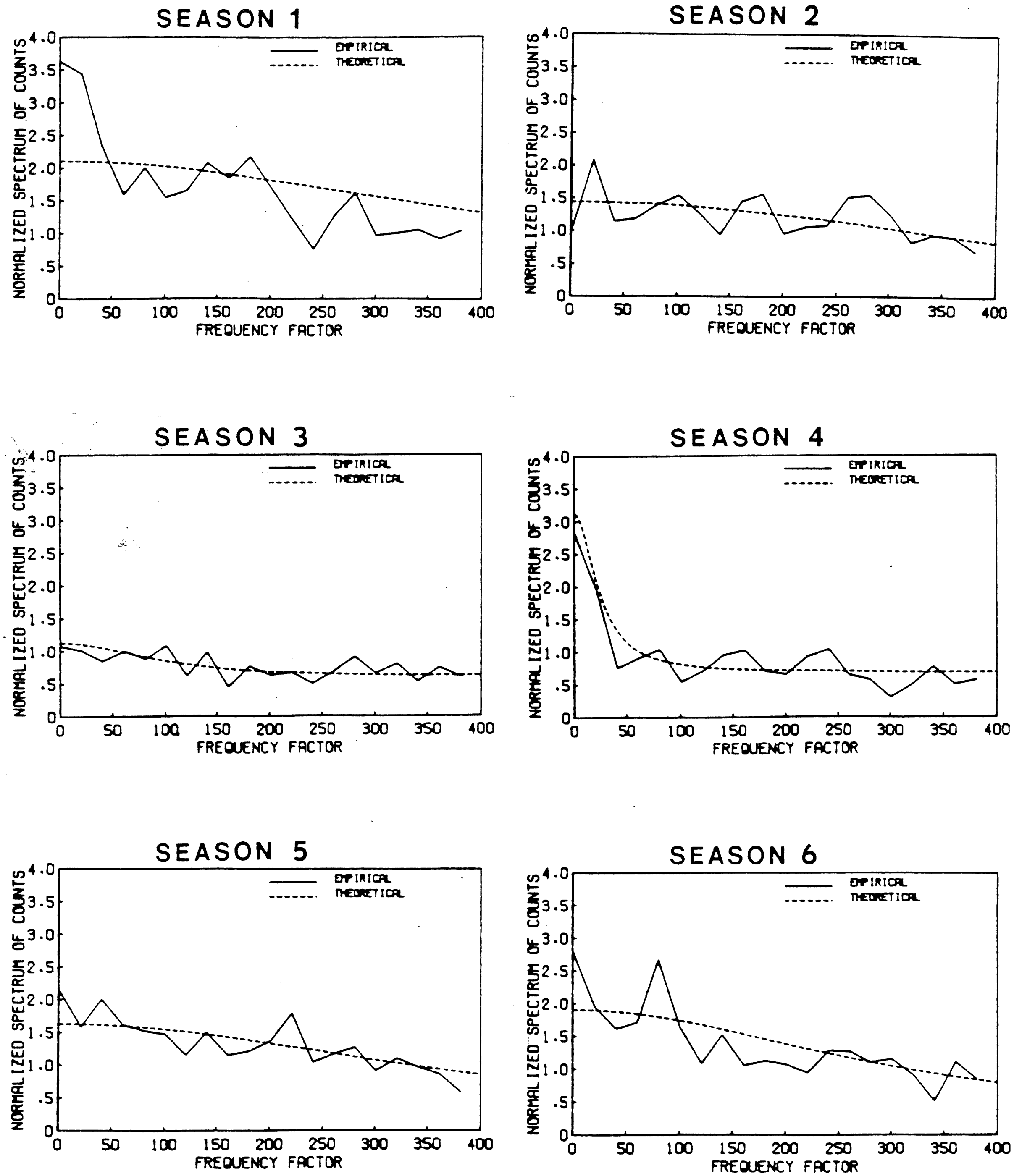


Figure 7.4 Comparison of empirical and theoretical spectra of counts for Roosevelt--seasonal analysis.

7.3 Modeling the Non-Zero Daily Rainfall Amounts

Based on previous research (see Chapter 2) and some exploratory analyses, the following three marginal distributions were selected as candidates to fit the daily rainfall amounts: Weibull, Gamma, and mixed exponential. The properties and fitting procedures for these distributions are discussed below.

Weibull distribution. The probability density function (pdf) of the Weibull distribution is:

$$f(x) = \frac{\alpha}{\beta-\gamma} \left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha-1} \exp\left[-\left(\frac{x-\gamma}{\beta-\gamma}\right)^{\alpha}\right], \tag{7.1}$$

where α, β, and γ are parameters to be estimated. The mean, standard deviation and skewness coefficient are given in terms of the parameters α, β, and γ as:

$$\mu = \gamma + (\beta-\gamma)\ \Gamma(1+1/\alpha), \tag{7.2}$$

$$\sigma = (\beta-\gamma)\ [\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)]^{1/2}, \tag{7.3}$$

$$c_s = \frac{\Gamma(1+3/\alpha) - 3\Gamma(1+2/\alpha)\ \Gamma(1+1/\alpha) + 2\Gamma^3(1+1/\alpha)}{[\Gamma(1+2/\alpha) - \Gamma^2(1+1/\alpha)]^{1/2}}, \tag{7.4}$$

(see Kite, 1978), where $\Gamma(\cdot)$ is the usual gamma function. A method of moments parameter estimation procedure was followed. This consists of solving (7.4) iteratively for α, and then solving (7.2) and (7.3) for the other two parameters, β and γ.

Gamma distribution. The pdf of the three parameter gamma distribution is:

$$f(x) = \frac{1}{\alpha\Gamma(\beta)} \left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp\left[-\left(\frac{x-\gamma}{\alpha}\right)\right], \qquad (7.5)$$

where α, β, and γ are parameters to be defined. The mean, standard deviation and skewness coefficient are given as

$$\mu = \alpha\beta + \gamma,$$

$$\sigma = a\sqrt{\beta}\ , \qquad (7.6)$$

and

$$c_s = 2/\sqrt{\beta}\ ,$$

from which method of moments estimates can be easily obtained.

Mixed exponential distribution. The pdf of a mixed exponential distribution is

$$f(x) = \alpha\lambda_1 \exp[-\lambda_1 x] + (1-\alpha)\lambda_2 \exp[-\lambda_2 x], \qquad (7.7)$$

where λ_1 and λ_2 are the parameters of the two exponential distributions and α is their mixing ratio. The mean and variance of this distribution are

$$\mu = \frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2}\ , \qquad (7.8)$$

$$\sigma^2 = \frac{\alpha}{\lambda_1^2} + \frac{1-\alpha}{\lambda_2^2} + \alpha(1-\alpha)\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)^2 .$$

Everitt and Hand (1981) suggest several methods of estimating the parameters α, λ_1, and λ_2. Here, the method of maximum likelihood was used.

The log-likelihood function of a mixed exponential distribution is

$$\begin{aligned} L'(x) &= \ln[\prod_{i=1}^{n} f(x_i)] \\ &= \sum_{i=1}^{n} \ln\{\alpha\lambda_1 \exp[-\lambda_1 x_i] + (1-\alpha)\lambda_2 \exp[-\lambda_2 x_i]\}. \end{aligned} \tag{7.9}$$

Estimates of the parameters were obtained using the Nelder and Mead (1965) simplex algorithm for the maximization of L'(x). Notice that the parameter α is a probability, and therefore should lie in [0,1]. The transformation (6.5) was used to constrain this parameter.

All three distributions were fitted to the non-zero daily rainfall amounts. A visual comparison indicated that the mixed exponential gave the best fit, and this distribution was subsequently used for all seasons. Figures 7.5 and 7.6 show the empirical and fitted mixed exponential cumulative probability plots for the fitted distributions. The parameters of the fitted distributions are given in Table 7.7.

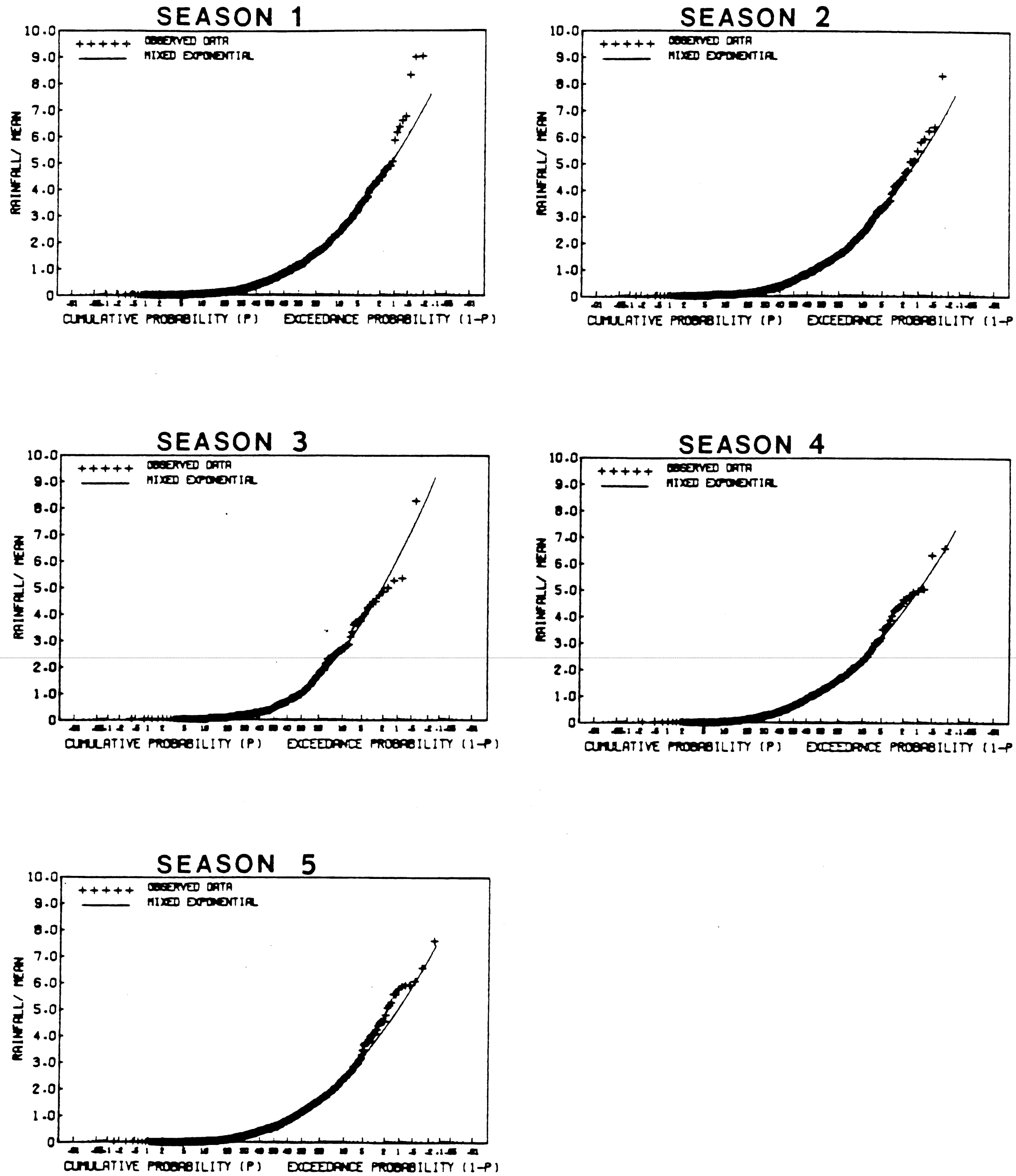


Figure 7.5 Empirical and theoretical cumulative probability functions of the non-zero daily rainfall amounts for Snoqualmie Falls--seasonal analysis.

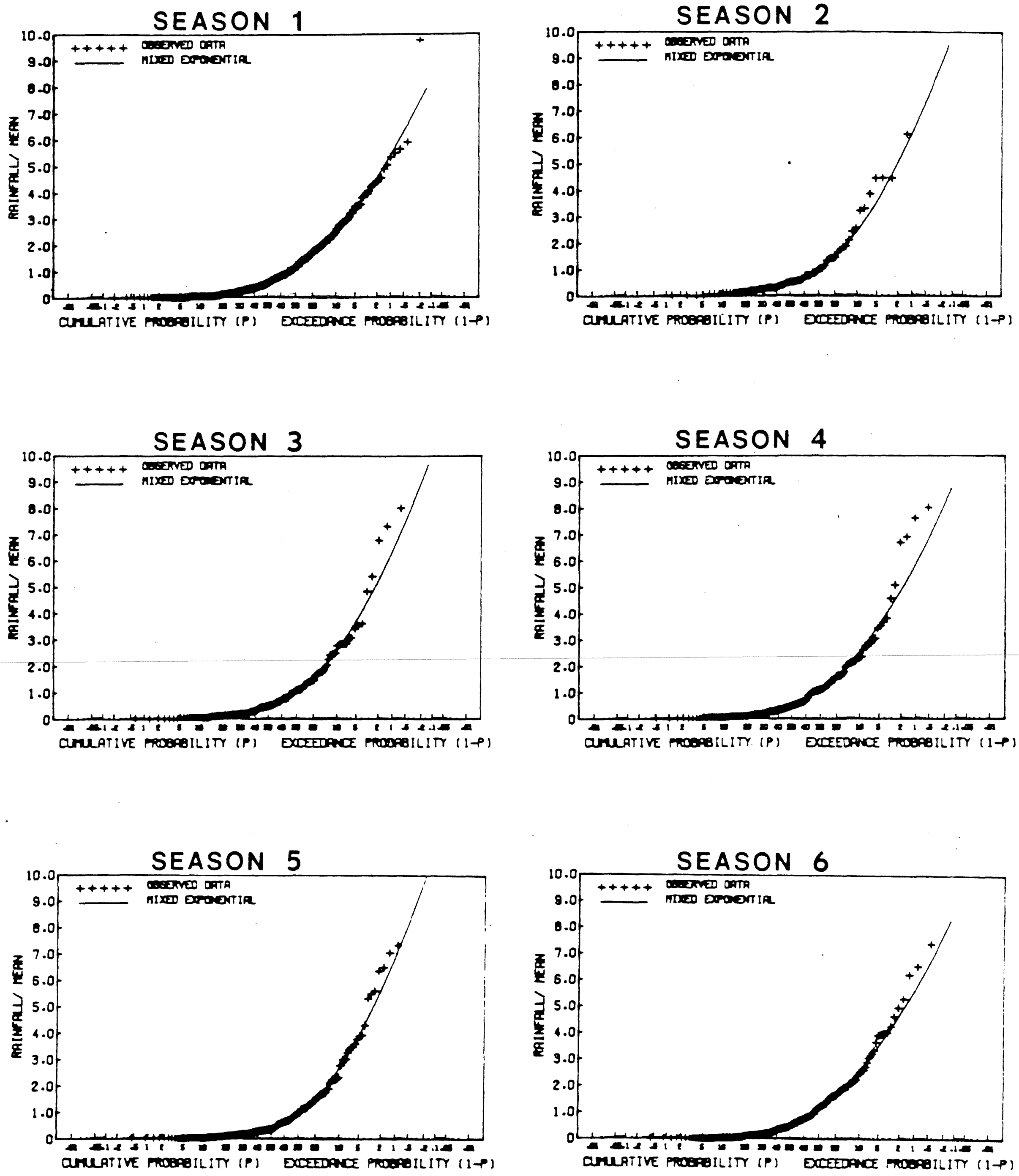


Figure 7.6 Empirical and theoretical cumulative probability functions of the non-zero daily rainfall amounts for Roosevelt--seasonal analysis.

Table 7.7 Parameters of the Mixed Exponential Distribution Fitted to the Non-Zero Daily Rainfall Amounts

Season	α	λ_1	λ_2
(a) Snoqualmie Falls			
1	0.182	17.627	2.257
2	0.201	17.033	3.504
3	0.412	17.500	3.065
4	0.120	26.743	2.855
5	0.152	19.654	2.123
(b) Roosevelt			
1	0.251	16.347	2.963
2	0.558	9.631	2.797
3	0.514	9.527	2.317
4	0.387	11.626	2.514
5	0.561	8.406	1.577
6	0.276	17.303	2.163

7.4 Coupling the Models for Occurrences and Amounts and Overall Model Performance

The cross correlation analysis of interarrival times and non-zero rainfall amounts (Table 7.5) indicated that no significant correlations were present for the Roosevelt station but that small, although significant, correlations were present for Snoqualmie Falls. The significance of the Snoqualmie Falls correlations may be due in part to the greater number of events at that station. If the small correlations are taken to justify an assumption of independence, this implies that given the occurrence of an event, the corresponding daily rainfall amount does not depend on whether or not the event was the first or last rainy day in a sequence of rainy days. If independence is assumed, the coupling of the rainfall occurrence model with the rainfall amounts model becomes easy, since the two processes are simply superimposed. For example, a generation scheme for daily rainfall sequences, would consist of generating the position of daily rainfall occurrences from a semi-Markov model, and then assigning to each rainy day a rainfall amount from the desired marginal distribution.

For the purposes of streamflow prediction, or other applications where a mass balance is desired, one is interested in the distribution of the total rainfall over the next t days. For example, for rainfall/runoff studies, an important property of a daily rainfall generation scheme is its ability to preserve the total rainfall amounts over periods of given length, i.e., one week or one month. The statistical properties of the accumulated rainfall process are given below.

Let R(t) denote the accumulated rainfall process over a period of length t. Then,

$$R(t) = \sum_{i=1}^{N_t} Y_i, \qquad (7.10)$$

where $\{Y_i\}$ is the process of the non-zero daily rainfall amounts and $\{N_t\}$ is the daily rainfall occurrence process. Making the assumption that the non-zero daily rainfall amounts $\{Y_i\}$, are independent and identically distributed, and that the daily rainfall occurrence process $\{N_t\}$ is independent of the rainfall amounts process $\{Y_i\}$, the mean and variance of R(t) are given as

$$E[R(t)] = \mu_y mt \qquad (7.11)$$

and

$$Var[R(t)] = \sigma_y^2 mt + \mu_y^2 V(t), \qquad (7.12)$$

where $\mu_y = E[Y_i]$, $\sigma_y^2 = Var(Y_i)$, V(t) is the variance time curve of the counting process $\{ N_t \}$, and m is its rate of occurrence. For a semi-Markov model, m and V(t) are given in terms of the parameters a_1, a_2, p_1, and p_2, from equations (5.29) and (5.34) of Chapter 5. For a mixed exponential distribution, μ_y and σ_y^2 are given in terms of the parameters a, λ_1, λ_2 by (7.8).

Table 7.8 shows the empirical seasonal means and standard deviations together with their theoretical counterparts for the fitted

Table 7.8 Comparison of the Empirical and Theoretical Seasonal Means and Standard Deviations

Season	Mean		Standard Deviation	
	Empirical	Theoretical	Empirical	Theoretical
(a) Snoqualmie Falls				
1	22.145	22.426	5.821	5.288
2	10.621	10.255	2.643	2.513
3	3.352	3.468	1.587	1.484
4	8.689	8.238	2.749	2.699
5	17.789	17.506	4.445	4.729
(b) Roosevelt				
1	4.481	4.086	2.326	1.688
2	0.542	0.676	0.649	0.629
3	1.341	1.767	0.841	1.156
4	1.768	1.369	1.654	0.964
5	2.570	2.511	2.188	1.644
6	2.698	2.573	2.620	1.519

model. The preservation of these seasonal statistics are very satisfactory for Snoqualmie Falls, whereas for the station of Roosevelt the results are not as good. This is not surprising given the small number of events available for the estimation of the model parameters.

CHAPTER 8
SUMMARY AND CONCLUSIONS

And so these men of Indostan
    Disputed loud and long,
Each in his own opinion
    Exceeding still and strong.
Though each was partly in the right
    And all were in the wrong !

John Godfrey Saxe (1816-1887)
Reprinted in Engineering Concepts
Curriculum Project (1971)

Several authors have recently had apparent success in applying continuous-time point process models to daily rainfall observation sequences. In this work we have shown that major problems arise when the observation sequence represents cumulative rainfall amounts over a period (e.g., one day) which is on the order of the process interarrival time. In particular, the use of continuous-time point process models for daily rainfall occurrences may result in incorrect inferences about the underlying rainfall generating mechanisms. Since daily rainfall occurrences form a discrete point process, it seems only natural that daily rainfall sequences should be compared with the discrete independent Bernoulli, and not with the continuous Poisson process. Daily rainfall structures that are underdispersed (more regular occurrences) relative to the independent Poisson process may in fact be overdispersed (more random occurrences) relative to the Bernoulli process.

The statistical analysis of six daily rainfall records from diverse climatologic regimes throughout the U.S. (Snoqualmie Falls, Washington; Roosevelt, Arizona; Austin, Texas; Miami, Florida; Philadelphia, Pennsylvania; and Denver, Colorado) has confirmed the inappropriateness of the continuous point process modeling approach for the daily rainfall occurrence process. The daily rainfall occurrences for some months and some stations are underdispersed relative to Poisson, a condition that is inconsistent with the continuous point process models used by other authors. However, comparison of the statistics of the rainfall occurrence processes at these stations with the Bernoulli indicated that all were clustered, that is, overdispersed, which is consistent with the underlying physical processes. A further disadvantage of continuous point process models is that they cannot be used for the generation of synthetic rainfall sequences. It has been shown that, using these models, generation of synthetic rainfall sequences leads to serious upward biases in the event interarrival times and in dependence structures which may be much different than those of the apparent generating process.

To meet these shortcomings of continuous point process models, a discrete point process model has been developed and its structural properties derived. The model belongs to the class of semi-Markov (or Markov renewal) processes and has a flexible structure. In the semi-Markov model the sequence of times between events is formed through sampling from two geometric distributions, according to transition probabilities specified by a Markov chain. In that sense, higher probabilities of transition from the geometric distribution with the smaller mean to the same geometric distribution, rather than to the one with the larger mean, incorporates a clustering structure in the process.

Several methods for fitting the proposed model have been studied. Due to the heavy-tailed distributions of the interarrival times, method of moment estimates do not perform well. An approximate likelihood method which estimates the equilibrium probabilities of the Markov chain of intervals, and subsequently the transition probabilities, has been proposed. This approximate maximum likelihood approach was found to perform adequately, especially for daily rainfall structures with small autocorrelations in the sequence of interarrival times.

The semi-Markov model was fitted to the daily rainfall occurrences of the Snoqualmie Falls and Roosevelt stations, both on a monthly and seasonal basis. Seasons were selected after a careful examination of all the statistical properties of intervals and counts. The fit of the model was assessed by the preservation of selected statistical properties of the series which were not used directly in the estimation. It was shown that the fitted model gave a theoretical spectrum of counts surprisingly close to the empirical one. The semi-Markov model of daily rainfall occurrences coupled with a mixed exponential distribution for the non-zero daily rainfall amounts preserved the seasonal means and standard deviations for the Snoqualmie Falls station, but not for the Roosevelt station. The preservation of cumulative rainfall amounts over longer periods (e.g., weeks or months) is an important property of a daily rainfall generation model, especially when the model is used for rainfall/runoff studies.

The proposed use of discrete-time point process models (including the semi-Markov approach) for daily rainfall occurrences opens a number of areas for future research. Among these are the following:

(1) The possible use of alternate discrete point process model structures for daily rainfall. For example, it seems feasible to derive discrete point process models with structures similar to the two-level hierarchical structures of the continuous Poisson cluster models, i.e., a discrete analogue of the Neyman-Scott model.

(2) Improved fitting techniques for discrete point process models. Although continuous point process models have been extensively studied statistically, not much work has been done on discrete point processes. Specifically, for daily rainfall, alternate fitting methods that explicitly preserve the monthly or seasonal rainfall statistics might be investigated. This would probably require an iterative estimation scheme to accommodate the trade-off between exact preservation of short and long term statistics.

(3) Application of the semi-Markov model to shorter time increment rainfall sequences, such as hourly. In particular, the compatibility of the semi-Markov model with continuous point process models applied to the unobserved continuous generalized stochastic process, $\xi(t)$ (see figure 1.1), could be investigated. One interesting question, related to ongoing work on process scale effects, is to determine whether the statistics for daily rainfall derived using a continuous Neyman-Scott model for $\xi(t)$ are in agreement with the statistics obtained by

modeling the daily rainfall sequences directly with a semi-Markov model.

(4) Improved methods for coupling rainfall occurrence models with rainfall amounts models. An area deserving further attention is the development of a model structure that accounts for cross-correlation between the occurrence and amounts processes.

(5) Extension of discrete point process models to multiple dimensions. This is essential generalization for rainfall-runoff studies and for the estimation of missing data in rainfall sequences.

APPENDIX A

STATISTICAL PROPERTIES OF THE SIX DAILY RAINFALL RECORDS ANALYZED

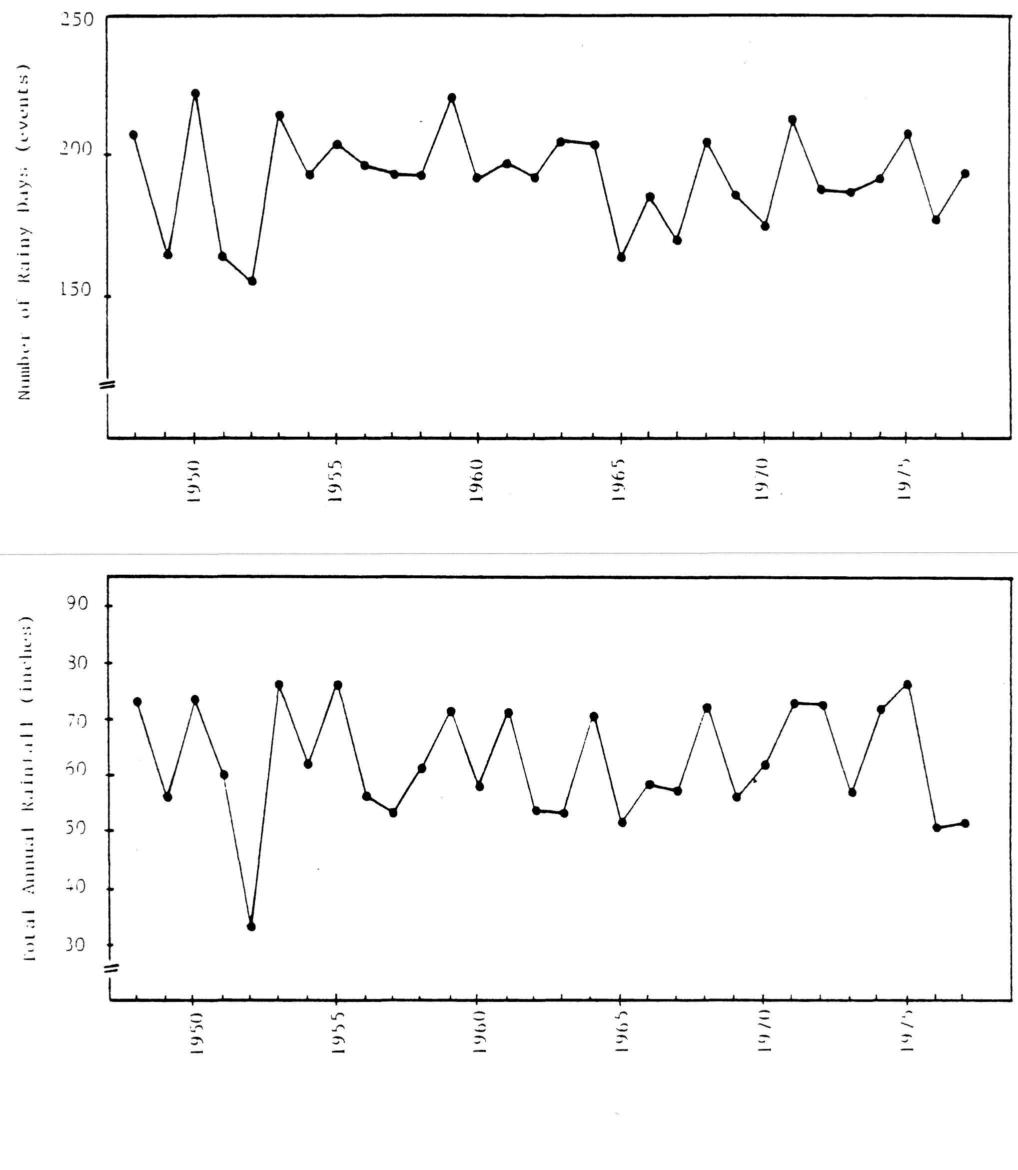


Figure A.1 Number of rainy days per year and total annual rainfall amounts for Snoqualmie Falls, Washington.

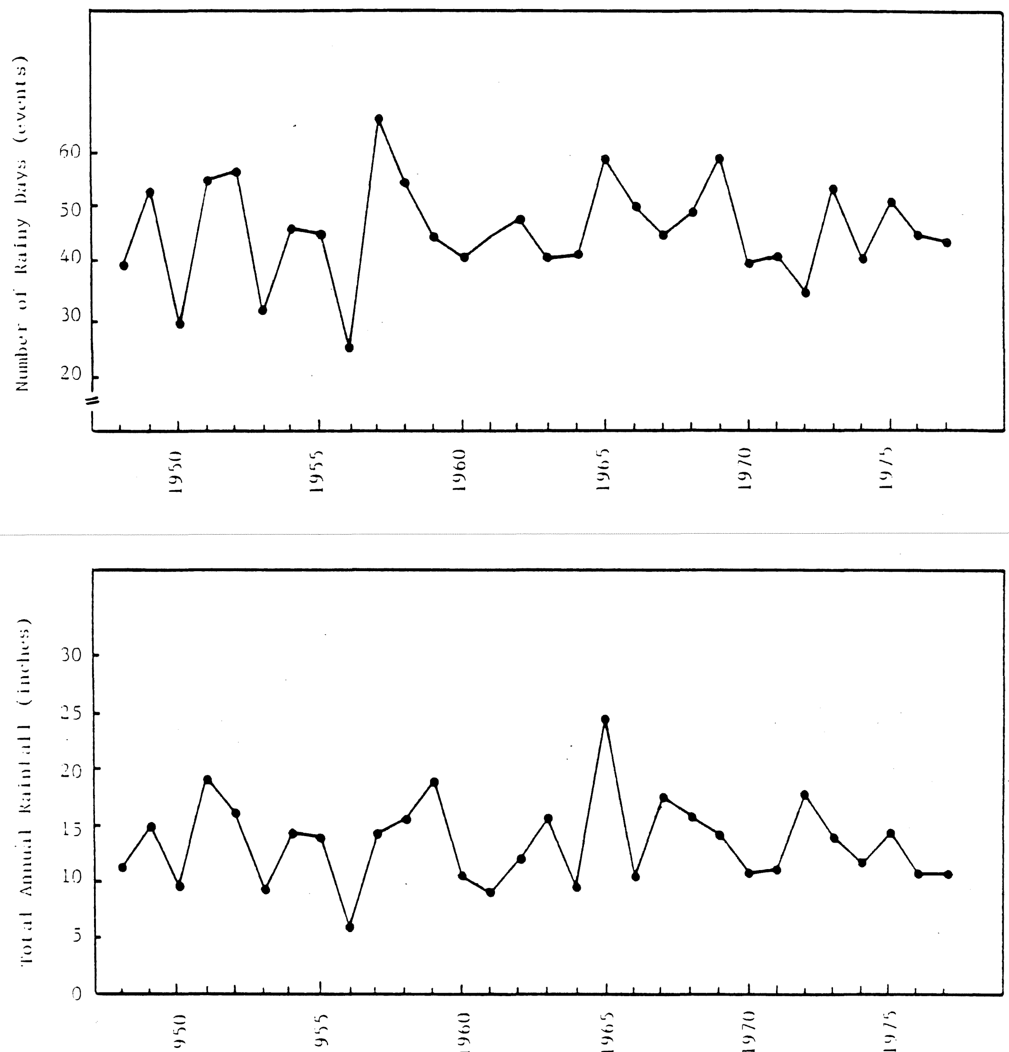


Figure A.2 Number of rainy days per year and total annual rainfall amounts for Roosevelt, Arizona.

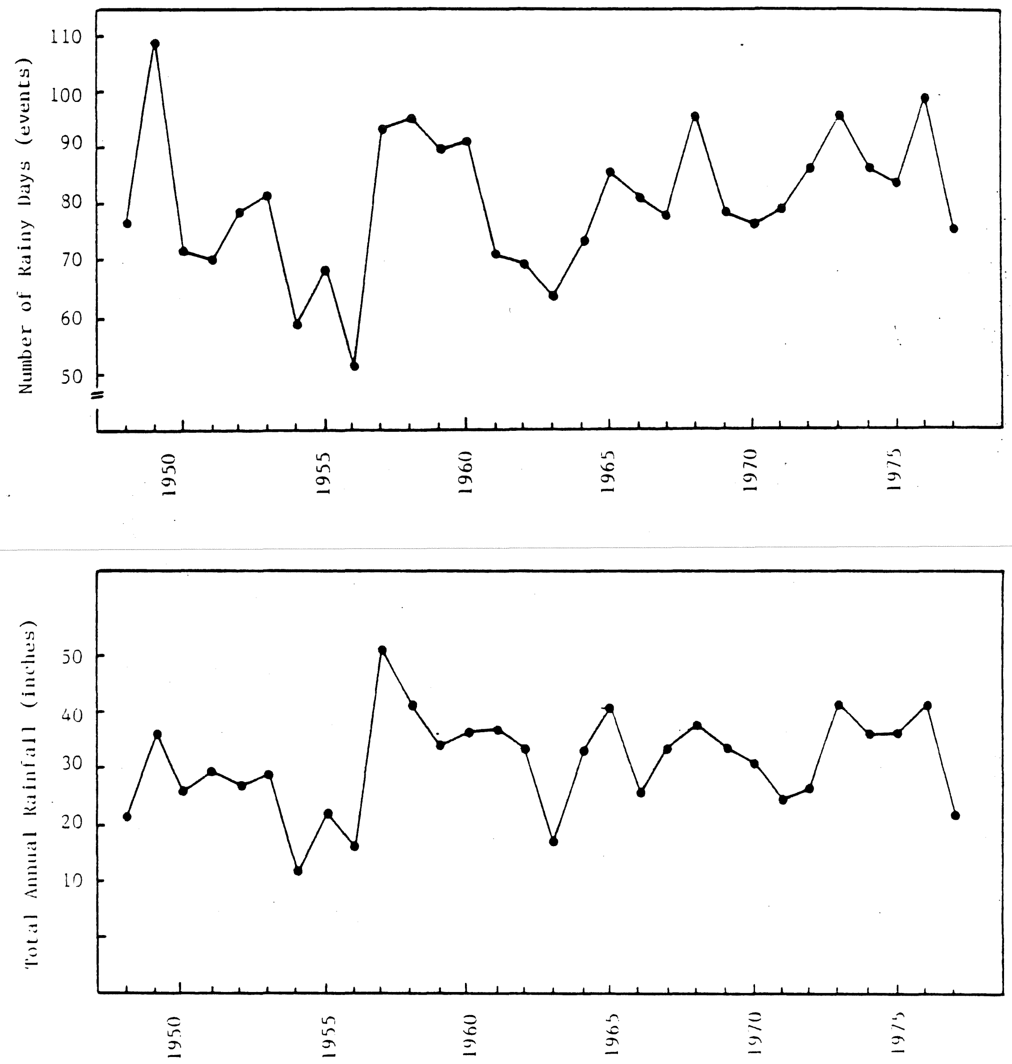


Figure A.3 Number of rainy days per year and total annual rainfall amounts for Austin, Texas.

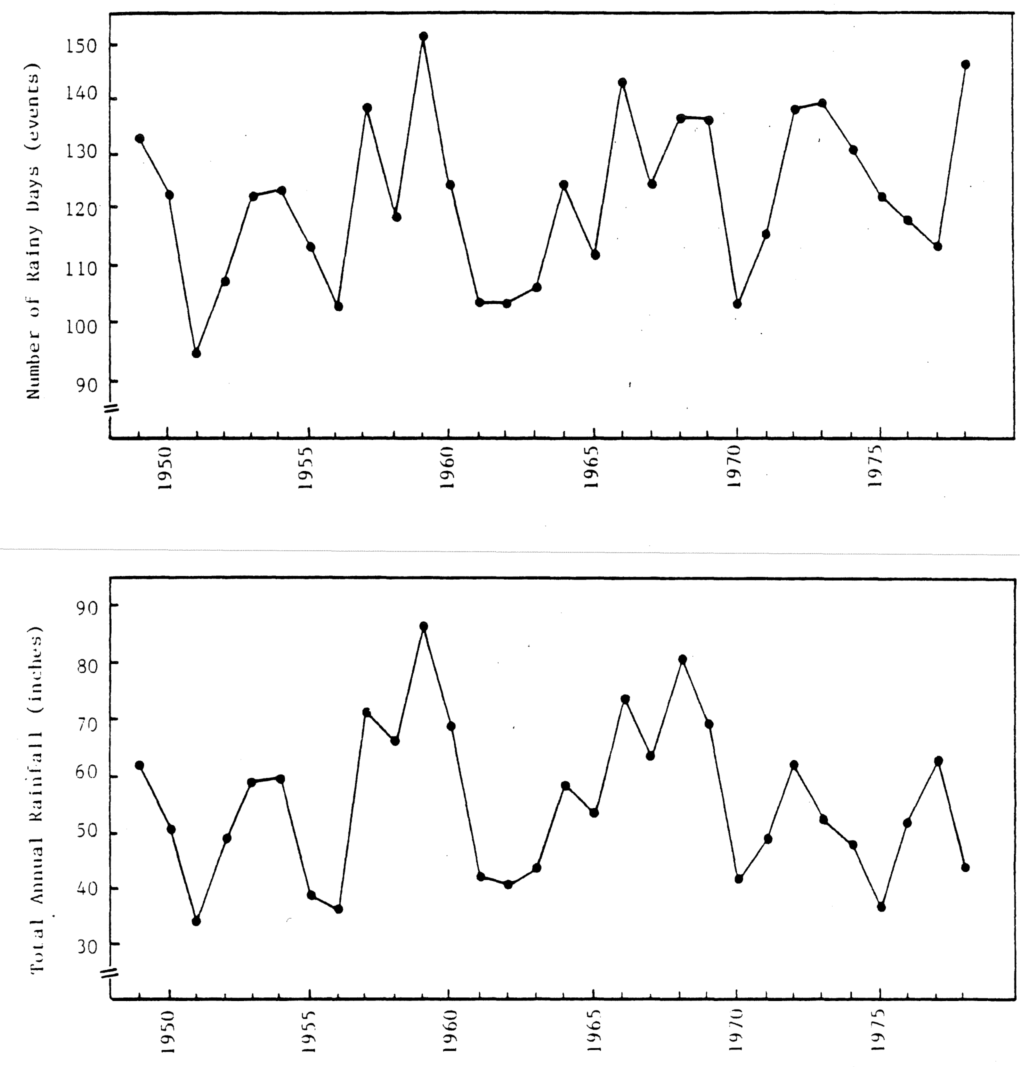


Figure A.4 Number of rainy days per year and total annual rainfall amounts for Miami, Florida.

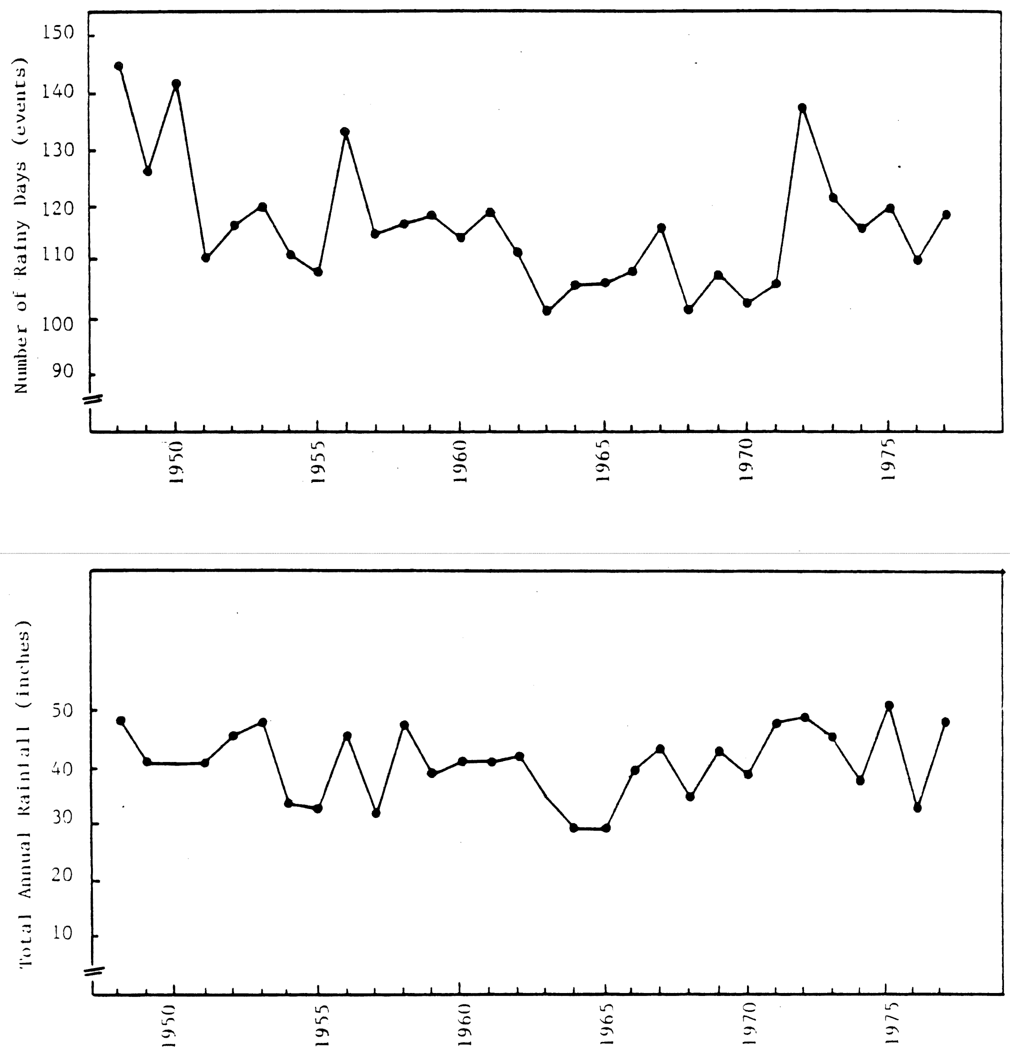
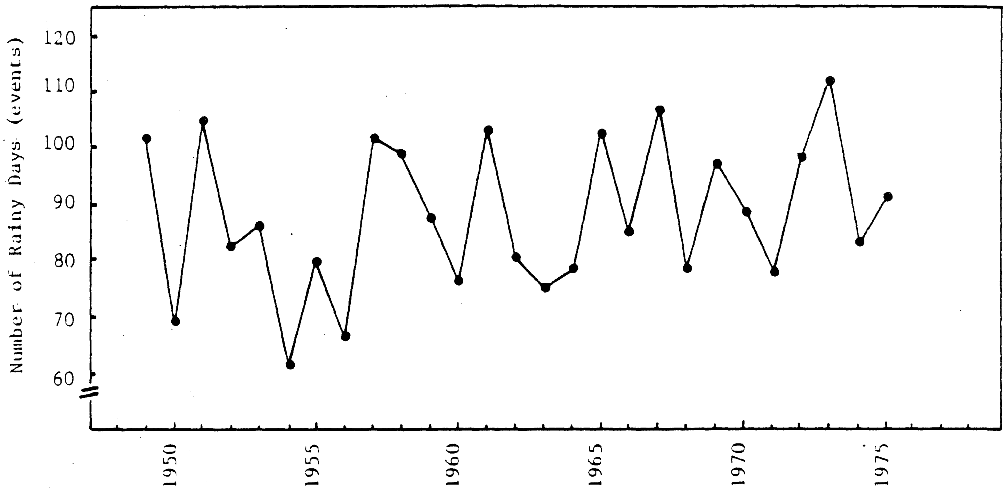


Figure A.5 Number of rainy days per year and total annual rainfall amounts for Philadelphia, Pennsylvania.



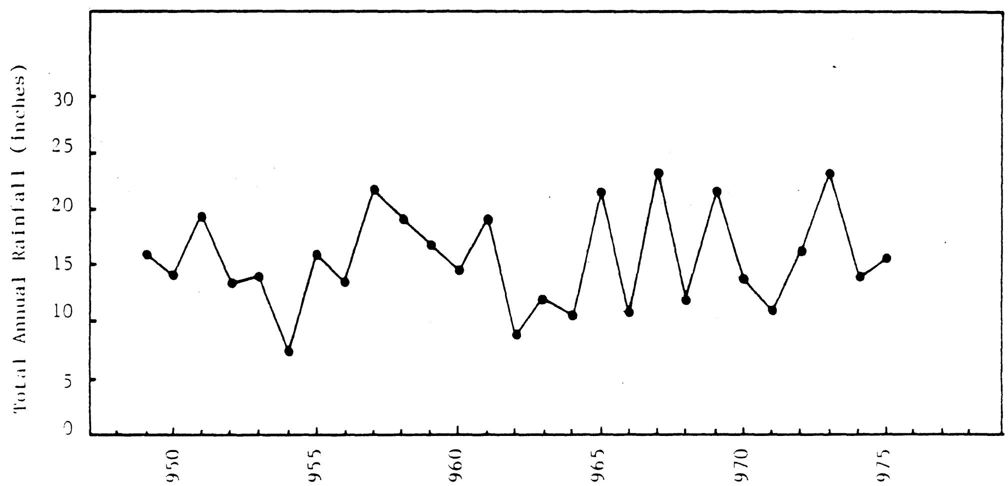


Figure A.6 Number of rainy days per year and total annual rainfall amounts for Denver, Colorado.

Table A.1 Autocorrelation Coefficients of Interarrival Times--Monthly Analysis

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
						(a) Snoqualmie Falls						
r_1	0.026	0.016	0.063	0.074	0.005	-0.077	0.038	0.096	0.072	0.001	-0.002	0.017
r_2	0.004	-0.062	0.046	0.012	0.017	-0.015	0.012	0.111	0.089	0.020	0.043	-0.047
r_3	-0.009	-0.074	-0.014	0.084	-0.012	0.016	-0.064	0.021	0.022	0.005	-0.016	-0.053
r_4	0.077*	-0.021	0.040	-0.056	0.029	0.014	-0.059	0.066	-0.020	0.011	-0.045	0.024
r_5	0.017	-0.010	0.017	-0.072	-0.017	-0.022	-0.127	-0.007	-0.026	0.001	0.019	-0.013
						(b) Roosevelt						
r_1	-0.006	0.008	-0.056	0.039	-0.046	-0.012	-0.112	-0.013	-0.143	0.010	0.025	-0.006
r_2	0.159	0.099	0.079	0.139	-0.206	-0.055	0.073	-0.056	0.052	0.003	0.063	-0.070
r_3	-0.015	-0.100	-0.043	0.135	-0.102	-0.200	-0.008	0.056	0.099	-0.042	0.183	-0.010
r_4	-0.091	-0.007	-0.046	0.045	0.111	-0.172	0.006	0.073	-0.092	0.140	-0.022	-0.018
r_5	-0.063	0.058	-0.031	0.028	-0.161	0.079	-0.098	-0.073	0.011	0.043	-0.065	0.043
						(c) Austin						
r_1	0.034	0.064	0.027	0.097	0.048	0.002	0.011	0.134	-0.055	0.042	0.159*	0.007
r_2	0.061	0.010	0.003	0.007	0.089	-0.006	0.187*	-0.003	-0.027	-0.039	0.070	0.023
r_3	0.084	-0.084	0.116	0.007	0.074	-0.043	-0.106	-0.097	-0.039	-0.062	0.076	-0.005
r_4	0.000	0.167*	0.002	0.158*	0.036	0.058	-0.056	0.001	-0.064	0.011	-0.009	0.154*
r_5	0.183*	0.081	0.000	-0.035	0.016	0.039	-0.047	-0.109	0.046	0.057	-0.006	0.090

Table A.1 (Continued)

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
(d) Miami												
r_1	0.100	-0.051	-0.007	0.031	-0.073	0.014	0.009	0.035	-0.019	0.039	0.026	0.069
r_2	-0.024	0.021	0.094	0.058	0.063	0.019	-0.063	-0.061	-0.006	-0.063	-0.024	0.121
r_3	-0.062	-0.098	0.029	0.064	-0.040	0.066	-0.015	-0.022	-0.011	-0.036	0.014	0.182
r_4	0.012	-0.074	0.027	0.022	-0.034	0.033	-0.013	0.019	-0.011	-0.057	-0.028	0.038
r_5	0.128	-0.051	0.030	-0.106	0.047	0.064	-0.006	0.049	0.001	-0.054	-0.054	0.006
(e) Philadelphia												
r_1	-0.036	-0.045	0.006	0.000	0.022	-0.039	-0.032	-0.050	-0.052	-0.006	-0.087	0.045
r_2	0.052	0.058	-0.060	0.024	0.068	0.018	0.093	0.002	0.080	0.015	-0.015	-0.011
r_3	0.037	-0.052	-0.060	0.008	0.015	-0.043	0.053	-0.080	0.014	-0.015	-0.055	-0.046
r_4	-0.048	0.079	-0.036	0.041	-0.051	-0.028	-0.059	0.114	-0.165*	-0.061	0.062	-0.042
r_5	0.028	-0.060	0.024	-0.032	0.087	0.111	0.040	-0.037	-0.031	0.064	0.024	0.031
(f) Denver												
r_1	-0.097	-0.110	-0.018	-0.146*	0.027	0.004	0.052	-0.014	0.136	0.094	0.044	-0.069
r_2	0.118	0.044	-0.055	0.128*	-0.005	0.014	-0.056	-0.011	-0.013	0.079	0.050	-0.057
r_3	0.063	-0.006	0.003	-0.012	-0.053	-0.032	0.037	0.041	-0.006	0.179*	-0.090	0.081
r_4	0.147	0.072	0.007	0.040	-0.057	0.000	-0.013	-0.014	0.014	0.057	-0.020	0.079
r_5	-0.037	0.012	0.029	0.079	-0.001	-0.095	0.033	-0.007	-0.007	0.179*	-0.024	-0.019

N = sample size
* = significance at the 5% level
** = significance at the 1% level

Table A.2 Statistics of Interarrival Times--Monthly Analysis

Month	$\bar{x}$	s_x	c_v	c_s	Number of events
(a) Snoqualmie Falls					
Jan	1.393	1.171	0.841	3.825	667
Feb	1.499	1.328	0.886	3.933	557
Mar	1.542	1.548	1.004	5.542	607
Apr	1.717	1.588	0.925	3.060	530
May	2.214	2.501	1.129	3.190	429
Jun	2.693	4.354	1.617	4.941	375
Jul	4.259	6.080	1.428	2.912	205
Aug	3.323	4.226	1.272	2.259	260
Sep	2.708	3.764	1.390	4.266	332
Oct	1.752	1.668	0.952	3.084	499
Nov	1.442	1.231	0.854	4.217	613
Dec	1.343	0.978	0.728	4.167	694
(b) Roosevelt					
Jan	6.241	9.059	1.452	2.500	145
Feb	6.224	10.174	1.635	2.978	125
Mar	7.604	13.184	1.734	3.366	139
Apr	12.693	21.596	1.701	2.399	88
May	22.524	21.968	0.975	0.601	42
Jun	14.455	13.661	0.945	0.480	33
Jul	4.671	4.759	1.019	1.739	152
Aug	6.052	10.880	1,798	4.345	191
Sep	7.319	9.675	1.322	2.533	116
Oct	9.066	13.439	1.482	1.961	106
Nov	8.216	11.108	1.352	2.138	97
Dec	7.761	16.275	2.097	6.345	134
(c) Austin					
Jan	3.513	3.973	1.313	2.017	240
Feb	3.881	5.315	1.369	4.277	227
Mar	4.480	4.484	1.001	1.585	200
Apr	3.808	4.048	1.063	2.329	224
May	4.232	5.258	1.242	2.656	241
Jun	5.197	6.972	1.342	2.038	178
Jul	6.553	8.724	1.331	2.466	141
Aug	5.278	6.531	1.237	2.251	162
Sep	4.613	6.471	1.403	3.195	212
Oct	4.654	5.988	1.287	3.254	188
Nov	4.615	6.059	1.313	3.613	192
Dec	4.359	5.924	1.359	3.069	209

Table A.2 (Continued)

Month	$\bar{x}$	s_x	c_v	c_s	Number of events
			(d) Miami		
Jan	4.640	4.891	1.054	2.243	197
Feb	6.160	6.372	1.034	1.631	131
Mar	5.802	6.199	1.068	1.866	162
Apr	4.989	5.751	1.153	1.958	174
May	2.261	2.615	1.157	3.333	375
Jun	1.989	1.974	0.993	3.248	444
Jul	2.157	1.919	0.889	2.476	439
Aug	1.829	1.472	0.805	2.396	502
Sep	1.840	1.625	0.884	2.642	486
Oct	2.650	3.068	1.158	3.244	366
Nov	4.741	5.135	1.083	1.994	197
Dec	4.913	5.521	1.124	2.158	195
			(e) Philadelphia		
Jan	2.711	2.517	0.928	2.107	332
Feb	2.978	2.555	0.858	1.865	278
Mar	2.854	2.521	0.883	2.118	323
Apr	2.888	2.782	0.963	2.275	322
May	2.879	2.930	1.018	2.726	323
Jun	2.993	2.680	0.895	2.076	297
Jul	3.395	2.843	0.837	1.391	266
Aug	3.502	3.845	1.098	2.638	273
Sep	3.737	3.867	1.035	2.210	243
Oct	4.221	4.486	1.063	2.130	213
Nov	3.251	3.245	0.998	2.239	287
Dec	2.807	2.503	0.892	1.706	316
			(f) Denver		
Jan	5.458	6.015	1.102	1.777	168
Feb	4.421	4.794	1.084	2.502	164
Mar	3.668	4.513	1.230	2.950	244
Apr	3.552	4.078	1.148	2.130	259
May	2.965	3.071	1.036	2.006	283
Jun	3.483	4.047	1.162	2.752	261
Jul	3.114	3.132	1.006	2.002	280
Aug	3.836	4.035	1.052	2.447	244
Sep	4.872	6.304	1.294	2.977	187
Oct	5.993	7.049	1.176	1.861	141
Nov	5.813	6.121	1.053	1.633	155
Dec	5.560	7.067	1.271	2.439	150

Table A.3 Autocorrelation Coefficients of Non-Zero Daily Rainfall Amounts--Monthy Analysis

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
(a) Snoqualmie Falls												
r_1	0.296**	0.216**	0.126**	0.090*	0.013	0.077	-0.067	0.024	0.057	0.053	0.065	0.136**
r_2	0.054	0.030	-0.014	-0.066	0.044	0.016	-0.074	0.109	-0.058	0.009	-0.017	0.007
r_3	0.039	-0.075	-0.017	0.096*	-0.044	-0.038	-0.051	0.086	-0.092	0.010	0.005	0.026
r_4	0.028	-0.018	-0.022	-0.020	-0.067	0.083	-0.114	0.070	-0.030	-0.067	0.112	-0.010
r_5	-0.006	0.045	0.017	-0.046	-0.089	-0.066	0.017	0.017	0.013	-0.076	0.030	-0.026
(b) Roosevelt												
r_1	0.113	0.002	0.062	0.020	0.032	0.410*	-0.072	0.049	-0.087	0.213*	0.065	0.136
r_2	-0.066	-0.075	0.022	-0.100	0.074	-0.146	-0.092	-0.030	-0.025	0.020	0.107	-0.010
r_3	-0.008	0.141	-0.071	0.034	0.120	-0.186	0.042	-0.030	0.009	0.095	-0.018	0.033
r_4	-0.074	-0.044	0.056	0.085	0.036	-0.070	0.023	0.015	-0.053	0.043	0.015	-0.001
r_5	0.016	-0.025	0.158	0.211	0.403*	0.045	-0.177*	0.074	0.040	-0.109	-0.114	0.027
(c) Austin												
r_1	0.244**	0.086	0.018	0.038	0.017	0.153*	0.132	-0.040	0.017	0.106	0.112	0.123
r_2	0.123*	-0.029	0.031	0.014	-0.035	0.100	-0.083	0.017	-0.108	-0.064	-0.007	0.002
r_3	0.098	-0.069	-0.016	0.123*	0.073	0.015	0.000	-0.103	-0.075	-0.039	-0.044	0.017
r_4	0.200**	-0.022	-0.090	-0.015	0.028	0.085	0.191*	0.078	-0.007	0.072	-0.020	-0.115
r_5	-0.016	-0.035	-0.081	-0.085	0.039	0.062	-0.028	-0.004	0.092	-0.093	-0.067	-0.099

Table A.3 (Continued)

						(d) Miami						
r_1	0.118	0.092	0.022	-0.076	0.149**	0.158	-0.011	0.026	0.143	0.108	0.076	0.102
r_2	0.129	0.066	-0.038	0.023	-0.052	0.019	0.038	-0.036	0.041	-0.081	0.010	-0.048
r_3	0.007	-0.019	-0.088	-0.041	-0.015	-0.093	-0.021	-0.010	0.015	0.030	-0.010	-0.029
r_4	-0.63	-0.047	-0.069	-0.005	-0.050	-0.113	0.081	-0.006	0.102	-0.011	0.007	0.127
r_5	-0.108	0.066	-0.061	-0.071	-0.009	0.005	-0.021	0.028	0.044	0.014	-0.055	-0.057
						(e) Philadelphia						
r_1	0.009	-0.125*	0.025	0.003	0.007	0.094	-0.033	0.125*	0.011	-0.017	0.019	-0.056
r_2	0.023	-0.018	0.013	0.121*	0.012	-0.040	-0.041	0.007	0.001	-0.067	-0.051	0.045
r_3	-0.050	0.094	-0.041	-0.060	0.120*	-0.014	-0.013	-0.046	0.033	-0.057	0.072	-0.016
r_4	0.016	-0.037	0.015	-0.073	-0.004	-0.060	0.012	-0.042	-0.035	0.101	0.003	0.041
r_5	-0.003	0.072	-0.150**	0.013	0.013	0.060	0.041	-0.028	0.004	0.052	0.088	-0.119*
						(f) Denver						
r_1	0.005	-0.143	0.057	-0.043	0.083	0.119*	-0.008	0.116	0.025	0.054	-0.055	0.345**
r_2	-0.047	-0.015	0.008	-0.042	-0.025	-0.022	0.037	-0.096	-0.080	-0.079	0.020	0.184**
r_3	-0.071	-0.054	0.028	-0.040	0.003	-0.061	0.126*	0.048	0.064	0.016	-0.003	0.029
r_4	-0.073	-0.030	-0.035	0.065	0.001	-0.070	-0.041	0.118	0.016	0.049	-0.027	-0.057
r_5	-0.043	0.077	-0.066	-0.043	0.008	-0.061	0.050	-0.091	-0.119	0.057	0.110	0.052

N = sample size
* = significance at the 5% level
** = significance at the 1% level

Table A.4 Statistics of Non-Zero Daily Rainfall Amounts--Monthly Analysis

Month	$\bar{x}$	s_x	c_v	c_s
		(a) Snoqualmie Falls		
Jan	0.415	0.462	1.112	2.060
Feb	0.360	0.449	1.247	3.196
Mar	0.306	0.371	1.212	3.914
Apr	0.253	0.264	1.041	1.695
May	0.228	0.267	1.174	2.640
Jun	0.231	0.300	1.301	2.641
Jul	0.217	0.281	1.296	2.099
Aug	0.213	0.273	1.282	2.417
Sep	0.227	0.335	1.208	1.922
Oct	0.337	0.362	1.073	1.789
Nov	0.408	0.459	1.125	1.940
Dec	0.408	0.480	1.177	2.575
		(b) Roosevelt		
Jan	0.308	0.347	1.125	1.622
Feb	0.239	0.270	1.130	1.941
Mar	0.305	0.353	1.158	2.632
Apr	0.185	0.215	1.162	1.930
May	0.169	0.205	1.217	2.157
Jun	0.278	0.327	1.174	1.658
Jul	0.262	0.352	1.343	2.757
Aug	0.278	0.367	1.322	2.801
Sep	0.319	0.451	1.412	2.565
Oct	0.366	0.577	1.578	3.521
Nov	0.289	0.379	1.314	2.627
Dec	0.395	0.471	1.192	1.973
		(c) Austin		
Jan	0.204	0.349	1.715	4.647
Feb	0.333	0.535	1.606	2.604
Mar	0.230	0.326	1.415	2.658
Apr	0.457	0.623	1.365	2.084
May	0.485	0.683	1.409	2.539
Jun	0.530	0.738	1.392	2.261
Jul	0.352	0.607	1.727	3.742
Aug	0.414	0.621	1.502	3.003
Sep	0.502	0.725	1.445	2.612
Oct	0.558	0.821	1.471	2.721
Nov	0.301	0.537	1.781	4.241
Dec	0.287	0.486	1.694	3.449

Table A.4 (Continued)

Month	$\bar{x}$	s_x	c_v	c_s
(d) Miami				
Jan	0.316	0.411	1.302	1.994
Feb	0.373	0.588	1.578	3.527
Mar	0.359	0.686	1.910	6.316
Apr	0.585	1.161	1.986	6.151
May	0.584	0.884	1.512	3.195
Jun	0.562	0.778	1.385	2.745
Jul	0.367	0.478	1.300	2.592
Aug	0.424	0.637	1.502	3.501
Sep	0.469	0.667	1.443	3.041
Oct	0.482	0.828	1.719	3.917
Nov	0.372	0.831	2.235	4.400
Dec	0.280	0.411	1.471	2.263
(e) Philadelphia				
Jan	0.265	0.334	1.260	2.416
Feb	0.300	0.338	1.128	1.758
Mar	0.349	0.393	1.125	1.861
Apr	0.317	0.388	1.227	2.169
May	0.307	0.387	1.259	2.095
Jun	0.387	0.600	1.551	2.985
Jul	0.423	0.587	1.388	2.293
Aug	0.449	0.618	1.337	2.707
Sep	0.421	0.650	1.545	3.410
Oct	0.381	0.512	1.345	2.488
Nov	0.355	0.525	1.479	3.309
Dec	0.335	0.394	1.176	1.604
(f) Denver				
Jan	0.093	0.137	1.480	3.117
Feb	0.120	0.151	1.257	2.471
Mar	0.145	0.182	1.253	2.904
Apr	0.205	0.369	1.805	4.948
Mar	0.249	0.435	1.748	3.526
Jun	0.194	0.365	1.878	4.499
Jul	0.198	0.304	1.534	2.679
Aug	0.161	0.288	1.788	3.502
Sep	0.202	0.283	1.402	2.358
Oct	0.198	0.258	1.301	2.518
Nov	0.146	0.157	1.075	1.455
Dec	0.102	0.141	1.386	4.267

Table A.5 Cross Correlation Coefficients of the Non-Zero Daily Rainfall Amounts with Preceding and Following Interarrival Times (X_i = interarrival time following the event P_i)

	(X_{i-2}, P_i)	(X_{i-1}, P_i)	(X_i, P_i)	(X_{i+1}, P_i)
		(a) Snoqualmie Falls		
Jan	-0.036	-0.080*	-0.089*	-0.041
Feb	0.033	-0.120**	-0.129**	-0.047
Mar	-0.063	-0.022	-0.011	-0.065
Apr	-0.068	-0.035	0.006	0.032
May	-0.019	-0.024	-0.117	-0.124*
Jun	-0.055	-0.037	-0.007	-0.051
Jul	-0.029	-0.037	-0.198**	-0.118
Aug	-0.001	-0.099	-0.132*	-0.168**
Sep	0.005	-0.086	-0.117*	-0.095
Oct	0.035	-0.064	-0.092*	-0.149**
Nov	0.033	0.030	-0.111**	-0.097**
Dec	-0.028	-0.134**	-0.132**	0.013
		(b) Roosevelt		
Jan	-0.046	-0.084	-0.112	-0.103
Feb	-0.056	-0.010	0.059	0.102
Mar	0.247**	0.026	-0.058	-0.188*
Apr	0.049	0.194	-0.166	0.044
May	0.164	-0.067	-0.201	-0.065
Jun	-0.102	-0.036	-0.011	0.248
Jul	-0.085	0.019	-0.014	-0.052
Aug	-0.035	-0.019	-0.007	-0.021
Sep	-0.025	-0.075	0.064	-0.006
Oct	-0.029	0.004	-0.157	-0.167
Nov	-0.066	0.013	0.039	0.048
Dec	-0.091	-0.007	-0.121	-0.130
		(c) Austin		
Jan	-0.021	0.070	-0.023	0.013
Feb	0.044	0.037	0.128	-0.016
Mar	0.165*	-0.001	-0.104	0.007
Apr	-0.016	-0.084	-0.070	-0.008
May	0.025	-0.039	0.054	0.017
Jun	-0.048	0.125	0.057	0.168*
Jul	-0.055	-0.031	-0.063	0.060
Aug	0.016	0.044	-0.085	-0.109
Sep	0.033	-0.006	0.041	0.105
Oct	-0.005	0.029	0.007	0.061
Nov	0.009	-0.027	-0.021	-0.089
Dec	0.068	0.276**	-0.046	-0.009

Table A.5 (Continued)

	(X_{i-2},P_i)	(X_{i-1},P_i)	(X_i, P_i)	(X_{i+1},P_i)
		(d) Miami		
Jan	0.017	0.159	0.061	-0.115
Feb	-0.048	-0.088	-0.033	-0.054
Mar	0.078	-0.008	-0.078	0.035
Apr	-0.041	-0.137	0.035	-0.143
May	-0.050	-0.081	-0.080	-0.033
Jun	-0.014	0.008	-0.019	-0.011
Jul	-0.029	-0.038	-0.029	-0.014
Aug	0.018	-0.071	-0.016	-0.004
Sep	0.037	-0.016	-0.080	-0.001
Oct	-0.052	0.033	0.043	0.033
Nov	0.020	-0.061	-0.126	-0.114
Dec	0.077	-0.094	-0.029	-0.047
		(e) Philadelphia		
Jan	0.073	0.019	0.057	0.139*
Feb	0.039	-0.087	0.008	0.045
Mar	0.076	0.062	0.079	0.008
Apr	0.044	0.047	-0.061	-0.034
May	-0.094	-0.061	-0.015	0.001
Jun	-0.029	0.130*	0.049	-0.048
Jul	-0.013	-0.008	-0.010	-0.028
Aug	0.095	0.142*	-0.075	-0.087
Sep	0.000	0.059	0.120	-0.035
Oct	0.101	0.050	0.064	0.143*
Nov	-0.026	-0.051	0.055	-0.003
Dec	0.112*	-0.045	-0.051	-0.030
		(f) Denver		
Jan	-0.048	-0.020	-0.112	0.000
Feb	-0.003	-0.056	0.050	-0.054
Mar	-0.010	0.009	0.045	0.015
Apr	-0.049	0.204**	-0.056	0.004
May	-0.052	0.019	0.071	-0.104
Jun	0.093	-0.010	-0.076	-0.060
Jul	-0.045	-0.035	-0.040	-0.002
Aug	-0.050	0.077	-0.007	-0.050
Sep	0.046	-0.031	-0.004	-0.049
Oct	0.035	0.092	-0.068	-0.103
Nov	0.127	-0.017	0.145	0.065
Dec	0.163*	0.017	0.051	-0.052

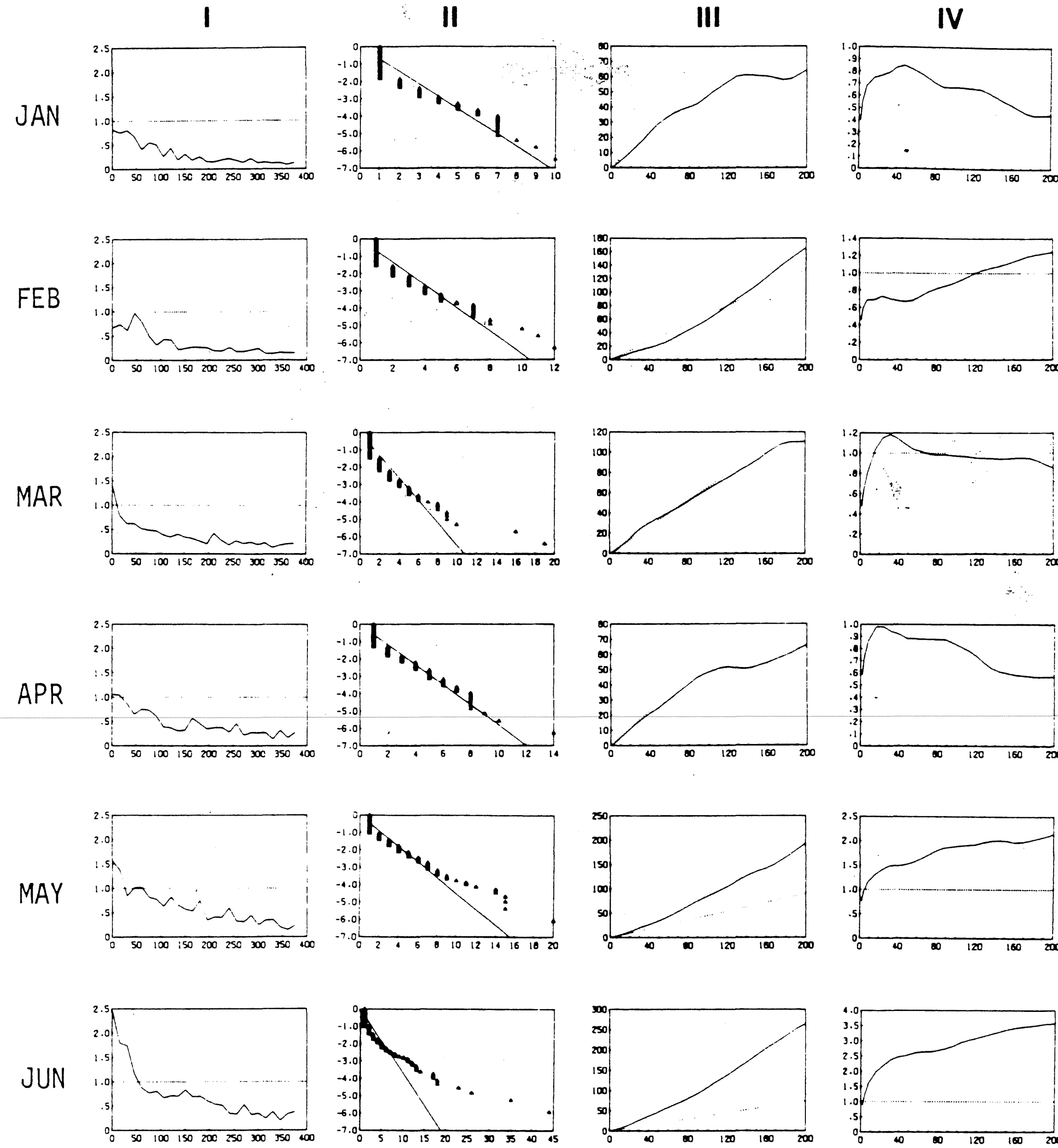


Figure A.7 Statistical properties of intervals and counts for Snoqualmie Falls--monthly analysis.

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

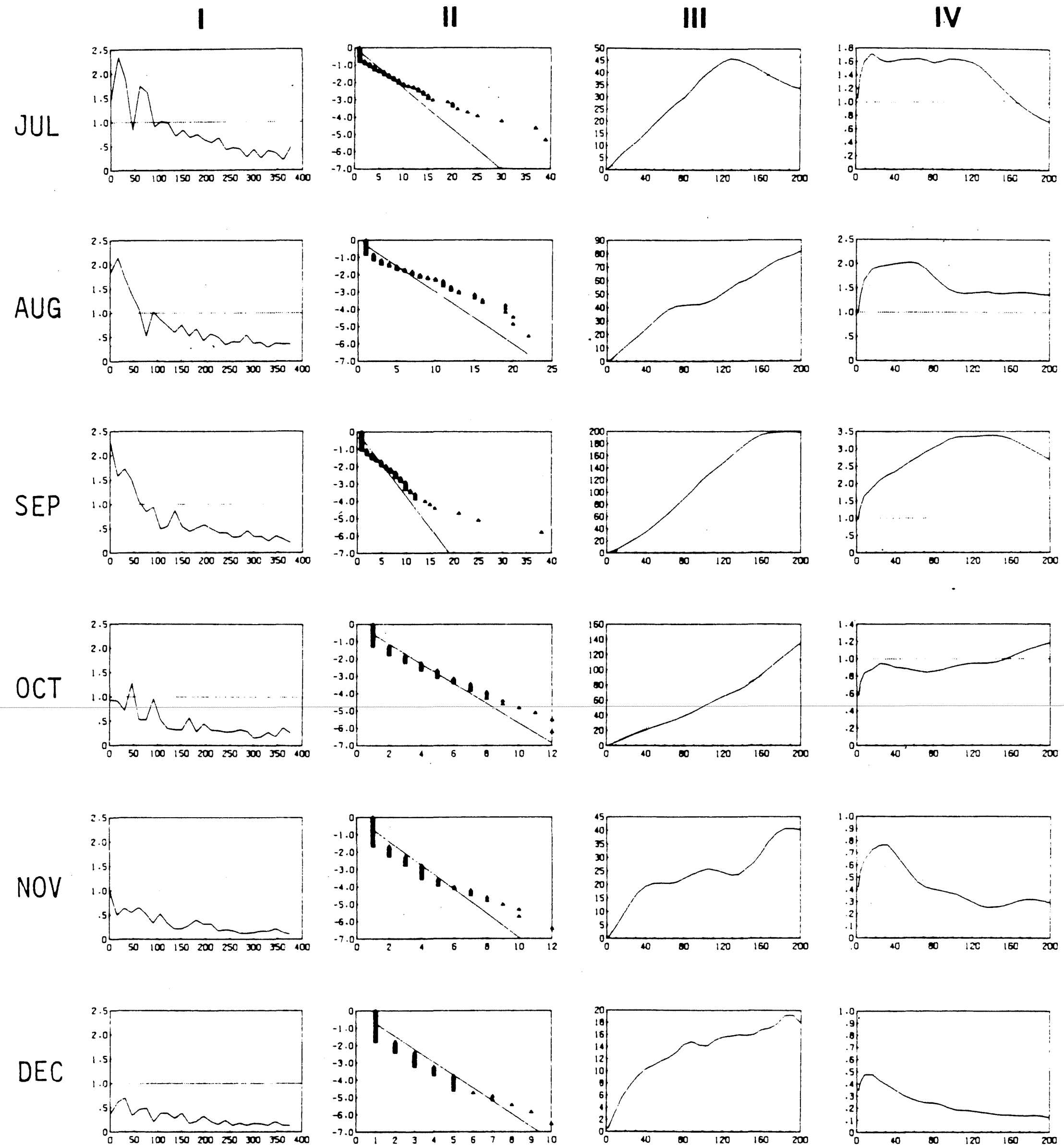


Figure A.7 (continued)

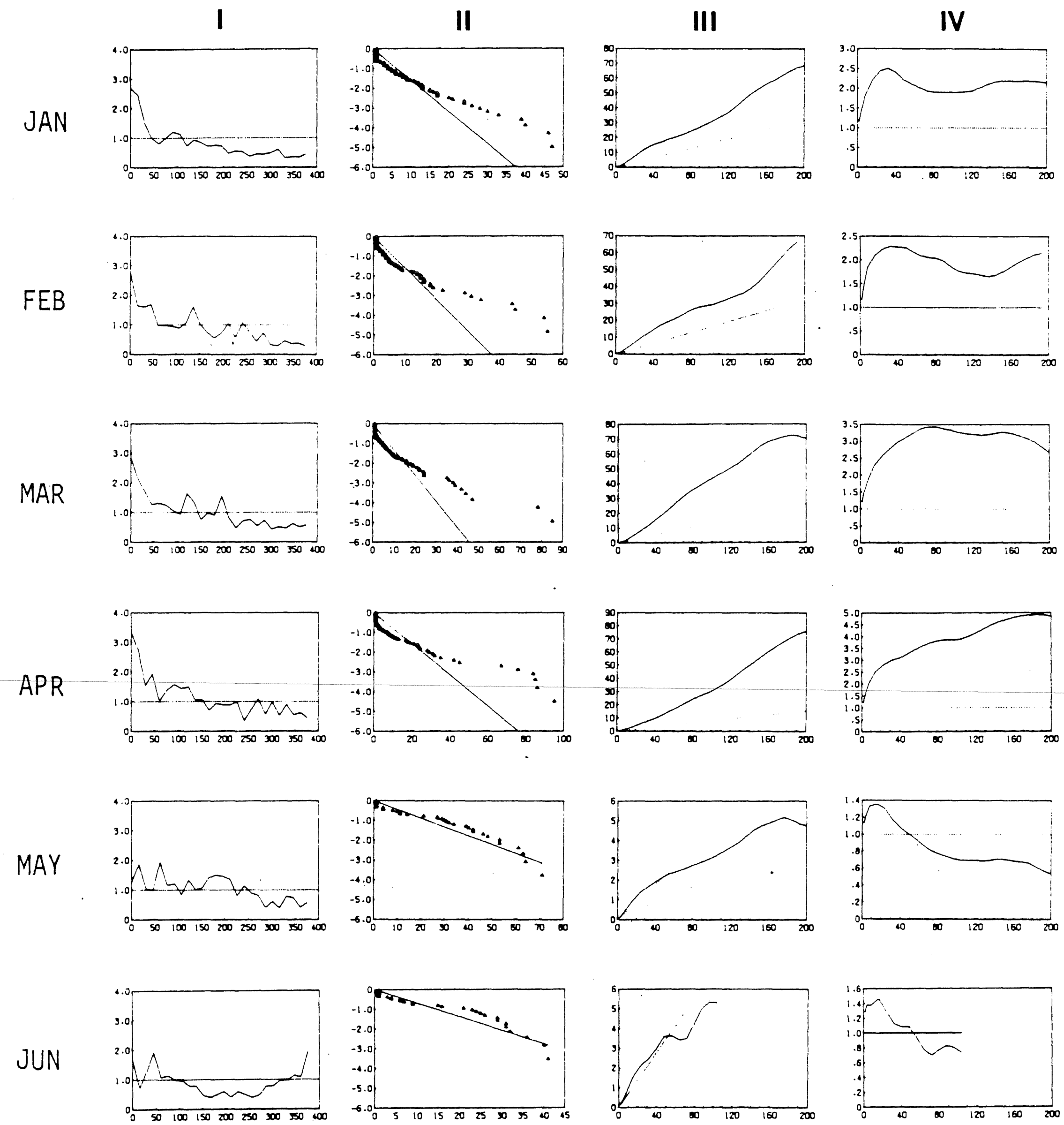


Figure A.8 Statistical properties of intervals and counts for Roosevelt--monthly analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

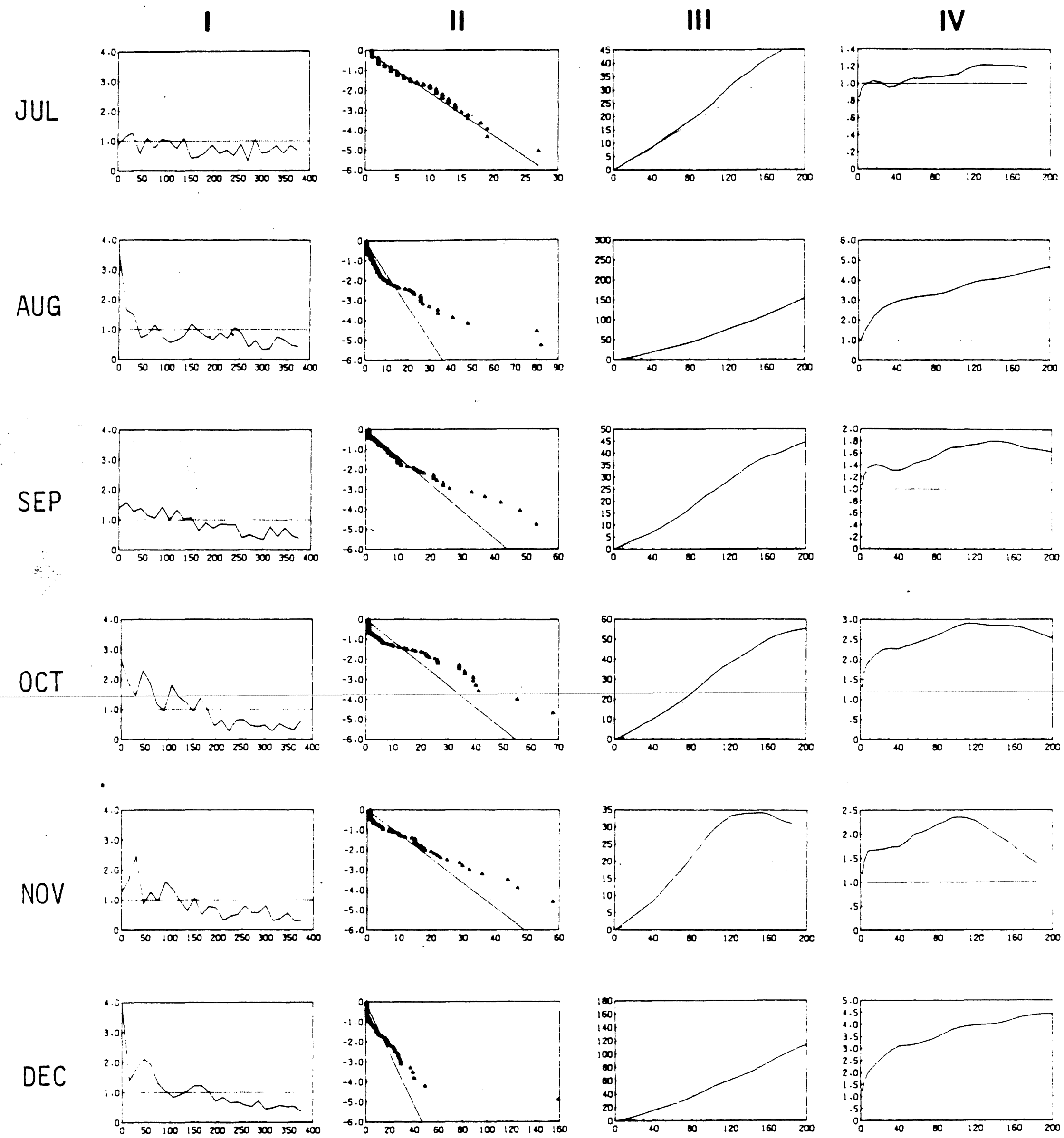


Figure A.8 (continued)

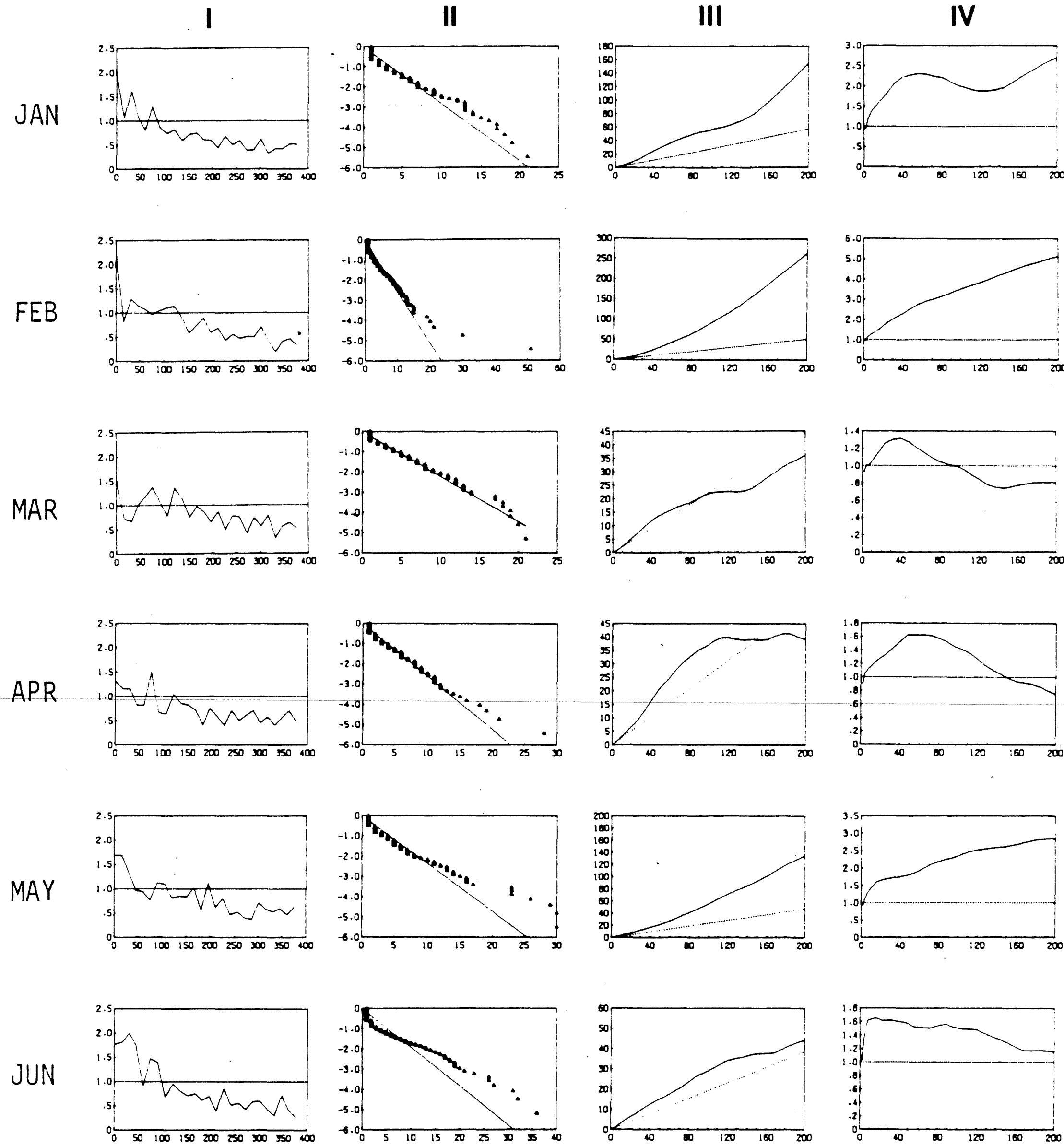


Figure A.9 Statistical properties of intervals and counts for Austin--monthly analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

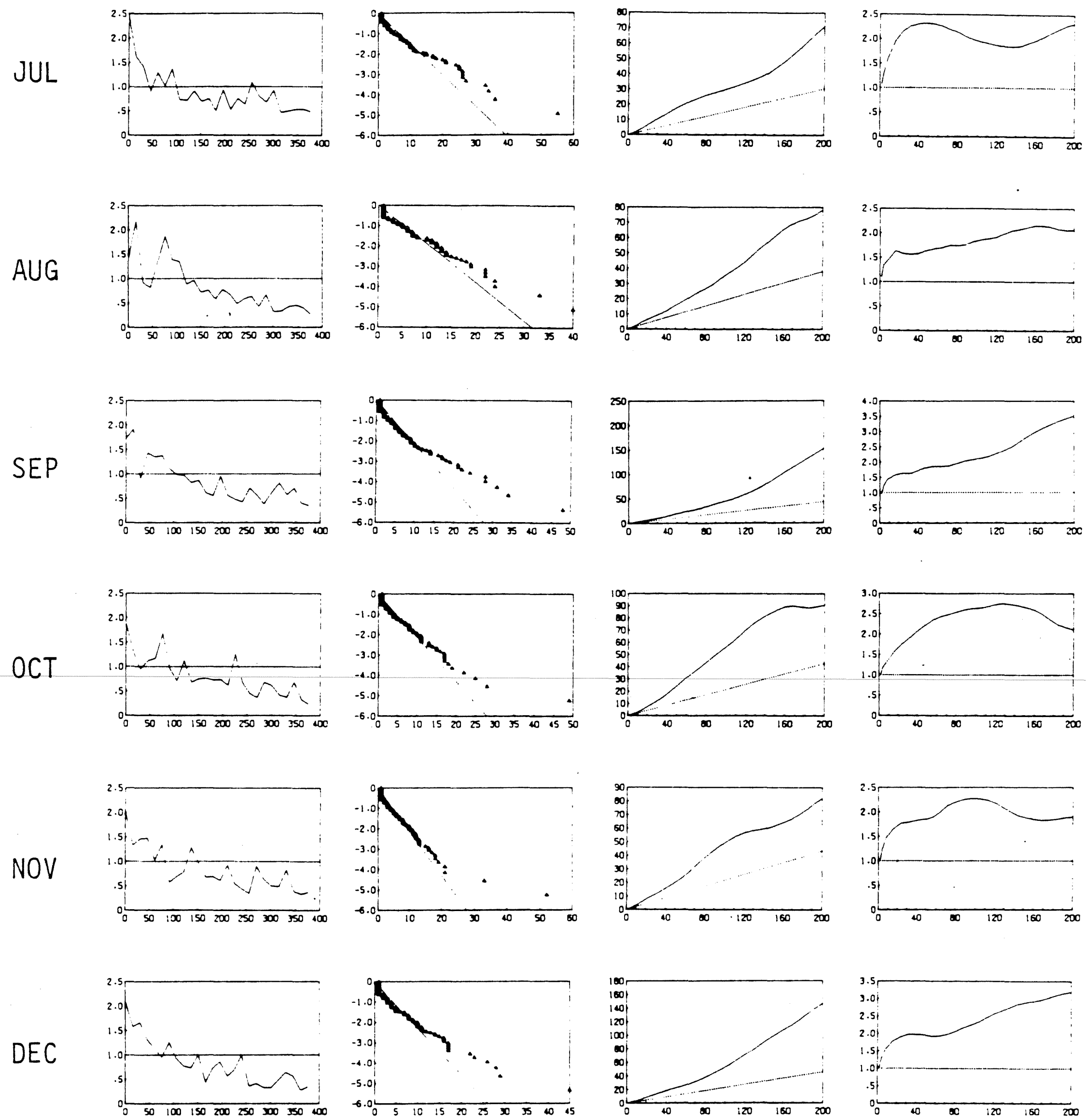


Figure A.9 (continued)

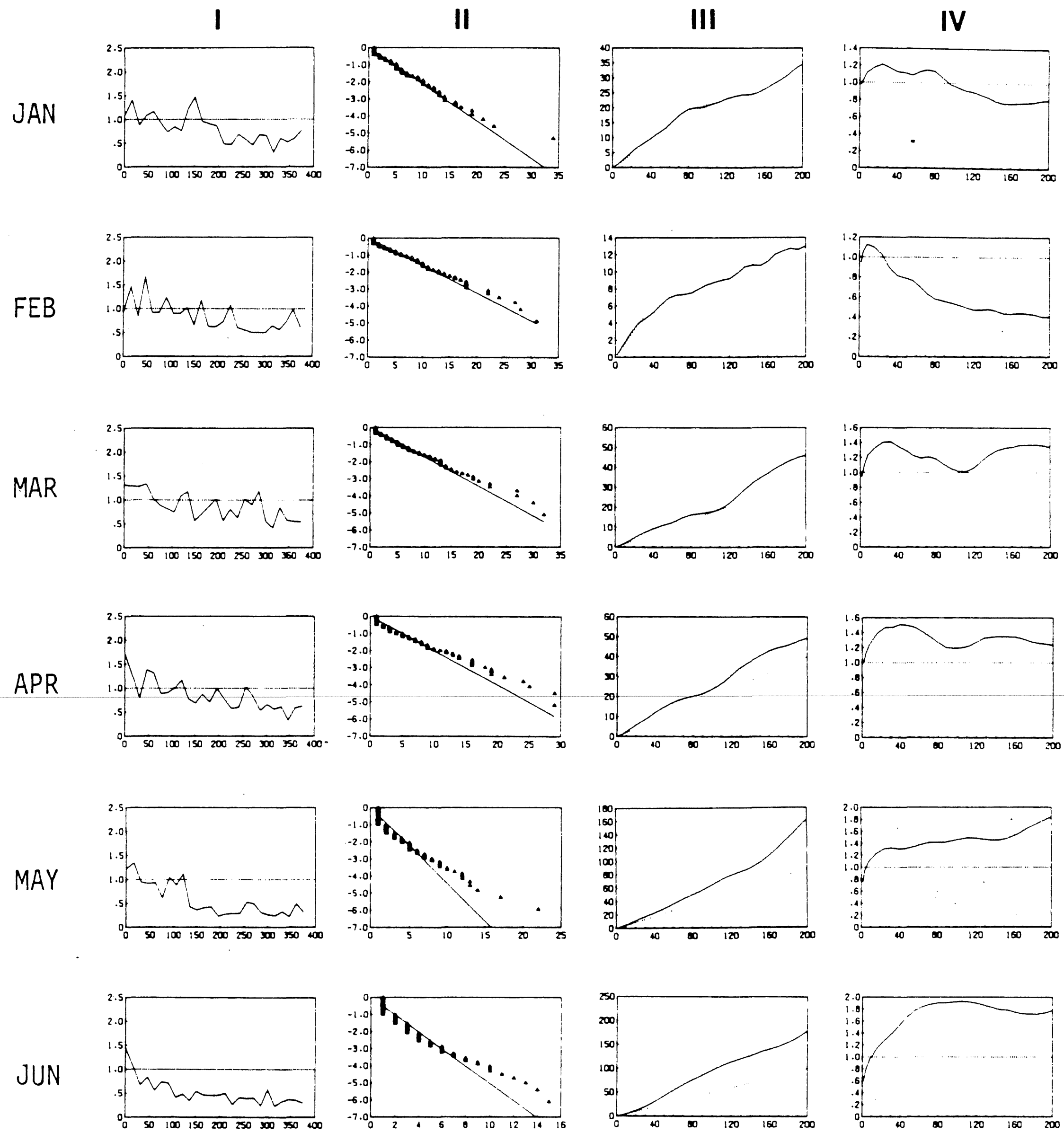


Figure A.10 Statistical properties of intervals and counts for Miami--monthly analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

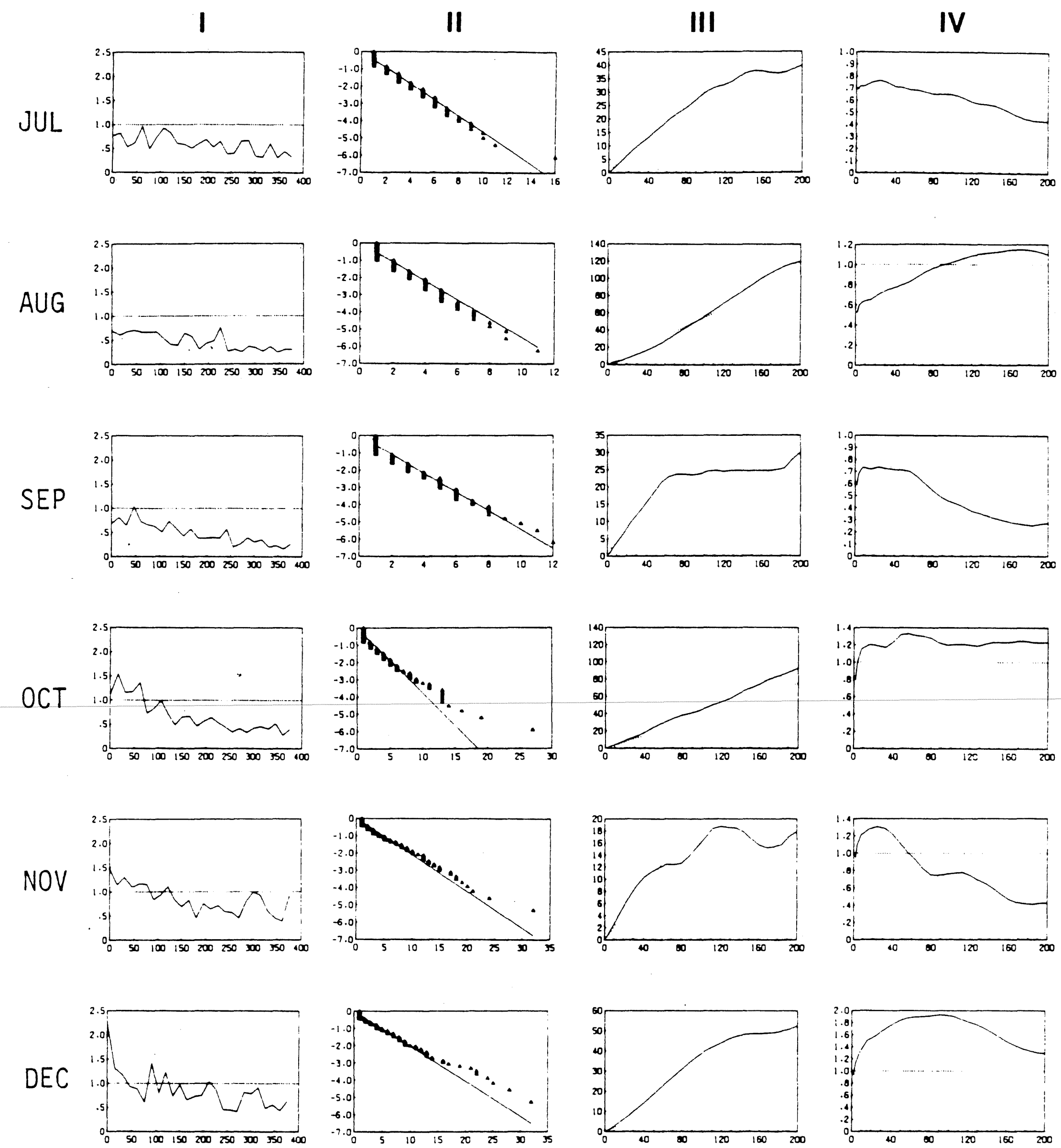


Figure A.10 (continued)

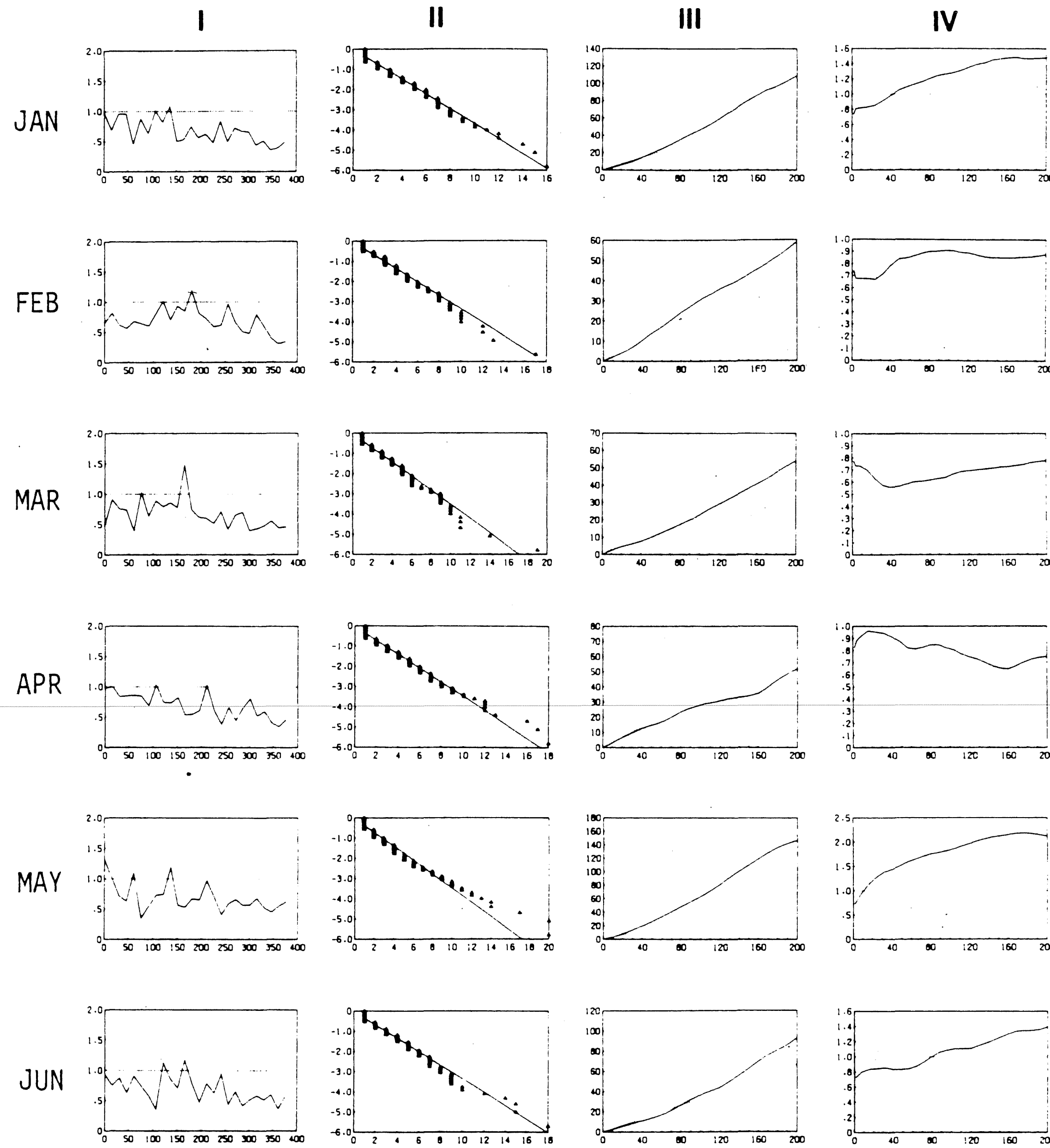


Figure A.11 Statistical properties of intervals and counts for Philadelphia--monthly analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

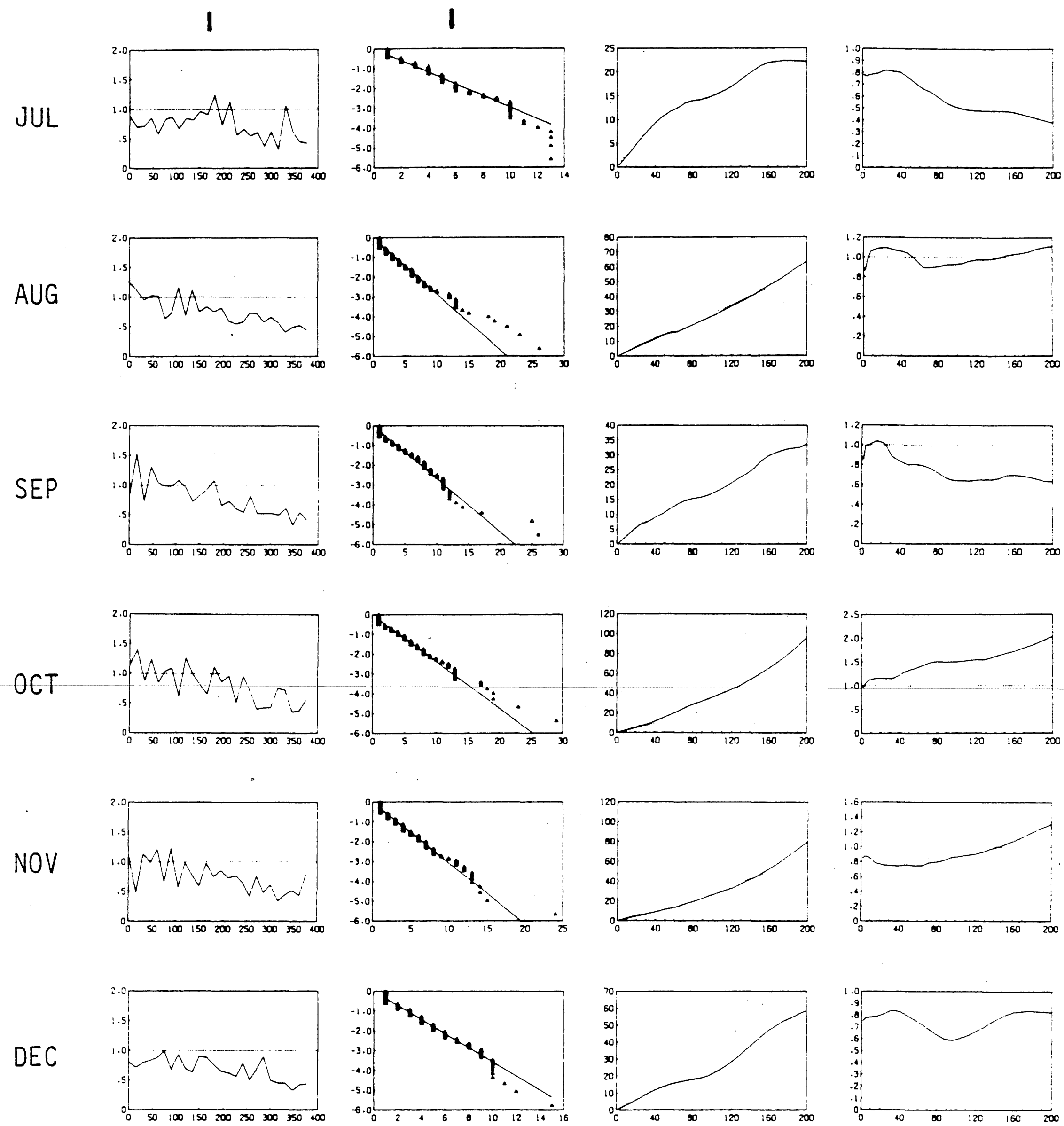


Figure A.11 (continued)

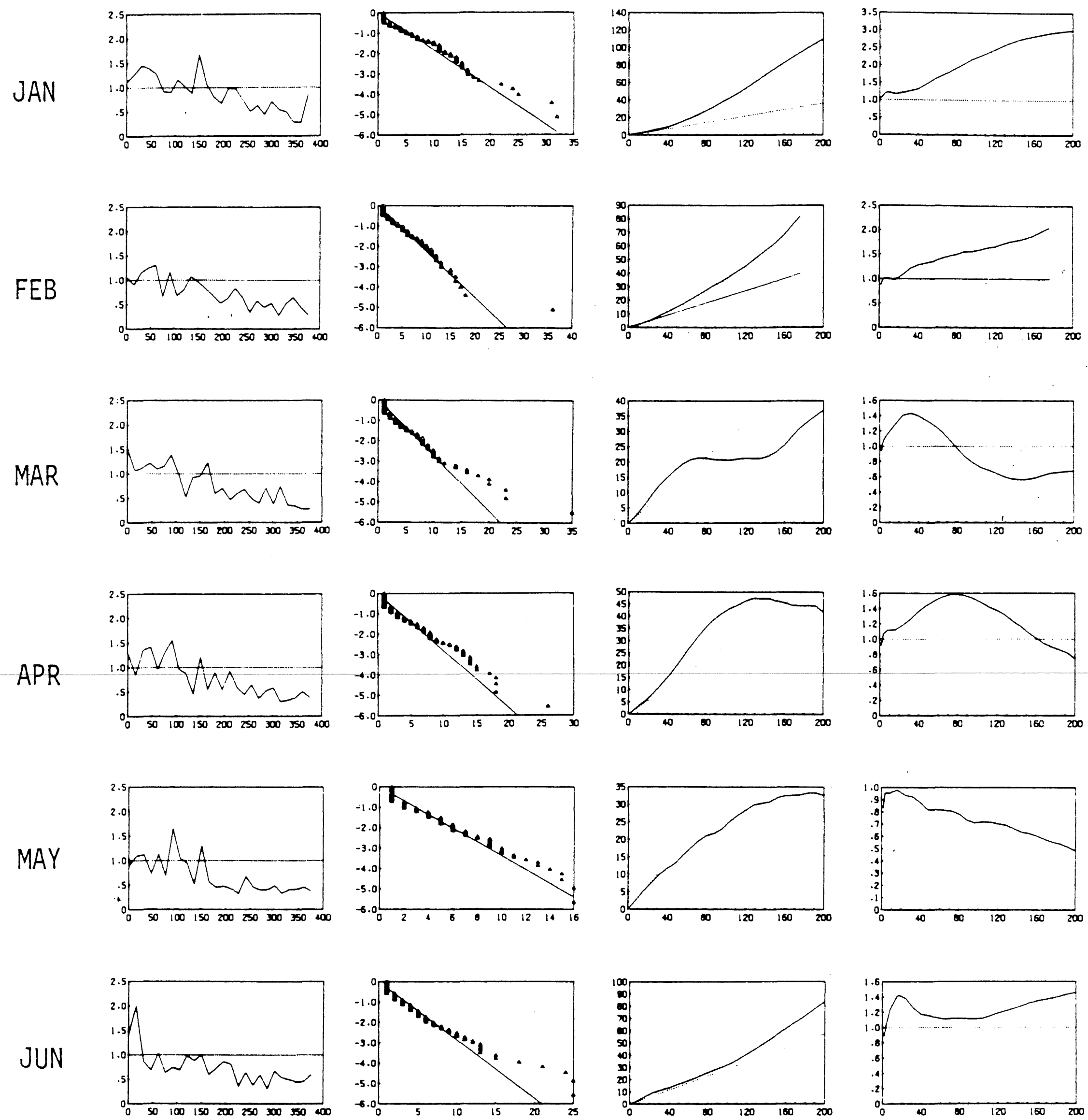


Figure A.12 Statistical properties of intervals and counts for Denver--monthly analysis

I: Normalized spectrum of counts vs. frequency factor
II: Log-survivor function vs. interarrival time (days)
III: Variance of counts vs. interval length (days)
IV: Index of dispersion vs. interval length (days)

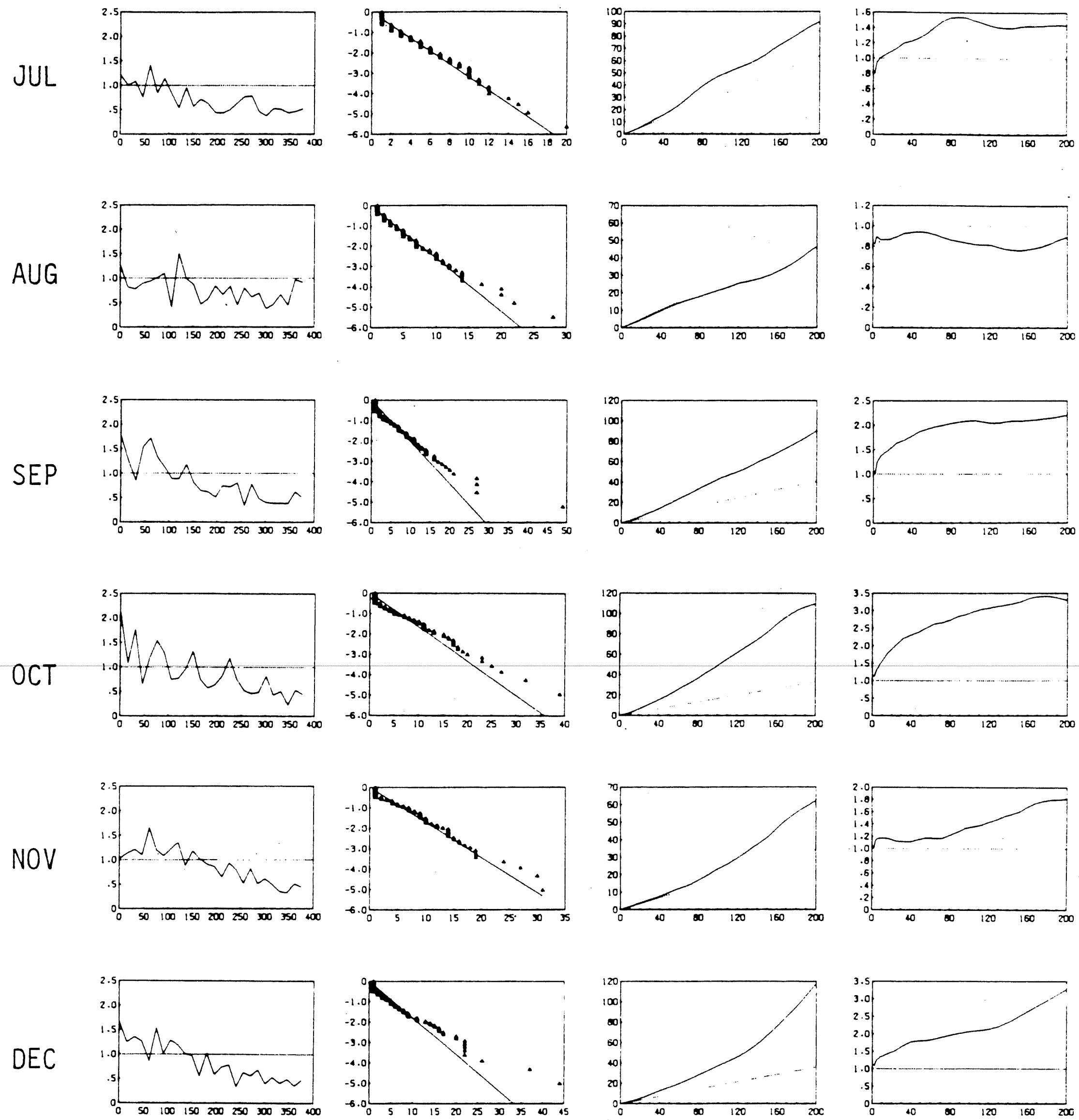


Figure A.12 (continued)

APPENDIX B

DETAILS ON THE DERIVATION OF THE CONDITIONAL INTENSITY FUNCTION OF THE SEMI-MARKOV PROCESS

The Laplace transform of the conditional intensity function of a general two-state semi-Markov process is given as

$$h^*(s) = \frac{(1-a_2)f_1^*(s) + (1-a_1)f_2^*(s) + (1-a_1-a_2)(2-a_1-a_2)f_1^*(s)f_2^*(s)}{(2-a_1-a_2)\ 1-a_1f_1^*(s) - a_2f_2^*(s) - (1-a_1-a_2)f_1^*(s)f_2^*(s)} , \tag{B.1}$$

where $f_i^*(s)$, i=1,2, are the Laplace transforms of the two probability density functions of the interarrival times (see Cox and Lewis, 1978). A geometric distribution with parameter p, can be written in the continuous-function form

$$f(t) = \sum_{k=1}^{\infty} p(1-p)^{k-1}\delta(t-k) , \tag{B.2}$$

where $\delta(\cdot)$ is the Dirac function. The Laplace transform of f(t) can be easily shown to be

$$f^*(s) = \frac{pe^{-s}}{1 - (1-p)e^{-s}} , \tag{B.3}$$

where we have made use of the fact that $\mathcal{L}(\delta(t-k)) = e^{-ks}$. Expressions for $f_1^*(s)$ and $f_2^*(s)$, analogous to (B.3), then substituted in (B.1) to give

$$(2-a_1-a_2)h^*(s)e^s$$

$$= \frac{p_1(1-a_2) + p_2(1-a_1) + [p_1p_2(2-a_1-a_2)^2 - p_1(1-a_2) - p_2(1-a_1)]e^{-s}}{1 + [p_1(1-a_1) + p_2(1-a_2) - 2]e^{-s} + [1 - p_1(1-a_1) - p_2(1-a_2)]e^{-2s}}$$

$$= \frac{\delta + [p_1p_2(2-a_1-a_2)^2 - \delta]e^{-s}}{1 + (\beta-2)e^{-s} + (1-\beta)e^{-2s}} \quad , \tag{B.4}$$

where

$$\delta = p_1(1-a_2) + p_2(1-a_1)$$

and (B.5)

$$\beta = p_1(1-a_1) + p_2(1-a_2) \quad .$$

To obtain h(t) the inverse Laplace transform of (B.4) is needed. The polynomial in the denominator of (B.4) has two real roots, one equal to 1 and the other equal to $(1-\beta)$, and therefore (B.4) can be written as

$$h^*(s)e^s = \frac{G}{1-e^{-s}} + \frac{A}{1-(1-\beta)e^{-s}} \quad , \tag{B.6}$$

where G and A are

$$G = \frac{p_1 p_2 (2-a_1-a_2)}{\beta} \tag{B.7}$$

$$A = \frac{\delta}{2-a_1-a_2} - G \quad . \tag{B.8}$$

By taking the inverse Laplace transform of (B.6)

$$\mathcal{L}^{-1}[h^*(s)e^s] = \sum_{k=0}^{\infty} [G + A(1-\beta)^k]\delta(t-k) \quad , \tag{B.9}$$

h(t) takes the form

$$h(t) = \sum_{k=1}^{\infty} [G + A(1-\beta)^k]\delta(t-k) \quad . \tag{B.10}$$

Comparing (B.10) with the discrete-analogue expression of the conditional intensity function, that is

$$h(t) = \sum_{k=1}^{\infty} h_k \delta(t-k) \quad ,$$

we can write

$$h_k = G + A(1-\beta)^k \quad . \tag{B.11}$$

Note that G of (B.7) reduces to the constant intensity, m, of the semi-Markov model. Therefore, (B.11) takes the form

$$h_k = m + AW^k \qquad (B.12)$$

where,

$$A = e_1p_1 + e_2p_2 - m \qquad (B.13)$$

and

$$W = 1 - p_1(1-a_1) - p_2(1-a_2) \quad . \qquad (B.14)$$

This completes the proof of PROPOSITION 3 of Chapter 5 which gives the conditional intensity function of a semi-Markov process.

REFERENCES

Austin, P.M., and R.A. Houze, Jr., Analysis of the structure of precipitation patterns in New England, J. Appl. Meteorology, 11, 926-935, 1972.

Bartlett, M.S., The spectral analysis of point processes, J.R. Statist. Soc., B 25, 264-296, 1963.

Billingsley, P., Statistical methods in Markov chains, Ann. Math. Statist., 32, 12-40, 1961.

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

Brillinger, D.R., Comparative aspects of the study of ordinary time series and of point processes, in Developments in Statistics, Vol. 1, 22-133, ed. P.R. Krishnaiah, Academic Press, New York, 1978.

Buishand, T.A., Some remarks on the use of daily rainfall models, J. Hydrology, 36, 295-308, 1978.

Carey, D.I., and C.T. Haan, Markov process for simulating daily point rainfall, Journal of Irrigation and Drainage Div., ASCE, 104 (IRI), 111-125, 1978.

Caskey, J.E., A Markov chain model for the probability of precipitation occurrence in intervals of various lengths, Mon. Weather Review, 91(6), 298-301, 1963.

Chang, T.J., M.L. Kavvas, and J.W. Delleur, Stochastic daily precipitation modeling and daily streamflow transfer model, Purdue Univ. Water Resource Center, Tech. Rep. No. 146, 1982.

Chang, T.J., M.L. Kavvas, and J.W. Delleur, Daily precipitation modeling by discrete autoregressive moving average processes, Water Resour. Res., 20(5), 565-580, 1984.

Chin, E.H., Modeling daily precipitation occurrence process with Markov chain, Water Resour. Res., 13(6), 949-956, 1977.

Çinlar, E., Introduction to Stochastic Processes, Prentice Hall, Englewood Cliffs, New Jersey, 1975.

Cole, J.A., and J.D.F. Sherriff, Some single- and multi-site models of rainfall within discrete time increments, Journal of Hydrology, 17(1/2), 97-113, 1972.

Court, A., Precipitation research, 1975-1978., Rev. Geophys. Space Phys., 17(6), 1165-1175, 1979.

Cox, D.R., Some models for series of events, Int. Statist. Inst., 40, 737-746, 1963.

Cox, D.R., The null distribution of the first serial correlation coefficient, Biometrika, 53, 623-626, 1966.

Cox, D.R., and V. Isham, Point Processes, Chapman and Hall, London, 1980.

Cox, D.R., and P.A.W. Lewis, The Statistical Analysis of Series of Events, Methuen, London, 1978.

Cox, D.R., and H.D. Miller, The Theory of Stochastic Processes, Methuen, London, 1965.

Cramer, H., Mathematical Methods of Statistics, Princeton University Press, 1946.

Daley, D.J., and D. Vere-Jones, A summary of the theory of point processes, in Stochastic Point Processes, Statistical Analysis, Theory and Applications, P.A.W. Lewis (ed.), 299-383, 1972.

Diggle, P.J., Discussion of the paper "A model fitting analysis of daily rainfall data" by Stern, R.D., and R. Coe, J.R. Statist. Soc., A 147, 1-34, 1984.

Eagleson, P.S., Climate, soil and vegetation, 2, The distribution of annual precipitation derived from observed storm sequences, Water Resour. Res., 14(5), 713-721, 1978.

Everitt, B.S., and D.J. Hand, Finite Mixture Distributions, Chapman and Hall, New York, 1981.

Feller, W., An Introduction to Probability Theory and Its Applications, Vol. 1, Wiley, New York, 1968.

Feyerherm, A.M., and L.D. Bark, Goodness of fit of a Markov chain model for sequences of wet and dry days, J. Appl. Meteorol., 6, 770-773, 1967.

Gabriel, K.R., and J. Neumann, On a distribution of weather cycles by lengths, Quart. J. R. Meteor. Soc., 83, 375-380, 1957.

Gabriel, K.R., and J. Neumann, A Markov chain model for daily rainfall occurrences at Tel Aviv, Quart. J. R. Meteor. Soc., 88, 90-95, 1962.

Galloy, E., A. Le Breton, and S. Martin, A model for weather cycles based on daily rainfall occurrences, in Cosnard, M., et al. (eds.): Rhythms in Biology and other fields of application, Springer Lecture Notes in Biomathematics, Vol. 49, 303-318, 1981.

Gaver, D.G., and P.A.W. Lewis, First order autoregressive Gamma sequences and point processes, Adv. Appl. Prob., 12(3), 1980.

Grace, R.A., and P.S. Eagleson, The Synthesis of Short-Time-Increment Rainfall Sequences, Hydrodynam. Lab. Rep. 91, Dept. of Civil Eng., Mass. Inst. of Techn., Cambridge, Massachussetts, 1966.

Green, J.R., A model for rainfall occurrence, J. R. Statist. Soc., B 26, 345-353, 1964.

Green, J.R., Two probability models for sequences of wet or dry days, Monthly Weather Review, 93, 155-156, 1965.

Gupta, V.K., and L. Duckstein, A stochastic analysis of extreme droughts, Water Resour. Res., 12(2), 221-228, 1975.

Guttorp, P., and M.L. Thompson, Estimating point process parameters from sampled counting processes, paper in Technical Rep. No. 35, Dept. of Statistics, Univ. of Washington, Seattle, 1983.

Haan, T.N., D.M. Allen, and J.O. Street, A Markov chain model of daily rainfall, Water Resour. Res., 12(3), 443-449, 1976.

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

Hoel, P.G., A test for Markov chains, Biometrika, 41, 430-433, 1954.

Hopkins, J.W., and P. Robillard, Some statistics from the Canadian Prairie provinces, J. Appl. Meteorology, 3, 600-602, 1964.

Ison, N.T., A.M. Feyerherm, and L.D. Bark, Wet period precipitation and the gamma distribution, J. Appl. Meteorology, 10(4), 658-665, 1971.

Jacobs, P.A., and P.A.W. Lewis, A mixed autoregressive-moving average exponential sequence and point process, EARMA (1,1), Adv. Appl. Prob., 9, 87-104, 1977.

Jacobs, P.A., and P.A.W. Lewis, A discrete time series generated by mixtures I: Correlational and run properties, J.R. Statist. Soc., B 40(1), 94-105, 1978a.

Jacobs, P.A., and P.A.W. Lewis, A discrete time series generated by mixtures II: Asymptotic properties, J.R. Stat. Soc., B 40(2), 222-228, 1978b.

Katz, R.W., An application of chain-dependent processes to meteorology, J. Appl. Prob., 14, 598-603, 1977.

Kavvas, M.L., and J.W. Delleur, The Stochastic and Chronologic Structure of Rainfall Sequences--Application to Indiana, Tech. Rep. 57, Water Resour. Res. Center, Purdue Univ., West Lafayette, Indiana, 1975.

Kavvas, M.L., and J.W. Delleur, A stochastic cluster model for daily rainfall sequences, Water Resour. Res., 17(4), 1151-1160, 1981.

Kavvas, M.L., and K.R. Herd, A radar-based stochastic model for short-time-increment rainfall, paper presented at the AGU Spring Meeting, Cincinatti, 1984.

Khanal, N.N., and R.L. Hamrick, A stochastic model for daily rainfall data synthesis, Proceedings of the symposium of Statistical Hydrology, Tucson, Arizona, U.S. Dept. of Agric., Publ. No. 1275, 197-210, 1974.

Khintchine, A.Y., The Theory of Queueing, Griffin's Statistical Monographs, No. 7, Hafner Publ. Co., New York, 1960.

Kite, G.W., Frequency and Risk Analysis in Hydrology, Water Resour. Publ., Fort Collins, Colorado, 1978.

Lawrance, A.J., Some models for stationary series of univariate events, in Stochastic Point Processes, Statistical Analysis, Theory and Applications, P.A.W. Lewis (ed.), 199-256, 1972.

Lawrance, A.J., The mixed exponential solution to the first-order autoregressive model, J. Appl. Prob., 17, 546-552, 1980.

Lawrance, A.J., and P.A.W. Lewis, An exponential moving-average sequence and point process, EMA(1), J. Appl. Prob., 14, 98-113, 1977.

Lawrance, A.J., and P.A.W. Lewis, An exponential auto-regressive-moving average process EARMA(p,q), J.R. Statist. Soc., B 42(2), 150-161, 1980.

LeCam, L., A stochastic description of precipitation, paper presented at the 4th Berkeley Symposium on Mathematics, Statistics, and Probability, Univ. of Calif., Berkeley, California, 1961.

Lewis, P.A.W., A branching Poisson process model for the analysis of computer failure patterns, J.R. Statist. Soc., B 26, 398-456, 1964.

Lewis, P.A.W., Remarks on the theory, computation and application of the spectral analysis of series of events, J. Sound Vib., 12(3), 353-375, 1970.

Lewis, P.A.W., (ed.), Stochastic Point Processes, Theory and Applications, Wiley-Interscience, 1972.

Lewis, P.A.W., Simple models for positive-valued and discrete-valued time series with ARMA correlation structure, in Multivariate Analysis V., P.R. Krishnaiah (ed.), North-Holland, Amsterdam, 151-166, 1978.

Lewis, P.A.W., A.M. Katcher, and A.H. Weiss, SASE IV -- A computer program for the statistical analysis of series of events, IBM Rep. RC 2365, 1969.

Mielke, P.W., Jr., Another family of distributions for describing and analyzing precipitation data, J. Appl. Meteorol., 10(2), 275-280, 1973.

Mielke, P.W., Jr., and E.S. Johnson, Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology, Water Resour. Res., 10(2), 223-236, 1974.

Moran, P.A.P., A note on serial correlation coefficients, Biometrika, 57, 670-673, 1970.

Nelder, J.A., and R. Mead, A Simplex method for function minimization, Computer Jour., 7, 308, 1965.

Neyman, J.E., and Scott, E.L., A statistical approach to problems of cosmology, J.R. Statist. Soc., B 20, 1-43, 1958.

Obeysekera, J.T.B., and Salas, R.D., GAR Modeling of streamflow series, presented at the AGU Fall meeting, San Francisco, AGU Abstract: EOS, Vol. 63, No. 45, 935, November, 1983.

Parzen, E., Modern Probability Theory and its Applications, John Wiley and Sons, New York, 1960.

Parzen, E., Stochastic Processes, Holden-Day, San Francisco, 1962.

Petterssen, S., Introduction to Meteorology, 3rd Ed., McGraw-Hill, New York, 1969.

Pyke, R., Markov renewal processes: Definitions and preliminary properties, Ann. Math. Statist., 32, 1231-1242, 1961a.

Pyke, R., Markov renewal processes with finitely many states, Ann. Math. Statist., 32, 1243-1259, 1961b.

Ramirez-Rodriguez, J.A., and R.L. Bras, Optimal irrigation control using stochastic cluster point processes for rainfall modeling and forecasting, MIT Ralph M. Parson Lab., Report No. 275, 1982.

Raudkivi, A.J., and N. Lawgun, Generation of serially correlated non-normally distributed rainfall durations, Water Resour. Res., 8(2), 398-409, 1972.

Richardson, C.W., Stochastic models of daily precipitation, temperature, and solar radiation, Water Resour. Res., 17(1), 182-190, 1981.

Rodriguez-Iturbe, I., V.K. Gupta, and E. Waymire, Scale considerations in the modeling of temporal rainfall, Water Resour. Res., 20(11), 1611-1619, 1984.

Roldan, J., and D.A. Woolhiser, Stochastic daily precipitation models, 1, A comparison of occurrence processes, Water Resour. Res., 18(5), 1451- 1460, 1982.

Skees, P.M. and L.R. Shenton, Comments on the statistical distribution of rainfall per period under various transformations, Proceedings of the Symposium on Statistical Hydrology, Tucson, Arizona, U.S. Dept. of Agr., Misc. Publ., No. 1275, 172-196, 1974.

Smith, E.R., and H.A. Schreiber, Point processes of seasonal thunderstorm rainfall, Part 1, Distribution of rainfall events, Water Resour. Res., 9(4), 871-884, 1973.

Smith, J.A., Point Process Models of Rainfall, Ph.D. Dissertation, The Johns Hopkins University, Maryland, 1981.

Smith, J.A., and A.F. Karr, A Point process model of summer season rainfall occurrences, Water Resour. Res., 19(1), 95-103, 1983.

Smith, W.L., Regenerative stochastic processes, Proc. Roy. Soc., A 232, 6-31, 1955.

Srinivasan, S.K., Stochastic Point Processes and their Applications, Griffin's Statistical Monographs, No. 34, Hafner Press, New York, 1974.

Teugels, J.L., A bibliography on semi-Markov processes, Journ. of Comp. and Appl. Mathematics, 2(2), 125-144, 1976.

Thom, H.G., A note on the gamma distribution, Monthly Weather Review, 86(4), 117-122, 1958.

Todorovic, P., and D.A. Woolhiser, Stochastic model of daily rainfall, Proc. of the USDA-IASPS Symposium on Statist. Hydrology, Tucson, Arizona, 1971.

Todorovic, P., and V. Yevjevich, Stochastic processes of precipitation, Colorado State Univ., Hydro. Paper 35, 1-61, 1969.

Tong, H., Determination of the order of a Markov chain by Akaike's information criterion, J. Appl. Prob., 12, 488-497, 1975.

Vere-Jones, D., Stochastic models for earthquake occurrence, J. R. Statist. Soc., B 32, 1-62, 1970.

Watson, G.S., and W.R. Wells, On the possibility of improving the mean useful life of items by eliminating those with short lives, Technometrics, 3, 281-298, 1961.

Waymire, E., and V.K. Gupta, The mathematical structure of rainfall representations, Part 1, A review of stochastic rainfall models, Water Resour. Res., 17(5), 1261-1271, 1981a.

Waymire, E., and V.K. Gupta, The mathematical structure of rainfall representations, Part 2, A review of the theory of point processes, Water Resour. Res., 17(5), 1273-1286, 1981b.

Waymire, E., and V.K. Gupta, The mathematical structure of rainfall representations, Part 3, Some applications of the point process theory to rainfall processes, Water Resour. Res., 17(5), 1287-1294, 1981c.

Waymire, E., V.K. Gupta, and I. Rodriguez-Iturbe, A spectral theory of rainfall intensity at the meso-b scale, Water Resour. Res., 20(10), 1453-1465, 1984.

Weiss, L.L., Sequences of wet and dry days described by a Markov chain model, Monthly Weather Review, 92, 169-176, 1964.

Wiser, E.H., Modified Markov probability models of sequences of precipitation events, Monthly Weather Review, 93, 511-516, 1965.

Woolhiser, D.A., and G.G.S. Pegram, Maximum likelihood estimation of Fourier coefficients to describe seasonal variations of parameters in stochastic daily precipitation models, J. Appl. Meteorol., 8(1), 34-42, 1979.

Woolhiser, D.A., and J. Roldan, Stochastic daily precipitation models, 2, A comparison of distributions of amounts, Water Resour. Res., 18(5), 1461-1468, 1982.

BIOGRAPHICAL SKETCH

Efi Foufoula-Georgiou received her Diploma in Civil Engineering from the Technical University of Athens, Greece, in 1979. After working for one year at the Ministry of Public Works, Athens, she came to the United States for graduate study. She was a graduate research assistant in the Department of Environmental Engineering Sciences, University of Florida, from which she received the Master of Engineering degree in December, 1982, and the Ph.D. degree in May, 1985. Most of her Ph.D. dissertation research was performed while a visiting graduate student at the Department of Civil Engineering, University of Washington, Seattle.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Wayne C. Huber, Chairman
Professor of Environmental
Engineering Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James P. Heaney, Co-Chairman
Professor of Environmental
Engineering Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Dennis P. Lettenmaier
Associate Professor of
Environmental Engineering Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Warren Viessman, Jr.
Professor of Environmental
Engineering Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Barry A. Benedict
Professor of Civil Engineering

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Mark C. K. Yang
Professor of Statistics

This dissertation was submitted to the Graduate Faculty of the College of Engineering and to the Graduate School, and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

May, 1985

Dean, College of Engineering

Dean for Graduate Studies and Research