5-1995

Nonparametric Stochastic Generation of Daily Precipitation and Other Weather Variables

Rajagopalan Balaji
Utah State University

NONPARAMETRIC STOCHASTIC GENERATION OF DAILY PRECIPITATION AND OTHER WEATHER VARIABLES

by

Rajagopalan Balaji

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

DOCTOR OF PHILOSOPHY

in

Civil and Environmental Engineering

UTAH STATE UNIVERSITY
Logan, Utah

1995

To,
my parents
Rajagopalan and Vagulavalli

ACKNOWLEDGMENTS

I express my deepest sense of gratitude to my major professor, Dr. David S. Bowles, for generously providing financial support throughout the period of this work. I am extremely thankful to him for giving me unconditional freedom to learn anything that I wished, and for all the support and encouragement in enabling me to attend various conferences. As a result, my stay at Utah State University has been a rich learning experience. Thank you, sir.

I am indebted to Dr. Upmanu Lall for all the advice and assistance he gave me throughout the course of my stay at Utah State University. This work owes him a lot for introducing me to the fascinating ideas of nonparametric functional estimation,and for his stimulating discussions and guidance. I have learned so much from him, which has played decisive roles in not only shaping my professional skills but also in forming my values and shaping my conduct. I am deeply moved by his genuine concern in my personal growth without expecting anything in return. My indebtedness to him is more than the words of thanks can ever convey.

I wish to express my sincere thanks to Dr. David G. Tarboton for his inimitable review of the the manuscript and critical remarks at all stages of work, which improved the content and style in many places. Working with him has been a rewarding experience

I want to thank Drs. Robert W. Gunderson, Michael Minnotte, and Gail E. Bingham for serving on my dissertation committee, and Dr. Esmaiel E. Malek for conducting my defense.

Whatever little I have done so far owes to my uncles, Drs. G. Srinivasan and G. Rajamannar, and to my cousin, Dr. R. Venkatesh, from whom I derived a tremendous amount of inspiration to pursue higher studies. Their constant encouragement and support

are gratefully acknowledged. Thanks are due to my cousins K. Vinodh and R. Gopalan, for all their financial and moral support, especially during the initial stages of my graduate program.

This work would not have been possible without the timely help of my uncle, G. Krishnan. His crucial financial help in enabling me to come to Utah State University for my graduate program will be remembered forever.

My gratitude goes to my parents for all their sacrifice, continued support, and understanding. From them I learnt that "work is worship," an attitude that helped me a lot in the success of my program. To other members of my family, who have provided support and encouragement throughout my career, I will be forever grateful.

Thanks are due to Unni, Ashish, Moon, and Aladdin (all of them soon to be PhDs) for all the stimulating discussions that helped this work in many ways. I have benefitted greatly from their insight and suggestions. Special mention to Unni, from whom I learnt a lot more things in life than he would ever imagine!

I wish to acknowledge the help of my friends, Raghu, Venky, Ashutosh, Umesh, Sanjay, Shaleen, Drs. Alok Sikka, Mohan Reddy, and Mrudula Reddy, for all their hospitality and help throughout my stay at Logan.

And finally, thanks to a large number of friends and well wishers who made my stay at Utah State University all these years a memorable and pleasant experience.

The funding for this study through the United States Forest Service/United States Department of Agriculture is thankfully acknowledged.

Rajagopalan Balaji

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Nonparametric Stochastic Generation of Daily Precipitation
and Other Weather Variables

by

Rajagopalan Balaji, Doctor of Philosophy
Utah State University, 1995

Major Professor: Dr. David. S. Bowles
Department: Civil and Environmental Engineering

Traditional stochastic approaches for synthetic generation of weather variables often assume a prior functional form for the stochastic process, are often not capable of reproducing the probabilistic structure present in the data, and may not be uniformly applicable across sites. In an attempt to find a general framework for stochastic generation of weather variables, this study marks a unique departure from the traditional approaches, and ushers in the use of data-driven nonparametric techniques and demonstrates their utility.

Precipitation is one of the key variables that drive hydrologic systems and hence warrants more focus. In this regard, two major aspects of precipitation modeling were considered: (1) resampling traces under the assumption of stationarity in the process, or with some treatment of the seasonality, and (2) investigations into interannual and secular trends in precipitation and their likely implications.

A nonparametric seasonal wet/dry spell model was developed for the generation of daily precipitation. In this the probability density functions of interest are estimated using nonparametric kernel density estimators. In the course of development of this model,

various nonparametric density estimators for discrete and continuous data were reviewed, tested, and documented, which resulted in the development of a nonparametric estimator for discrete probability estimation.

Variations in seasonality of precipitation as a function of latitude and topographic factors were seen through the nonparametric estimation of the time-varying occurrence frequency. Nonparametric spectral analysis, performed on monthly precipitation, revealed significant interannual frequencies and coherence with known atmospheric oscillations. Consequently, a nonparametric, nonhomogeneous Markov chain for modeling daily precipitation was developed that obviated the need to divide the year into seasons.

Multivariate nonparametric resampling technique from the nonparametrically fitted probability density functions, which can be likened to a smoothed bootstrap approach, was developed for the simulation of other weather variables (solar radiation, maximum and minimum temperature, average dew point temperature, and average wind speed). In this technique the vector of variables on a day is generated by conditioning on the vector of these variables on the preceding day and the precipitation amount on the current day generated from the wet/dry spell model.

(241 pages)

CHAPTER 1

GENERAL INTRODUCTION

Problem Statement and Research Relevance

Synthetic sequences of daily precipitation and other weather variables (e.g. dew point temperature, maximum temperature, minimum temperature, solar radiation) are often used for investigating likely scenarios for agricultural water requirements, reservoir operation for analyses of antecedent moisture conditions, and erosion prediction, and for runoff generation in a watershed. Traditional statistical approaches for this purpose typically fit a parametric statistical model to the data at a site. Organizations (e.g., United States Forest Service, United States Department of Agriculture) specify one such model for applications from site to site. A problem with this setup is that "models" that work well in some regions/sites fail at others. This suggests the need for developing data-adaptive statistical methods that can be applied uniformly across regions.

Precipitation is known to be one of the key variables that trigger several hydrologic processes (e.g., runoff, erosion, floods, droughts, etc.) and affect various weather variables. There is increasing recognition of strong links between precipitation and other hydrologic processes with atmospheric circulation, especially at interannual and interdecadal time scales. Consequently there is a need to better understand precipitation fluctuations with respect to some known atmospheric oscillations (e.g., El-Niño, QBO). Another important aspect of precipitation is its seasonality. The timing and duration of the "seasons" of high precipitation at a site are important since they indicate the form (rain or snow) of precipitation as well as the nature of the input "signal" for the surface hydrologic system.

Robust nonparametric techniques that are guided only by information in the observed data offer an attractive alternative that can better address such nonstationarities. These techniques are relatively new and have only had limited application in hydrology. In this regard this study represents a point of departure from the traditional attempts at stochastic generation of precipitation and other weather variables.

Objectives and Scope of Study

Given the need for simulation of synthetic daily weather sequences and nonstationary nature of the precipitation as briefed in the introduction, the major questions that come to mind are (1) How best can one simulate daily precipitation and other weather variables, while preserving the relative frequencies of events, without prior assumptions as to the parametric form of the underlying probability models? and (2) How can one identify systematic variations in precipitation patterns with known atmospheric oscillation?

As mentioned at the end of the previous section, nonparametric techniques are used to address these questions. The specific objectives of this work were to: (1) develop nonparametric stochastic models for the generation of sequences of daily precipitation and Other weather variables; (2) develop nonparametric procedures to identify seasonal variability in precipitation and their implications to precipitation modeling; (3) identify variability in precipitation patterns, with regard to low frequency atmospheric variability through nonparametric spectral analysis techniques.

Outline

This study is presented in a multiple-paper format, and is a compilation of investigations related to the objectives indentified in the previous section. This work is divided into nine chapters including the introductory chapter. Various tasks related to the objectives are outlined here.

The development of a nonparametric model seasonal wet/dry spell model for daily precipitation is described in Chapter II. As the name suggests, the model operates on seasons. In the course of developing this model, various techniques of nonparametric probability density estimation of continuous and discrete variables were reviewed and tested with various data sets. These experiences are documented in Chapter III. Comparisons of various techniques lead to the development of a new estimator for discrete probabilities in Chapter IV.

Seasonality of daily precipitation, an important aspect of precipitation, was next studied, using a fully data-adaptive nonparametric technique and was applied to data from various sites. This is presented in Chapter V. Significant changes in seasonality across the various sites were found. Also the timing varied across the years. This indicated plausible low frequency variability in precipitation at the interannual time scales. Subsequently, investigations into low frequency variability of precipitation associated with known atmospheric oscillations like El Niño/QBO were embarked upon using nonparametric spectral analysis, which is described in Chapter VI. Strong consistent signals that correspond to these oscillations were found across all the sites analyzed.

Results from seasonality and spectral studies motivated a need to improve the seasonal wet/dry spell model so that the partitioning of the year into rigid seasons is obviated. It was then found that a nonhomogeneous Markov chain model using nonparametric estimators could address this effectively. Consequently a nonparametric, nonhomogeneous Markov chain was developed in Chapter VII.

One of the major objective of this study was to develop procedures for synthetic generation of other weather variables. In this regard a multivariate nonparametric model that generates a daily vector of weather variables conditioned on the vector of values on the previous day and the precipitation amount of the current day of interest was developed in

Chapter VIII. Precipitation is generated from the wet/dry spell model. Summary of the this research is outlined in Chapter IX. This study resulted in the development of two different nonparametric approaches for generating daily precipitation. The attributes of these models are also discussed in Chapter IX.

CHAPTER II

A NONPARAMETRIC WET/DRY SPELL MODEL FOR RESAMPLING DAILY PRECIPITATION[1]

Abstract

A nonparametric wet/dry spell model is developed for resampling daily precipitation at a site. The model considers alternating sequences of wet and dry days in a given season of the year. All marginal, joint, and conditional probability densities of interest (e.g., dry spell length, wet spell length, precipitation amount, wet spell length given prior dry spell length) are estimated nonparametrically using at-site data and kernel probability density estimators. Procedures for the disaggregation of wet spell precipitation into daily precipitation, and for the generation of synthetic sequences are proffered. An application of the model for generating synthetic precipitation traces at a site in Utah is presented.

Introduction

Synthetically generated sequences of daily precipitation are often used for investigating likely scenarios for agricultural water requirements, reservoir operation for analyses of antecedent moisture conditions and for runoff generation in a watershed. Preserving the characteristics of multi-day wet and dry spells is often important in these applications. This chapter presents a stochastic model for resampling daily precipitation where the probability distributions functions (PDFs) of alternating wet and dry spell lengths and of rainfall amount are estimated nonparametrically using kernel density estimators. This procedure is equivalent to a bootstrap or sampling with replacement of the observed data sequence of spell lengths and precipitation amounts. It differs from the

[1]Coauthored by Rajagopalan Balaji, Upmanu Lall, and David G. Tarboton

for resampling, and sequential attributes of spells may be preserved. Necessary calibration parameters are chosen automatically from the data set using measures aimed at providing a good fit to the unknown underlying PDFs.

Our particular interest was in developing a scheme for synthetic simulation of daily precipitation in mountainous regions in the western United States. Precipitation in these areas is in the form of snow in the winter with orographic and frontal mechanisms dominant. Convective rainfall processes occur in other seasons. Marked differences in the storm tracks and moisture sources over the seasons are observed. A mixture of markedly different mechanisms (some related to the El Niño Southern Oscillation) leads to the precipitation process in the western U.S. [Webb and Bettencourt, 1992; Cayan and Riddle, 1992]. Recognition of such heterogeneities has led to efforts at regime identification and modeling of rainfall conditional on weather types [e.g., Katz and Parlange, 1993; Wilson and Lettenmeier, 1993; Bogardi et al., 1993]. While this is an attractive and necessary approach, deconvolution of mixtures is not always easy from a finite data set and the weather type designations used can be subjective. Traditionally parametric probability models (e.g., exponential distribution), whose functional form is completely specified by a small set of parameters, are used to fit the relevant frequency distributions. Selecting the best such model is tenuous [see Vogel and McMartin, 1991] even where mixtures are not of concern.

The work presented here was motivated by the following questions:

1. Is it possible to resample the data while preserving the relative frequencies and conditional relative frequencies of wet and dry spells and precipitation amounts, without prior assumptions as to the parametric forms of the underlying probability models ?
2. What is a good way to empirically model the relevant PDFs for resampling and to guide development of statistical models?

3. Can such a data-based assessment of probabilities or relative frequencies be used to judge the adequacy of conceptual and statistical models posed for daily rainfall?

The first question is relevant not only from a conceptual standpoint but also because organizations (e.g., United States Forest Service, United States Department of Agriculture) specify a uniform procedure for applications from site to site. Where parametric distributions or procedures are used, "models" that work well in some regions/sites fail at others. In our view it is unlikely that a robust parametric framework for model specification and selection can be devised for uniform application given the likely heterogeneity in precipitation generation mechanisms. Here we sidestep such issues by using a resampling strategy that honors at-site data directly.

The second question is addressed in a companion paper [Rajagopalan et al., 1995] where we document our investigations into developing appropriate kernel density estimators for resampling continuous (e.g., precipitation amount) and discrete (e.g., spell length in days) random variables.

As regards the third question, we argue that the answer is likely to be yes, given that the relevant probability densities can be estimated reliably from the data. However, this is an area that we expect to research formally in the future, and discuss only generally here.

We begin with a brief review of available models for simulating daily precipitation and an introduction to the central ideas in kernel density estimation. The nonparametric, alternating wet/dry spell model is presented next and the resampling/simulation strategy is indicated. Results from an application to a Utah data set follow. The performance of the nonparametric scheme is compared with a simple, parametric alternative. A discussion of applicability, limitations of the approach, and musings on pointers to related work in progress concludes the presentation.

Background

Reviews of stochastic precipitation models are offered by Waymire and Gupta [1981 a,b,c], Georgakakos and Kavvas [1987] and Foufoula-Georgiou and Georgakokas [1988]. The reader is referred to these papers for an appreciation of the literature and the central issues perceived in the field. While we are aware of the need to look at the concurrent representation of the precipitation process at different time scales, our focus here will only be on daily precipitation. Precipitation models have two components: (1) a model for precipitation occurrence, usually formulated as a Markov process, and (2) a model for precipitation amount, once a precipitation event has been generated. In the latter case, typically a parsimonious member of the Exponential family that best fits a given data set is used. A firm basis for such a choice has yet to emerge, and typical tests for selecting between parametric distributions, such as the chi-square test, often lack the power to discriminate between different candidate distributions, since most of the mass of the PDF is concentrated near the origin. This practice is also questionable given our earlier comments that a mix of generating processes likely governs precipitation. A brief discussion of the attributes of some models for daily precipitation occurrence follows.

Markov chain models

The most popular approach is to consider the precipitation occurrence process to be described by a finite state (typically 2, a day is wet or dry) Markov chain (MC) of finite order (typically 1), with seasonally (or time varying) transition probabilities. The basic assumption is that the present state (wet or dry) depends only on the immediate past. The transition probabilities for transitions (i.e., WW, WD, DW, DD) between the two states (W or D) are estimated directly from the data through a counting process. Fourier series

methods [Feyerharm and Bark, 1965; Woolhiser et al., 1988] may be used to parameterize seasonal variations in the transition probabilities. The degree of dependence in time is limited by the order of the MC. Feyerharm and Bark [1967] and Chin [1977] suggest that the order may need to be seasonally variable as well. Lack of parsimony is a drawback of MC models as the order is increased. A number of researchers [Hopkins and Robillard, 1964; Haan et al., 1976; Srikanthan and McMahon, 1983; Guzman and Torrez, 1985] have also stressed the need for multistate MC models that consider the dependence between transition probabilities and rainfall amount.

Chang et al. [1984] and Foufoula-Georgiou and Georgakakos [1988] argue that Markov chain models do not reproduce long-term persistence and event clustering very readily. Markov chain models can be attractive because of their largely nonparametric nature, ease of application and interpretability, and well developed literature. Wilson and Lettenmeier [1993] pursue a hierarchical MC model to describe the daily precipitation process given the heterogeneous generating mechanisms prevalent in the western United States. While this approach addresses the heterogeneity issue, the relative lack of parsimony and shortcomings of the MC model identified above detract from the formulation.

The spell approach

In probabilistic terminology, this approach is also called the alternating renewal model (ARM). The term renewal stems from the implied independence between the dry and wet period length while the term alternating refers to the fact that wet and dry states alternate. No transition to the same state is possible. An advantage of this representation is that it allows direct consideration of a composite precipitation event, rather than its discontinuous truncation into arbitrary daily segments.

A Geometric or a negative Binomial distribution [Roldan and Woolhiser, 1982]

may be used as a model for spell length, where a daily time step is of interest. A probability distribution for wet spell precipitation amount also needs to be developed, as does a procedure for the disaggregation of wet spell precipitation to daily precipitation, for wet spells that are longer than one day.

The primary difficulties cited with the spell approach for daily rainfall modeling are (1) the need for disaggregation of wet spell precipitation into daily or event precipitation (this is not an issue if independence in daily precipitation amounts is assumed, since that is typically assumed by Markov Chain models), (2) justification of the independence between the dry and wet spell lengths at short time scales, and (3) the effective reduction in the sample size by considering spells rather than days. We also find the usual parametric specifications for probability distributions, and assumptions of independence of spells in such models objectionable in light of the likely heterogeneous nature of the data of interest to us. However, we do find this structure plausible and address some of the difficulties cited here in our development.

Model Formulation

For the nonparametric, seasonal wet/dry spell model (NSS) presented here, the random variables of interest are the wet spell length, w days; dry spell length, d days; daily precipitation amount, p inches; and the wet spell precipitation amount, p_W inches. Note that throughout the chapter, wet day precipitation is referred to as daily precipitation. Variables w and d are defined through the set of integers greater than 1 (and less than season length), and p and p_W are defined as continuous, positive (actually greater than a measurement threshold, e.g., 0.01 inches rather than 0) random variables. A mixed set of discrete and continuous random variables is thus considered. Appropriate season definitions are prescribed by the model user, and model definitions that follow pertain to a

given season of the year. The natural sequence of seasons is maintained, and spells in progress at the end of a season are allowed to terminate in the succeeding season.

The general structure of the model is similar to that of a wet/dry spell model. Our model differs from the traditional wet/dry spell model in a number of ways, as illustrated in Figure 2.1. The dry and wet spell lengths in a season may be dependent. The data are allowed to indicate whether such an assumption is necessary. Rather than fitting parametric probability densities to the data, we consider kernel estimators of the probability mass/density function (PMF/PDF) of wet spell length f(w), dry spell length f(d), wet day precipitation amount f(p), wet spell precipitation amount $f(p_W)$, the joint PMF of wet and dry spell length f(w,d), the joint PDF. of wet spell length and wet spell precipitation $f(w,p_W)$, and the conditional PDFs of wet spell length given dry spell length f(w|d), dry spell length given wet spell length f(d|w), and wet spell precipitation given wet spell length $f(p_W|w)$.

First, the significance of the dependence between successive wet and dry spell lengths is assessed by computing their sample correlation for each season. The precipitation occurrence process in a given season is described through the conditional PMFs f(w|d) and f(d|w) if the correlation is significant and the marginal PMFs f(w) and f(d) otherwise. The latter with parametrically specified PMFs corresponds to the traditional alternating renewal model. The former is a more general dependence structure. Next, one estimates, for each season, the autocorrelation function for precipitation amounts p_i, i=1...w for each spell length. If these correlations are not significant, it is assumed that there is no "statistical structure" in the within spell precipitation, at least for daily precipitation amounts. In this case, daily precipitation is modeled directly through an estimate of the PDF f(p). If there is evidence for structure in wet spell precipitation, wet spell precipitation p_W becomes the primary variable, and a disaggregation approach that preserves the within spell structure is

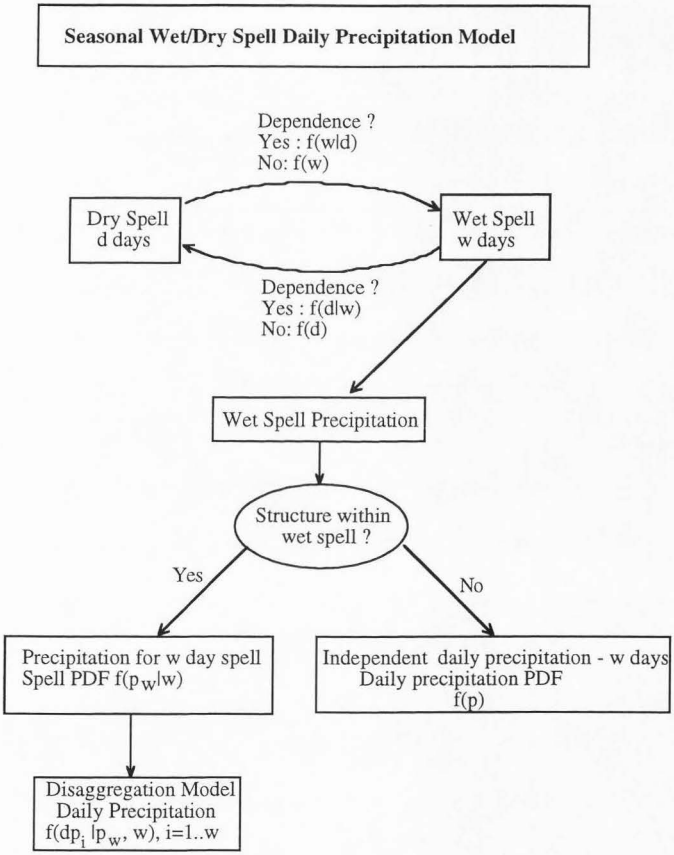


Figure 2.1. Structure of the wet/dry spell precipitation model.

used to disaggregate p_W to daily precipitation amounts. In most applications using traditional wet/dry spell models or the one presented here, the disaggregation approach is eschewed in favor of treating daily precipitation as an independent random variable.

The decisions on model structure as well as the relevant PDFs for each variable for each season are different and are independently estimated. To save on notation, we have chosen not to index any of our variables by season. In summary, the primary differences with the traditional wet/dry spell model are (1) the relevant probability functions are estimated without recourse to prior assumptions as to the parametric form of the model, and (2) a more general conditional dependence structure is admitted.

We stress that while we are ultimately interested in developing a nonparametric model for generating daily precipitation sequences, the nonparametric density estimates generated en route are interesting since they reveal tendencies or structure in the precipitation process. We now describe how the PDFs and PMFs are estimated. The univariate cases are discussed first followed by the bivariate/conditional cases. The disaggregation approach is presented last.

Nonparametric kernel function estimation

Nonparametric estimation of probability and regression functions is an emerging area in stochastic hydrology. A review of recent applications is offered by Lall [1994]. A function approximation method is considered nonparametric if (1) it is capable of approximating a large number of target functions, (2) it is "local," in that estimates of the target function at a point use only observations located within some small neighborhood of the point, and (3) no prior assumptions are made as to the overall functional form of the target function. A histogram is a familiar example of such a method. Such methods do have parameters (e.g., the bin width of the histogram) that influence the estimate at a point.

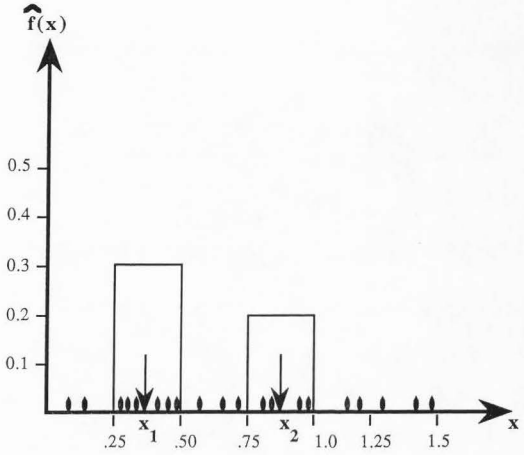
However, they are different from "parametric" methods where the entire function is indexed by a finite set of parameters (e.g., mean and standard deviation), and a prescribed functional form.

Kernel density estimation is a nonparametric method of estimating PDFs from data that is related to the histogram. Recent expository monographs that develop these ideas include [Silverman, 1986; Scott, 1992; Härdle, 1991]. Given a set of observations $x_1,..,x_n$ (in general x may be a scalar or a vector), the kernel density estimate is defined as:

$$\hat{f}(x) = \frac{1}{h\,n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \qquad (2.1)$$

where K(.) is a weight or kernel function, and h is a bandwidth.

The idea is illustrated through Figure 2.2. Consider the definition of probability as a relative frequency of event occurrence. Now an estimate of the probability density at a point x (refer to points x_1 and x_2 in Figure 2.2) may be obtained if we consider a box or window of width 2h centered at x and count the number of observations that fall in such a box. The estimate $\hat{f}(x)$ is then (number of x_i that lie within [x-h,x+h])/(2hn)). In this example, we have used a rectangular kernel (K(t)=1/2 for |t|<1, 0 else; t=(x-x_i)/h) for the estimate in the locale of x. As the sample size n grows, one could shrink the bandwidth h such that asymptotically the underlying PDF is well approximated. Note that for a finite sample, this is much like describing a histogram, except that the "bins" are being centered at each observation or at each point of estimate, as desired. From the point of view of resampling, one can treat each observation (x_i) as being equally likely to occur in the window $x_i \pm h$ and resample it uniformly in that interval, for this example. Clearly, one is not restricted to rectangular kernels.



(x_1 and x_2 are points of estimate, bandwidth, h = 0.125)

Figure 2.2. Example of kernel density estimation using 20 points with an histogram.

The "parameters" of this method are the kernel function or "local density" and the bandwidth h. A valid PDF. estimate is obtained for any K(.) that is itself a valid PDF. Symmetry of K(.) is assumed for unbounded data to ensure pointwise unbiasedness of the estimate. For bounded data, special boundary kernels that correspond to the interior kernels are used in the boundary region, to assure unbiasedness. Finite variance of K(.) is assumed to ensure that $\hat{f}(x)$ has finite variance. This still leads to a wide choice of functions for K(.). It turns out that in terms of the mean square error (MSE) of $\hat{f}(x)$ the choice of K(.) is not crucial. Different kernels can be made equivalent under rescaling by choosing appropriate bandwidths. A Gaussian kernel with a large bandwidth can give MSE of $\hat{f}(x)$

comparable to that using a rectangular kernel with a smaller bandwidth. Thus, given a Kernel function, the focus shifts to appropriate specification or estimation of the bandwidth.

It is important to note that specifying a kernel function does not have the same implications as choosing a parametric model for the whole density because the focus remains on a good pointwise or local approximation of the density rather than on fitting the whole curve directly. Different choices of K(.) still yield a local approximation of the underlying curve point by point. One can understand this by thinking of a weighted Taylor series approximation to f(x) at a point x. The interplay between the h and K(.) can be thought of in terms of the interval of approximation and a weight sequence used to localize the approximation. The length of the interval (or bandwidth in this case) is more important in terms of approximation error. However, the tail behavior of the K(.) is important in the resampling context since it relates to the likely degree of extrapolation of the process. Some typically used kernels are listed in Table 2.1.

The feeling in the statistics literature [e.g., Silverman, 1986] is that the choice of kernel is secondary in estimating f(x) and research has focused on choosing an appropriate bandwidth optimally (in a likelihood or MSE sense) from the data. The bandwidth may vary by location (i.e., value of x) being larger where the data is sparser. Bandwidth and kernel selection issues and the success of the kernel scheme for approximating discrete, continuous and bivariate PDF.'s are discussed in Rajagopalan et al., [1995]. Here, we present the estimators that we recommend be used for the NSS model.

Kernel estimation of continuous, univariate PDFs

The continuous, univariate PDFs of interest to us are f(p), the PDF of daily precipitation, and $f(p_w)$, the PDF. of wet spell precipitation for a season. The data set in the first case is composed of n_p days of daily precipitation values, p_i, for all days with

Table 2.1. Examples of Kernel Functions

Note $t = (x-x_i)/h$

Continuous Random Variables, Univariate

Kernel

Normal $K(t) = (2\pi)^{-1/2} e^{-t^2/2}$

Epanechnikov $K(t) = 0.74(1- t^2) \quad |t| \leq 1$

Bisquare $K(t) = 0.9375(1-t^2)^2 \quad |t| \leq 1$

Discrete Random Variables, Univariate (DK) estimator

Note t = (L-j)/h, and L is point at which density is estimated

Interior region (i.e., $L \geq h+1$)

Quadratic $K(t) = at^2 + b \quad$ for $|t| \leq 1$

$$a = \frac{-3h}{(1-4h^2)} \quad \text{and} \quad b = \frac{3h}{(1-4h^2)}$$

Left Boundary

for the case $1 < L < h+1$

Quadratic $K(t) = at^2 + b \quad$ for $|t| \leq 1$

$$a = \frac{-D}{2h(h+L)} \left[\frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^3(h+L)}\right)} \right] \quad \text{and} \quad b = \left[1 - \frac{aC}{6h^2}\right] \frac{1}{(h+L)}$$

where, $C = h(h-1)(2h-1) + (L-2)(i-1)(2L-3)$; $D = -h(h-1) + (L-2)(L-1)$; $E = -(h(h-1))^2 + ((L-2)(L-1))^2$

for the case $L = 1$

Quadratic $K(t) = at^2 + b \quad$ for $|t| \leq 1$

$$a = \frac{-D}{2h^2} \left[\frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^4}\right)} \right] \quad \text{and} \quad b = \left[1 - \frac{aC}{6h^2}\right] \frac{1}{h}$$

where, $C = h(h-1)(2h-1)$; $D = -h(h-1)$; $E = -(h(h-1))^2$

measurable precipitation, in season s for the y year record. For p_w, the data set is composed of n_w wet spells with total precipitation $p_{w,j}$ for each spell of length w, in season s for the y year record.

A logarithmic transform of the precipitation data prior to density estimation is often considered. Such a transformation is also attractive in the kernel density estimation context. It can provide an automatic degree of adaptability of the bandwidth (in real space), thus alleviating the need to choose variable bandwidths with heavily skewed data, and also alleviates problems that the kernel density estimation has with PDF estimates near the boundary (e.g., the origin) of the sample space. The resulting kernel density estimator can be written as:

$$\hat{f}(r) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{hr} K\left(\frac{\log(r) - \log(r_i)}{h}\right) \qquad (2.2)$$

where h is the bandwidth of the log transformed data, r is p or p_w, and n is correspondingly n_p or n_w.

The bandwidth h is chosen using a recursive method of Sheather and Jones [1991] (SJ) that minimizes the average mean integrated square error (MISE) of $\hat{f}(\log(r))$. Figures 2.3a and b provide an illustration of the kernel estimated PDF and the underlying true PDF for two situations described in Table 2.2

Kernel estimation of discrete univariate PMFs

In this section, we present procedures for the estimation of the discrete, univariate probability mass functions f(d) and f(w) for each season s. This corresponds to the assumption of independence between w and d in a traditional alternating renewal model.

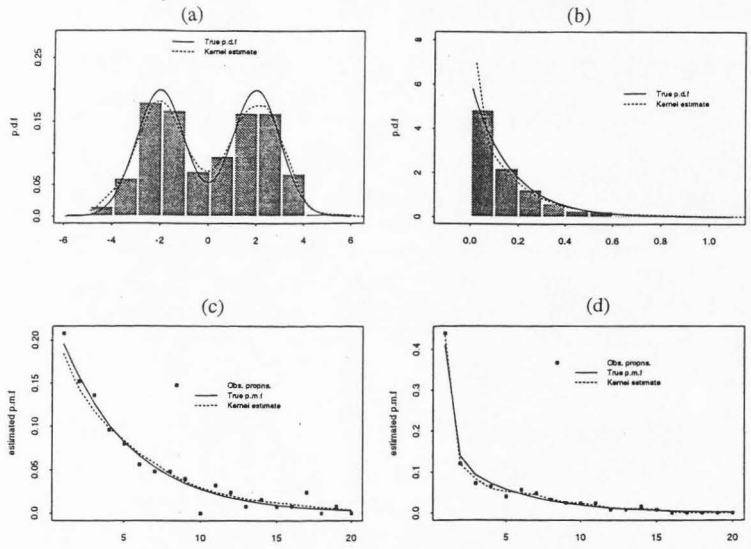


Figure 2.3. True PDF, kernel estimated PDF, and histogram of the data generated from (a) 0.5[N(-2,1) + N(2,1)], (b) Exp(0.15), (c) Geom(0.2) and (d) 0.3Geom(0.9) + 0.7Geom(0.2).

Table 2.2. Statistics of Known Distributions from Which a Sample of Size 250 Was Taken to Test Kernel Density Estimation Methods

Figure	Parent	Method	Sample Mean	Sample St. Dev	Kernel Bandwidth
2.3a	{0.5N(-2,1) + 0.5N(2,1)}	*Epanechnikov kernel, SJ bandwidth*	-0.00	2.26	1.22
2. 3b	Exp(0.15)	*Log transform, Epanechnikov kernel, SJ bandwidth*	0.16	0.18	0.94
2.3c	Geom(0.2)	*Quadratic kernel, DK estimator LSCV bandwidth*	5.11	4.19	6
2.3d	{0.3Geom(0.9) + 0.7Geom(0.2)}	*Quadratic kernel, DK estimator LSCV bandwidth*	3.92	4.02	2

We adopt the discrete kernel (DK) estimator developed in Rajagopalan and Lall [in press] for PMF estimation. The DK estimator for the PMF $\hat{f}(L)$, where L is either w or d, and n is the corresponding sample size is given as:

$$\hat{f}(L) = \sum_{j=1}^{L_{max}} K_d\left(\frac{L-j}{h}\right) \tilde{p}_j \qquad (2.3)$$

where $\tilde{p}_j$ is the sample relative frequency (n_j/n) of spell length j, n_j is the number of spells of length j, L_{max} is the maximum observed spell length (note that $\sum_{j=1}^{L_{max}} \tilde{p}_j = 1$), $K_d(.)$ is a discrete kernel function, and L, j, and h are positive integers. The kernel function $K_d(.)$ is given as:

$$K_d(t) = at_j^2 + b \quad \text{for } |t| \leq 1 \qquad (2.4)$$

The expressions for a and b for the interior of the domain, L > h+1 and the boundary region L < h are given in Table 2.1.

The bandwidth h is estimated by minimizing a least squares cross validation (LSCV) function given as:

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}(j))^2 - 2\sum_{j=1}^{L_{max}} \hat{f}_{-j}(j)\ \tilde{p}_j \qquad (2.5)$$

where, $\hat{f}_{-j}(j)$ is the estimate of the PMF of spell length j, formed by dropping all the spells of length j from the data. This method has been shown by Hall and Titterington [1987] to automatically adapt the estimator to an extreme range of sparseness types. Monte Carlo results showing the effectiveness of the DK estimator with bandwidth selected by LSCV are presented in Rajagopalan and Lall [in press]. Figures 2.3c and d show examples of the DK estimator for two situations described in Table 2.2.

Kernel estimation of bivariate and conditional PDFs.

The bivariate PDFs of interest to us are f(w,d) and $f(w,p_w)$. The conditional PDFs of interest are f(w|d), f(d|w), and $f(p_w|w)$. It is important to note that the order in f(w|d) and f(d|w) is important, f(w|d) is estimated from data pairing wet spells following dry spells and vice versa for f(d|w). Recall that the conditional PDF f(y|x) of a random variable y given x is given as f(x,y)/f(x), where f(x,y) is the joint PDF of x and y, and f(x) is the unconditional PDF of x. Since we have discussed univariate kernel density estimation, the key step is to show how the bivariate density may be evaluated.

Bivariate kernel density estimators may be constructed in much the same manner as their univariate counterparts, i.e., through the convolution of appropriate kernel functions. Two types of bivariate kernel functions -- radially symmetric and product kernels--are

popular. Wand and Jones [1992] argue that for typical generalizations of the univariate kernels, there is little to choose between these representations. They point out that it is more important to choose bandwidths in each direction appropriately. We chose to use a product of univariate kernels for the bivariate kernel to allow a natural extension of the univariate kernel density estimators presented to discrete, bivariate or mixed (continuous and discrete) bivariate situations. The joint PDFs are estimated as follows:

$$\hat{f}(w,d) = \frac{1}{n_{sp}} \sum_{i=1}^{n_{sp}} K_d\left(\frac{w-w_i}{h_w}\right) K_d\left(\frac{d-d_i}{h_d}\right) \tag{2.6}$$

$$\hat{f}(p_w,w) = \frac{1}{n_w\, p_w\, h_{p_w}} \sum_{i=1}^{n_w} K\left(\frac{\log(p_w) - \log(p_{w_i})}{h_{p_w}}\right) K_d\left(\frac{w-w_i}{h_w}\right) \tag{2.7}$$

where n_{sp} is the number of consecutive wet and dry spells on record for season s, over the y year record, n_w is the number of wet spells.

The conditional PDFs are given by:

$$\hat{f}(w|d) = \sum_{i=1}^{n_{sp}} K_d\left(\frac{w-w_i}{h_w}\right) K_d\left(\frac{d-d_i}{h_d}\right) \Big/ \sum_{i=1}^{n_{sp}} K_d\left(\frac{d-d_i}{h_d}\right) \tag{2.8}$$

$$\hat{f}(d|w) = \sum_{i=1}^{n_{sp}} K_d\left(\frac{w-w_i}{h_w}\right) K_d\left(\frac{d-d_i}{h_d}\right) \Big/ \sum_{i=1}^{n_{sp}} K_d\left(\frac{w-w_i}{h_w}\right) \tag{2.9}$$

$$\hat{f}(p_w|w) = \frac{1}{p_w\, h_{p_w}} \sum_{i=1}^{n_w} K\left(\frac{\log(p_w) - \log(p_{w_i})}{h_{p_w}}\right) K_d\left(\frac{w-w_i}{h_w}\right) \Big/ \sum_{i=1}^{n_w} K_d\left(\frac{w-w_i}{h_w}\right) \tag{2.10}$$

We see from equations (2.8) to (2.10) that the kernel density estimator of the conditional PDF represents a weighted average of the relative frequency of values of the

dependent variable that correspond to a "weighted" neighborhood of the conditioning point. It will be seen in the section under generation of synthetic sequences, that for simulation it is not necessary to compute the joint and conditional PDFs, estimation of the bandwidths alone is sufficient.

McLachlan [1992] discusses the simultaneous selection of bandwidths in each coordinate, versus the use of the optimal univariate bandwidths in each direction. It is not clear that the additional effort of simultaneous selection of the two bandwidths is justified. Consequently, we choose the bandwidths h_w, h_d and h_{p_w} by the methods described for the univariate case.

As an illustration, a sample of size 250 is generated from a bivariate geometric distribution Geom(0.6,0.2) were used to test this procedure. The surface of the observed proportions is plotted in Figure 2.4a, the true density surface is shown in Figure 2.4b, the kernel estimated density surface is in Figure 2.4c, and the difference between the true and kernel estimates is plotted in Figure 2.4d. The bandwidth was 3 in the x direction and 6 in the y direction. To illustrate the conditional kernel density estimation, a slice is taken from the joint density in Figure 2.4c and presented in Figure 2.5.

In the precipitation data sets we have investigated thus far, the correlation between w and d is generally weak, and the serial correlation between daily precipitation for fixed spell length w is also weak. Thus, in most cases, the univariate PDFs, suffice. However, for the sake of completeness we describe a nonparametric, kernel-based disaggregation strategy for disaggregating a w day precipitation p_w into w daily precipitation amounts p_i.

Wet spell precipitation disaggregation

We follow the approach of Aitchison and Lauder [1985] for analyzing compositional data. A basic requirement for the disaggregation process is that $\sum_{i=1}^{w} p_i = p_w$.

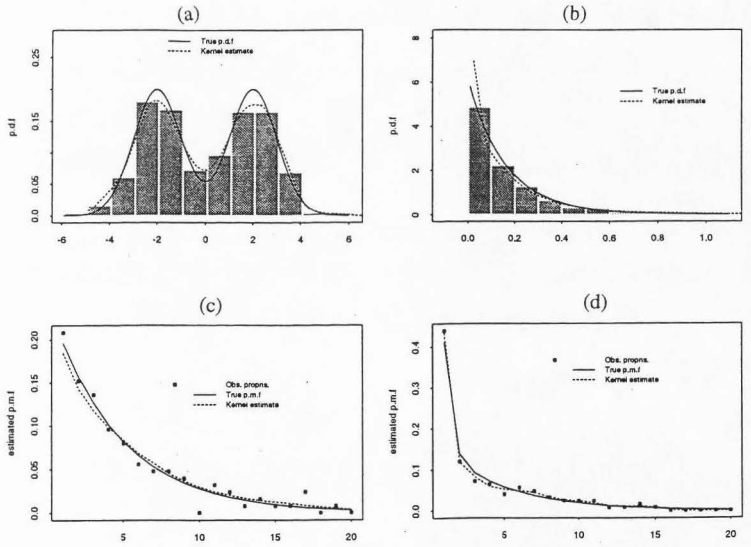


Figure 2.4. Surface plots for data generated from Geom(0.6,0.2) (a) observed proportions, (b) true PMF, (c) kernel estimated PMF, and (d) difference between kernel estimated and true PMF.

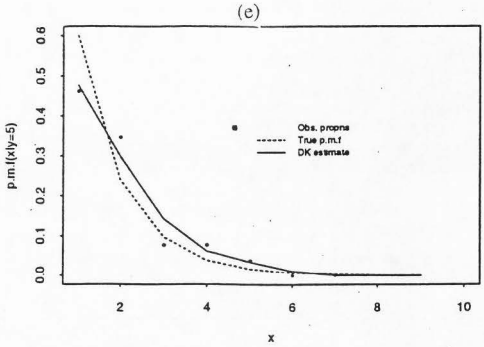


Figure 2.5. Conditional slice from Figure 2.4 (b) and Figure 2.4 (c), conditioned at y=5, along with the observed proportions at y=5.

Consider the rescaling $x_i = p_i/p_w$, so that $0< x_i < 1$, and $\sum x_i = 1$. Recognizing that the effective degrees of freedom are (w-1), we can write $x_w = 1 - \sum_{i=1}^{w-1} x_i$. Aitchison and Lauder [1985] now apply the transform

$$y_i = \log(x_i/x_w) \qquad i = 1,...,w-1 \tag{2.11}$$

The multivariate PDF f(**x**), where **x** is a vector of length (w-1) representing the first (w-1) proportions, is then estimated using the kernel method with a logistic normal kernel and n_w wet spells of length w as:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \sum_{i=1}^{n_w} \frac{1}{n_w} L(\mathbf{x},\mathbf{x}_i,\mathbf{y},\mathbf{y}_i,h) \\ &= \sum_{i=1}^{n_w} \frac{e^{-0.5\ (\mathbf{y}-\mathbf{y}_i)^T \mathbf{S}_y^{-1} (\mathbf{y}-\mathbf{y}_i)\ /h^2}}{n_w\ (2\pi)^{(w-1)/2}\ h^{(w-1)}\ \det(\mathbf{S}_y)^{1/2} \prod_{j=1}^{w} x_{ji}} \end{aligned} \tag{2.12}$$

where i is a spell index, $\mathbf{y}$ is a vector of length (w-1) as defined in Equation 2.11, x_{ji} represents the value of the j^{th} component of $\mathbf{x}$ for the i^{th} spell, $L(\mathbf{x},\mathbf{x}_i,\mathbf{y},\mathbf{y}_i,h)$ is the logistic-normal kernel, h is a bandwidth, and $\mathbf{S}_\mathbf{y}$ is the sample covariance matrix of $\mathbf{y}$, estimated using a robust method [see Huber, 1981]. The bandwidth h is selected using maximum likelihood cross validation, i.e., by choosing h to maximize $\prod_{i=1}^{Dw} f_{-i}(\mathbf{x}_i)$ where $f_{-i}(\mathbf{x}_i)$ is the estimate of $f(\mathbf{x})$ at $\mathbf{x}_i$ obtained by dropping the i^{th} point. Aitchison and Lauder [1985] demonstrated that performance of this algorithm is comparable to parametric alternatives with sample sizes ranging from 23 to 95 for 2 to 10 components.

The use of the sample covariance matrix $\mathbf{S}_\mathbf{y}$ of $\mathbf{y}$ as the covariance matrix for the kernel function for $\mathbf{y}$, leads to some degree of preservation of the covariance structure of the components of $\mathbf{y}$ and hence of the disaggregated daily precipitation amounts p_i. It also mitigates the effect of choosing x_w, rather than say x_1 as the normalizing variable in the transformation of Equation (2.11).

Using Equation (2.12), one can evaluate the PDF of the first (w-1) ratios x_i of daily precipitation to wet spell precipitation. A stochastic realization of these ratios can then be generated. The last ratio x_w is obtained by noting that all the ratios have to sum to one. Daily precipitation values are then obtained by multiplying x_i by p_w. This disaggregation procedure generalizes the logistic normal based disaggregation procedure through the use of the kernel method and admits multimodality and heterogeneity in the PDF of daily rainfall in a wet spell. A problem with any wet/dry spell model is that as w increases, n_w typically decreases. Consequently, this disaggregation scheme may not be practical for large w unless long records are available. Also, it fails to "borrow" information from spells of length other than the one generated. However, that can be a problem even for the usual parametric schemes.

Generation of synthetic sequences

Since our goal here is to generate random samples that are similar to the observed sequence, a "raw" bootstrap or resampling of the data with replacement could be considered as an alternative to sampling from the kernel density estimate. Such a strategy would be equivalent to sampling from the empirical distribution function of the data. The kernel density estimation can be thought of as a smoothed (moving average) estimate of the derivative of the empirical distribution function. Sampling from the kernel density estimate. can lead to a reduced variance of the Monte Carlo design [Silverman, 1986]. It also avoids the problem with the bootstrap where a number of the historical values are repeated in a generated sample, and provides an ability to fill in and extrapolate to a limited extent beyond the observed values.

Synthetic precipitation sequences at a site are generated continuously from season to season. A dry spell is first generated using $\hat{f}(d)$. By following the strategy indicated in Figure 2.1, a wet spell is generated using $\hat{f}(w)$ or $\hat{f}(w|d)$. Precipitation for each of w days is then generated using $\hat{f}(p)$ or $\hat{f}(p_w|w)$ followed by $\hat{f}(p_i|p_w)$. A dry spell is then generated using $\hat{f}(d)$ or $\hat{f}(d|w)$, and the process repeats. If a season boundary is crossed, the PDFs used switch to those for the new season.

For the univariate continuous case ($\hat{f}(r)$), the random variate (r) of interest can be generated readily from the kernel density following a two-step procedure [Devroye, 1986]. Consider the original sample (r_i, i=1...n) from which the kernel density (that depends on r, r_i and h) was constructed using a Kernel function K(.). To generate a random number r that follows the estimated distribution, first sample a random integer j uniformly between 1 and n, i.e., identify the historical data point to perturb. Now generate a random variate U from the probability density corresponding to the kernel function K(.), (e.g., $K(u)= 3/4(1-u^2)$ for the Epanechnikov kernel). The random variate r is then given by (r_j+Uh). This

reinforces the notion that the kernel density estimator is formed as a convolution of local densities centered at each observation, and that the generated sequence will constitute a smoothed bootstrap of the data. Any of a number of standard procedures (e.g., based on order statistics or rejection) for sampling from a density may be used to generate U from the density K(.). Devroye [1986] provides examples for the Epanechnikov kernel. The discrete random variables (w and d) are generated directly from the estimated cumulative mass function.

A similar strategy is possible for sampling from the conditional PDFs as well. Consider two continuous variables x and y. The conditional kernel density $\hat{f}(y|x)$ is given by:

$$\hat{f}(y|x) = \frac{1}{h_y} \sum_{i=1}^{n} K\left(\frac{y-y_i}{h_y}\right) K\left(\frac{x-x_i}{h_x}\right) / \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)$$
$$= \frac{1}{h_y} \sum_{i=1}^{n} wt_i \, K\left(\frac{y-y_i}{h_y}\right) \qquad (2.13)$$

where $wt_i = K\left(\frac{x-x_i}{h_x}\right) / \sum_{i=1}^{n} K\left(\frac{x-x_i}{h_x}\right)$. Now note that $\sum_{i=1}^{n} wt_i$ is equal to 1, and hence we can view the wt_i values as providing the probability metric with which the i^{th} point should be selected. Define F as the set of probabilities wt_i. Sample an integer $j \in [1,n]$ using F. Now sample a variate U from the density corresponding to the kernel function for y. The variate of interest is then $y = Uh + y_j$. The discrete variate case follows as before.

Model Application

The model described was applied to daily rainfall data from the Silver Lake station in Utah. Forty-four years of daily rainfall data were available from 1948-1992. For this

application we have divided the year into four seasons: season 1 (Jan - Mar), season 2 (Apr - Jun), season 3 (Jul - Sep), season 4 (Oct - Dec). Silver Lake is one of the higher elevation stations in Utah, situated at 40°36'N latitude, 111°35'W longitude, and at an elevation of 8740 ft. Most of the precipitation comes in the form of winter snow and season 4 rainfall. We see from Table 2.3 that season 4 (fall) has the highest mean wet day precipitation and maximum wet day precipitation, while season 1 (winter) has the highest percentage of yearly precipitation. Season 1 (winter) has the highest average wet spell length and the longest wet spell length. For the dry spells, season 3 (summer) has the highest average dry spell length and the longest dry spell length.

The successive wet and dry spells and the dry and wet spell length correlations for the data from Silver Lake, Utah were all near zero for each season. We present a representative scatter plot of the length of successive wet and dry spells for season 1 in Figure 2.6. The line in this figure is the LOWESS smooth [Cleveland, 1979]. There is little evidence of even nonlinear structure in the relationship. The correlations between daily precipitation amount on successive days within a spell were also found to be near 0. Consequently, we simulated the wet and dry spells alternately using the unconditional densities ($\hat{f}(w)$ and $\hat{f}(d)$), and used $\hat{f}(p)$ to describe the daily precipitation process. We also performed conditional simulations using the densities $\hat{f}(w|d)$ and $\hat{f}(d|w)$ for each season. The results of these simulations were very similar in terms of the performance measures (see the following section) to those from the unconditional simulations. As is to be expected, the conditional simulations exhibit slightly greater variability. Results for the conditional simulations are not presented here for the sake of brevity.

We first list some measures of performance that were used to compare the historical record and the model simulated record, and then outline the experimental design. As

Table 2.3. Statistics from the Historical Precipitation Record at Silver Lake, UT, 1948-1992

Statistic	Season 1 (Jan - Mar)	Season 2 (Apr - Jun)	Season 3 (Jul - Sep)	Season 4 (Oct - Dec)
Avg. wet spell length	2.6 days	2.2 days	1.85 days	2.5 days
Std. dev. of wet spell length	2.2 days	1.7 days	1.2 days	1.9 days
Fraction of wet days	0.62	0.44	0.36	0.55
Longest wet spell length	21 days	11 days	10 days	18 days
Avg. dry spell length	3.0 days	5.1 days	6.0 days	4.0 days
Std. dev. of dry spell length	2.80 days	6.0 days	6.0 days	4.0 days
Fraction of dry days	0.38	0.56	0.64	0.45
Longest dry spell length	19 days	42 days	45 days	24 days
Avg. wet day precip.	0.37 in.	0.33 in.	0.26 in.	0.40 in.
Std. dev. of wet day precip.	0.37 in.	0.33 in.	0.30 in.	0.42 in.
Fraction of yearly precip.	0.35	0.20	0.12	0.30
Max. wet day precip.	3.7 in.	3.0 in.	1.90 in.	3.5 in.

emphasized earlier in the manuscript our goal is to reproduce the frequency structure (i.e., the underlying PDF). One would then expect that the usual statistics are reproduced.

Performance measures

1. Probability distribution function of wet spell length, dry spell length, and wet day precipitation.

2. Mean of wet spell length, dry spell length, and wet day precipitation in each season.

3. Standard deviation of wet spell length, dry spell length, and wet day precipitation in each season.

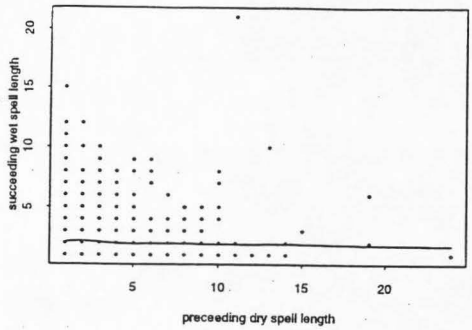


Figure 2.6. Scatter plot of preceeding dry spell length and following wet spell length in season 1, along with the LOWESS smooth (solid line).

4. Length of longest wet spell and dry spell in each season.

5. Maximum wet day precipitation in each season.

6. Percentage of yearly precipitation in each season.

7. Fraction of wet and dry days in each season.

Experiment design

The resampling process proceeded as follows:

1. Wet and dry spells for each season are determined from the daily precipitation data. Spells that cross seasonal boundaries are truncated at the season boundary and included in the appropriate seasons. We recognize that this could have the effect of

introducing a small bias in the spell characteristics for a given season. Missing data are skipped, and the spell count is restarted with the next event.

2. Probability density/mass functions are fitted for the wet day precipitation, wet spell lengths, and dry spell lengths for each season using the recommended kernel estimators.

3. Twenty-five synthetic records of 44 years each (i.e., the historical record length) are simulated using the NSS model.

4. The statistics of interest are computed for each simulated record, for each season and compared to statistics of the historical record using boxplots.

Results

In this section we present comparative results of the NSS model for the Silver Lake data. The statistics (PDFs) of the simulated records are compared with those for the historical record using boxplots. A box in the boxplots (e.g., Figure 2.8) indicates the interquartile range of the statistic computed from twenty -five simulations, the line in the middle of the box indicates the median simulated value. The solid lines correspond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics. The plots of the PDFs are truncated to show a common range across seasons and to highlight differences near the origin (mode).

Wet day precipitation

Figure 2.7 shows that the fitted kernel densities for wet day precipitation amount are similar to the histogram of the recorded data in all four seasons. They differ from the fitted Exponential and Gamma distribution, particularly in seasons 3 (summer) and 4 (fall). The kernel estimated PDFs of the simulated data reproduce the PDFs of the historical data

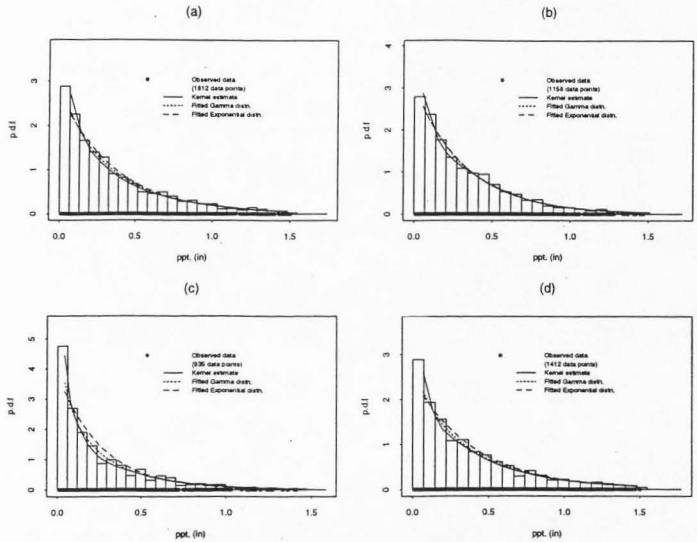


Figure 2.7. Plots of PDFs of wet day precipitation at Silver Lake, UT, estimated using SJL procedure, the fitted Exponential distribution, fitted Gamma distribution and histogram of the observed data (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

quite well, as can be seen in Figure 2.8. The other statistics are reproduced well by the model, as can be seen from the boxplots in Figure 2.9.

Wet spell length

Figure 2.10 shows that the PMF.'s of wet spell length estimated by DKE and the fitted geometric distribution are very close (except perhaps for season 1 (winter)). In this case one could argue for using the Geometric distribution rather than DKE. But the "loss" in using DKE is small and for uniform application across sites, DKE may still be a better choice. The PMFs of wet spell length from the simulations reproduce the historical PDF very well in all the seasons as can be noted from Figure 2.11, suggesting that the model is performing well in reproducing the underlying frequency structure. Figure 2.12 shows that the mean, standard deviation, fraction of wet days, and longest wet spell length are all well reproduced by the model.

Dry spell length

Figure 2.13 shows that the dry spell length PMFs estimated by DKE and the fitted Geometric distribution are generally similar with the most difference in season 3 (summer), which we noted as being the most "active" with regard to dry spell length extremes. Observationally, we know that there are dry summers with little rainfall activity and other summers with intermittent, stagnating precipitation systems in this area. Thus we would expect a mixture of mechanisms generating dry spells to show up in this.

The PMFs of wet spell length from the simulations reproduce the historical PDF very well in all the seasons as can be noted from Figure 2.14, suggesting that the model is performing well in reproducing the underlying frequency structure. Figure 2.15 shows that the statistics of the dry spell length are also well reproduced.

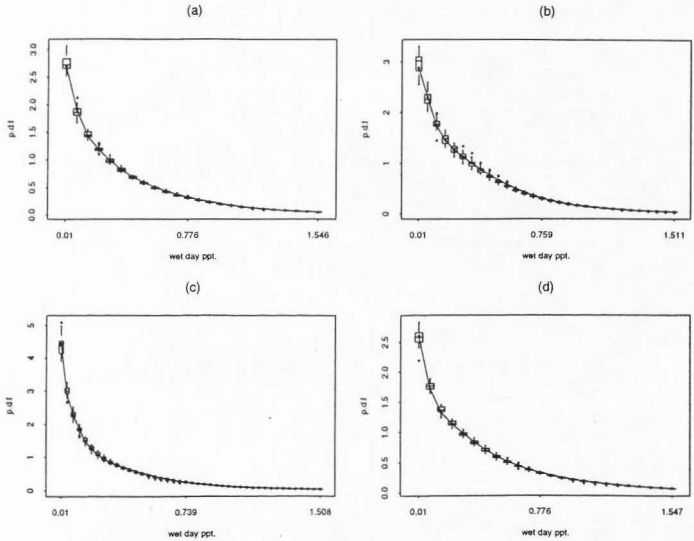


Figure 2.8. Boxplots of PDF of wet day precipitation for model simulated records along with the historical values (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

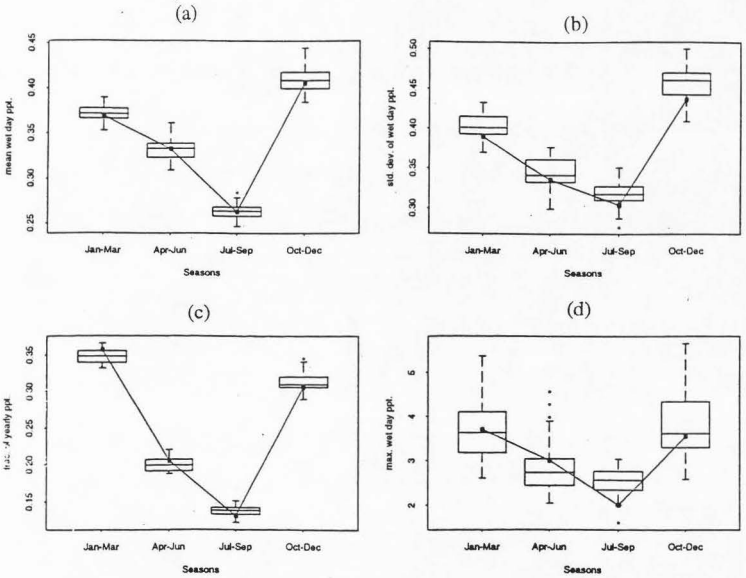


Figure 2.9. Boxplots of statistics of wet day precipitation (a) mean wet day precipitation, (b) standard deviation of wet day precipitation, (c) fraction of yearly wet day precipitation, and (d) maximum wet day precipitation for model simulations along with the historical values for the four seasons.

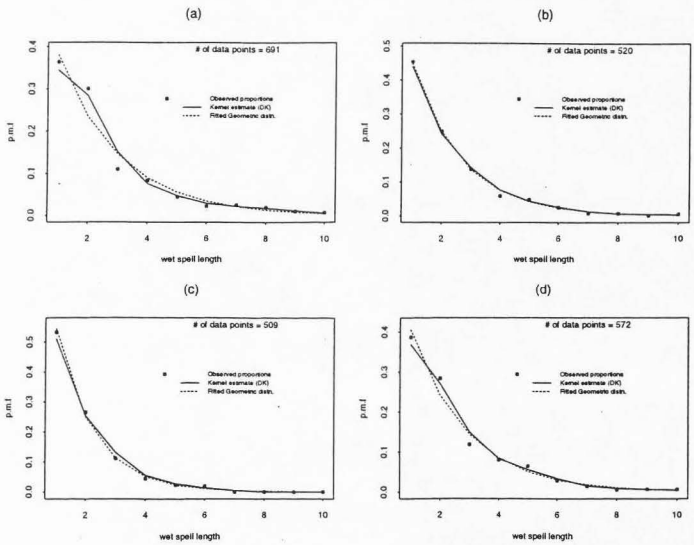


Figure 2.10. Plots of PMF of wet spell length at Silver Lake, UT, estimated using DK estimator. Along with the fitted Geometric distribution and observed proportions (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

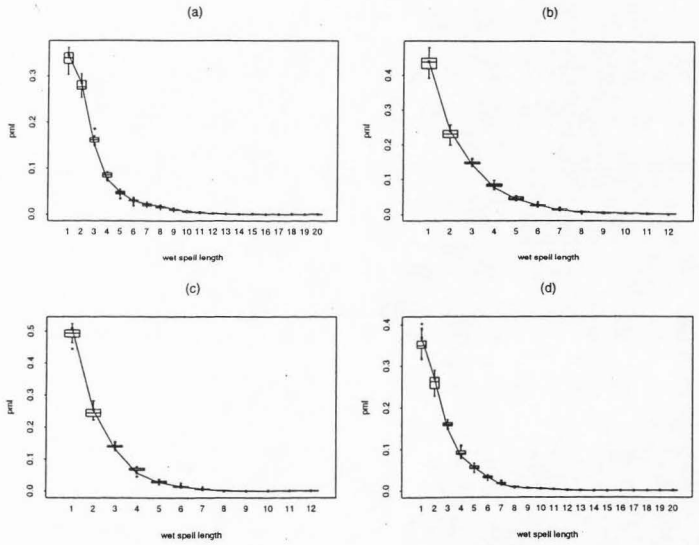


Figure 2.11. Boxplots of PMF of wet spell length for model simulated records along with the historical values (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

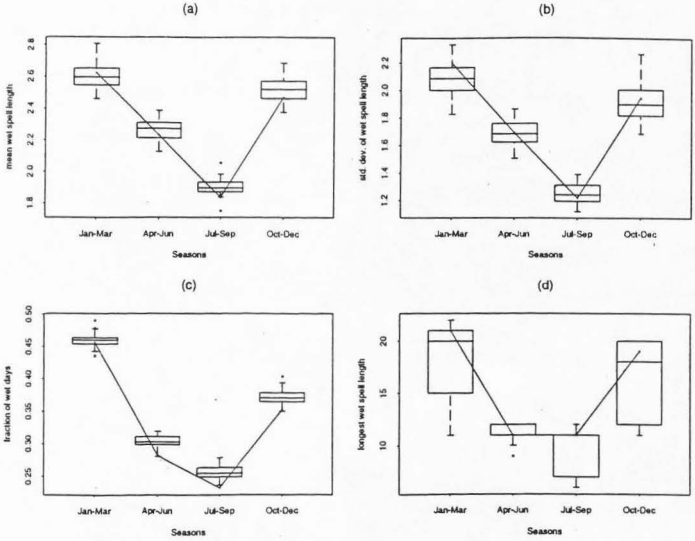


Figure 2.12. Boxplots of statistics of wet spell length (a) mean wet spell length, (b) standard deviation of wet spell length, (c) fraction of wet days, and (d) longest wet spell length for model simulations along with the historical values for the four seasons.

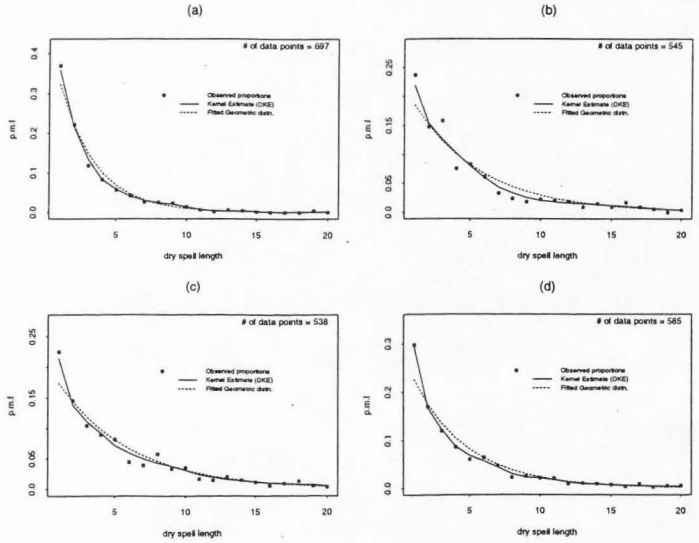


Figure 2.13. Plots of PMF of dry spell length at Silver Lake, UT, estimated using DK estimator. Along with the fitted Geometric distribution and observed proportions (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

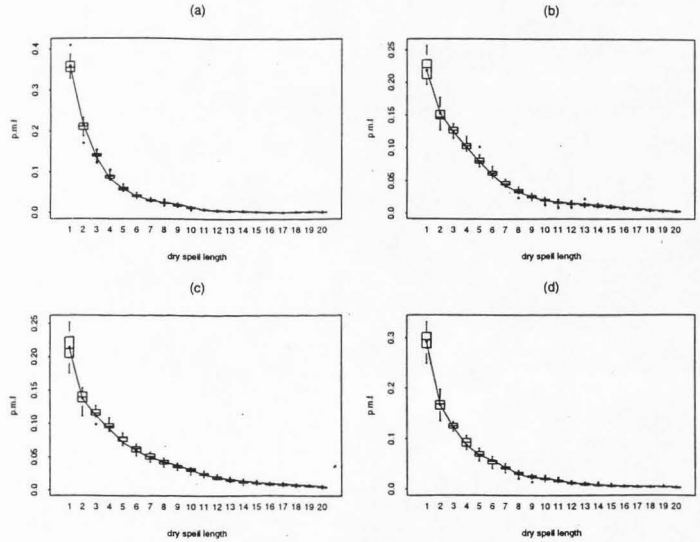


Figure 2.14. Boxplots of PMF of dry spell length for model simulated records along with the historical values (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

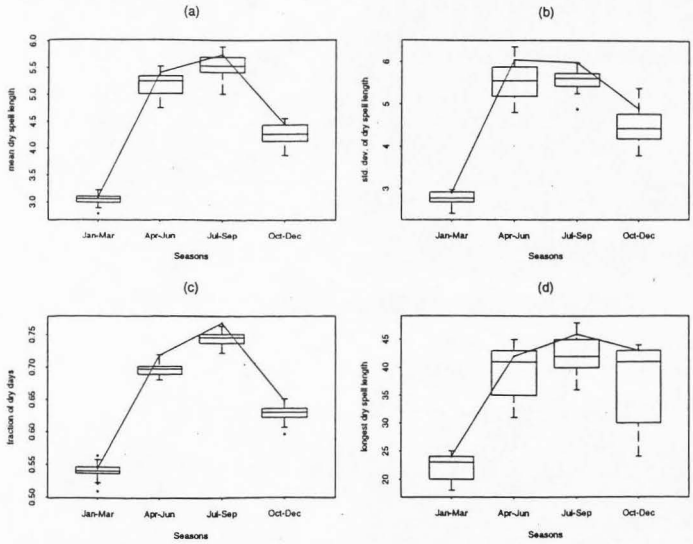


Figure 2.15. Boxplots of statistics of dry spell length (a) mean dry spell length (b) standard deviation of dry spell length, (c) fraction of dry days, and (d) longest dry spell length for model simulations along with the historical values for all the four seasons.

The reader may be tempted to suggest formal tests to check for a mixture of the geometric distributions in this case as an alternative to the kernel density estimate. While this may be a fruitful activity (we did consider it), it gets harder to perform and/or justify as we consider arbitrary, finite component mixtures. An advantage of the kernel density of the kernel density estimator (DKE) employed here is that it readily admits such mixtures without requiring that they be hypothesized or formally identified. We feel that this provides a more direct and parsimonious representation of this sort of structure if present in the data.

Summary and Conclusions

A nonparametric methodology for simulating daily precipitation is presented in this chapter. The traditional wet/dry spell model is extended to (1) consider heterogeneity in the PDF of precipitation or wet/dry spell length, and (2) consider dependence between wet/dry spell length, and between wet spell length and spell precipitation. All functions of interest are estimated nonparametrically. The primary intended use of the model is as a simulator that is faithful to the historical data sequence. The PDFs evaluated are also likely to be of use for justifying the use of other formal, parametric models of the underlying process.

While a rather flexible framework is provided by the model proposed, it is not without a price. Sample sizes needed for estimating the PDFs of interest are likely to be larger than for parametric estimation. However, the nonparametric specification of the PDFs leads to robustness with respect to the misspecification of the parametric model which may be valuable if the use of a particular model is to be legislated across a variety of sites and regions with different attributes. Only a crude treatment for seasonal nonstationarity is offered. This is something we expect to address in the future.

A number of issues of interest to stochastic precipitation modelers were not discussed here. The foremost is the behavior of the proposed model at different time scales.

We view our developments as "operational" and relevant to the time scale of the data, which was daily. Spell definitions are tenuous at best at finer time scales and sample sizes drop rapidly as longer time scales (e.g., monthly or annual) are considered. Thus while the scaling issue is of theoretical and practical interest, it is difficult to formally assess how such a model may fit in. It is an issue we expect to explore in due course. A second issue is the need to incorporate climatic or precipitation "types" [e.g., Bogardi et al., 1993, Wilson and Lattenmaier, 1993] into the daily precipitation model. We feel that implicit consideration of some of these factors is provided by our model by admitting an arbitrary mixture of generating mechanisms. Transitions between generating mechanisms are not explicitly modeled. However, their relative frequencies ought to be reproduced. Given limited data sets and the potentially large number of generating mechanisms, this may be all that is reliably feasible in a number of cases. Finally, there is the question of regionalization and/or portability of the method. The nonparametric approach clearly enjoys broader applicability than its parametric competitors. On the other hand, it may be less amenable to direct regionalization as is sometimes done in terms of the parameters of a parametric distribution. It is meaningless to talk of a regional bandwidth. It may be more fruitful to develop a space-time nonparametric precipitation model with a nonhomogeneous point process structure that is inferred from the data.

References

Aitchison, J., and I.J. Lauder, Kernel density estimation for compositional data, *Applied Statistics,* 34(2), 129-137, 1985.

Bogardi, I., I. Matyasovszky, A. Bardossy, and L. Duckstein, Application of space-time stochastic model for daily precipitation using atmospheric circulation patterns, *Journal of Geophysical Research*, 98, D9, 16653-16667, 1993.

Cayan, D., and L. Riddle, Atmospheric circulation and precipitation in the Sierra Nevada, Managing water resources during global change, *Conference Proceedings of American Water Resources Association*, Tucson, AZ, 1992.

Chang, T.J., M.L. Kavvas, and J.W. Delleur, Daily precipitation modeling by discrete autoregressive moving average process, *Water Resources Research*, 17, 1261-1271, 1984.

Chin, E.H., Modeling daily precipitation occurrence process with Markov Chain, *Water Resources Research*, 13, 949-956, 1977.

Cleveland, W.S., Robust locally weighted regression and smoothing scatter plots, *Journal of American Statistical Association*, 74, 829-836, 1979.

Devroye, L., *Non-Uniform Random Variate Generation*, Springer-Verlag, New-York, 1986.

Feyerherm, A.M., and L.D. Bark, Statistical methods for persistent precipitation patterns, *Journal of Applied Meteorology*, 4, 320-328,1965.

Feyerherm, A.M., and L.D. Bark, Goodness of fit of a markov chain model for sequences of wet and dry days, *Journal of Applied Meteorology*, 6, 770-773, 1967.

Foufoula-Georgiou, E., and K.P. Georgakakos, Recent advances in space-time precipitation modeling and forecasting, *NATO ASI on Recent Advances in the Modelling of Hydrologic Systems,* Sintra, Portugal, July, 1988.

Georgakakos, K.P., and M.L. Kavvas, Precipitation analysis, modeling, and prediction in hydrology, *Reviews of Geophysics*, 25(2), 163-178, 1987.

Guzman, A.G., and C.W. Torrez, Daily rainfall probabilities: conditional upon prior occurrence and amount of rain, *Journal of Climate and Applied Meteorology*, 24(10), 1009-1014, 1985.

Haan, C.T., D.M. Allen, and J.O. Street, A Markov chain model of daily rainfall. *Water Resources Research*, 12(3), 443-449, 1976.

Hall, P., and D.M. Titterington, On smoothing sparse multinomial data, *Australian Journal of Statistics*, 29(1), 19-37, 1987.

Härdle, W., *Smoothing Techniques with Implementation in S*, Springer-Verlag, New York, 1991.

Hopkins, J.W., and P. Robillard, Some statistics of daily rainfall occurrence for the canadian prairie provinces, *Journal of Applied. Meteorology*, 3, 600-602, 1964.

Huber, P. J., *Robust Statistics*, John Wiley, New York, 1981.

Katz, R.W., and M.B. Parlange, Effects of an index of atmospheric circulation on stochastic properties of precipitation, *Water Resources Research*, 29(7), 2335-2344, 1993.

Lall, U., Nonparametric function estimation: Recent hydrologic applications, *US National Report, 1991-1994, International Union of Geodesy and Geophysics*, 1994.

McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992.

Rajagopalan, B., and U. Lall, A kernel estimator for discrete distributions, *Journal of Nonparametric Statistics*, (in press).

Rajagoplan, B., U. Lall, and D.G. Tarboton, Evaluation of kernel density estimation methods for daily precipitation resampling, Working Paper WP-95-HWR-UL/007, In Utah Water Research Laboratory, Utah State University, Logan, UT, 1995

Roldan J., and D.A. Woolhiser, Stochastic daily precipitation models 1. A comparison of occurrence processes, *Water Resources Research*, 18(5), 1451-1459, 1982.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in probability and mathematical statistics, John Wiley, New York, 1992.

Sheather, S.J., and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society*, B. 53, 683-690, 1991.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

Srikanthan, R., and T.A. McMahon, Stochastic simulation of daily rainfall for Australian stations. *Transactions of the ASAE*, 754-766, 1983.

Vogel, R.M., and D.E. McMartin, Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type 3 distribution, *Water Resources Research*, 27(12), 3149-3158, 1991.

Wand, J.S., and M.C. Jones, Comparison of smoothing parameterizations in bivariate kernel density estimation, *Journal of American Statistical Association*, 88(422), 520-528, 1992.

Waymire, E., and V.K. Gupta, The mathematical structure of rainfall representations. 1. A review of the stochastic rainfall models, *Water Resources Research*, 17(5), 1261-1272, 1981a.

Waymire, E., and V.K. Gupta, The mathematical structure of rainfall representations. 2. A review of the theory of point processes, *Water Resources Research*, 17(5), 1273-1285, 1981b.

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

Webb, R.H., and J.L. Bettencourt, Climatic variability and flood frequency of the Santa Cruz river, Pima County, Arizona, U.S. Geological Survey Water-Supply, *Paper* 2379, 1992.

Wilson, L.L., and D.P. Lettenmaier, A hierarchical stochastic-model of large-scale atmospheric circulation patterns and multiple station daily precipitation, *Journal of Geophysical Research-Atmospheres*, 97(ND3), 2791-2809, 1993.

Woolhiser, D.A., C.L. Hanson, and C.W. Richardson, Microcomputer program for daily weather simulation, United States Department of Agriculture, Agricultural Research Service - 75, 49p, 1988.

CHAPTER III

EVALUATION OF KERNEL DENSITY ESTIMATION METHODS FOR DAILY PRECIPITATION RESAMPLING[1]

Abstract

Issues related to the selection and design of appropriate nonparametric estimators for the nonparametric wet/dry spell model developed in Lall et al.[1995] are examined. Here we present results of our investigations into selected aspects of kernel density estimation for both continuous and discrete variables with reference to the nature of data typically available for wet/dry spell modeling of daily precipitation.

Introduction

In a companion paper Lall et al. [1995], a nonparametric approach to a stochastic model for daily precipitation was presented. The salient features of this model were the consideration of alternating wet and dry spells and of a daily rainfall structure within the wet spell. Kernel density estimates were espoused as effective methods for recovering univariate, multivariate or conditional, discrete and/or continuous probability densities that were needed directly from the historical record. In the process of developing the nonparametric wet/dry spell model in Lall et al. [1995], kernel density estimators of continuous and discrete variables were reviewed and tested with various data sets. Our aim here is to present some of this experience, specifically with the type of data available for modeling daily precipitation as a wet/dry spell model.

Here, we shall explore some of the issues relevant to the implementation of the kernel density estimators proposed in Lall et al. [1995]. These are (1) the specification of the bandwidth of the kernel estimator for the continuous case, (2) the role of boundary

[1]Coauthored by Rajagopalan Balaji, Upmanu Lall and David G. Tarboton.

effects in kernel estimation, and (3) the selection of the estimator in the discrete case. The intent is to justify our recommended procedures by example, and to provide a comparison of some of the estimation schemes available in the literature.

Investigations for estimating the probability density function (PDF) of continuous random variables (here, it is the precipitation amount for a day or for a wet spell) are first presented followed by comparisons of methods for the estimation of the probability mass function (PMF) of discrete random variables (here, it is the length of a wet spell or dry spell in days).

Kernel Density Estimation of Continuous Random Variable

We start with the introduction of some basic ideas of kernel density estimation for continuous univariate case. The kernel density estimation for univariate, continuous random variates was reviewed recently by Lall et al. [1995], in the flood frequency estimation context. The presentation here adds a few recent bandwidth estimation methods, and a discussion of the possible utility of boundary kernels with precipitation data. The interested reader is referred to Silverman [1986] for a pragmatic treatment of Kernel density estimation, to Devroye and Györfi [1985] for a rigorous treatment using L1 (absolute value) methods, and to Scott [1992] for a recent monograph with an excellent treatment of multivariate estimation. Hydrologic applications are reviewed in Lall [1994].

Basic ideas

Hydrologists are familiar with the frequency histogram as an estimator of the PDF While the histogram is capable of estimating the relative frequency distribution of the data, it has several drawbacks. It is difficult to manipulate analytically. It is not easy to visualize for multivariate situations, and it allows for no extrapolation beyond the data. The

indicated frequency distribution is sensitive to the class width, as well as the origin of each class. Silverman [1986] illustrates these problems graphically. One can improve the histogram by centering bins at each observation (to gain independence from choice of origin) and then using boxes of shapes other than a rectangle (to get differentiable and continuous densities). This is precisely what the kernel density estimator introduced by Rosenblatt [1956] does. Given observations $x_1, x_2, ..., x_n$, the kernel density estimator at any point x is $\hat{f}_n(x)$ is defined as:

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - x_i}{h}\right)) \tag{3.1}$$

where K(.) is a kernel function centered on the observation x_i, which is usually taken to be a symmetric, positive, probability density function that satisfies conditions in Equation (3.2) (positivity, integrates to unity, first moment equal to zero and finite variance) and h is a bandwidth or "scale" parameter of the kernel.

$$\text{(a) } K(t) > 0 \text{ ; (b) } \int K(t)dt = 1 \text{ ; (c) } \int tK(t) = 0;$$

$$\text{(d) } \int t^2K(t) = k2 \neq 0 \tag{3.2}$$

From Equation (3.1), we can see that the estimator $f_n(x)$ is a local weighted average of the relative frequency of observations in the neighborhood of x. The kernel function K(.) prescribes the relative weights, and the bandwidth h prescribes the range of x values over which the average is computed. If K(.) integrates to unity, and is positive, the basic conditions for a valid probability density are satisfied by $\hat{f}_n(x)$. Symmetry of K(.) leads to equal weighting of observations on either side of x_i, and helps reduce the asymptotic bias

of $\hat{f}_n(x)$, while finite variance of K(.) ensures that the variance of $\hat{f}_n(x)$ is finite. Examples of kernel functions that are often used are provided in Table 3.1. In this work, we have used the Epanechnikov and the Bisquare kernels.

Table 3.1. Examples of Continuous Variable Kernel Functions

Note $t = (x-x_i)/h$

Kernels

Normal	$K(t) = (2\pi)^{-1/2} e^{-t^2/2}$
Epanechnikov	$K(t) = 0.74(1- t^2) \quad \lvert t \rvert \leq 1$
Bisquare	$K(t) = 0.9375(1-t^2)^2 \quad \lvert t \rvert \leq 1$

Continuous (Left) Boundary Kernels, Univariate (Müller, 1991)

Note that $q = x/h$, $0 \leq q \leq 1$ and x is the point at which the density is estimated, and h is the bandwidth.

for Epanechnikov $K(q,t) = 6(1+t)(q-t)\frac{1}{(1+q)^3}\{1 + 5(\frac{1-q}{1+q})^2 + 10\frac{1-q}{(1+q)^2}t\}$

One can also see from Equation (3.1) that the kernel density estimator is a convolution estimator, i.e., it results from the convolution of local densities across the data set. This interpretation is illustrated in Figure 3.1 using the Bisquare kernel. Note that specification of the kernel function K(.) and the bandwidth completely describes the above estimator. These are the parameters of the method. Nevertheless, such an estimator is called *nonparametric* because the resulting estimate is "local," i.e., defined over a neighborhood (parametrized by K(.) and h) of the point of estimate, and no assumptions have been made about the "global" underlying form of the probability density function. Since the PDF is estimated piece by piece (essentially as a moving average), a large class of underlying PDFs can be estimated by the kernel density estimator. This is a key feature of a *nonparametric* estimator. By contrast, a parametric estimator would have the entire function of specified form and could be indexed by a finite number of parameters.

Consequently, a parametric model addresses a much more restrictive set of target probability models.

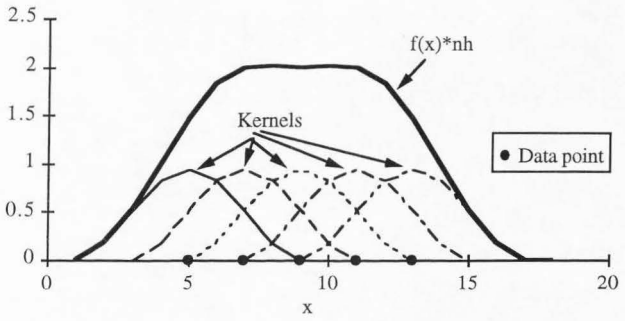


Figure 3.1. Example of kernel density estimation using 5 equally spaced values (5-13) with Bisquare kernel, h = 4.

Other examples [see Devroye and Györfi, 1985; Silverman, 1986; Scott, 1992] of nonparametric estimators are the k nearest neighbor density estimator, Fourier series estimators, adaptive shifted histograms, frequency polygons, penalized likelihood estimators, and orthogonal series estimators. All these methods can be shown to be equivalent to kernel density estimators with special kernels.

The goal of nonparametric density estimation is to obtain a good pointwise estimate of the underlying PDF. Consequently, the performance of the estimator is judged by the pointwise error. The choice of the estimator and the bandwidth is motivated through an analysis of mean squared error (MSE) in estimating the density at a point x, given as

$$MSE(\hat{f}_n(x)) = E\{[f(x) - \hat{f}_n(x)]^2\} \qquad (3.3)$$

where E[.] denotes the expectation operator. Härdle [1991] provides the asymptotic mean square error of the kernel density estimate (Equation [3.1]) for differentiable f(x) and for kernels satisfying (2), through a Taylor Series expansion of the MSE as:

$$MSE(\hat{f}_n(x)) = \{ 0.5\, h^2 f''(x) \int t^2 K(t)\, dt\}^2 + (nh)^{-1} f(x) \int K^2(t)\, dt + O(h^2) \qquad (3.4)$$

The first term in Equation (3.4) is the bias squared and the second is the variance of the estimate at x. Since it is a weighted moving average, the kernel density estimator typically underestimates the density at the modes, and overestimates it at the antimodes, corresponding to the bias term that is proportional to f''(x). The mean integrated squared error (MISE $= \int MSE(\hat{f}_n(x)\, dx$) and related measures of performance can be developed from Equation (3.4). For kernels satisfying (3.2), and an optimally selected, fixed bandwidth h, the rate of convergence in terms of MISE of the kernel density estimate is proportional to $n^{-4/5}$ (compare with $n^{-2/3}$ for the histogram), see Silverman [1986, sec. 3.7.2] for details. The best rate for a parametric estimator is proportional to n^{-1}. If higher order kernels (these are symmetric kernels with the first [p-1] moments zero, the p^{th} moment finite, nonnegativity is not enforced) are used and/or variable bandwidths are employed, higher convergence rates ($n^{-2p/(2p+1)}$ for an order p kernel) can be achieved, Scott [1992]. However, for p>2, the resulting estimate $\hat{f}_n(x)$ may not be positive, and may not constitute a valid probability density. Kernels satisfying (2) are of order 2.

Epanechnikov [1969] showed that the MSE optimal kernel (among the class of kernels that are positive everywhere and have first moment and second moment finite) for

density estimation is the quadratic kernel bearing his name given in Table 3.1. He also showed that the asymptotic relative MSE efficiency (MSE($\hat{f}_n(x)$ using kernel/MSE($\hat{f}_n(x)$ using optimal kernel) of any other admissible kernel function (even the rectangular kernel) was always close to one. The reason for this is that different kernels can be made equivalent in this sense through appropriate choices of the bandwidth [Scott, 1992]. Consequently it is generally believed that the choice of a kernel function is not very important for density estimation as far as the asymptotic MSE is concerned. However, there are other factors that are important for choosing a kernel function. The differentiability of the kernel function is inherited by the resulting density estimate. The Epanechnikov kernel is not differentiable at the ends of its support. The Bisquare kernel (Table 3.1) is to be preferred in this regard. Where the random variable is bounded (e.g., precipitation is defined only over $[0,\infty]$), a kernel with bounded support is to be preferred (e.g., Epanechnikov or Bisquare) over one with infinite support (e.g., Normal) to minimize boundary effects (which will be discussed in the following section).

Typically the bandwidth and the kernel are selected by minimizing the estimated average mean integrated square error (AMISE=E[$\int$ MSE($\hat{f}_n(x)$ dx)]). Methods for bandwidth selection are described in the section under bandwidth selection schemes and are summarized in Table 3.2.

Since kernel density estimation is a local averaging process, estimates in the tail (especially for data from long tailed distributions) can be rough (have high variance of estimate) because there will be fewer and fewer data points to average for a fixed bandwidth. A natural way to deal with with such situations is to use a larger h in regions of low density (e.g., tails) and smaller h in regions of high density (e.g., near the modes). The bandwidth may thus vary over the range of the data. The estimator in this case is called a variable kernel density estimator and is given as:

Table 3.2. Choices of Bandwidth Selection for Kernel Estimators of Continuous Variables

Method	Equation	Criteria/Remarks
PR-M	$h_{opt} = 3.03n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming underlying probability density function to be 0.5N(-2,1)+0.5N(2,1).
PR-N	$hopt = 2.13\widehat{\sigma}n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $N(0,\widehat{\sigma}^2)$. $\widehat{\sigma}$ is the sample standard deviation.
PR-E	$hopt = 1.97\widehat{\sigma}n^{-0.2}$	Based on minimization of MISE, given Epanechnikov kernel and assuming the underlying probability density function to be $Exp(\widehat{\alpha})$. $\widehat{\sigma}$ is the sample standard deviation.
LSCV	$LSCV(h) = \int \hat{f}^2 - 2n\text{-}1 \sum_{i=1}^{n} \hat{f}_{-i}(x_i)$	Choose h to minimize LSCV(h) function. $\hat{f}_{-i}$ represents the kernel density estimate constructed by dropping the i^{th} observation.
MLCV	$MLCV(h) = n^{-1}\sum_{i=1}^{n} \log(\hat{f}_{-i})$	Choose h to maximize MLCV(h) function.
SJ	refer to Equations, 3.13-3.15,	Based on recursive estimation of MISE
SJL	Same as SJ, but applied to log transformed data	

Note:

PR Parametric reference

LSCV Least squares cross validation

MLCV Maximum likelihood cross validation

SJ Sheather and Jones [1991] procedure

SJL Sheather and Jones [1991] procedure applied to log transformed data

MISE Mean integrated squared error

$$\hat{f}_n(x) = \sum_{i=1}^{n} \frac{1}{n\, h_i} K\left(\frac{x - x_i}{h_i}\right) \qquad (3.5)$$

where h_i is the bandwidth prescribed at the observation x_i.

Estimation of a variable bandwidth h_i is more difficult than the estimation of the global bandwidth h. A practical approach is a procedure suggested by Silverman [1986] based on recommendations by Abramson [1982], who showed that choosing h_i proportional to $\hat{f}_n(x_i)^{-1/2}$ could improve the MSE rate of convergence of $\hat{f}_n(x)$ from $O(n^{-4/5})$ to $O(n^{-8/9})$. Here O(.) refers to "terms of the order of," and for comparison the optimal convergence rate for a parametric density estimate is usually $O(n^{-1})$. The strategy is to perturb an appropriate fixed or global bandwidth h into a sequence of bandwidths h_i at each observation x_i as:

$$h_i = h(\hat{f}_n(x_i)) / g)^{-1/2} \qquad (3.6)$$

where g is the geometric mean of $\hat{f}_n(x_i)$. One can iteratively re-estimate $\hat{f}_n(x_i)$ and hence h_i using the latest kernel density estimate. Two to three such iterations were found to be sufficient to achieve pointwise convergence to a fractional tolerance of 0.001 in the resulting density estimate.

Boundary effects and their treatment

An annoying aspect of kernel estimators of probability densities (both continuous and discrete) is the increased bias within one bandwidth of the boundary (e.g., 0) of the sample space. The bias is a consequence of the increasingly asymmetric distribution of the random variable as one approaches the boundary. Modifications to kernel density estimate

are necessitated within a bandwidth of the boundary (e.g., 0 for data from exponential distribution) of the sample space. Two problems are faced for estimation in the boundary region.

The first is that a kernel can extend past the boundary if the bandwidth is larger than the observation at which a kernel function is centered. This leads to a leakage of probability mass, and the resulting $\hat{f}_n(x)$ will not integrate to 1 over the sampling domain. Clearly this problem is aggravated if a kernel with infinite support is used (such as the Gaussian kernel, see Table 3.1). The boundary problem is illustrated in Figure 3.2. Consider the continuous univariate random variable $x \in [0,\infty]$, and a fixed bandwidth (h=0.1). For the point of estimation in the Figure 3.2 (i.e., x = 0.01), which is within one bandwidth of the boundary, the interior Epanechnikov kernel is truncated at the boundary (x=0.0), resulting in the leakage of probability mass. Boundary kernels developed by Müller [1992] alleviate this problem.

The second problem is increased bias that results from the asymmetric distribution of observations around the point of estimate. Let us say that the smallest sample value is x_1, and that x_1 is greater than h. Now if a kernel estimate of $\hat{f}_n(x)$ is needed for x<h, i.e., in the boundary region, all the sample values are to the right of x, leading to an increased bias in the estimate $\hat{f}_n(x)$. Attempts to overcome this bias typically lead to an increased variance due to the relatively few points caught in a bandwidth of the kernel.

A number of methods for dealing with the boundary problems mentioned above have been proposed. We investigated four methods for boundary modification of the kernel estimator.

The first method is "cut and normalize." One computes the area of each kernel that lies within the sample space, and normalizes the truncated kernel to have unit area, by dividing the kernel function by this area. Bias reduction issues are not addressed.

The second method, reflection, augments the data set by reflection of the real data across the boundary. The assumption is that $f'(x) = 0$. There is no basis for this assumption and it is unlikely that it holds for the precipitation data sets.

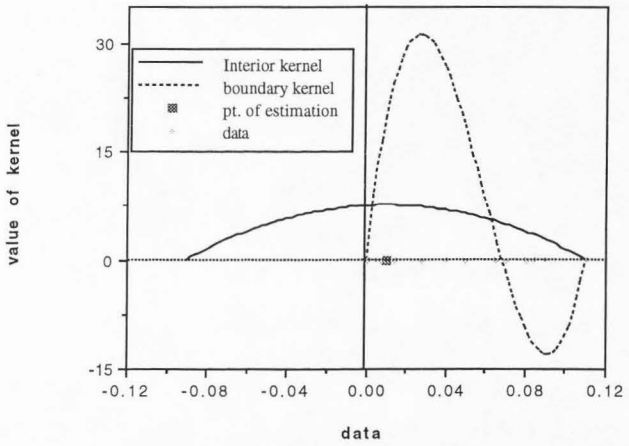


Figure 3.2. Conceptual figure of the boundary problem in kernel density estimation.

The third method, which is more general, considers the development of special boundary kernels [see Müller, 1988, 1992; and Table 3.1] that are asymmetric, unbiased, and minimum variance but are not nonnegative. These kernels are modified versions of the kernels used in the interior of the sample space, and are derived from variational conditions

[see Müller, 1992 for details]. We have investigated such kernels in the univariate case with reasonably good results. Bias of the density estimate is reduced in the boundary region, typically with some increase in the variance of estimate. For the type of data we were dealing with (precipitation or spell length), the density is high near the origin (i.e., 0.01 and 1, respectively), and the possible negative values of the boundary kernel function near the origin do not translate into negative density estimates. For the discrete case, Dong and Simonoff [1994] have developed boundary kernels for the Epanechnikov and Bisquare kernels (see Table 3.1 for boundary kernels for Epanechnikov kernel).

A fourth method relevant for data concentrated near the boundary (e.g., Exponential, log normal) is a logarithmic transform of the data prior to density estimation. Such a transformation can also provide an automatic degree of adaptability of the bandwidth (in real space), thus alleviating the need to choose variable bandwidths with heavily skewed data, and also alleviates problems that the kernel density estimator has with PDF estimates near the boundary (e.g., the origin) of the sample space. The resulting kernel density estimator can be written as:

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_x x} K\left(\frac{\log(x) - \log(x_i)}{h_x}\right) \qquad (3.7)$$

where h_x is the bandwidth of the log transformed data. The above estimator worked well for data concentrated near the origin (e.g., Exponential type) and hence is recommended.

Bandwidth selection schemes

In this section we review some choices of bandwidth selection for kernel density estimation for continuous variables. Comparisons of these alternatives with synthetic data are presented next. Rather than reproducing a variety of statistical results, we shall focus

on getting the basic ideas across through a brief review of the univariate, continuous random variable case.

Four methods for selecting the optimal global bandwidth were considered.

1. Parametric reference (PR) procedure.

The optimal bandwidth h_{opt} and kernel are selected by first minimizing the mean integrated squared error (MISE), Equation (3.4) integrated with respect to h. The result is the optimal bandwidth h_{opt} and then solving for the optimal kernel [see Silverman, 1986].

The MISE of the fixed, univariate, continuous, kernel density estimator and the corresponding optimal global bandwidth h_{opt} are given by Silverman [1986] as:

$$MISE(f_n(p)) \approx (nh)^{-1}R(K) + 0.25h^4\sigma_K^2R(f'') \tag{3.8}$$

$$h_{opt} = \{R(K)/(\sigma_K^2R(f''))\}^{1/5}\, n^{-1/5} \tag{3.9}$$

where $R(g) = \int g^2(x)dx$ and $\sigma_g^2 = \int x^2g(x)dx$. The terms R(K) and σ_K^2 depend only on the known kernel K(.). Consequently, the unknown term in Equations (3.8) and (3.9) is R(f"), which depends on the unknown density f(x). Now one could fit the "best" parametric model for precipitation, e.g., the exponential, and then "knowing" f(x) compute R(f") and thereby evaluate h_{opt}. Silverman [1986] provides h_{opt} using the normal distribution as a reference. We investigated such schemes, and found that bandwidths selected in this manner can be quite sensitive to the choice of the reference distribution. For example, for a Gaussian kernel, the h_{opt} for a Normal parent PDF is 1.33 times the h_{opt}

for an Exponential parent. The need to refer to a parametric model detracts from the utility of this method, but the method is less sensitive to boundary effects while selecting h_{opt}.

From Equation (3.9) observe that knowing the optimal bandwidth h_N for the Normal kernel, the optimal bandwidth h_K for a kernel different from the Normal kernel can be readily evaluated as:

$$h_K = \{(R(K)\sigma_N^2)/(\sigma_K^2 R(N))\}^{1/5} h_N \tag{3.10}$$

where "N" identifies the Normal kernel, and "K" the kernel of interest. Different kernels can thus be made equivalent.

2. Least squares cross validation (LSCV) [see Silverman, 1986, section 3.4]. The optimal bandwidth is solved by the minimization of

$$LSCV(h) = \int f^2 - 2n^{-1} \sum_{i=1}^{n} f_{-i}(x_i) \tag{3.11}$$

where f_{-i} represents a kernel density estimate constructed by dropping the i^{th} observation.

LSCV is prone to undersmoothing where the data exhibits fine structure, and also suffers from a high degree of sampling variability, leading to rather poor MISE convergence rates ($O(n^{-1/10})$ [see Hall and Marron, 1987]. The computational burden and poor convergence rate of this method are discouraging. However, its broad applicability to a wide class of situations renders it popular.

3. Maximum likelihood cross validation (MLCV) [see Silverman, 1986, section 3.4]. The optimal bandwidth is solved by the maximization of a pseudo-likelihood criteria given as:

$$MLCV(h) = n^{-1}\sum_{i=1}^{n} \log(f_{-i}(x_i)) \tag{3.12}$$

MLCV leads to degenerate solutions if the data is long tailed, and also suffers from the same low convergence rate that characterizes LSCV. The degeneracy can be corrected [Schuster, 1985] by excluding a fraction of the right tail data from the MLCV score (not from the density estimate). The subjectivity of the choice of such a cutoff point and the computational burden of the scheme detract from its usage.

4. Direct minimization of estimated MSE/MISE.

"Plug-in" or recursive estimators are methods that use data-driven kernel estimates of f(x) and R(f") (or equivalent measures in the discrete case). Such methods were originally proposed by Woodroofe [1970], and pursued by Scott et al. [1977], Scott and Factor [1981], and Sheather [1983, 1986]. Improvements by Park and Marron [1990] and Sheather and Jones [1991] (hereafter, SJ), among others, have lent stability to these methods and have led to a MISE convergence rate of h_{opt} of the order of $n^{-5/14}$, as well as a reduction in the size of the constants associated with this rate.

A summary of the SJ procedure for the continuous, univariate kernel density estimator follows. They developed a kernel estimate $S(\alpha)$ for R(f") as :

$$S(\alpha) = \{n(n-1)\}^{-1}\alpha^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{iv}((x_j-x_i)/\alpha) \tag{3.13}$$

$$\alpha(h) = 1.357\ \{S(a)/T(b)\}^{1/7}\ h^{5/7} \tag{3.14}$$

$$T(b) = -\{n(n-1)\}^{-1}b^{-7}\sum_{i=1}^{n}\sum_{j=1}^{n} K^{vi}((x_i-x_j)/b) \tag{3.15}$$

$a = 0.92\lambda n^{-1/7}$ and $b = 0.912\lambda n^{-1/9}$

where α is a bandwidth (not equal to h), and $K^{iv}(.)$ is a special kernel for estimating fourth derivative of the density, $K^{vi}(.)$ is a special kernel for estimating the sixth derivative of the density, and λ is the sample interquartile range ($x_{0.75} - x_{0.25}$), T(b) is an estimate of R(f''') and a, b are bandwidths that are evaluated with reference to a Normal distribution for the derivative kernels considered.

Relatively crude estimates (with reference to a known distribution) of the bandwidths used in estimating R(f'') and R(f''') suffice given that the dependence of the MISE expression Equation (3.8) on these expressions is successively weaker (note the exponents). The optimal bandwidth h_{opt} is now evaluated by computing a and b from the data, evaluating S(a) and T(b), and substituting the Equation (3.15) into Equation (3.14), and Equation (3.14) into Equation (3.13). This leads to a nonlinear expression in terms of h, which is solved using the Newton Raphson method. Sheather and Jones specify the normal kernel for K(.) and evaluate the derivative kernels as the appropriate derivatives of this kernel. While this is the most attractive data-based approach that we tested, it does not consider the boundary behavior of the kernel estimator. In the case where the data is positive and heavily concentrated near the origin, the SJ procedure tends to grossly undersmooth relative to the theoretical optimal bandwidth.

Comparative results of various bandwidth selection schemes

The most critical aspect of developing the kernel density estimator is the specification of the bandwidth. A second factor is the need for specialized treatment near x=0 (i.e., the boundary problem). We compare the different methods outlined in sections 2.2 and 2.3 with two synthetic data sets.

First we sample (C1) from a Gaussian mixture(0.5N(-2,1)+0.5N(2,1)), to demonstrate estimability with location mixtures. The second sample (C2), was generated from an Exponential distribution with mean 0.15, to demonstrate the boundary effect. In each case a sample of size 250 was used. Sample statistics and values of the key parameters in each case are summarized in Table 3.3. The corresponding PDFs estimated by selected methods are shown in Figures 3a to 3e.

We consider six estimators for density estimation for the above mentioned data sets. These are (1) (PR-N) parametric reference assuming the underlying probability density function to be $N(0,\widehat{\sigma}^2)$; (2) (PR-M), parametric reference assuming the underlying probability density function to be a Gaussian mixture 0.5N(-2,1)+0.5N(2,1); (3) (PR-E) parametric reference assuming the underlying probability density function to be $Exp(\widehat{\alpha})$; (4) (LSCV) least squares cross validation; (5) (MLCV) maximum likelihood cross validation; (5) (SJ) Sheather and Jones [1991] procedure; and (6) (SJL) Sheather and Jones [1991] procedure applied to log transformed data. Table 3.2 summarizes the bandwidth estimation procedures. In the first three methods the term *parametric reference* means the bandwidth is chosen to be optimal with reference to an assumed underlying parametric distribution. The first five methods, which consider untransformed real space data, also use Silverman's method (discussed in the section titled basic ideas) to specify a local rather than a fixed global bandwidth. Boundary kernels as defined by Müller [1991] were used to adjust the density estimates near the lower boundary ($x \geq 0$), but were not used during bandwidth estimation. The SJL procedure eliminated the boundary problem and provided some local bandwidth adaption, so no local bandwidth adjustment and no boundary kernels were used.

For data set C1 we used methods PR-N, PR-M,LSCV, MLCV, and SJ, while for data set C2 we used PR-E, LSCV, MLCV, SJ, and SJL.

The following observations are apparent from the figures:

1. The parametric reference (PR) procedures work very well as expected when the assumed PMF matches the underlying PDF. However, under misspecification, performance suffers. In case of C1, the bandwidth from the true reference (PR-M) is 1.0, while from using the normal distribution (i.e., misspecification) as the reference (PR-N) the bandwidth is 1.76. This results in gross oversmoothing of the two modes present in C1 (see Figure 3.3a). The parametric reference bandwidth is the best possible estimate of h provided f(x) is known. Of course, one reason we pursue nonparametric estimates of the PDF is lack of knowledge of the underlying model. In this context, PR estimates with the correct f(x) are useful as a benchmark to compare the performance of fully data driven methods.

2. LSCV and MLCV are prone to undersmoothing especially when the data exhibits fine structure (e.g multiple modes) and is long tailed [see Hall and Marron, 1987). Also the cross-validation functions (which are minimized for the bandwidth estimation) have spurious local optima (corresponding to clustering of data at different scales) at small bandwidths [see Hall and Marron, 1987]. Thus, we expect small bandwidths from LSCV and MLCV which leads to an undersmoothed density estimate. This can be seen from Figures 3.3b and 3.3d, where the estimates from LSCV and MLCV are very rough, suggesting that the variance is high.

3. SJ has been shown to have a better mean integrated square error (MISE) convergence rate than cross validation methods [see Sheather and Jones, 1991] and hence should lead to a better estimate. This is borne out in Figures 3.3 a through d, Figure 3.4, and Table 3.3. Note that the SJ optimal bandwidth for C1 is close to the optimal bandwidth based on the Gaussian mixture as reference (PR-M). However, for C2 the SJ optimal bandwidth is much smaller than the optimal bandwidth for the exponential distribution.

Table 3.3. Statistics (Sample size =250 for each) and Methods for Figures 3.3 and 3.4

Data	Method (corresponding to Appendix 2)	Global Bandwidth
C1 (Gaussian mixture)	PR-M	1.00
($\bar{x}$ = 0.00, s = 2.26)	PR-N	1.76
	LSCV	0.48
	MLCV	0.53
	SJ	1.03
C2 (Exponential)	PR-E	0.11
($\bar{x}$ = 0.16, s = 0.18)	LSCV	0.015
	MLCV	0.02
	SJ	0.04
	SJL	0.77 (in log space)

Note:

$\bar{x}$ is sample mean and s is sample standard deviation

The SJL estimator is, (Equation 3.2)

$f_n(p) = \frac{1}{n}\sum_{1}^{n} \frac{1}{hp} K(\frac{\ln(p) - \ln(p_i)}{h})$ with Epanechnikov kernel

The Parametric reference, LSCV, MLCV and SJ all use

$f_n(p) = \sum_{i=1}^{n} \frac{1}{nh_i} K(\frac{x-x_i}{h_i})$ with Epanechnikov kernel and Müller boundary kernels.

Local bandwidths h_i are given by, $h_i = h(f(p_i)/g)^{-1/2}$, where h is global bandwidth, (p_i) is the kernel density estimate at p_i using the global bandwidth h, and g is the geometric mean of $f(p_i)$. These estimators only differ in the procedure used to obtain global bandwidth.

This is due to the fact that the boundary effect is not considered while estimating the SJ bandwidth, which is a problem in case C2 but not in C1. In both cases the SJ bandwidth is superior to those chosen by MLCV and LSCV.

Note that in all these cases, the optimal h is determined without using the boundary kernels, and is perhaps smaller than it would be (to reduce the effect of leakage across the boundary) if boundary kernels were used during bandwidth estimation. This emphasizes the need for proper treatment of the boundary of the domain during all phases of kernel density estimation. We expect to pursue modifications of the SJ estimator to account for boundaries during bandwidth selection.

4. For C2, in Figures 3.3c and 3.3d, we use the Müller boundary kernels (except when using SJL) to reduce the bias at the boundary. Despite this a considerable bias can be observed near the origin in these figures, for each of these estimators. This is a consequence of the high curvature of the target density near the origin, and the "leakage" from the kernels across the boundary at x=0. Figure 3.4 for the case C2 includes a PDF estimated without using boundary kernels (SJ-NBK) along with those from SJ and SJL. The inclusion of boundary kernels in SJ offers only a marginal improvement over SJ in this case, since it still suffers from a bias due to the high curvature of f(x) in this area. SJL, on the other hand, does not suffer as much from this problem and hence performs better.

5. For data sets with a heavy concentration of data near the origin, a log transformation is an attractive choice. We see from Figure 3.4 that the SJL procedure provides a very competitive kernel density estimate in this situation. Note that SJL provides local bandwidth adaptation in real space. For the wet day precipitation data, which is usually modeled using an Exponential, or a Gamma distribution, this may be a natural transformation to consider.

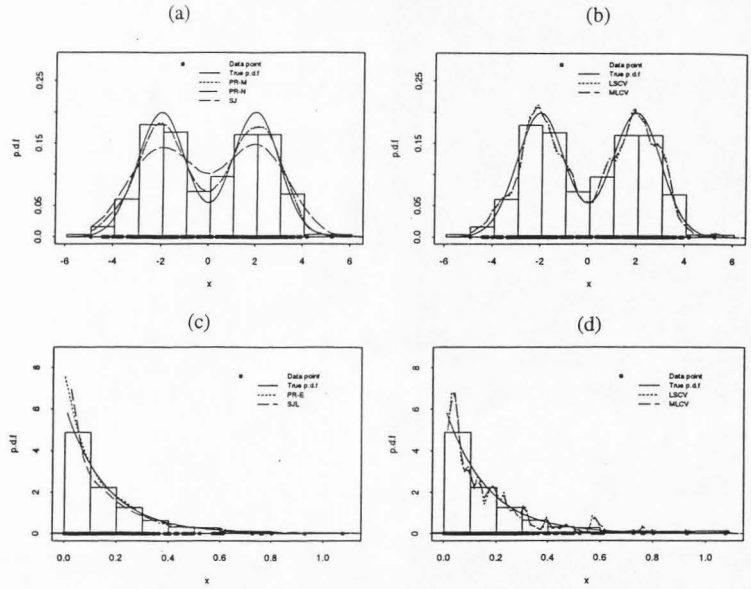


Figure 3.3. Plots of data, histogram of data, true underlying PDFs and PDFs estimated from (a) PR-M (h=1), PR-N (h=1.76), SJ (h=1.03) for the data set C1, (b) LSCV (h=0.48), MLCV (h=0.53) for the data set C1, (c) PR-E (h=0.11), SJ (h=0.04), SJL (h=0.77), for the data set C2, (d) LSCV (h=0.015), MLCV (h=0.02), for the data set C2.

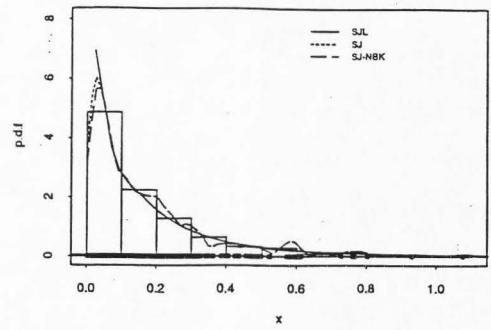


Figure 3.4. Plot of PDFs estimated from SJ, SJL, and SJ-NBK (bandwidth chosen from SJ procedure but boundary kernels are not used). Along with observed data and histogram of observed data, for the data set C2.

Our recommendation of SJL is motivated largely by a desire to deal with the boundary effects and local bandwidth adaptation in a natural way given the nature of the precipitation data. Where boundary effects are not of concern (e.g., C1), a direct application of SJ would be preferred. Once a modification of SJ to account for boundary effects during bandwidth estimation is successful, SJL need not be the method of choice even in this situation.

Kernel Density Estimation for Discrete Random Variables

Wet spell and dry spell lengths are treated as an integer number of days in our rainfall model Lall et al. [1995], consequently estimators for discrete data are reviewed here. The presentation of discrete kernel estimators is new to the hydrologic literature, and includes a new estimation method we developed [Rajagoplan and Lall, in press]. For a discussion of the methods for discrete data refer to Hand [1982], Bishop et al. [1975], and Coomans and Broeckaert [1986].

Basic ideas

The basic concepts of kernel estimation of PDFs in the continuous case introduced earlier hold for the discrete case as well. In the discrete case one can first estimate the sample relative frequencies. These relative frequencies or multinomial cell proportions can then be "smoothed" using a kernel estimator. The problem of nonparametric smoothing of the multinomial cell proportions has not been studied as extensively as nonparametric density estimation, its counterpart in the continuous case. Here we have a sample $y_1,y_2,..,y_n$ for n multinomial trials with possible outcomes $1,2,..L..,L_{max}$ with probabilities of occurrence $f_1,f_2,...,f_{L_{max}}$ that are unknown. Estimates $\hat{f}_n(L)$ for any cell L may be obtained as sample relative frequencies ($\tilde{p}_L = n_L/n$), or by smoothing the $\tilde{p}_L$. Hall and Titterington [1989] note that smoothing can be beneficial when there are many cells with small or zero frequencies, i.e., the data are sparse. This is the case with the wet and dry spell length data.

A kernel estimator $\hat{f}_n(L)$ is given as:

$$\hat{f}_n(L) = \sum_{i=1}^{L_{max}} K_d(L, i,h))\tilde{p}_i \qquad (3.16)$$

where h is the bandwidth, L_{max} is the maximum observed spell length and $K_d(.)$ is a discrete kernel (or weight function).

A nonparametric estimator of the discrete probabilities of the wet or dry spell lengths (w or d) would be the maximum likelihood estimator that yields directly the relative frequencies (e.g., (number of w_i)/n_w, for the i^{th} wet spell length w_i in a sample of size n_w). The kernel method is superior to this approach, because (1) it allows extrapolation of probabilities to spell lengths that were unobserved in the sample, and (2) it has higher MSE efficiency (Hall and Titterington, 1987). Three major estimators identified in literature and a fourth one developed by Rajagopalan and Lall [in press] for smoothing probabilities of discrete data, are described. Their performance with synthetic data sets is compared in the following sections.

Choice of discrete kernel estimators

The estimators considered are (1) the Geometric kernel estimator developed by Wang and Van Ryzin [1981], hereafter WV; (2) maximum penalized likelihood estimator (MPLE) developed by Simonoff [1983]; (3) the estimator by Hall and Titterington [1987], hereafter HT; and (4) the discrete kernel (hereafter DK) estimator developed by Rajagopalan and Lall [in press]. These are summarized in Table 3.4.

1. Wang and Van Ryzin [1981] estimator (WV)

The kernel estimator of the probability mass function (PMF) of a discrete variable L, (here the length of wet or dry spell with n sample values) given by Wang and Van Ryzin [1981] uses Equation (3.16) with the geometric kernel given as:

$$K_d(L,i,h) = 0.5(1-h)h^{|L-i|} \quad \text{if } |L-i| \geq 1 \qquad h \in [0,1]$$

$$= (1-h) \qquad \text{if } L = i \qquad (3.17)$$

The bandwidth h can be global or local.

Wang and Van Ryzin [1981] derive optimal global and local bandwidths to minimize the MSE (mean square error = $E[(f(L)-\hat{f}_n(L))^2]$). They estimate the local bandwidths h(i) by minimizing the approximate MSE of $f_n(i)$, while truncating the geometric kernel at i±2. The resulting expressions are in terms of the unknown true probabilities f(i). They show that substitution of the relative frequencies of i, estimated from the sample as $\tilde{p}_i$ ($\tilde{p}_i = n_i/n$) in the expressions, leads to a strongly consistent procedure. An optimal global bandwidth is obtained by minimizing the average MSE (i.e., $1/n\sum_i MSE(i)$) over the data. Expressions for the optimal global and local bandwidths are given in Table 3.4.

Note that for small values of h, the estimator is close to the naive maximum likelihood estimator (MLE) (i.e., $\tilde{p}_i$), and for $\tilde{p}_i$ small, h is larger, leading to a higher smoothing, or larger "smearing" of the relative frequencies. An improved extrapolation in the tail of the density can result through the use of the local bandwidths.

2. Maximum penalized likelihood estimator (MPLE)

The MPLE was first introduced by Good and Gaskins [1971] for continuous variables, and was later extended to the density estimation for discrete variables by Simonoff [1983]. Simonoff [1983] proposes a solution for the "category" probabilities $\hat{f}_i$ that maximizes a penalty function given by:

LFN = Log likelihood - roughness penalty (3.18)

The idea is to balance the goodness-of-fit of the estimate (i.e., likelihood) with its smoothness (i.e., roughness penalty). The smoothest estimate is obtained if all cell probabilities are equal over the range of cells considered. With this in mind, the penalized likelihood function is defined as:

$$LFN = \sum_{i=1}^{L_{max}} n_i \log(\hat{f}_i) - \beta \sum_{i=1}^{L_{max}} \{\log(\hat{f}_i/\hat{f}_{i+1})\}^2 \quad (3.19)$$

$$\text{where} \sum_{i=1}^{L_{max}} \hat{f}_i = 1 \quad (3.20)$$

$\beta \geq 0$ is a smoothing parameter, and L_{max} is the largest cell considered (or the longest spell considered). The smoothing parameter β controls the relative weight assigned to smoothness and consequently has the same role as the bandwidth used in kernel estimation. Here a data-dependent β is used through the following procedure which minimizes asymptotic mean square error.

1. An initial β is chosen as $0.009N(L_{max})^{0.6}(\log(L_{max}))^{0.4})$, where N is the sample size.

2. Given this β, the penalized likelihood (Equation 3.19) is maximized with respect to $\hat{f}_i$, $i = 1,..,L_{max}$ using the method of Lagrange multipliers.

Table 3.4. Examples of Discrete Kernel Estimators

Wang and VanRyzin [1981] (WV) Geometric Kernel estimator

Geometric kernel $K(x) = \begin{cases} 0.5(1-h)h^{|x-x_i|} & \text{if } |x - x_i| \geq 1 \\ (1-h) & \text{if } x = x_i \end{cases}$ $h \in [0,1]$

Global bandwidth $h = \beta_1\{3/2 + B_1 - B_2 + (n-1)\beta_{10}\}^{-1}$

Local bandwidth $h(i) = d_i\{p_i + \frac{1}{4}E_i + F_i - G_i + (n-1)e_i\}^{-1}$

where,

$$\beta_1 = 1 - \sum_{i}^{n} \tilde{p}_i^2 + \frac{1}{2}B_1,\ B_1 = \sum_{i=1}^{L_{max}} \tilde{p}_i(\tilde{p}_{i-1} + \tilde{p}_{i+1}),\ B_2 = \sum_{i=1}^{L_{max}} \tilde{p}_i(\tilde{p}_{i-2} + \tilde{p}_{i-2}),$$

$$\beta_{10} = \sum_{i=1}^{L_{max}} \tilde{p}_i^2 - B_1 + \frac{1}{4}B_0$$

$$G_i = \tilde{p}_i(\tilde{p}_{i-2} + \tilde{p}_{i+2}),\ F_i = \tilde{p}_i(\tilde{p}_{i-1} + \tilde{p}_{i+1}),\ E_i = (\tilde{p}_{i-1} + \tilde{p}_{i+1}),\ d_i = \tilde{p}_i(1 - \tilde{p}_i) + \frac{1}{2}F_i,$$

$$e_i = (\tilde{p}_i - \frac{1}{2}E_i)^2,\ B_0 = \sum_{i=1}^{L_{max}} (\tilde{p}_{i-1} + \tilde{p}_{i+1})^2$$

where, $\tilde{p}_i$ ($\tilde{p}_i = n_i/n$) are the sample relative frequencies

Maximum Penalized Likelihood Estimator (MPLE) of Simonoff [1983]

$$LFN = \sum_{i=1}^{L_{max}} n_i \log(\hat{f}_i) - \beta \sum_{i=1}^{L_{max}} \{\log(\hat{f}_i/\hat{f}_{i+1})\}^2$$

where $\sum_{i=1}^{L_{max}} \hat{f}_i = 1$, $\beta \geq 0$, is a smoothing parameter, and L_{max} is the largest cell (e.g. longest spell length) considered

The smoothing parameter β controls the relative weight assigned to smoothness and consequently has the same role as the bandwidth used in kernel estimators. The LFN function is minimized to solve for each $\hat{f}_i$s (the required cell probability estimates).

Hall and Titterington [1987] HT estimator

$$W(t) = K(t)/\sum_{j=-\infty}^{\infty} K(j/h) \qquad K(t) \text{ is a continuous r.v. kernel, j is integer}$$

$1/h \in [0,1]$

Discrete (Left) Boundary Kernels, Univariate, Dong and Simonoff, [1994]

Note that $q = (x-1)/h$, $0 \leq q \leq 1$ and x is the point at which the density is estimated

for Epanechnikov $K(q,t) = \frac{-6}{(1+q)^3}t^2 + \frac{3(q^2+1)}{(1+q)^3}$

DK estimator

Note $t = (L-j)/h$, and L is point at which density is estimated

Table 3.4 (contd.)

Interior region (i.e. $L \geq h+1$)

Quadratic kernel $K(t) = at^2 + b$ for $|t| \leq 1$

$$a = \frac{-3h}{(1-4h^2)} \quad \text{and} \quad b = \frac{3h}{(1-4h^2)}$$

Left Boundary (i.e. $1 < L < h+1$)

for Quadratic kernel $K(t) = at^2 + b$ for $|t| \leq 1$

$$a = \frac{-D}{2h(h+L)} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^3(h+L)}\right)} \quad \text{and } b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{(h+L)}$$

where, $C = h(h-1)(2h-1) + (L-2)(i-1)(2L-3)$; $D = -h(h-1) + (L-2)(L-1)$; $E = -(h(h-1))^2 + ((L-2)(L-1))^2$

Left Boundary (i.e. $L = 1$)

for Quadratic kernel $K(t) = at^2 + b$ for $|t| \leq 1$

$$a = \frac{-D}{2h^2} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^4}\right)} \quad \text{and } b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{h}$$

where, $C = h(h-1)(2h-1)$; $D = -h(h-1)$; $E = -(h(h-1))^2$

3. An optimal β is now estimated by minimizing an asymptotic MSE, defined as an asymptotic approximation to $\sum_{i=1}^{L_{max}} (\hat{f}_i - \pi_i)^2$, where π_i is the unknown probability of cell i. Simonoff [1983] develops this asymptotic MSE expression in terms of the sample relative frequencies $\tilde{p}_i$ ($\tilde{p}_i = n_i/n$), β and the unknown probability π_i. For π_i he uses the estimates $\hat{f}_i$ from step 2.

4. Steps 2 and 3 are repeated till convergence is achieved. Simonoff [1983] argues that although a formal proof for the convergence of this procedure is not available, extensive computations have indicated that the scheme does converge. The need to specify L_{max} (in excess of the longest observed spell) detracts from the use of this method. We would prefer a natural extension of the tail of the PMF by the method used, rather than a prior specification of its extent.

3. Hall and Titterington [1987] estimator (HT)

The HT estimator developed by Hall and Titterington [1987] uses a discrete kernel function formed from a continuous kernel as:

$$K_d(L,j,h) = \frac{K((L-j)/h)}{s(h)} \tag{3.21}$$

where $h > 1$ and $s(h) = \sum_{j=L-h}^{j=L+h} K(j/h)$. $K(.)$ is any suitable continuous univariate kernel function, with compact support, satisfying properties in Equation (3.2). The bandwidth h is selected as a minimizer of a least squares cross validation (LSCV) function suggested in Hall and Titterington [1987] over a suitable range for h given as:

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}_n(j))^2 - 2\sum_{j=1}^{L_{max}} \hat{f}_{n,-j}(j)\ \tilde{p}_j \tag{3.22}$$

where $\hat{f}_{n,-j}(j)$ is the estimate of the PMF of spell length j, by dropping all the spells of length j from the data. This method has been shown by Hall and Titterington [1987] to automatically adapt the estimator to an extreme range of sparseness types.

Note that this estimator has the same convolution structure as the kernel density estimator in the continuous case. The HT estimator uses a standard continuous variate kernel function rescaled by the sum of the weights applied to an integer set of points. This estimator is defined over the set of integers. However, wet spell and dry spell lengths are counting numbers (integers greater than 1). To avoid the problem of the estimator assigning probability to integers less than 0 (the boundary problem), Dong and Simonoff [1994] developed boundary kernels for Epanechnikov and Bisquare kernels, which are given in Table 3.4. By HT we refer to the boundary modification of Dong and Simonoff [1994].

For finite samples, some disquieting aspects of the HT estimator become apparent. The noninteger bandwidth leads to an effective kernel that also varies with h in a manner quite different from that prescribed by Equation (3.2). The effective integer support of $K_d(L,j,h)$ in Equation (3.21) is [(L-h*),(L+h*)], where h* is the closest integer greater than or equal to h. HT kernels are defined as quadratics or other polynomials over [L-h,L+h]. Since this is not the effective integer support of the kernel, the effective kernel over the space of integers is not the quadratic defined.

Alternatively, it is possible to develop a kernel that recognizes the data to be in integer space, has an integer bandwidth and satisfies all the required conditions in the integer space. This also obviates the need for normalization of the kernel weights as done in HT. We explored this line of thought and sought a direct, discrete analog of the continuous kernel density estimator, which led to the development of the discrete kernel (DK) estimator [Rajagopalan and Lall, in press].

4. Discrete Kernel Estimator (DK)

Our estimator $\hat{f}n(L)$ uses Equation (3.16) with discrete quadratic kernel (QK) is given as:

$$K_d(L,i,h) = at_i^2 + b \qquad (3.23)$$

here $t_i = \frac{L - i}{h}$. Epanechnikov [1969] showed that the MSE optimal kernel of second order, is the quadratic kernel (QK), also known as the Epanechnikov kernel. Here we need to specify the constants a and b for the interior ($i > h+1$) and the boundary region ($1 \le i \le h+1$). The constants a and b are solved to satisfy: (1) the kernel function goes to zero for $|i-j| \ge h$, i.e $K(t_j) = 0$ for $|t_j| \ge 1$, (2) sum of the weights is unity, i.e.,

$\sum_{j=i-h}^{j=i+h} K(\frac{i-j}{h}) = 1$ and (3) the first moment of the kernel function is zero, i.e $\sum_{j=i-h}^{j=i+h} K(\frac{i-j}{h})t_j = 0$

These three conditions are the discrete versions of the conditions given in Equation (3.2) for continuous variable kernels. One could choose higher order Beta kernels and derive results similar to these that follow for DQ. The resulting kernels for the interior and the boundary are given in Table 3.4. Derivations of these kernels are presented in Rajagopalan and Lall [in press].

Note that the kernel and hence the estimator $\hat{f}_n(L)$ are expressed strictly in terms of the bandwidth h. An optimal choice of h then completes the definition of the estimator. The bandwidth is selected by minimizing the least squared cross validation function given as:

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}_n(j))^2 - 2\sum_{j=1}^{L_{max}} \hat{f}_{n,-j}(j)\ \tilde{p}_j \qquad (3.24)$$

where $\hat{f}_{n,-j}(j)$ is same as defined in earlier. Hall and Titterington [1987] also show that cross-validation automatically adapts the estimator to an extreme range of sparseness types. If the multinomial is only slightly sparse, cross validation will produce an estimator which is virtually the same as the cell-proportion estimator. As sparseness increases, cross validation will automatically supply more and more smoothing, to a degree which is asymptotically optimal.

Comparative results of various discrete kernel estimators

The four methods (WV, MPLE, HT, and DK) are compared with two synthetic data sets generated from long-tailed distributions (e.g., Geometric distribution). First we use a sample (D1) from a geometric distribution with π=0.2. The second sample (D2) was generated from a mixture of two Geometric distributions defined as (0.3G(π=0.9)+

$0.7G(\pi=0.2)$). In each case a sample of size 250 was used. We also fitted a geometric distribution (GP) to D1 and D2 using the method of moments. Sample statistics and values of the key parameters in each case are summarized in Table 3.5. The corresponding probabilities estimated by each method for D1 and D2 are shown in Figures 3.5 and 3.6.

1. The WV procedure does not smooth the sample proportions ($\tilde{p}_i$) properly. In most cases, there is very little smoothing. In cases where there is some smoothing (e.g., Figure 3.5a, in the range x=4 to 6), the resulting estimate is rather unsatisfactory, and is inconsistent with the underlying population. We feel that part of this behavior is due to the rapid "drop off" of weight associated with the Geometric kernel, and part due to the method used for selecting the bandwidth h.

2. On the other hand, since the roughness penalty tries to make the PMF uniform, MPLE emphasizes smoothness.

Table 3.5. Statistics (Sample size =250 for each) and Methods for Figures 3.5 and 3.6

Figure	Data	Estimator	Kernel Used	Method of Bandwidth Selection
3.4a	D1 ($\bar{x}$ = 5.11, s = 4.19)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
3.4b	D1	HT	Epanechnikov kernel	LSCV
		DK	Quadratic kernel	LSCV
3.4c	D2 ($\bar{x}$ = 3.92, s = 4.02)	WV	Geometric kernel	MSE
		MPLE	- - -	- - -
3.4d	D2	HT	Epanechnikov kernel	LSCV
		DK	Quadratic kernel	LSCV

Note: $\bar{x}$ is sample mean and s is sample standard deviation
Quadratic kernel is the discrete equivalent of the Epanechnikov kernel.

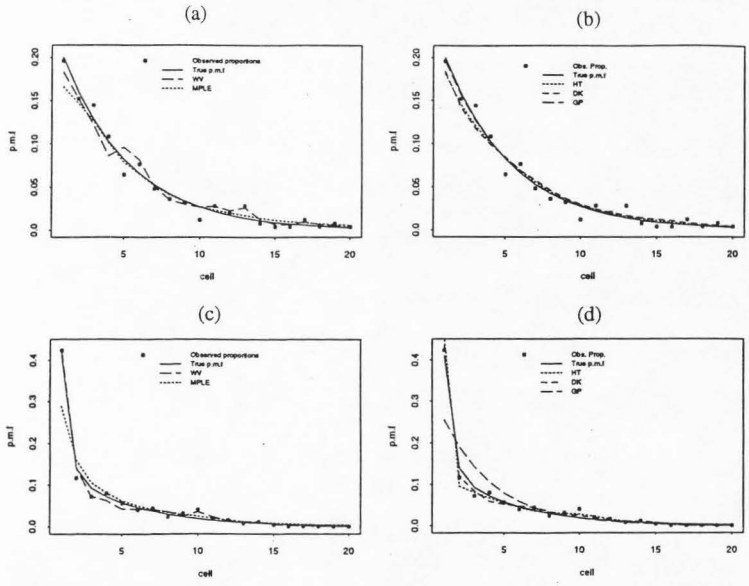


Figure 3.5. Plots of data, observed proportions, true underlying PMFs and PMFs estimated from (a) WV (h = 0.43), MPLE (β=30.25), for the data set D1, (b) HT (h=5), DK (h=6), GP (p=0.1956), for the data set D1, (c) WV (h=0.08), MPLE (β=28.25), for the data set D2, (d) HT (h=3), DK (h=2), GP (p=0.2554), for the data set D2.

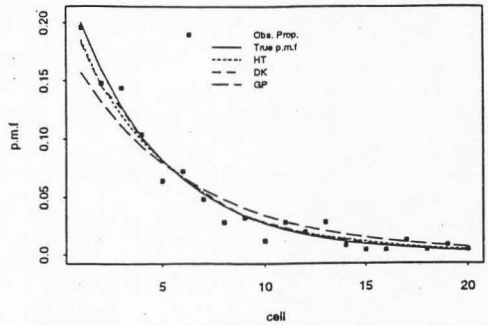


Figure 3.6. Plot showing the effect of outliers on fitted Geometric distribution (GP), HT and DK estimate. Outliers at 45, 50, 75, 100 in the data set D1.

Consequently, when the true PMF has a high second derivative (e.g., near the origin), MPLE has difficulty distinguishing between "true" curvature and observed variation. The resulting estimate often has a strong downward bias near the origin (Figure 3.5a). The MPLE is also sensitive to the value specified for L_{max}, the longest spell length considered. As L_{max} is increased, the downward bias at the origin is increased and the entire PMF is "flattened."

3. The GP fit is very good (estimated π=0.1956) for D1 where the true distribution was geometric. As expected, a large bias is incurred near the origin for D2 (see Figures 3.5b and 3.5d), where the estimated π was 0.2554.

4. Figures 3.5b and 3.5d indicate that HT and DK perform comparably and are the best among the estimators considered. As both these estimators are quite similar in construction this is expected. The estimated PMF is smooth, and it also exhibits the least pointwise bias. The HT and DK estimators automatically adapts to a large range of density variation, providing optimal smoothness in finite samples. Unlike parametric fits, the HT and DK estimates are robust to certain kinds of outliers, as shown in Figure 3.6. Outliers were added at 45, 50, 75, and 100. These could be generated if the data were contaminated by a few large values (e.g., from a Geometric distribution with π=0.01). The fitted Geometric distribution, i.e., (GP) is very much affected by the outliers and deviates from the true distribution, especially near the mode (i.e., 1). The HT and DK estimators still follow the data closely.

It is apparent from the figures that the HT and DK estimators perform the best. Rajagopalan and Lall [in press] found in their Monte Carlo comparisons of HT and DK that they gave comparable results with better approximation of the tail and the modes by DK. DK was also computationally faster, and had a lower variance of optimal bandwidth selection that HT. Consequently it is recommended.

Summary and Conclusions

Issues in estimating parameters for continuous and discrete kernel density estimators were discussed and recommended procedures were developed through examples.

In summary, we recommend using the SJL procedure for estimating the PDF of wet day precipitation amount. This entails the use of a Epanechnikov (or quadratic kernel)

with log transformed precipitation data with bandwidth chosen in log space using the Sheather and Jones [1991] recursive procedure. The resulting density estimate is then transformed to real space. Generally this may be the method of choice for data sets that exhibit a high density near the origin. For discrete data such as spell lengths, we recommend the DK procedure with discrete quadratic kernels in the interior and boundary regions and bandwidth chosen by least squared cross validation.

We found that where the parametric procedure was appropriate, the nonparametric procedure worked nearly as well. Where the parametric model was inappropriate, the nonparametric kernel density estimators were superior. Given that the nonparametric procedures are robust and reproduce different parametric alternatives without prior assumptions, they offer a very general procedure for uniform application across a variety of sites and processes.

Problems with kernel density estimates are high relative bias and variance in the tail of the density if local adaption of the bandwidth is not used. Ability to extrapolate is limited to one bandwidth of the maximum observed value. Where a local bandwidth is used, the local bandwidth at the extreme point of observation is usually quite large and this problem is ameliorated.

The nonparametric modeling framework provides a promising alternative to the parametric approach. The assumption free, data adaptiveness and robust nature of the nonparametric estimators makes the model attractive in a broad class of situations.

References

Abramson, I.S., On bandwidth variation in kernel estimates-a square root law, *The Annals of Statistics*, 10(4),1217-1223, 1982.

Bishop, Y.M., S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.

Coomans. D., and I. Broeckaert, *Potential Pattern Recognition in Chemical and Medical Decision Making*, John Wiley , New York, 1986.

Devroye, L., and L. *Györfi, Nonparametric Density Estimation: The L1 View*, John Wiley, New York, 1985.

Dong J., and J. Simonoff, The construction and properties of boundary kernels for sparse multinomials, *Journal of Computational and Graphical Statistics*, 3, 1-10, 1994.

Epanechnikov, V.A., Nonparametric estimation of a multidimensional probability density, *Theory of Probability and Applications*, 14, 153-158,1969.

Good, I.J., and R.A. Gaskins, Nonparametric roughness penalties for probability densities, *Biometrika*, 58, 255-277, 1971.

Hall, P., and J.S. Marron, Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation, *Probability Theory Related Fields*, 74, 567-581, 1987.

Hall, P., and D.M. Titterington, On smoothing sparse multinomial data, *Australian Journal of Statistics*, 29(1), 19-37, 1987.

Hand D.J., *Kernel discriminant analysis*, Pattern recognition and image processing research studies series, vol. 2, series Ed. J. Kittler, Research Studies Press: Chichester, UK, 1982.

Härdle, W., *Smoothing techniques with implementation in S*, Springer Verlag, New York, 1991.

Lall,U., Nonparametric function estimation: Recent hydrologic applications, *US National report, 1991-1994, International Union of Geodesy and Geophysics*, 1994.

Lall, U., B. Rajagopalan, and D.G. Tarboton, A nonparametric wet/dry spell model for resampling daily precipitation, Working Paper WP-95-HWR-UL/006, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Müller, H.G., *Nonparametric Regression Analysis of Longitudinal Data*, Springer-Verlag, New York, 1988.

Müller, H.G., Smooth optimum kernel estimators near endpoints, *Biometrika.*, 78(3), 521-530, 1992.

Park, B.U., and J.S. Marron, Comparison of data-driven bandwidth selectors, *Journal of American Statistical Association*, 85, 66-72, 1990.

Rajagopalan, B., and U. Lall, A kernel estimator for discrete distributions, *Journal of Nonparametric Statistics*, (in press).

Rosenblatt, M., Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27, 832-837, 1956.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in Probability and Mathematical Statistics, John Wiley, New York, 1992.

Scott, D.W., and L.E. Factor, Monte carlo study of three data-based nonparametric density estimators, *Journal of American Statistical Association*, 76, 9-15, 1981.

Scott, D.W., R.A. Tapia, and J.R. Thompson, Kernel density estimation revisited, *Nonlinear Analysis*, 1, 339-372, 1977

Sheather, S.J., A data-based algorithm for choosing the window width when estimating the density at a point, *Computational Statististics and Data Analysis,* 1, 229-238, 1983.

Sheather, S.J., An improved data-based algorithm for choosing the window width when estimating the density at a point, *Computational Statististics and Data Analysis,* 4, 61-65, 1986.

Sheather, S.J., and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society*, B. 53, 683-690, 1991.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

Simonoff, J., A penalty function approach to smoothing large sparse contingency tables, *Annals of Statistics*, 11, 208-218, 1983.

Wang, M. C., and J. Van Ryzin, A class of smooth estimators for discrete distributions, *Biometrika*, 68(1), 301-9, 1981.

Woodroofe, M. On choosing a delta-sequence, *Annals of Mathematical Statistics* 41, 1665-1671, 1970.

CHAPTER IV

A KERNEL ESTIMATOR FOR DISCRETE DISTRIBUTIONS[1]

Abstract

We present a discrete kernel estimator appropriate for estimating probability mass functions (PMFs) for integer data. Discrete kernel functions analogous to the Beta functions used as kernels in the continuous case are derived for the interior and for the boundary of the domain. An integer bandwidth is considered. Cross validation is used for bandwidth selection. The estimator was motivated by the need to characterize processes (e.g., mixtures of Geometric distributions) with long tailed distributions with high mass near the origin, and integer arguments of the random variable. Monte Carlo comparisons with the Hall and Titterington [1989] (HT) estimator are offered. An application for estimating the PMFs of wet and dry spell lengths for a nonparametric renewal model of daily rainfall is also presented. Other possible methods for obtaining discrete weight sequences are also presented.

Background

The problem of nonparametric smoothing of the empirical discrete PMF (or multinomial cell proportions) has been of interest in recent years. However, it has not been studied as intensively as nonparametric density estimation, its counterpart in the continuous case. Hall and Titterington [1989] mention that smoothing can be beneficial when there are many cells with small or zero frequencies, i.e., the data are sparse. Here we consider that we have a sample $x_1,...,x_n$ for n multinomial trials with possible outcomes $1,2,..,k_{max} \in V$ with probabilities of occurrence $p_1,...,p_{kmax}$ that are unknown. Estimates $\hat{p}_i$ of the

[1] Coauthored by Rajagopalan Balaji and Upmanu Lall.

probabilities p_i may be obtained as sample relative frequencies ($\tilde{p}_i = n_i/n$) or cell proportions, or by smoothing the $\tilde{p}_i$. In the latter case we presume that V is an ordered set and that "distance" between its members is definable through a standard Lebesgue measure. We consider cases where the set V may be bounded or unbounded, and focus on developing an appropriate smoother for the sample relative frequencies that properly deals with the discrete nature of the process.

Our practical interest lay in developing a discrete, nonparametric PMF for data on the length (in days) of dry or wet spells of rainfall. The shortest spell considered is 1 day. In general, the longest possible spell is not known a priori. Data suggests long right-tailed distributions for dry spell length that may correspond to a mixture of Geometric PMFs [see Rajagopalan et al., 1993].

The concept of smoothing in the context of multinomial cell probability estimation was introduced by Good [1965; 1967]. This was later studied and improved by Fienberg and Holland [1973], Stone [1974], Titterington [1980], Titterington [1976], Aitchison and Aitken [1976], and Titterington and Bowman [1985], among others. Bishop et al. [1975] show that these estimators are often better than the cell proportion estimate under squared error loss. Hall and Titterington [1989] argue that $\tilde{p}_i$ may not be consistent in data sparse situations. The smoothing estimators developed by Wang and Van Ryzin [1981], Simonoff [1983], and Hall and Titterington [1989] formed a starting point for our work.

The general form of smoothing estimators in this context is given by

$$\hat{p}_i = \sum_{j=-\infty}^{j=\infty} K(i,j,h)\, \tilde{p}_j \qquad i,j \in I, \text{ the set of integers} \tag{4.1}$$

K(i,j,h) is a weight function or kernel, $\tilde{p}_j$ is the relative frequency of cell j, and h is called the bandwidth or window width.

Wang and Van Ryzin [1981] developed a class of estimators of the form (4.1), using a Geometric kernel (WV) ($K(i,j,h) = 0.5h(1-h)^{|i-j|}$ if $|i-j| \geq 1$; $K(i,j,h) = (1-h)$ if $i=j$ and $h \in [0,1]$). The "drop off" of weights associated with the Geometric kernel is rapid. Wang and Van Ryzin [1981] estimate h under an approximate (MSE) criterion formed by truncating the Geometric kernel beyond two cells. As a result, very little smoothing is obtained in most cases and not much may be gained for sparse data.

By imposing a smoothness constraint on the cell probabilities, Simonoff [1983] obtained relative consistency results for an estimator based on a maximum penalized likelihood criterion (MPLE). In this approach, the estimates $\hat{p}_i$ are solved by maximizing a penalized likelihood function defined as:

$$L = \sum_{i=1}^{k_u} n_i \log(\hat{p}_i) - \beta \sum_{i=1}^{k_u} \{\log(\hat{p}_i/\hat{p}_{i+1})\}^2$$

such that

$$\sum_{i=1}^{k_u} \hat{p}_i = 1 \qquad (4.2)$$

$\beta \geq 0$, is a smoothing parameter, and $V : [1,k_u]$

The estimates from MPLE depend significantly on the extent of estimation required (i.e., k_u) beyond the maximum observed cell (i.e., k_{max}). This is of concern, because we would prefer a natural extension of the tail of the PMF by the method used, rather than a prior specification of its extent.

The estimator developed by Hall and Titterington [1989] (hereafter referred to as HT) is given as:

$$\hat{p}_i = \sum_{j=-\infty}^{j=\infty} W(i,j,h)\ \tilde{p}_j \qquad (4.3)$$

where $W(i,j,h) = \dfrac{K((i-j)/h)}{s(h)}$, $h > 1$ and $s(h) = \sum_{j=-\infty}^{j=\infty} K(j/h)$. K(.) is any suitable continuous univariate kernel function, with compact support satisfying the conditions of positivity, integration to unity, symmetry, and finite variance, which are:

$$\text{(a) } K(u) > 0;\ \text{(b) } \int K(u)du = 1;\ \text{(c) } \int uK(u)du = 0;\ \text{(d) } \int u^2K(u)du = k2 \neq 0 \qquad (4.4)$$

where $(u = (i-j)/h)$, and s(h) is a multiplicative factor required to normalize the continuous variable kernel function for use with discrete data, such that the desired conditions on W(.) viz., $\sum_{j=-\infty}^{j=\infty} W(i,j,h) = 1$ and $\sum_{j=-\infty}^{j=\infty} j\,W(i,j,h) = 0$ are satisfied. Hall and Titterington [1989] proposed a cross-validatory procedure for selecting h. This was later studied by Dong and Simonoff [1994] who extended this estimator to boundary kernels.

It is well known that kernel estimators suffer from increased bias in the boundary region (i.e., $1 \leq i \leq h+1$ in our situation of interest). For the estimates of cells in the boundary there is a lack of full complement of observations on either side of the cell of estimate. As a result, the desired conditions on W(i,j,h) mentioned above will not be preserved. To correct this, special boundary kernels that satisfy the required conditions are used [see Müller, 1991]. Müller [1991] formally developed special boundary kernels in the continuous case. Dong and Simonoff [1994] developed special boundary kernels in the continuous case. Dong and Simonoff [1994] developed boundary kernels (condition 4.4 [a] is relaxed) that could be used in the HT estimator for the discrete case. We refer to the HT estimator with the boundary modification of Dong and Simonoff [1984] as HT/DS.

We performed comparisons of these three estimators (viz., WV, MPLE, and HT/DS) on data generated from long tailed distributions [see Rajagopalan et al., 1993] and found HT/DS to be the best. Hence, we compare the relative performance of the estimator we develop later in this paper with HT/DS.

For finite samples, some disquieting aspects of the HT estimator become apparent. The noninteger bandwidth leads to an effective kernel that also varies with h in a manner quite different from that prescribed by (4.4). The effective integer support of W(i,j,h) is $[(i-h^*),(i+h^*)]$, where h* is the closest integer less than or equal to h. HT/DS kernels are defined as quadratics or other polynomials over [i-h,i+h].

Alternatively, it is possible to develop a kernel that recognizes the data to be in integer space, has an integer bandwidth, and satisfies all the required conditions in the integer space. This also obviates the need for normalization of the kernel weights as done in HT/DS. We explored this line of thought and sought a direct, discrete analog of the continuous kernel density estimator.

The estimator is first presented. Bandwidth estimation is described next. Monte Carlo comparisons with HT/DS are then presented. Comparisons with real data sets follow. Discussion of the new estimator and other possible discrete estimators conclude the chapter.

The Discrete Kernel Estimator (DKE)

We define our estimator $\hat{p}_i$ for cell i through a weighted linear combination of the sample relative frequencies, $\tilde{p}_i$, as:

$$\hat{p}_i = \sum_{j=1}^{k_{max}} K(\iota_j)\ \tilde{p}_j \qquad (4.5)$$

where i,j and h are positive integers, $t_j = (i-j)/h$, K(t) is a kernel function, and V : [1,∞]. In the continuous case, Epanechnikov [1969] showed that the MSE optimal kernel of second order is the quadratic kernel (QK), also known as the Epanechnikov kernel. The general form of the QK is:

$$K(u) = au^2 + b \quad \text{for } |u| \leq 1 \tag{4.6}$$

In the continuous case, a=-0.75, b = 0.75. Scott [1992, p. 140, Equation 6.25] points out that this corresponds to a Beta density function, defined for t ∈ [-1,1]. Other members of this class can be used if additional smoothness is desired.

Here, we chose a discrete quadratic (DQ) kernel of the form $K(t_j) = at_j^2 + b$, where $t_j = (i-j)/h$. The main focus then is to specify the constants a and b for the interior (i > h+1) and the boundary region (1 ≤ i ≤ h+1). The constants a and b are solved to satisfy : (A) the kernel function goes to zero for |i-j| ≥h, i.e., $K(t_j) = 0$ for $|t_j| \geq 1$, (B) sum of the weights is unity, i.e., $\sum_{j=i-h}^{j=i+h} K(\frac{i-j}{h}) = 1$ and (C) the first moment of the kernel function is zero, i.e., $\sum_{j=i-h}^{j=i+h} K(\frac{i-j}{h})t_j = 0$. Note that the above conditions are the discrete versions of the conditions are the discrete versions of the conditions given in Equation (4.3) for continuous variable kernels. One could choose higher order Beta kernels and derive results similar to these that follow for DQ.

For the interior region (i > h+1) using conditions (A) and (B) gives Equations (4.7) and (4.8) as:

$$K(t_{i+h}) = K(t_{i-h}) = 0 \tag{4.7}$$

$$\sum_{j=i-h}^{j=i+h} (at_j^2 + b) \quad = \quad 1, \quad \text{where } t_j = (i-j)/h \tag{4.8}$$

Condition (C) is satisfied if a=-b. The coefficients a and b can now be expressed in terms of the bandwidth h as:

$$a = \frac{-3h}{(1-4h^2)} \quad \text{and} \quad b = \frac{3h}{(1-4h^2)} \tag{4.9}$$

For the boundary region ($1<i\leq h+1$) condition A is modified as:

$$K(t) = 0 \text{ for } t \leq -1 \text{ and } t \geq q \text{ where } q = (i-1)/h. \tag{4.10}$$

Applying conditions (B) and (C), we get Equations (4.11) and (4.12) as:

$$\sum_{j=1}^{j=i+h} (at_j^2 + b) \quad = \quad 1 \tag{4.11}$$

$$\sum_{j=1}^{j=i+h} t_j(at_j^2 + b) \quad = \quad 0 \tag{4.12}$$

Solving for a and b we get:

$$a = \frac{-D}{2h(h+i)} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^3(h+i)}\right)}, \quad b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{(h+i)} \tag{4.13}$$

where

$C = h(h+1)(2h+1) + (i-2)(i-1)(2i-3)$; $D = -h(h+1) + (i-2)(i-1)$; $E = -(h(h+1))^2 + ((i-2)(i-1))^2$

From Equation (4.10) it can be seen that at the boundary (i.e., i = 1) the weight associated with the kernel is zero. This is not desirable because, for longtailed distributions defined on the interval $[1,\infty)$, most of the mass is concentrated right at i=1. Clearly, using the boundary modification in Equation (4.13) for estimation of PMF at the boundary (i.e., i=1) will introduce a large bias in the estimate. Therefore, we need a further modification for estimation at i=1. By not enforcing the $K(t) = 0$ at i = 1, we modify (A) to be:

$$K(t) = 0 \text{ for } t \leq -1 \tag{4.14}$$

while Equation (4.11) and (4.12) remain the same. Solving Equations 4.14, 4.11 and 4.12 for a and b we get:

$$a = \frac{-D}{2h^2} \times \frac{1}{\left(\frac{E}{4h^3} - \frac{CD}{12h^4}\right)}, \quad b = \left[1 - \frac{aC}{6h^2}\right]\frac{1}{h} \tag{4.15}$$

where

$C = h(h-1)(2h-1)$; $D = -h(h-1)$; $E = -(h(h-1))^2$

From Equations (4.9), (4.13) and (4.15), note that the kernels and hence the estimator $\hat{p}_i$ are expressed strictly in terms of the bandwidth h. An optimal choice of h then completes the definition of the estimator.

Three criteria often used for bandwidth estimation are (1) direct minimization of average mean square error (MSE), (2) maximum likelihood cross validation (MLCV), and

(3) least squares cross validation (LSCV). These could be optimized over a discrete set of h values.

We tested all the three methods and found LSCV to be the best. Hall and Titterington [1989] and Dong and Simonoff [1994] also argue in favor of LSCV. The bandwidth is selected by minimizing the LSCV function given as:

$$\text{LSCV}(h) = \sum_{i=1}^{k_{max}} (\hat{p}_i)^2 \; - \frac{2}{n} \sum_{i=1}^{k_{max}} \hat{p}_{-i} \; n_i \qquad (4.16)$$

where $\hat{p}_{-i}$ is the estimate of the i^{th} cell, by dropping the i^{th} cell and n. In a related context, Hall and Titterington [1989] also show that cross validation automatically adapts the estimator to an extreme range of sparseness types. If the multinomial is only slightly sparse, cross validation will produce an estimator which is virtually the same as the cell-proportion estimator. As sparseness increases, cross validation will automatically supply more and more smoothing, to a degree which is asymptotically optimal.

An example application comparing DKE (with DQ kernel) to HT/DS with QK-based kernels for four data sets is shown in Figures 4.1, 4.2, 4.3, and 4.4. The data in Figure 4.1a were sampled from a Geometric distribution (G1) defined as $G(\pi=0.2)$. The data in Figure 4.1b was sampled from a mixture of two Geometric distributions (G2) defined as $(0.3G(\pi=0.9) + 0.7G(\pi=0.2))$. The sample sizes for G1 and G2 are 250. Figure 4.1c shows the PMF estimates estimated for the mines data of sample size 55, analysed by Dong and Simonoff [1994]. Figure 4.1d shows the estimated PMF from both estimators of dry spell length data, for season 3 (i.e., Jul - Sep) for the Woodruff station in Utah. The sample size in this case was 539. All four figures indicate that both DKE and HT/DS perform comparably. Because both the estimators are similar, this is expected. Through Monte Carlo simulations we investigate the behavior of these estimates for

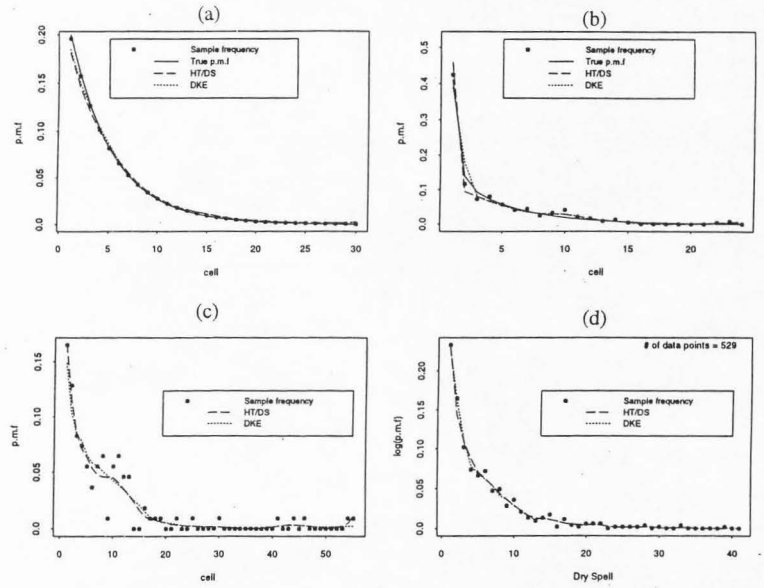


Figure 4.1. True PMF, estimated PMF from HT/DS and DKE, along with the sample frequency (a) of a sample of size 250, generated from Geometric (π=0.2), (b) of a sample of size 250, generated from 0.7* Geometric (π=0.2)+0.3* Geometric (π=0.9), (c) of mines data, of sample size 55, and (d) of dry spell length data at Woodruff, UT, of sample size 529.

selected situations. The behavior of the weight sequence from both the estimators is also probed. The results are discussed in the following section.

Monte Carlo Comparisons

We present results from Monte Carlo simulations, comparing our estimator with the HT/DS estimator using QK. Data sets were generated from situations that may be of interest in our particular context (e.g., Geometric distribution, with a considerable boundary region). We generated 500 realizations from the two populations G1 and G2. Sample sizes chosen were n = 50, 100, 200, 300, 500.

The statistical measures computed to assess the relative performance of DKE and HT/DS estimators are:

1. Average sum of squared errors (ASSE) $\left(\sum_{j=1}^{j=nsim} \left(\sum_{i=1}^{i=k_u} (\hat{p}_{ij} - p_i)^2 \right) / nsim \right)$

across all realizations for each sample size.

2. Sum of squared error (SSE_j) $\left(\sum_{i=1}^{i=k_u} (\hat{p}_{ij} - p_i)^2 \right)$ for each realization j = 1,..,nsim

3. Average sum of absolute error (ASAE) $\left(\sum_{j=1}^{j=nsim} \left(\sum_{i=1}^{i=k_u} abs(\hat{p}_{ij} - p_i) \right) / nsim \right)$

across all realizations for each sample size.

4. Cell root mean square error (CRMSE) $\{ \sum_{j=1}^{j=nsim} ((\hat{p}_{ij} - p_i)^2) / nsim \}^{0.5}$ across all realizations for each sample size and for each cell $i = 1,..,k_u$

5. Fractional cell root mean square error : $FCRMSE_i = CRMSE_i/p_i$

6. Average cell bias $(CBIAS_i)$ $\sum_{j=1}^{j=nsim} ((\hat{p}_{ij} - p_i) / nsim)$ across all realizations for each sample size and for each each cell $i = 1,..,k_u$

7. Fractional cell bias: $FCBIAS_i = CBIAS_i/p_i$

8. Coefficient of variation of bandwidth $C_v = s/\bar{h}$ for each sample size, where s and $\bar{h}$ are the standard deviation and mean of the bandwidths obtained for all the nsim realizations.

Note that we chose k_u to be 30 in this case, and p_i's are the true PMFs obtained from the known underlying distributions from which the samples were generated, nsim is the number of simulations, in our case 500.

Table 4.1 shows the ASSE and ASAE for the two estimators for the two populations G1 and G2 considered. It can be observed from Table 4.1 and Figures 4.2a and 4.2b that the performance of the two estimators over these two measures is quite close. Figures 4.2a and 4.2b indicate that the ASSE appears to decrease with n at rates -1.03 and -0.86 for HT/DS and -0.85 and -0.9 for DKE, for G1 and G2, respectively. These rates are very similar, and are close to the rate n^{-1} as anticipated in Hall and Titterington's [1989] Theorem 2.1. However, the SSE for HT/DS has a larger spread that DKE as can be seen from Figures 4.3a and 4.3b for G1 and G2, respectively, for a sample size of 50. The results were generally similar for other sample sizes.

As mentioned earlier we are interested in the behavior of these estimators at the boundary (left boundary) and in the tails. To assess this, $CRMSE_i$ and $FCRMSE_i$ for different sample sizes n were estimated. As an illustration we present the estimates of $FCRMSE_i$ for sample sizes 50 and 500 for G1 in Figures 4.4a and 4.4b, respectively. Figures 4.5a and 4.5b are corresponding figures for G2. These figures suggest that DKE performs better than HT/DS in the tail region for all sample sizes, more so for smaller sample sizes. The results for other sample sizes were intermediate.

From Figures 4.6a and 4.6b we see that part of the poorer performance of HT/DS in the tails is due to higher bias.

The MSE expression of the estimate $\hat{p}_i$ as given by Wang and Van Ryzin [1981] is

Table 4.1. Comparison of ASSE and ASAE

	ASSE			ASAE		
	DKE	PAR	HT/DS	DKE	PAR	HT/DS
Samples generated from G1 (Geometric (π=0.2))						
n = 50	0.0058	0.0008	0.0084	0.2032	0.0816	0.2737
n = 100	0.0032	0.0006	0.0038	0.1558	0.0599	0.1814
n = 200	0.0019	0.0003	0.0019	0.1183	0.4250	0.1264
n = 300	0.0013	0.0002	0.0012	0.1000	0.0323	0.0987
n = 500	0.0008	0.0000	0.0008	0.0780	0.0226	0.0797
Samples generated from G2 (0.7* Geometric (π=0.2)+0.3* Geometric (π=0.9))						
n = 50	0.0080	—	0.0081	0.2300	—	0.2481
n = 100	0.0039	—	0.0038	0.1676	—	0.1638
n = 200	0.0021	—	0.0022	0.1261	—	0.1194
n = 300	0.0016	—	0.0016	0.1071	—	0.0978
n = 500	0.0010	—	0.0011	0.0855	—	0.0785

Note:

PAR is the fitted parametric (in this case the fitted Geometric distribution)

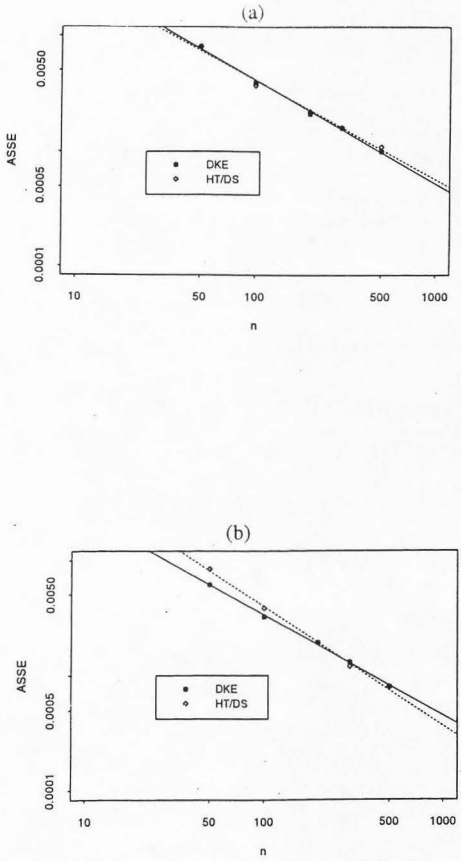


Figure 4.2 Log-log plot of ASSE with sample size n, along with the fitted lines (a) of samples generated from Geometric (π=0.2), and (b) of samples generated 0.7*Geometric (π=0.2) + 0.3*Geometric(π=0.9).

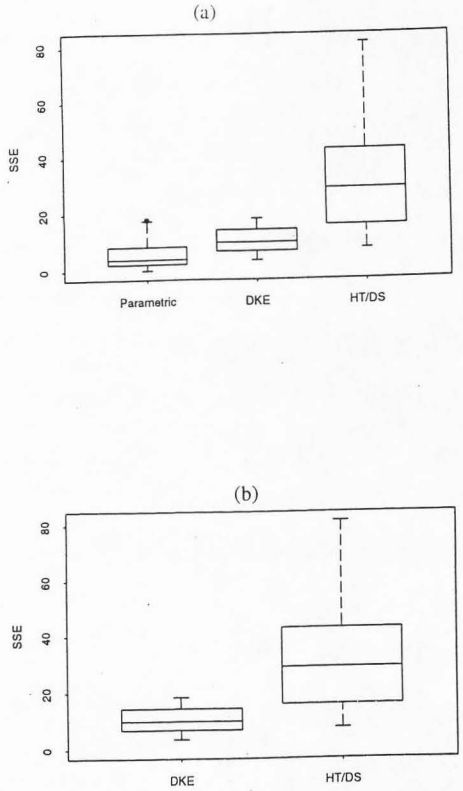


Figure 4.3 Boxplots of SSE_j (a) HT/DS, DKE and fitted Parametric distribution, of samples generated from Geometric (π=0.2) of sample size 50, and (b) HT/DS and DKE of samples generated from 0.7*Geometric (π=0.2)+0.3* Geometric (π=0.9) of sample size 50.

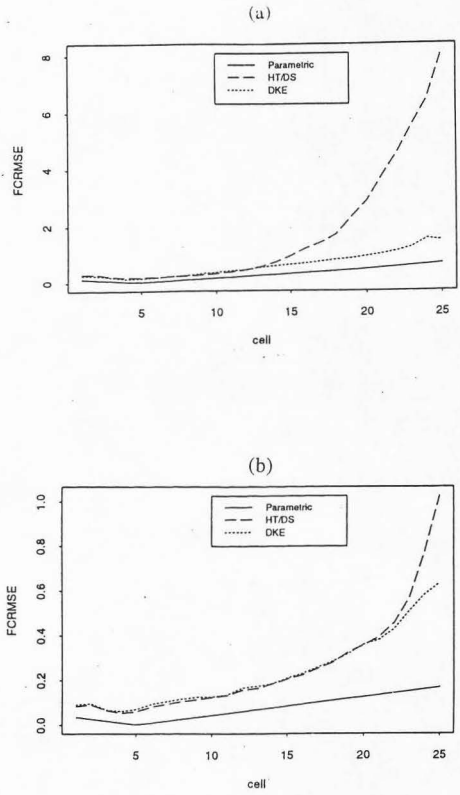


Figure 4.4 $FCRMSE_i$ from HT/DS, DKE and fitted Parametric distribution, of samples generated from Geometric (π=0.2) (a) of sample size 50, and (b) of sample size 500.

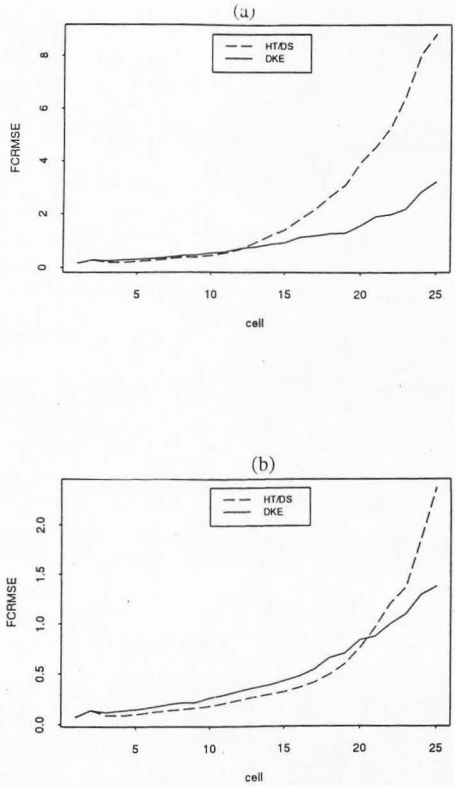


Figure 4.5 $FCRMSE_i$ from HT/DS and DKE, of samples generated from 0.7*Geometric (π=0.2)+0.3* Geometric (π=0.9) (a) of sample size 50, and (b) of sample size 500.

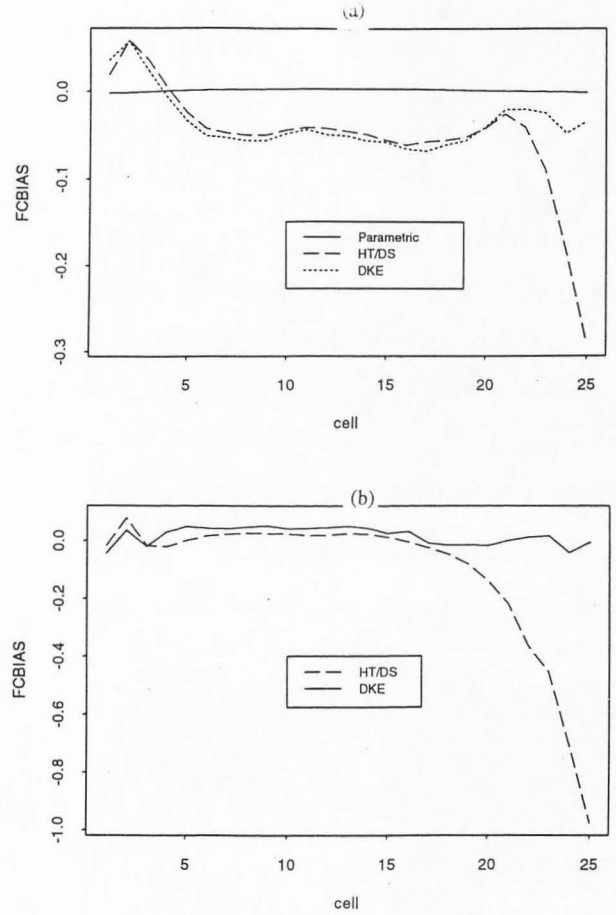


Figure 4.6 $FCBIAS_i$ from HT/DS and DKE, of samples of size 500 (a) generated from Geometric (π=0.2), and (b) generated from 0.7*Geometric (π=0.2)+0.3* Geometric (π=0.9).

$$E[\sum_{i=1}^{k_{max}} \{ \hat{p}_i - p_i \}^2] = \sum_{i=1}^{k_{max}} \sum_{j=1}^{k_{max}} W^2(i,j,h)p_j/n - \sum_{i=1}^{k_{max}} \{ \sum_{j=1}^{k_{max}} W(i,j,h)p_j \}^2/n + \sum_{i=1}^{k_{max}} \{ \sum_{j=1}^{k_{max}} W(i,j,h)p_j - p_i \}^2 \quad (4.17)$$

where p_i is the true PMF, W(i,j,h) is the weight function, h is the bandwidth, and n is the sample size. For the the two populations considered, viz., G1 and G2, we know the true PMF. Substituting this for p_i in the above equation, the optimal bandwidth can be determined for various sample sizes. These bandwidth values are then compared with the corresponding average bandwidths obtained from the simulations. These along with the coefficient of variance of bandwidth C_v are summarized in Table 4.2. It can be observed that C_v is smaller for DKE for all the sample sizes for G1 and G2. Note that DKE smooths the Geometric distribution data (G1) more than HT/DS, and smooths the mixture data (G2) less than HT/DS. Also the average bandwidths from DKE are close to the MSE optimal bandwidths. This suggests that the bandwidth from DKE is more stable than from HT/DS.

The behavior of HT/DS in these simulations is interesting. There is a tendency to undersmooth relative to the optimal bandwidth. As a result the boundary bias decreases with n, while the tail bias may be high. The higher coefficient of variance of the HT/DS bandwidth suggests a higher degree of adaptation to sample attributes. However, this fails to consistently provide a lower bias on MSE than DKE.

The need to choose a bandwidth in the boundary region that is different from the interior has been recognized by several researchers [e.g., Müller, 1991]. Generally variation in h across the range of the data, and especially in the tails is needed. The selection of a "local" bandwidth considering boundary kernels and tail regions remains an area of research.

Table 4.2. Bandwidth Statistics

	Coefficient of Variation		Average Bandwidth		Optimal Bandwidth from MSE Criteria	
	DKE	HT/DS	DKE	HT/DS	DKE	HT/DS
Sample from G1						
n = 50	0.349	0.442	6.73	5.48	7.00	8.06
n = 100	0.305	0.401	6.13	4.97	6.00	8.06
n = 200	0.316	0.361	4.96	4.36	5.00	7.14
n = 300	0.290	0.314	4.51	4.21	4.00	6.25
n = 500	0.275	0.341	4.00	3.47	4.00	5.56
Sample from G2						
n = 50	0.309	0.291	2.844	3.067	3.00	4.10
n = 100	0.210	0.220	2.280	2.931	2.00	4.03
n = 200	0.007	0.213	2.020	2.902	2.00	4.03
n = 300	0.000	0.212	2.000	2.912	2.00	4.03
n = 500	0.000	0.214	2.000	2.844	2.00	4.03

Other Possible Estimators

Müller [1991] shows how one can develop minimum variance kernels and kernels belonging to different smoothness classes for continuous variates. Extensions of these ideas to the discrete case are also feasible. Here we outline two such extensions.

A discrete, minimum variance (DMV), second-order kernel can be developed as the solution to:

$$\text{Minimize} \sum_{j=q}^{i+h} w_j^2 \tag{4.18}$$

Subject to:

$$w_q = w_{i+h} = 0 \tag{4.19}$$

$$\sum_{j=q}^{i+h} w_j = 1 \tag{4.20}$$

$$\sum_{j=q}^{i+h} t_j w_j = 0 \tag{4.21}$$

where $t_j=(i-j)/h$, i,j,h are integers, and $q=\max(i-h,1)$ recognizes whether we are in the boundary region or the interior.

A smooth, discrete ($DS\mu$) kernel of smoothness μ can be defined by solving the problem:

$$\text{Minimize} \sum_{j=q}^{i+h-\mu} (w_{j+\mu} - w_j)^2$$

subject to the conditions (4.19) through (4.21) above.

Solutions to the two problems defined above can be readily obtained by defining the associated Lagrangian problems and solving them for the weights w_j that define the kernel sequence over the appropriate span of integers.

The weight sequences resulting for DMV and DS1 (μ=1) for selected values of h and i are compared with the DQ, and HT/DS weight sequences in Table 4.3. In the interior, the HT/DS, DQ, and DS1 weight sequences coincide. This is to be expected since they all converge to the quadratic kernel. The DMV sequence degenerates to uniform weights as expected. An examination of the weight sequences in the boundary region shows that the DQ sequences stay closer to the DS1 sequences than the HT/DS ones. Thus if a computationally fast approximation to the DS1 sequences was desired in the boundary region, DQ would be preferred. Note that the DMV sequences in the boundary region are still generally closer to the DS1 than the HT/DS.

An interesting aspect of the HT/DS sequence is the adaptation of the weight sequence as h varies between two integers. We observe that the weight sequences at the intermediate h value are not strictly in between the weight sequences at the end points. While this may lead to a high degree of adaptability of the HT/DS procedure, it makes it rather difficult to assess its impact on the estimation procedure. The high coefficient of variation of the bandwidth selected by HT/DS may be related to the nature of the resulting weight sequence.

The boundary kernels developed by Dong and Simonoff [1994] do not correspond to the ones presented by Müller [1991] for the continuous case. It may be interesting to try the Müller [1991] boundary kernels, possibly with a floating boundary value, directly with the HT procedure.

Computational considerations have restricted our Monte Carlo investigations thus far to DQ and HT/DS. The relative utility of DMV and DS may be investigated subsequently. Except in the boundary region, our limited investigations show that differences between the different kernels may not be large. Consequently, kernels that are easier to compute are expedient. In this respect the DQ kernels are to be preferred.

Table 4.3. Comparison of Weight Sequences

	h = 2	h = 2.5	h = 3
Interior			
DQ	0,.3,.4,.3,0	---	0,.14,.23,.26,.23,.14,0
HT/DS	0,.3,.4,.3,0	0,.11,.25,.29,.25,.11,0	0,.14,.23,.26,.23,.14,0
DMV	0,.33,.33,.33,0		0,.2,.2,.2,.2,.2,0
DS1	0,.28, .44,.28,0		0,.14,.23,.26,.23,.14,0
Boundary			
i = 1			
DQ	1,0,0	–	.75,.5,-.25,0
HT/DS	0,1,0	0,1.7,-.7,0	0,.88,.12,0
i = 2			
DQ	0,1,0,0	---	0,.75,.5,-.25,0
HT/DS	0,.63,.37,0	0,.62,.45,-.07,0	0,.5,.4,.1,0
DMV	0,1,0,0		0,.83,.33,-.16,0
DS1	0,1,0,0		0,.8,.4,-.2,0
i = 3			
DQ		---	0,.3,.4,.3,0,0
HT/DS		0,.28,.35,.28,.08,0	0,.28,.32,.28,.12,0
DMV			0,.4,.3,.2,.1,0
DS1			0,.34,.37,.23,.06,0

Notes:

i is the point of estimate, on which the kernel is placed, h is the bandwidth.

DQ, DMV and DS1 do not admit non integer bandwidths.

The HT/DS weights correspond to a quadratic kernel, and admits noninteger h.

Summary and Conclusions

The estimator presented here was motivated by practical considerations. We offer this work in the hope that it will stimulate interest and theoretical development. We show that the discrete kernel procedure advocated can give results comparable to those from the HT/DS procedure. Computational advantages of the DKE procedure and the similarity of its properties to kernel sequences based on smoothness criteria were demonstrated. The relative stability of the bandwidth selection procedure and the DQ weight sequence also recommend it as an alternative to the HT/DS method.

We present only one special case (a quadratic kernel in the interior and in the boundary region). Clearly other similar higher order kernels can be derived. However, as is typical in the kernel smoothing literature, bandwidth selection is likely to be a more tenuous issue than kernel specification. The LSCV choice of h appears to perform quite satisfactorily for the test cases. Extensions to the multivariate case are being investigated.

References

Aitchison, J. and C.G. Aitken, Multivariate binary discrimination by the kernel method, *Biometrika*, 63, 413-420,1976.

Bishop, Y.M., S.E. Fienberg and P.W. Holland, *Discrete multivariate analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.

Dong. J. and J.S. Simonoff, The construction and properties of boundary kernels for sparse multinomials, *Journal of Computational and Graphical Statistics*, 3, (1), 57-66, 1994.

Epanechnikov, V.A., Nonparametric estimations of a multivariate probability density, *Theoretical Probability and Applications*, 14, 153-158, 1969.

Fienberg, S.E and P.W. Holland, Simultaneous estimation of multinomial cell probabilities, *Journal of American Statististical Association*, 68 , 683-691,1973.

Good, I.J., *The estimation of probabilities*, MIT Press, Cambridge, Mass, 1965.

Good, I.J., A Bayesian significance test for multinomial distributions (with discussion), *Journal of Royal Statistical Society, Section B.*, 29, 399-431,1967.

Hall, P. and D.M. Titterington, On smoothing sparse multinomial data, *Australian Journal of Statistics* 29, 19-37, 1989.

Müller, H.G., Smooth optimum kernel estimators near endpoints, *Biometrika* 78(3), 521-530, 1991.

Rajagopalan, B., U. Lall, D.G. Tarboton, *Simulation of daily precipitation from a nonparametric renewal model*, Working Paper WP-93-HWR-UL/003. In Utah Water Research Laboratory, Utah State University, Logan, UT, 1993.

Scott, D.W., *Multivariate density estimation*, Wiley Series in Probability and Mathematical Statistics, John Wiley, New York, 1992.

Simonoff, J.S., A penalty function approach to smoothing large sparse contingency tables, *Annals of Statist.*, 11, 208-218, 1983.

Stone, M., Cross-validation and multinomial prediction, *Biometrika*, 61, 509-515,1974.

Titterington, D.M., Updating a diagnostic system using unconfirmed cases, *Applied Statistics* , 25, 238-247,1976.

Titterington, D.M., A comparative study of kernel-based density estimates for categorical data, *Technometrics*, 22, 259-268, 1980.

Titterington, D.M. and A.W. Bowman, A comparative study of smoothing procedures for ordered categorical data, *Journal of Statistical Computation and Simul*ation 21, 291-312,1985.

Wang, M.C. and J. Van Ryzin, A class of smooth estimators for discrete distributions, *Biometrika*, 68(1), 301-309,1981.

CHAPTER V

SEASONALITY OF PRECIPITATION ALONG A MERIDIAN IN THE WESTERN U.S.[1]

Abstract

We investigate seasonality of daily precipitation along a meridian in the western U.S. using a nonparametric technique. The occurrence of daily precipitation is treated as a nonhomogeneous Poisson process and the time-varying intensity function is estimated for every calendar day using a kernel estimator. The technique is fully data adaptive. We apply this technique to selected long record stations along a meridional transect spanning from Tuscon, Arizona to Priest River, Idaho. Differences in the seasonality of precipitation occurrence and magnitude are revealed as a function of latitude and topographic factors. A monotonic trend in the seasonality of precipitation over the length of record is also observed.

Introduction

Seasonality in hydroclimatic variables is usually related to the unequal heating of the earth's surface over the year, particularly as one moves to higher latitudes. Precipitation is an important hydrologic variable since it is a primary input into surface hydrologic models. The timing and duration of the "seasons" of high precipitation at a site are important since they indicate the form (rain or snow) of precipitation as well as the nature of the input "signal" for the surface hydrologic system.

Here we were interested in dynamically visualizing how the seasonality of rainfall varies by latitude along a transect in the western U.S. (approx. longitude 112° W). Long record precipitation stations that had essentially complete records were selected from

---

[1]Coauthored by Rajagopalan Balaji and Upmanu Lall.

latitude 48° 17' N to latitude 32° 15' N . We were interested in daily precipitation because of its use for agriculture, crop management, and forest management. The attributes of interest considered are precipitation "magnitude" and "relative frequency of occurrence."

Stochastic precipitation models as well as other hydrologic models often deal with the nonstationarity in precipitation and other climatic inputs by dividing the year into a number of seasons and then fitting model parameters independently for each season. The leading terms (one or two) of a Fourier series representation of the precipitation data are commonly used to identify seasonality, for time-varying parameter description, and for delineating seasons.

An attractive alternative to Fourier series methods is provided in this chapter. We focus first on the rate of occurrence of precipitation as a function of calendar date (1 to 366) within the year. A kernel estimator is used to estimate the "rate" of rainfall occurrence of precipitation by calendar day, by "smoothing" a binary (1 or 0) indicator sequence that represents precipitation occurrence on a given day in the historical record. This rate is interpretable as the time varying rate parameter of a nonhomogeneous Poisson process. Variation in precipitation magnitude over a 90-day moving window is also investigated.

An interesting trend in seasonality is exhibited by the stations we analyzed. There appears to be a consistent shift in the seasons identified on the basis of precipitation rate. The calendar dates associated with the highest and the lowest precipitation rates for a given year appear to move forward each year of the record.

Methodology

Precipitation is an intermittent process. For understanding climatic variations it is often useful to consider adaptive representations that allow a smooth, continuous time interpretation of precipitation. The Poisson process has been used to describe rainfall

occurrence as a point process [Waymire and Gupta, 1981; Cox and Isham, 1980]. In the stationary point process, the number of events (e.g., the events are occurrence of wet days) n(T) occurring in a duration T is a random variable with a Poisson distribution with mean λT:

$$p(n(T) = k) = (\lambda T)^k e^{-(\lambda T)}/k! \quad k = 0,1,2.. \tag{5.1}$$

where λ is called the rate or intensity parameter. Often, it is hard to distinguish between changing intensity of the process and event clustering. This situation can be addressed by explicitly allowing changing event intensity in the model, and consequently modeling the daily precipitation as a nonhomogeneous Poisson process (same as Equation (5.1) but with a time-varying rate parameter λ, i.e., $\lambda(\tau)$, $\tau = 1,..,366$) to capture the changing precipitation pattern over the year. Our thesis here is that this time varying rate parameter is a useful indicator of precipitation seasonality at a site.

Kernel intensity estimators [see Diggle, 1985; Solow, 1991] can be used to estimate $\lambda(\tau)$ from the record, through an optimal, weighted moving average of the rate of rainfall occurrence over time. To form such an estimate, we need to define an appropriate weight function, a span over which to average and a criterion for choosing the weight function and span in an optimal way. Our presentation here is informal and is restricted to a description of the estimation process used.

Daily precipitation data from about a dozen sites spread along Arizona, Utah, and Idaho were used to estimate the intensity parameter for each day of the historical record. Table 5.1 summarizes the site and data information.

Table 5.1. Data Sets Analyzed

	Latitude	Longitude	Elevation (ft. above MSL)	Record Length
Priest River, Idaho (PRR)	48° 21' N	116° 50' W	2380	1911-1992
Sandpoint, Idaho [SNP]	48° 17' N	116° 34' W	2100	1910-1992
Laketown, Utah [LAK]	41° 49' N	111° 19' W	5980	1948-1992
Logan, Utah [LOG]	41° 45' N	111° 48' W	4790	1928-1992
Woodruff, Utah [WOD]	41° 32' N	111° 09' W	6320	1948-1992
Silverlake, Utah [SIL]	40° 36' N	111° 35' W	8740	1948-1992
Snake Creek, Utah [SNC]	40° 33' N	111° 30' W	6010	1928-1992
Heber, Utah [HEB]	40° 30' N	111° 25' W	5630	1928-1992
Spanish Fork, Utah [SPF]	40° 05' N	111° 36' W	4720	1932-1992
Alton, Utah [ALT]	37° 26' N	112° 29' W	7040	1929-1992
Miami, Arizona [MIA]	33° 24' N	110° 53' W	3560	1914-1992
Tucson, Arizona [TUS]	32° 15' N	110° 57' W	2440	1901-1992

Estimation Procedure

We considered the estimation of $\lambda(\tau)$, for each calendar day τ (1,2,..,366), for each year of record y. The average across years of the estimates of $\lambda(\tau)$ provides a measure of the typical seasonality at the site.

The kernel estimator used for $\lambda_y(\tau)$, the rate on calendar day τ, in year y is

$$\widehat{\lambda_y(\tau)} = \frac{1}{h_y} \sum_{i=1}^{n_y} K\left(\frac{\tau - \tau_{i,y}}{h_y}\right) \qquad (5.2)$$

In Equation (5.2), τ (1,2,..,366) is the calendar day on which the estimate is required; $\tau_{i,y}$ is the index of a calendar day on which there was rain in year y; K(.) is a kernel function which is taken to be a positive function that integrates to unity, is symmetric, and has finite variance; and h_y is a bandwidth or "scale" parameter (for year y) of the kernel function, which controls the smoothness of $\widehat{\lambda_y(\tau)}$.

The estimator in Equation (5.2) is very similar to a kernel density estimator [see Silverman, 1986; Scott, 1992]. The choice of a kernel function is considered secondary [Silverman, 1986; Scott, 1992] to the choice of the bandwidth in terms of the mean square error (MSE) of the resulting estimate $\widehat{\lambda_y(\tau)}$. Different kernels can be made equivalent in this sense through an appropriate choice of the bandwidth. Diggle and Marron [1988] show the equivalence between density and intensity (or rate) estimation and show that the same bandwidth is optimal in both cases under a mean square error criterion. The "plug-in" or recursive bandwidth estimator due to Sheather and Jones [1991] has worked the best in our tests for kernel density estimation [Rajagopalan et al., 1995]. This procedure strives to minimize the average mean integrated square error in density estimation through a data-

driven estimate of the pointwise bias and variance of the estimate. We used this procedure to select the bandwidth h_y. For this study we used the Epanechnikov kernel, given as:

$$K(x) = \frac{3}{4}(1-x^2)^2 \quad |x| \leq 1 \text{ where } x = \frac{\tau-\tau_i}{h_y} \tag{5.3}$$

Periodic boundaries are used for the estimation process by (1) recognizing that dates from the end of one year can be within a bandwidth h_y of dates in the beginning of the next year, and (2) using data from year (y-1) or (y+1) for estimates on days within such a bandwidth in year y.

The intensity parameter of the nonhomogeneous Poisson process is estimated for each calendar day (τ = 1,..,366) of each year (1,..,y) in the historical record using the estimator in Equation (5.2). Weighted average precipitation for each calendar day of each year in the historical record is also estimated using the Epanechnikov weight function with a bandwidth of 90 days.

Results

The average rate across years and the average weighted precipitation for each calendar day, estimated as described above, are plotted for all twelve stations. The x-axis on all the figures is the calendar day (i.e., 1 to 366), where 1 corresponds to January 1 and 366 to December 31, respectively. In all these figures the solid line denotes the average daily rate, and the dotted lines indicate the average weighted precipitation. The following observations are offered from the figures.

1. The average daily rate and the average weighted precipitation fluctuate in about the same way at all the stations (see Figures 5.1a through 5.1L). Thus, the use of the rate to describe seasonality seems to be a useful notion.

2. Stations in the north of the meridional transect (namely, SNP, PRR, LAK, LOG, SIL,SNC, HEB, and SPF) have similar shape of the rate and precipitation curves as can be seen from Figures 5.1a, 5.1b, 5.1c, 5.1d, 5.1f, 5.1g, 5.1h, and 5.1i. These stations seem to have higher than average values of the rate function around the first 70 to 100 days and the last 70 to 100 days of the year, with the exact number of days varying from station to station. A similar trend is seen in the precipitation.

3. The curves of rate and precipitation are similar for stations near the southern end of the meridional transect (namely, ALT, MIA, and TUS) as seen from Figures 5.1j, 5.1k, and 5.1l. These stations appear to have high rates during the middle 100 days of the year and increased rates during the first and last 30 to 60 days of the year. This is prominent at ALT, and is subdued in MIA and TUS. The "wet" seasons in the north appear to correspond to "dry" seasons in the south and vice versa. This observation corresponds to the largely zonal flow driven winter/spring precipitation in the north, as opposed to the largely convective summer precipitation in the south [Ropelewski and Halpert, 1986, 1987].

4. Station WOD exhibits an interesting pattern (see Figure 5.1e). The rate appears to be high during day 70 to 130 of the year (i.e., in spring) and is low the rest of the time. WOD lies in a rain shadow region with respect to the large-scale atmospheric flow and hence gets very little precipitation during the general wet period and gets all its precipitation during the spring time due to local orographic/convective effects. There are two periods with higher than average daily precipitation at this station. One that corresponds to the high rate (day 70 through 130) and another during day 190 to 290. Apparently this station can receive high convective rainfall in the summer/fall even though the number of rainy days is low then.

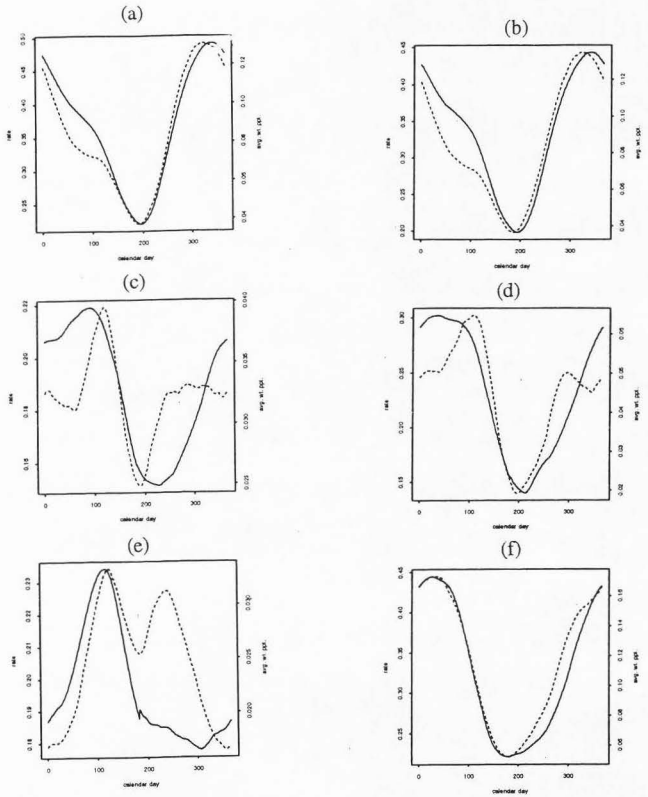


Figure 5.1. Average daily rate (solid line) and average weighted precipitation (dotted line) for each calendar day, at (a) Priest River, ID, (b) Sandpoint, ID, (c) Laketown, UT, (d) Logan, UT, (e) Woodruff, UT, (f) Silverlake, UT, (g) Snake Creek, UT, (h) Heber, UT, (i) Spanish Fork, UT, (j) Alton, UT, (k) Miami, AZ, and (L) Tucson, AZ.

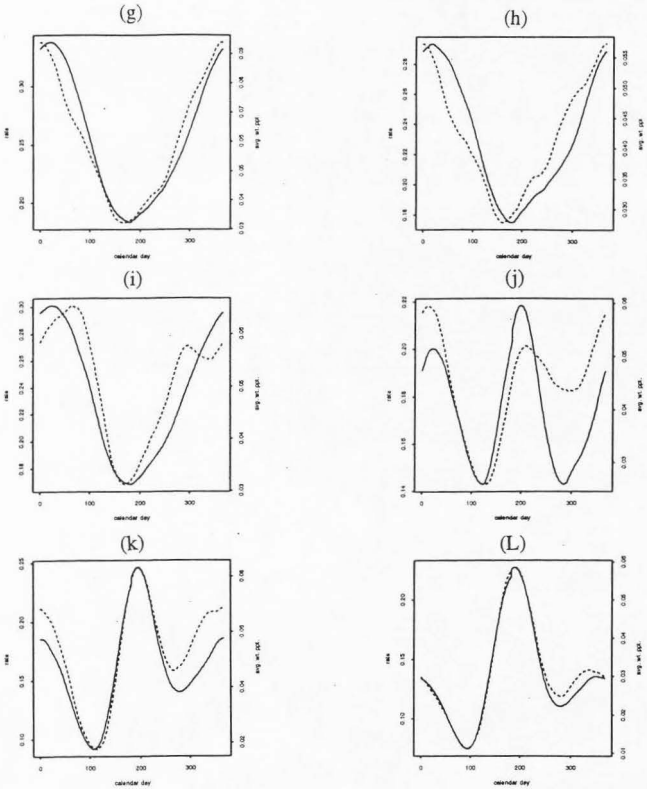


Figure 5.1 (contd.)

Seasonality Trends Over This Century

Schneider [1995] reports that D. J. Thomson found significant changes in the timing of seasons since around 1940 in the northern hemisphere by analyzing the 1651-1991 central England temperature record. The seasonality of temperature in the northern hemisphere is determined by radiative heating which peaks on June 22, and transport of heat from other parts of the globe. The peak temperature occurs later in the year as one moves to higher latitudes in the Northern hemisphere reflecting the delay in transport of heat. Thomson's thesis is that in an atmosphere enriched by carbon dioxide, heating and transport of heat are more efficient, and the advance in the seasons in the northern hemisphere is evidence of global warming.

Consequently, it was of interest to examine changes in the seasonality of precipitation along our meridional transect, as reflected by the estimated rate and average weighted precipitation amounts. We estimate the average rate for the periods before and after 1950 (a time approximately in the middle of the data sets) at four stations with long records, which are PRP, SAN, MIA, and TUS, and plot them in Figures 5.2a, 5.2b, 5.2c, and 5.2d, respectively. In these four figures the thick line is the average rate from the entire historical record, the dotted line is the average rate from the historical record before 1950 and the dashed line is the average rate from the historical record after 1950. The average rate curves for the periods before and after 1950 are shifted from the average rate curve estimated from the entire historical record. It can be seen that the average rate after 1950 is shifted to the left (i.e., the peaks and valleys are shifted left) relative to the average rate before 1950. Similar observations can be seen from the above analysis on the average weighted precipitation amounts, in Figures 5.3a, 5.3b, 5.3c, and 5.3d at the four stations PRN, SAN, MIA, and TUS, respectively.

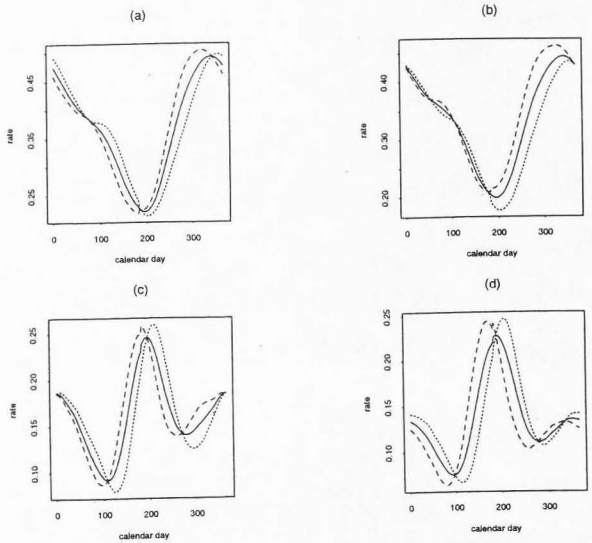


Figure 5.2. Average daily rate from the entire historical record (solid line), from the historical record before 1950 (dotted line) and from the historical record after 1950 (dashed line), at (a) Priest River, ID, (b) Sandpoint, ID, (c) Miami, AZ, and (d) Tucson, AZ.

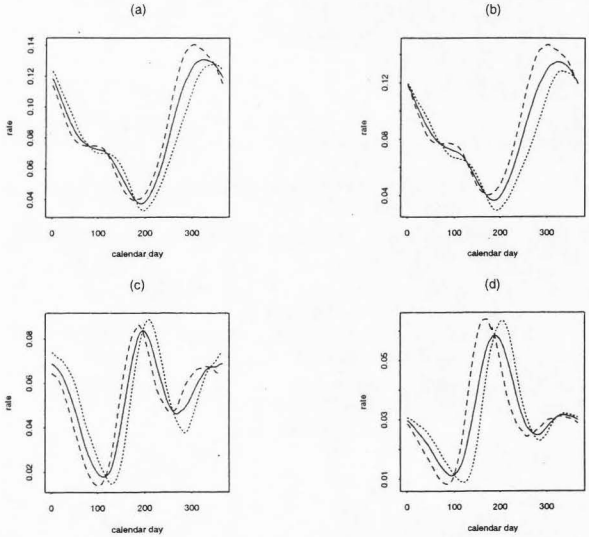


Figure 5.3. Average weighted precipitation from the entire historical record (solid line), from the historical record before 1950 (dotted line) and from the historical record after 1950 (dashed line), at (a) Priest River, ID, (b) Sandpoint, ID, (c) Miami, AZ, and (d) Tucson, AZ.

On observing these patterns in seasonality, we decided to analyze the records to see how this shift was occurring over time, i.e., is it a sudden or continuous trend. The calendar day in each year on which the estimated rate was maximum and the date on which it was a minimum were selected. The maximum (minimum) rate at PRR/TUS occur near the end (or beginning) of the calendar year. Thus a change in seasonality could move this date across calendar year boundaries. It is easier to analyze the transition in the date of the maximum rate at PRR and the minimum rate at TUS if we change the year boundaries away from these dates. Consequently, the date associated with maximum rate at PRR and the minimum rate at TUS is computed on a calendar year that runs from July 1 to June 30, rather than Jan. 1 to December 31. The dates for the minimum rate at PRR and the maximum rate at TUS are computed using the standard calendar.

These dates are plotted for two stations, PRR and TUS (the northern and the southern extremes of our data set), in Figures 5.4a and 5.4b for maximum rate and Figures 5.5a and 5.5b for minimum rates, respectively. The line in these figures is a nonparametric smooth fitted by LOWESS [Cleveland, 1979]. One can see that the date for both the maximum and minimum rates has a decreasing trend with year. The nonparametric Mann-Kendall test [Gilbert, 1987] for monotonic trend showed that these trends were significant (p-values in all cases were of the order of e^{-10}). Robust estimates of the Sen slopes [see Gilbert, 1987] range from -0.33 to -1 days per year. We performed the above analysis with the average weighted precipitation and a similar behavior was observed. Results are not presented for brevity. It is rather curious that the march of seasons as measured by the precipitation rate and also the average weighted precipitation is advancing at these sites at roughly a constant rate over the whole record.

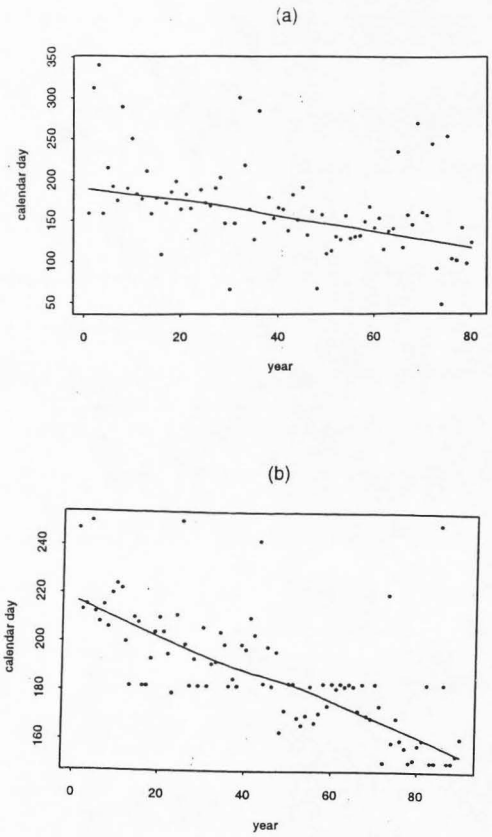


Figure 5.4. Calendar date of maximum estimated average daily rate in each year (dots), along with a LOWESS smooth (thick line), at (a) Priest River, ID, and (b) Tucson, AZ.

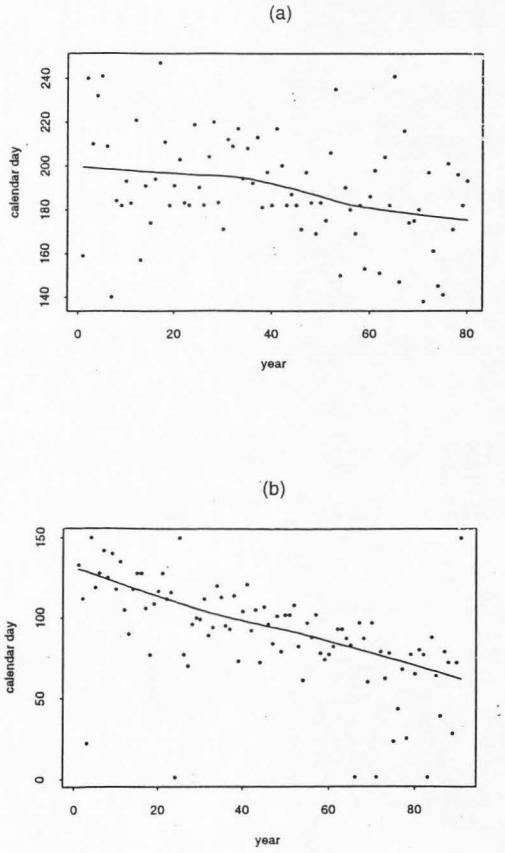


Figure 5.5. Calendar date of minimum estimated average daily rate in each year (dots), along with a LOWESS smooth (thick line), at (a) Priest River, ID, and (b) Tucson, AZ.

Closure

The nonparametric methods presented here were shown to be useful for identifying seasonal variations in precipitation occurrence as a function of latitude and also for variations in seasonality across years. For the data sets we analyzed, remarkable differences were seen in the timing and duration of the precipitation seasons along the meridional transect selected west of the Rockies. An interesting trend in the seasonality across the sites was also identified. If this trend is related to global warming, it has important implications for the form of precipitation in these areas, and also for crop water requirements in the growing season. Further investigation of such trends and their relationship to atmospheric circulation is warranted.

References

Cleveland, W.S, Robust locally weighted regression and smoothing scatter plots, *Journal of American Statistical Association,* 74, 829-836, 1979.

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

Diggle, P.J, A kernel method for smoothing point-process data, *Applied Statistics*, 34, 138-147, 1985.

Diggle, P.J., and J.S. Marron, Equivalence of smoothing parameters selectors in density and intensity estimation, *Journal of American Statistical Association*, 83, 793-800, 1988.

Gilbert, R.O, *Statistical Methods for Environmental Pollution Monitoring*, Van Nostrand Reinhold Company, New York, 1987.

Rajagoplan, B., U. Lall, and D.G. Tarboton, Evaluation of kernel density estimation methods for daily precipitation resampling, Working Paper WP-95-HWR-UL/003, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Ropelewski, C.F., and M.S. Halpert, North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO), *Monthly Weather Review*, 114, 2352-2362, 1986.

Ropelewski, C.F., and M.S. Halpert, Global and regional scale precipitation patterns associated with the El Niño/southern oscillation, *Monthly Weather Review*, 115, 1606-1626, 1987.

Schneider, D., Global warming is still a hot topic: Arrival of the seasons may show greenhouse effect, *Scientific American*, 272(2), 13, 1995.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in probability and mathematical statistics, John Wiley, New York, 1992.

Sheather, S.J., and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Section B.* 53, 683-690, 1991.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

Solow, A.R., The nonparametric analysis of point process data: the freezing history of lake konstanz, *Journal of Climate*, 4, 116-119, 1991.

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

CHAPTER VI

LOW FREQUENCY VARIABILITY IN WESTERN U.S. PRECIPITATION[1]

Abstract

Low frequency (interannual or longer period) climatic variability is of interest because of its significance for the understanding and prediction of protracted climatic anomalies. Since precipitation is one of the key variables driving various hydrologic processes, it is useful to examine precipitation records to better understand long term climate dynamics. Here we use multi-taper spectral analysis (MTM) to analyze the monthly precipitation time series (both occurrence and amount) at a few stations along a meridional transect from Tucson, Arizona to Sandpoint, Idaho. We also examine spectral coherence between monthly precipitation and widely used atmospheric indices like Central Northen Pacific (CNP) and Southern Oscillation Index (SOI). This analysis reveals strong "signals" in 3-7-year frequency bands and 2-year frequency bands, which seem to be consistent across time series. These interannual signals are consistent with those related to El Niño Southern Oscillation (ENSO) and quasi-biennial variability identified by others.

Introduction

The search for hidden order is one of science's aspirations. The identification and explanation of recurrent climatic patterns can have significant implications for long-term climatic forecasts. Though the variation in climate from year to year may seem random, careful examinations of historical data can sometimes reveal a remarkably coherent global pattern of oceanic and atmospheric anomalies that reappear every few years in approximately the same sequence and form. There is growing evidence to this effect and also to the fact that global and regional climate variability is well organized on interannual

[1]Coauthored by Rajagopalan Balaji and Upmanu Lall.

and interdecadal time scales [Mann and Park, 1993; in press]. Two modes of low frequency variability (at the interannual time scales) are the El Niño/Southern Oscillation (ENSO) and the Quasi-Biennial Oscillation (QBO) [Ropelewski and Halpert, 1986; Burroughs, 1992; Peixoto and Oort, 1992; Rasmusson and Wallace, 1983]. ENSO-related events can have major impacts on U.S. atmospheric weather patterns, which in turn modulate the surface climate (i.e., wind, temperature, and precipitation) and consequently the streamflow [Kahya and Dracup, 1993; in press; Cayan and Peterson, 1989; Cayan and Webb, 1992].

Recognition of low frequency variability leads to changes in the interpretation and utility of hydro-climatic records. The impact of climate variability on the hydrologic cycle is also important from the point of view of understanding the underlying dynamics of the system. The identification of coherent, low frequency patterns may also be relevant to interpretation of long-range persistence or the Hurst effect.

From the recent works of Klein and Bloom [1987], Kiladis and Diaz [1989], Cayan and Peterson [1989], Leathers et al. [1991], and Lins [1993], among numerous others, it is clear that atmospheric oceanic conditions in the Pacific basin exert considerable influence on the low frequency patterns of North American climatic and hydrologic variability.

In this study we focus on connections between two atmospheric indices and variability in precipitation along a meridional transect in the western U.S. Past studies include simply examining the historical records for subtle changes in climatic patterns [Rasmusson and Wallace, 1983; Rasmusson and Carpenter, 1983], using correlation type of analysis to find strong statistical relationship between atmospheric indices versus precipitation, temperature, and streamflow [e.g., Bradley et al., 1987; Yarnal and Diaz, 1986; Cayan and Peterson, 1989], and a harmonic analysis to examine the climate

anomalies [Ropelewski and Halpert, 1986, 1987, 1989; Piechota and Dracup, in press], on a case-by-case basis.

In this chapter we use the nonparametric multi-taper method (MTM) of spectral analysis due to Thomson [1982] on the time series of monthly precipitation and monthly rates (defined as number of wet days in the month divided by the number of days in the month). The monthly rate is used as a proxy for the occurrence process. A significance testing of peaks is done as part of the MTM procedure.

In what follows, a brief description of the data sets is first provided. The MTM procedure of spectral analysis is next outlined. Results from the analysis are then summarized and discussed.

Data Sets

We chose seven stations at approximately 112-116°W longitude going from Arizona (AZ) to Idaho (ID). In order to look for connections in precipitation with large-scale atmospheric fluctuation (ENSO, QBO), we chose two atmospheric indices, namely, the Southern Oscillation Index (SOI) and Central North Pacific index (CNP), which have been shown as good indicators for western U.S. atmospheric variability [Cayan and Peterson, 1989]. The hydrologic impact of variability in atmospheric circulation is strong in this arid region. The station and data information (latitude, longitude, elevation, and source of data) are given in Table 6.1. To keep the length of the record common across the various data sets, we chose a common period of 1932-1992, during which all the data sets were available.

From the daily precipitation data, total monthly precipitation and the monthly rate (i.e., number of wet days in the month divided by number of days in the month) were first calculated for each station.

The SOI data are a time series of monthly mean difference in sea level pressure (SLP) at Tahiti (approximately 150°W, 18°S) and Darwin (approximately 130°E, 13°S). ENSO is an identified family of atmospheric and oceanic variations. ENSO is a warm event in the tropical Pacific Ocean and is considered a significant perturbation of general atmospheric circulation. The ENSO has typically a life cycle of about 22 months and recurrence interval of about 3-8 years. When SOI is a low negative value, a strong El Niño event is in progress, the atmospheric pressure in the eastern Pacific decreases, and the trade winds usually weaken. Then the warm water pool extends eastward, piling up off the coast of Peru and southern Ecuador.

Table 6.1. Data Sets Analyzed

	Latitude	Longitude	Elevation (ft. above MSL)
Priest River, Idaho [PRR]	48° 21' N	116° 50' W	2380
Sandpoint, Idaho [SNP]	48° 17' N	116° 34' W	2100
Logan, Utah [LOG]	41° 45' N	111° 48' W	4790
Snake Creek, Utah [SNC]	40° 33' N	111° 30' W	6010
Alton, Utah [ALT]	37° 26' N	112° 29' W	7040
Miami, Arizona [MIM]	33° 24' N	110° 53' W	3560
Tucson, Arizona [TUS]	32° 15' N	110° 57' W	2440
South Oscillation Index [SOI]	SLP(Tahiti) - SLP(Darwin)		
Central Northern Pacific [CN]	average SLP(170E-150W, 35N-55N)		

Note:
SLP = sea level pressure
All the data except SOI and CNP were obtained from Earth Info, CD-ROM
SOI and CNP data were obtained from Dr. Cayan.
All the data sets were of the same length, i.e., 1932-1992.

The CNP index [Cayan and Peterson, 1989] is constructed by averaging the sea level pressure (SLP) over the region 35°N - 55°N and 170°E - 150°W. This index is similar to the Pacific North America (PNA) index and is available for a longer period than PNA. The CNP index has been shown to be more strongly tied to the precipitation in the north-west than SOI [Cayan and Webb, 1992]. The SOI and CNP data were obtained from Dr. D.R. Cayan at the Scripps Institution of Oceanography, San Diego.

Multi-taper Method of Spectral Analysis (MTM)

We performed spectral analysis using the multi-taper method on each of the time series given in Table 6.1 and identified significant frequency peaks. Next, we estimated the spectral coherence between the precipitation series and the atmospheric indices (SOI and CNP) to identify the significant coherent frequencies. Lastly, we bandpassed the time series at a few significant frequencies and examine the bandpassed time series for variability in the amplitudes.

The description of the multi-taper method of spectral analysis is abstracted from Lall and Mann [1994]. Thomson [1982] provides the following motivation for the MTM algorithm. He points out that (1) the classical periodogram is an inconsistent estimator of the spectrum, (2) without a taper window, it may be too biased to be useful, (3) usual tapers can reduce variance efficiency, (4) smoothing the periodogram is unsatisfactory for spectra with large range and line and broadband components, since the true spectrum is not smooth, and (5) since the periodogram-based spectral estimator does not directly use phase information, line detection is poor. He sets his sights on developing an estimator (MTM) that (1) is consistent, (2) has good small sample performance in terms of variance

efficiency, (3) is data adaptive, (4) is nonparametric, i.e., locally approximates the spectrum using information only from neighboring frequencies, (5) works well with spectra with a high dynamic range, (6) is computationally easy, and (7) has statistics that can be estimated, and hence significance tests for line components and coherence can be provided. We outline the aspects of the MTM algorithm relevant to our presentation and refer the reader to Thomson [1982] for details.

The finite discrete Fourier transform (DFT) of the data, x(0), .x(t),..x(n-1) is given by:

$$y(f)=\sum_{t=0}^{n-1} e^{-i2\pi f(t-(n-1)/2)}x(t) \tag{6.1}$$

For a finite data set, the DFT is related to the spectrum as:

$$y(f)=\int_{-1/2}^{1/2}\frac{\sin n\pi(f-\nu)}{\sin\pi(f-\nu)}dZ(\nu)=\int_{-1/2}^{1/2}G(n,f,\nu)dZ(\nu) \tag{6.2}$$

where the spectrum S(f) is defined through $\{S(f)\,df = E[|dZ(f)|^2]\}$, where E[.] denotes expectation.

The periodogram estimate $S_p(f)$ is simply $|y(f)|^2$, whose properties will not correspond to those of S(f), since the term G(n,f,v) in Equation (6.2) poorly approximates a Dirac delta function. This term is a consequence of a rectangular window of width n placed on the underlying process. Given the estimate y(f), one can seek a solution for $dZ(f_0)$ in Equation (6.2) in some locale (f_0-W, f_0+W) of a frequency f_0. This is an inverse problem parameterized by G(n,f,v). Thomson pursues a least squares solution by considering a weighted eigen function expansion in this locale, and then an appropriate

combination of the resulting estimates. Consider the K term (k=0...K-1) eigenfunction expansion:

$$\lambda_k(n,W)\cdot U_k(n,W;f)=\int_{-W}^{W} G(n,f,v)U(n,W;v)dv \tag{6.3}$$

where $U_k(n,W;f)$ is the k^{th} eigen function centered at f, with window width W, and $\lambda_k(n,W)$ is the corresponding eigen value.

The eigen functions (called discrete prolate spheroidal wave functions) are ordered by decreasing eigenvalue, with the first nW eigenvalues close to 1. Consequently, of all functions that are DFTs of some discrete sequence, these leading eigen functions have a maximum energy concentration in the interval (f_0-W, f_0+W). This implies that the tapers are leakage resistant. The window width W is 0<W<1/2, and is usually of order 1/N to retain high resolution of the resulting estimate. The idea here is that if the K term approximation in Equation (6.3) is "good," then a good solution to the estimation of S(f) is available. Thomson derives such a solution by first considering K spectral estimates corresponding to each of the eigen functions and then combining them using an optimality criterion derived from estimates of the mean square error of estimate of the spectrum in the locale of interest. The K eigen spectra $S_k(f)$, k=0,...K-1, are defined through:

$$y_k(f)=\sum_{t=0}^{n-1} x(t)\frac{v_{t,k}(n,W)}{\varepsilon_k}e^{-i2\pi f(t-(n-1)/2)} \tag{6.4}$$

$$S_k(f) = |y_k(f)|^2 \tag{6.5}$$

where ε_k is 1 for k even, and i for k odd; and $v_{t,k}(n,W)$, the k^{th} discrete prolate spheroidal sequence (DPSS) is defined such that its Fourier transform gives $U_k(n,W;f-f_0)$.

The MTM estimate is obtained as:

$$S_M(f)=\sum_{k=0}^{K-1} w_k(f)S_k(f) \qquad (6.6)$$

and $w_k(f)$ is a weight associated with the k^{th} eigen spectrum estimate at frequency f.

The windows $U_k(.)$ are positive everywhere, and hence the problem of getting negative estimates of S(f) resulting from traditional higher order spectral windows is averted. The combined estimate from K orthogonal tapers also circumvents the loss of resolution and variance efficiency problems endemic to periodograms smoothed with a single taper. The orthogonality of the eigen functions leads the S_k to be approximately uncorrelated. MTM recovers information lost by using a single taper and by ignoring the phase information in the periodogram. A number of strategies for choosing the weights $w_k(f)$ at each frequency f are indicated by Thomson. These range from a simple average, to weights proportional to the eigenvalues λ_k, to a fully data adaptive and recursive procedure that internally estimates the bias and variance of the local estimate. We used the last two strategies in our work. The latter allows improved separation of the line and broad band spectral components. We refer the reader to Thomson for details of the DPSS and the w_k, and discuss the choice of W and K, the user-selected parameters of the model.

The half bandwidth W is usually specified in terms of the Rayleigh frequency $f_R = (n\Delta t)^{-1}$, where Δt is the sampling frequency, as pf_R, where p is usually a small integer. The corresponding DPSS is called a $p\pi$ taper. The corresponding spectral estimate averages in the frequency band $f\pm pf_R$. For example, a 2π taper, for a 100-year annual data set,

would average over f±0.02 cycles/year. Note that this would correspond to periods of 1.92 to 2.08 years for a band centered at f=0.5, and 14.28 to 33.33 years for a band centered at f=0.05. We see from this example that it is desirable to use a small value of p to get higher resolution in the low frequency range. On the other hand, a small value of p can lead to peak splitting in the high frequency range. Comparing estimates obtained by varying p over a small range is consequently desirable. As K increases, the variance of S_M decreases; however, the broad band bias can increase. S_M is distributed as χ^2_{2K}, rather than as χ^2_2 for the periodogram, and the increased degrees of freedom correspond to reduced variance. The first (2p-1) tapers are leakage resistant, so K is usually taken to be 2p-1. As p increases, the number of leakage resistant tapers increases. Note that, as n increases, one can increase p while retaining the same spectral resolution. The estimate $S_M(f)$ is unbiased, but its local features (amplitude) will depend on p and K. Consequently, it is desirable to also look at a significance test for line components based on the ratio of variance explained by a peak at f_0 to unexplained variance in a band centered at f_0.

Thomson shows that an F variance ratio test with 2, and 2K-2 degrees of freedom can be constructed for significance of line components through the statistic F(f):

$$F(f) = \frac{(K-1)|\mu(f)|^2 \sum_{k=0}^{K-1} U_k(n,W;0)^2}{\sum_{k=0}^{K-1} |y_k(f) - \mu(f)U_k(n,W;0)|^2} \qquad (6.7)$$

$$\text{where } \mu(f) = \frac{\sum_{k=0}^{K-1} U_k(0)y_k(f)}{\sum_{k=0}^{K-1} U_k^2(0)} \qquad (6.8)$$

Vautard et al. [1992] point out that the maxima of $S_M(f)$ and F(f) do not always coincide, and suggest using the maxima of F(f) for peak identification. We examined $S_M(f)$ for the different time series analyzed to identify any clear-cut bands with high values of $S_M(f)$. Then we assessed the total power (integral of $S_M(f)$) in each such band, and ranked the importance of each such band for each time series. Finally, we examined F(f) to identify any peaks that passed the 95% significance test in each frequency band where $S_M(f)$ is large.

The coherence C(f) across two time series $x_t^{(1)}$, t=0...,n-1, and $x_t^{(2)}$, t=0,...n-1, is estimated as:

$$C(f) = \frac{\sum_{k=0}^{K-1} y_k^{(1)*}(f) y_k^{(2)}(f)}{\left(\sum_{k=0}^{K-1} y_k^{(1)*}(f) y_k^{(1)}(f) \sum_{j=0}^{K-1} y_j^{(2)*}(f) y_j^{(2)}(f) \right)^{1/2}} \qquad (6.9)$$

where * represents a complex conjugate.

A confidence test [see Brillinger, 1981] similar to the F variance ratio test is used to test for the significance of the coherence amplitude.

Our experience with synthetic data suggested that the MTM procedures were very reliable and were not as sensitive to signal-to-noise ratio, or to the memory in the broadband noise process. MTM is generally superior for identifying phase coherent frequency structure.

Results from Spectral Analysis

The results from the spectral analysis are summarized in Table 6.2. After a

preliminary screening of the spectral output, it was clear that one could designate bands in which there was power. The bands are wider near the lower frequencies, recognizing the increasing effect of the averaging window in frequency space. The approximate spectral

Table 6.2. Results from Spectral Analysis

Data Set	Period 2-3 yr	3-5 yr	5-8 yr	8-10 yr	10-12 yr
PRR-R M^1	**2**(2.3,3.0)	**1**(3.3,4.7)			**3**[11.8,12.6]
PRR-P M^1		**1**(3.3,3.7,4.0,4.6)			**2**[11.6]
SNP-R M^1	**1**[2.5]	**2**(3.2,4.5)	**3**(5.6,7.3)		
SNP-P M^1	**1**[2,2.2,2.5]	**2**(3.3,3.7)	**3**(6.8,7.1)		
LOG-R M^1	**1**(2.0)	**2**[3.3,3.6,4.6]	**3**[5.3,5.4,8.8]		
LOG-P M^1	**1**(2.0,2.4)		**2**(5.2,6.6)		
SNC-R M^1	**2**(2.0,2.8)	**1**(3.3,3.7)	**3**(6.3)		
SNC-P M^1	**2**(2.0)	**2**[4.6)	**3**(5.5,7.0)		
ALT-R M^1	**1**(2.0,2.3,2.6)	**2**(5.0)			
ALT-P M^1	**1**(2.1,2.6)	**2**(3.5)	**3**(5.2)	**4**[9.5]	
MIM-R M^1	**3**(2.1,2.6)	**2**(3.0,3.4,4.2)	**1**(5.3)		
MIM-P M^1	**1**(2.1,2.6)	**2**(3.4,4.0)	**3**(5.3,6.4)	**4**[8.3]	
TUS-R M^1	**1**(2.1)	**2**(3.3)	**3**(5.3)	**4**[8.8]	
TUS-P M^1	**1**(2.1,2.6)	**2**(3.3,2.9)	**3**(5.3)		

Legend: R refers to the Rate of occurrence and P refers to the Precipitation.

M^1 =based on MTM with 3, 2π & $2,1\pi$ tapers. Frequencies significant at both the tapers are reported here. For MTM the entries for each band represent rank of spectral power for the band (#,#,...=peaks significant from Ftest at 95%).

The rank is based on the integral of the spectrum over the band. The band with the most power is ranked 1.

power in each band was ranked for each time series, and any spectral peak in that band that met the F variance ratio test for significance at the 0.95 level, for MTM was also recorded. Features that are resistant to the indicated variations in MTM parameters are reported in Table 6.2. The sites are arranged from north to south (downwards).

The behavior of precipitation amount and rate was found to be very similar. Representative MTM spectra of precipitation amount and rate (the spectrum of precipitation amount and rate is plotted) for three stations PRR, LOG and TUS are presented in Figures 6.1a, b, and c, respectively. The thick line in these figures indicate the spectra of precipitation amount and the dotted lines of precipitation rate, respectively.

The following observations are offered.

1. There is significant power in these series at selected bands, particularly in 2-3-yr, 3-5-yr and 5-8-yr frequency bands. These features seem to be consistent across sites.

2. Coherent cyclic activity with periods around 2., 2.3, 2.5, 3.0, 3.3, 4.6, 5.0, and 5.3 years shows up in the MTM analysis of virtually all the series. These frequencies are also found in the analyses of Mann et al., [1994]; Lall and Mann [1994]; Mann and Park, [1993, in press]; Keppene and Ghil [1992]; and Dettinger and Ghil [1991] to name a few.

3. Periods less than 3 years may relate to the QBO, which is observed in stratospheric winds. Those in the 3-5-yr range may be related to ENSO.

4. Representative MTM estimates of coherence and phase of the precipitation amount (Figures 6.2a, b, and c) and rate time series (Figures 6.3a, b, and c) at these three stations with SOI and CNP are presented. The thick line in Figures 6.2 and 6.3 corresponds to the squared coherence and the dotted line to phase, and the dashed horizontal line in all these figures shows the 95% confidence level of squared coherence. It can be seen from these figures that the coherence and phase of the precipitation amount and rate with SOI and CNP are quite consistent at the three stations. Table 6.3 presents the

Table 6.3. Results from Coherence Analysis

Data Set	Period 2-3 yr	3-5 yr	5-8 yr	8-10yr	10-12yr
Coherence with CNP					
PRR-R	**2**(2.3,3.0)	**1**(3.4,3.7,4.)			
PRR-P	**1**[2.2]	**2**(3.4,3.7,4.0)			
SNP-R		**1**(3.2,3.3,3.4,3.5)	**2**[5.2,5.9,6,6.9]		
SNP-P	**1**[2.62]	**2**(3.2,3.4)			
LOG-R	**2**[2.2,2.3,3]	**1**(3.3)		**3**[7.8]	
LOG-P	**1**[2.2,2.3]	**2**(3.7)			
SNA-R	**2**[2.2,2.3,3]	**1**(3.3)			
SNA-P	**1**[2.2,2.3,2.8]	**2**(3.6)	**3**[6.8]		
ALT-R					**1**[10.9]
ALT-P	**1**[2.6,2.7]	**2**(4.3)			
MIA-R	**1**[2.1,2.6]	**2**[3.2.4.2.4.3.4.5]		**3**[9]	
MIA-P		**2**(3.1,3.2)	**1**[7.5]		
TUS-R	**1**[2.1]	**2**[3.2,4.3,4.4]	**3**[7.4,7.5,7.7,8]	**4**[8.1,8.3,8.5,8.7,9]	
TUS-P					
Coherence with SOI					
PRR-R		**1**(4.)			
PRR-P		**1**(4.0)			
SNP-R			**1**[9.4]		
SNP-P					
LOG-R	**1**[2.5]				**2**[12]
LOG-P	**2**[2.5]		**1**[7.5,7.9]		
SNA-R	**1**[2.2,2.3]		**2**[7.1,8]		
SNA-P	**1**[2.6,2.7]	**2**[3.2]		**3**[8.9]	
ALT-R					
ALT-P	**2**[2.2,2.5]	**1**[4.1,4.7]			
MIA-R	**1**[2.1]	**2**[3.8,3.9,4.2,4.6,4.7]		**3**[9.0]	
MIA-P	**1**[2.2,2.4,2.5]	**2**[4.6,4.7]	**3**[7.2,7.3,7.5,7.6]		
TUS-R	**1**[2,2.1]	**2**[4.5,4.6,4.7]			
TUS-P			**2**[7.3,7.4]		**1**[12]

Legend: R refers to the Rate of occurrence and P refers to the Precipitation.
the results are based on MTM with 3, 2π

The entries for each band represent rank of squared coherence for the band(#,#,...= coherence peaks significant from Ftest at 95%)

The band with the highest squared coherence is ranked 1.
for 3, 2π tapers the F value for the squared coherence at 95% confidence is 0.79.

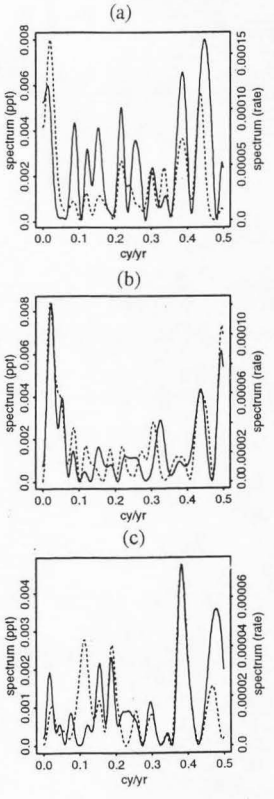


Figure 6.1. Spectra of precipitation amount (thick line) and rate (dotted line) from data at (a) Priest River, ID, (b) Logan, UT, and (c) Tucson, AZ.

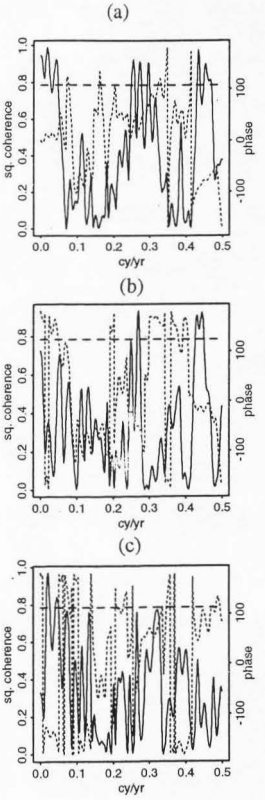


Figure 6.2. Squared coherence between precipitation amount and CNP (thick line), and the phase angle (dotted line) from data at (a) Priest River, ID, (b) Logan, UT, and (c) Tucson, AZ (dashed lines denote the 95% confidence level for the squared coherence).

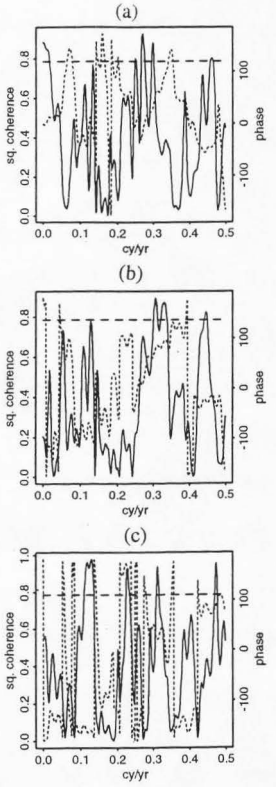


Figure 6.3. Squared coherence between precipitation rate and CNP (thick line), and the phase angle (dotted line) from data at (a) Priest River, ID, (b) Logan, UT, and (c) Tucson, AZ (dashed lines denote the 95% confidence level for the squared coherence).

frequencies with significant coherence between the precipitation time series and SOI and CNP. The frequencies with significant coherence are seen to be mainly in the 2-3 yr band and 3-5 yr band. These are consistent with the frequencies that are significant in the analysis of the individual time series (see Table 6.2). Note that the rates appear to be more coherent than the precipitation amount with the atmospheric indices, and are hence better indicators of the atmospheric variability-precipitation connection. This could be (1) because of nonlinearity in the generation of precipitation as a function of atmospheric flow and (2) because precipitation occurrence may have a larger coherent spatial "signal" than precipitation amount, which may fluctuate quite a bit due to local influences. Also note that the spectral coherence with SOI appears to increase as we move southwards, and the spectral coherence with CNP increases as we move northwards. This phase reversal is consistent with those observed by Kahya and Dracup [in press]; Cayan and Webb [1992], Cayan and Peterson [1989], and others in western U.S. using streamflow data, and precipitation and temperature data [Ropelewski and Halpert, 1986; Yarnal and Diaz, 1986].

5. Noting that a number of significant frequencies from the MTM spectra (Table 6.2) and from coherence analysis (Table 6.3) are in the 3-5 yr band, we bandpassed each of the time series to retain only this frequency band. Bandpassing can be thought of as filtering using the desired frequency band. The amplitude of the bandpassed series of SOI and CNP is similar as can be seen from Figure 6.4. Consequently, representative bandpassed series of precipitation amount and CNP (Figures 6.5a, b, and, c), and rate and CNP (Figures 6.6a, b, and, c) at the three stations are presented. Note that for the station PRR (Figures 6.5a and 6.6a) the amplitudes of precipitation amount and rate are in phase with CNP, station LOG (Figures 6.5b and 6.6b) also exhibits similar behavior. As we move to TUS (Figures 6.5c and 6.6c) there appears to be a considerable phase shift. For the stations in between TUS and PRR, transitional behavior was observed.

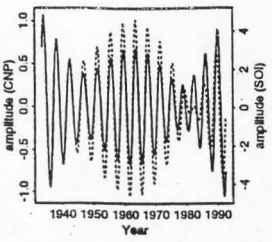


Figure 6.4. Bandpassed series of CNP (thick line) and SOI (dotted line), (bandpassed at 3-5 yr frequency band).

The phase lag between the band passed series of SOI and CNP corresponding to this frequency band is 1.6 months. The phase lag and coherence between the band passed series of CNP and the precipitation amount at the southernmost station TUS is 2.9 months and 0.7; for the station LOG in the middle of the transect it is 2.7 months and 0.78; while it is 2.1 months and 0.84 for the northernmost station PRR, respectively. The coherence with SOI was 0.81, 0.6, 0.5, respectively, at TUS, LOG, and PRR. As can be seen, the phase lags of the precipitation amount with CNP increase and the coherence decreases moving south, and with SOI the coherence increases moving south. This observation is consistent with our expectation, since the CNP is a more direct measure of the atmospheric flow (jet stream behavior) in the northern end of the domain, while the SOI may more directly measure the modulation of the atmospheric flow in the lower latitudes by tropical variability. Of course, the SOI and the CNP may reflect related modes of atmosphere-ocean variability as well.

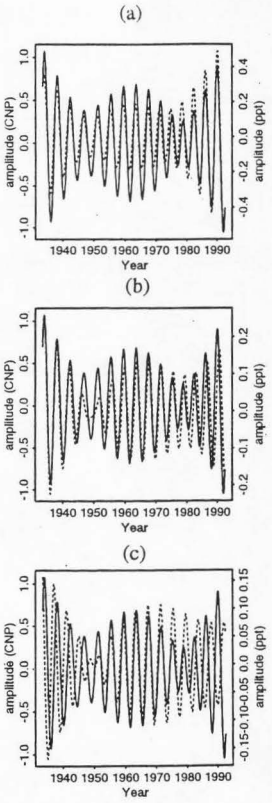


Figure 6.5. Bandpassed series of precipitation amount (thick line) and CNP (dotted line), at (a) Priest River, ID, (b) Logan, UT, and (c) Tucson, AZ (bandpassed at 3-5 yr frequency band).

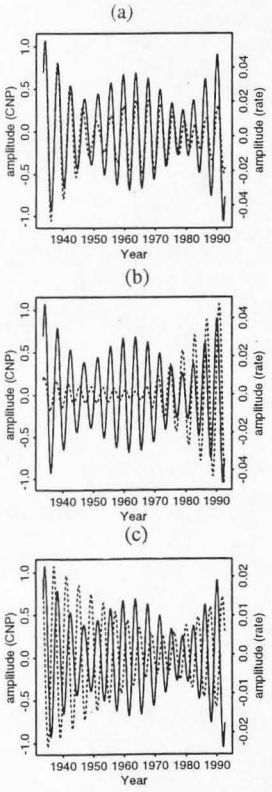


Figure 6.6. Bandpassed series of precipitation rate (thick line) and CNP (dotted line), at (a) Priest River, ID, (b) Logan, UT, and (c) Tucson, AZ (bandpassed at 3-5 yr frequency band).

6. Correlation between the bandpassed series of precipitation amount and CNP were estimated and are reported in Table 6.4, and correlations with SOI were estimated and reported in Table 6.5. Note from Tables 6.4 and 6.5 that the maximum and minimum correlations occur at a lag of approximately 24 months. Also note that the maximum correlations occur at a lower lag in the north and at a higher lag in the south (i.e., from Alton onwards) and vice versa for the minimum correlation, with the exception of Tuscon, Arizona. This suggests that in the 3-5 yr frequency band the modulation in the precipitation due to CNP appears to be opposite while going from north to south, which again corroborates the findings of various others mentioned in observation 4 above.

From this study we found that the precipitation pattern along the meridional transect that we chose seems more influenced by CNP. However, the bandpassed SOI and CNP series are highly correlated, suggesting that tropical atmospheric variation as represented by SOI is manifested in the western U.S. through its modulation of north Pacific atmospheric circulation.

Closure

Spectral analysis was performed on time series of precipitation amount and rates at seven stations along a meridional transect from Arizona to Idaho. We find consistent evidence for structured low frequency variability from the spectral analysis. Strong signals in 3-7 and 2-year frequency bands were revealed from the analysis, which seem to be consistent across time series. These interannual signals are consistent with El Niño Southern Oscillation (ENSO) and quasi-biennial variability identified by others. Spectral coherence between the precipitation amounts and rates with CNP and SOI were also shown to be significant in the above frequency range.

Table 6.4. Correlation Between Bandpassed Precipitation Amounts and CNP (Bandpassed at 3-5 yr Band)

	PRR (ID)	SNP (ID)	LOG (UT)	SNA (UT)	ALT (UT)	MIA (AZ)	TUS (AZ)
Max. Corrln.	0.9	0.8	0.8	0.6	0.5	0.5	0.3
	(2)	(4)	(4)	(4)	(28)	(31)	(9)
Min. Corrln.	-0.8	-0.8	-0.8	-0.5	-0.5	-0.5	-0.2
	(23)	(28)	(28)	(27)	(5)	(5)	(33)

Table 6. 5. Correlation Between Bandpassed Precipitation Amounts and SOI (Bandpassed at 3-5 yr Band)

	PRR (ID)	SNP (ID)	LOG (UT)	SNA (UT)	ALT (UT)	MIA (AZ)	TUS (AZ)
Max. Corrln.	0.7	0.6	0.7	0.3	0.5	0.3	0.5
	(1)	(1)	(4)	(5)	(24)	(33)	(7)
Min. Corrln.	-0.6	-0.5	-0.7	-0.2	-0.4	-0.3	-0.4
	(24)	(24)	(28)	(27)	(1)	(7)	(31)

The high coherence between precipitation amount and rates with SOI and CNP and also the significant frequencies in the ENSO band as suggested by the analyses here have directed our efforts into seeking an understanding of the coherent spatial variability of these variables at the chosen locations. We anticipate publishing that work as Rajagopalan et al, [1995].

A number of authors [Ropelewski and Halpert, 1986, 1987, 1989; Cayan and Webb, 1992; Cayan and Peterson, 1989; Kahya and Dracup, in press] have looked for connections between El Niño and La Niña events and precipitation, temperature, and streamflow series in the western United States, by focusing on first identifying El Niño/La Niña years in the record and then looking for evidence of anomalous behavior in the at-site hydrological variables over a time window centered at each such year. Such an approach is attractive, because it is easily understood and communicated. One can even visually present the results of such an analysis to show spatial patterns quite effectively [e.g., Kahya and Dracup, 1993]. Given the anharmonic nature of the ENSO, such analyses are justified.

The MTM-based approach presented here allows one to go beyond such analyses--one can identify frequency bands where there is structure in individual series, check to see if such structure is phase coherent (the F test) across the series analyzed and directly assess the associated phase lags, and finally bandpass the series at selected frequency bands to examine connections between the different time series. The most striking example of the utility of such an analysis is the suggestion of a meridional (south to north) pattern in the interaction of tropical atmospheric variability (as represented by ENSO) with continental precipitation. It is also interesting that the connections seem to manifest themselves more clearly through a North Pacific index of atmospheric circulation than the SOI directly. Is this simply because the CNP index is defined at a geographically closer location? Or, is there a suggestion that the high latitude North Pacific atmospheric flow is more directly modulated by the tropical variability? The latter is an area of active research.

Results from a rather limited data analysis were presented here. The meridional transect west of the Rockies was chosen on purpose. Given limited resources, we chose to analyze selected long record stations with minimal to no missing data. We feel that the results presented here are quite interesting and suggestive, and provide a useful illustration

of the MTM methodology in this context. We hope to perform a more comprehensive analysis of precipitation and streamflow data sets once requisite computational and financial resources are available.

References

Bradley, R.S., H.F. Diaz, G.N. Kiladis, and J.K. Eischeid, ENSO signal in continental temperature and precipitation records, *Nature*, 327(11), 497-501, 1987.

Brillinger, D. R., *Time Series, Data Analysis and Theory*, Holden-Day, San Francisco, 1981.

Burroughs, W. J., *Weather Cycles: Real or Imaginary?*, Cambridge University Press, 1992.

Cayan, D. R, and D. H. Peterson, The influence of north Pacific atmospheric circulation on streamflow in the west. aspects of climate variability in the Pacific and the Western America, *AGU, Geophysics Monogram,* 55, 75-397, 1989.

Cayan, D., and R. Webb, El Niño/Southern oscillation and streamflow in the western United States, in El Niño : historical and paleoclimatic aspects of the Southern Oscillation, edited by H. F. Diaz, and V. Markgraf. Cambridge University Press., 29-68, 1992.

Dettinger, M. D., and M. Ghil, Interannual and interdecadal variability of surface-air temperatures in the United States", in *Proc. XVIth annual climate diagnostics workshop*, U.S. department of commerce, NOAA, Los Angeles, CA, 209-214, 1991.

Kahya, E. and J.A. Dracup, U.S. Streamflow patterns in relation to the El Niño/Southern Oscillation, *Water Resources Research*, 29(8), 2491-2503, 1993.

Kahya, E., and J.A. Dracup, The influences of type 1 El Niño and La Niña events on streamflows in the southwestern U.S., *Journal of Climate*, in press.

Keppene, C.L., and M. Ghil, Adaptive filtering and prediction of the Southern Oscillation Index, *Journal of Geophysical Research*, 97, 20449-20454, 1992.

Kiladis, G.N., and H.F. Diaz, Global climatic anomalies associated with extremes in the Southern Oscillation, *Journal of Climate*, 2, 1069-1090, 1989.

Klein, W.H., and H.J. Bloom, Specification of monthly precipitation over the United States from the surrounding 700mb height field, *Monthly Weather Review*, 115, 2118-2132, 1987.

Lall, U., and M. Mann, The great Salt Lake: A barometer of low frequency climatic variability, Working Paper WP-94-HWR-UL/005, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1993. 1994.

Leathers, D.J., B. Yarnal, and M. Palecki, The Pacific/North American teleconnection pattern and United States climate. Part I: Regional temperature and precipitation associations, *Journal of Climate and Applied Meteorology*, 24, 463-471, 1991.

Lins, H.F., Streamflow variability in the United States: 1931-1979, *Journal of climate and Applied Meteorology*, 29, 463-471, 1993.

Mann, M. E., U. Lall, and B. Saltzman, Low frequency climate variability: understanding the rise and fall of the great Salt Lake, Working Paper WP-94-HWR-UL/009, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1993. 1994.

Mann, M. E., and J., Park, Spatial correlations of interdecadal variation in global surface temperatures, *Geophysical Research Letters*, 20, 1055-1058, 1993.

Mann, M.E. and J.,Park, Global modes of surface temperature variability on interannual to century time scales, *Journal of Geophysical Research*, in press.

Peixoto, J.P., and A.H. Oort, *Physics of Climate*, AIP, New York, 1992

Piechota, T.C, and J.A. Dracup, Precipitation and temperature patterns in the United States associated with El Niño/Southern Oscillation, *Journal of Geophysical Research*, in press.

Rajagopalan, B., Mann, M.E. and U. Lall, Spatial correlations of low frequency variability in precipitation along a meridion in western U.S., Working Paper WP-95-HWR-UL/013, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Rasmusson, E.M., and T.H. Carpenter, The relationship between eastern equatorial Pacific sea surface temperatures and rainfall over India and Sri Lanka, *Monthly Weather Review*, 111, 517-528, 1983.

Rasmusson E.M., and J.M. Wallace, Meteorological aspects of theEl Niño/Southern Oscillation, *Science*, 222, 1195-1202, 1993.

Ropelewski, C. F., and M. S. Halpert, North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO), *Monthly Weather Review*, 114, 2352-2362, 1986.

Ropelewski, C. F., and M. S. Halpert, Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation, *Monthly Weather Review*, 115, 1606-1626, 1987.

Ropelewski, C. F., and M. S. Halpert, Precipitation patterns associated with the high index phase of the Southern Oscillation, *Journal of Climate*, 2, 268-284, 1989.

Thomson, D. J., Spectrum estimation and harmonic analysis., *IEEE Proceedings*, 70, 1055-1096, 1982.

Vautard R., P. Yiou, and M. Ghil, Singular Spectrum Analysis: A toolkit for short, noisy and chaotic series, *Physica D*, 58, 95-126, 1992.

Yarnal, B., and H.F. Diaz, Relationships between extremes of the Southern Oscillation and the winter climate of the Anglo-American Pacific Coast, *Journal of Climatology*, 6, 197-219, 1986.

CHAPTER VII

A NONHOMOGENEOUS MARKOV MODEL FOR DAILY PRECIPITATION SIMULATION[1]

Abstract

We present a one step nonhomogeneous Markov model for describing daily precipitation at a site. Daily transitions between wet and dry states are considered. The one-step, 2x2 transition probability matrix is presumed to vary smoothly day by day over the year. The daily transition probability matrices are estimated nonparametrically. A kernel estimator is used to estimate the transition probabilities through a weighted average of transition counts over a symmetric time interval centered at the day of interest. The precipitation amounts on each wet day are simulated from the kernel probability density estimated from all wet days that fall within a time interval centered on the calendar day of interest over all the years of available historical observations. The model is completely data driven. An application to data from Utah is presented. Wet and dry spell attributes (specifically the historical and simulated probability mass functions (PMFs) of wet and dry spell length) appear to be reproduced in our Monte Carlo simulations. Precipitation amount statistics are also well reproduced.

Introduction

Markov chains [Gabriel and Neumann, 1962; Todorovic and Woolhiser, 1975; Smith and Schreiber, 1973] have been a popular method for modeling daily precipitation occurrence. Typically a two-state (wet or dry), one-step model is used, and the state transition probabilities (e.g., transition from wet a day to a wet day, wet day to a dry day) are estimated from the data. One problem with such a description is that the transition

[1]Coauthored by Rajagopalan Balaji, Upmanu Lall and David G. Tarboton.

probabilities may vary over the year, i.e., the process of precipitation occurrence is nonstationary.

Two approaches are commonly used to address this problem. In the first approach, the year is divided into periods (or seasons) and the transition probabilities are estimated separately for each period. There is an implicit assumption that the occurrence process is stationary over the period. This assumption may not be tenable. The second approach is to consider essentially a nonhomogeneous Markov process by allowing the transition probabilities to vary sytematically over the year, and to model such a variation through a Fourier series expansion [Feyerherm and Bark, 1965; Woolhiser et al., 1973; Woolhiser and Pegram, 1979]. This can be an effective approach where adequate data are available, and the seasonality in the precipitation process can be captured by a few Fourier series terms. Our nonparametric analyses [Rajagopalan and Lall, 1995] of the seasonality of precipitation for stations along a meridional transect in the western United States suggest that sometimes the number of Fourier series terms needed may be large relative to the amount of data available.

In this chapter, a nonhomogeneous Markov (NM) model is presented that uses kernel methods to estimate a nonhomogeneous transition probability matrix, and to estimate a corresponding nonstationary probability density function (PDF) of daily precipitation amount. Kernel methods are local, weighted averages of the target function (relative frequency of occurrence in this case). Since they are capable of approximating a wide variety of target functions with asymptotically vanishing error, and use only data from a "small" neighborhood of the point of estimate, they are considered nonparametric. Fourier series methods are shown to be a subset of kernel methods by Eubank [1988, secs. 3.4 and 4.1]. A review of hydrologic applications of nonparametric function estimation methods is

provided by Lall [1994].

A brief description of the Markov chain and its terminology is first presented as a background to motivate our formulation. The general structure of the NM model proposed is next outlined with the nonparametric estimators for the transition probabilities. The simulation procedure is then outlined. Results from an application of the model to a precipitation data from Utah follow. Musings on the results and discussion on limitations of the approach conclude the paper.

Background

The basic assumption in a two-state Markov chain model is that the present state (wet or dry) depends only on the immediate past. The transition probabilities for transitions (i.e., WW, WD, DW, DD) between the two states (W or D) are estimated directly from the data through a counting process. Two elements of the transition probability matrix are the probability of a dry day following a wet day, $P_{WD} = a_1$, and the probability of a wet day following a dry day, $P_{DW} = a_2$. The other probabilities, probability of a wet day following a wet day, P_{WW}, and the probability of a dry day following a dry day, P_{DD}, are $(1 - a_1)$ and $(1 - a_2)$, respectively.

Seasonal variations in the transition probabilities can be accounted for by expressing the changing transition probabilities through a Fourier series [Woolhiser and Pegram, 1979; Roldan and Woolhiser, 1982]. As an illustration, the transition probability P(WD) can be expressed as:

$$P_{WD}(t) = \overline{P}_{WD} + \sum_{k=1}^{m} c_k \sin(2\pi tk/365 + \theta_k); \quad t = 1,2,..,365 \qquad (7.1)$$

where m = the maximum number of harmonics required to describe the seasonal variability of the transition probability, $\overline{P}_{WD}$ is the annual mean value of the parameter, c_k is the amplitude, and θ_k is the phase angle in radians for the kth harmonic.

The means, amplitudes, and phase angles are estimated by numerical optimization of the log likelihood function, as described by Woolhiser and Pegram [1979] and Roldan and Woolhiser [1982]. Fourier series representations of parameters of a first-order Markov chain for precipitation have been used (among others) by Feyerherm and Bark [1965], who used least squares techniques for parameter estimation, and by Stern and Coe [1984], who formulated the estimation problem as a generalized linear model to obtain maximum likelihood estimators.

The degree of dependence in time is limited by the order (i.e., the number of past days the present state is presumed to depend on) of the Markov chain. Feyerharm and Bark [1967] and Chin [1977] suggest that the order may need to be seasonally variable as well. Lack of parsimony is a drawback of Markov chain models as the order is increased. A number of researchers [Hopkins and Robillard, 1964; Haan et al., 1976; Srikanthan and McMahon, 1983; Guzman and Torrez, 1985] have also stressed the need for multistate MC models that consider the dependence between transition probabilities and rainfall amount. In this paper, we shall consider only a two state, first order Markov chain. Extensions to other situations follow in the same spirit.

Model Formulation

The NM model that we present allows the one-step transition probability matrix to change over each day, thus capturing the day-to-day variation in the occurrence process in a natural manner. The daily transition probability matrices are estimated using a discrete

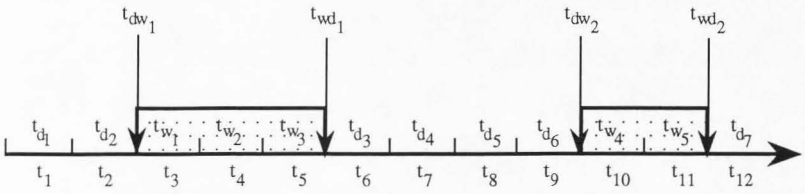
kernel estimator, which we describe in the following section. Daily precipitation occurrence sequences are then simulated using the transition probability matrices. To complete the model, precipitation amounts on each wet day are simulated from the nonparametric probability density estimated from all wet days that fall within a time interval or bandwidth centered on the calendar day of interest over all the years of available historical record. The model is completely data driven.

Transition probabilities and their estimation

The precipitation occurrence process is shown in Figure 7.1. From the daily precipitation record we can obtain four types of data (for illustration refer to Figure 7.1), which are (1) the day indices $t_{w_1}, t_{w_2},..,t_{w_{nw}}$ of nw wet days; (2) the day indices $t_{d_1}, t_{d_2},..,t_{d_{nd}}$ of nd dry days; (3) the day indices $t_{wd1}, t_{wd_2},..,t_{wd_{nwd}}$ of the nwd days on which a transition occurs from wet to dry, meaning days t_{wd_1}, t_{wd_2}, ... are wet and days $t_{wd_1}+1, t_{wd_2}+1$ ... are dry; (4) the day indices $t_{dw_1}, t_{dw_2},..,t_{dw_{ndw}}$ of the ndw days on which a transition occurs from dry to wet, meaning days t_{dw_1}, t_{dw_2}, ... are dry and days $t_{dw_1}+1, t_{dw_2}+1$ ... are wet. A day index refers to a number between 1 to 366, representing the calendar day of the observation. From these we estimate the transition probabilities $P_{wd}(t)$ (probability of transition from a wet day on calendar day t to a dry day on calendar day t+1), $P_{dw}(t)$ (probability of transition from a dry day on calendar day t to a wet day on calendar day t+1). The other two transition probabilities (namely $P_{ww}(t)$ and $P_{dd}(t)$) can be estimated directly from the relations $P_{wd}(t) + P_{ww}(t) = 1$ and $P_{dw}(t) + P_{dd}(t) = 1$. The transition probabilities for calendar day t are estimated from the data using discrete nonparametric kernel estimators.

For a traditional Markov chain the transition probabilities are estimated simply as the ratio of the number of transitions in the historical record to the number of wet or dry days in the historical record, as appropriate. Here,we try to localize such estimates about

the calendar day of interest using kernel estimators. The general idea is that the events (i.e.,



$t_1, t_2, \ldots$ are the day indices
$t_{w_1}, t_{w_2}, \ldots$ are wet day indices
$t_{d_1}, t_{d_2}, \ldots$ are dry day indices
$t_{dw_1}, t_{dw_2}, \ldots$ are day indices of transition from a dry day to wet day
$t_{wd_1}, t_{wd_2}, \ldots$ are the day indices of transition from a wet day to dry day

Figure 7.1. Precipitation occurrence process.

a wet or dry day, or a state transition) occurring near the calendar day of interest should be given more weightage while the ones further away should be given a lower weightage. The resulting kernel estimators for the transition probabilities $P_{wd}(t)$ and $P_{dw}(t)$ are given as:

$$\hat{P}_{wd}(t) = \frac{\sum_{i=1}^{nwd} K\left(\frac{t - t_{wd_i}}{h_{wd}}\right)}{\sum_{i=1}^{n_w} K\left(\frac{t - t_{w_i}}{h_{wd}}\right)} \tag{7.2}$$

$$\hat{P}_{dw}(t) = \frac{\sum_{i=1}^{ndw} K\left(\frac{t - t_{dw_i}}{h_{dw}}\right)}{\sum_{i=1}^{nd} K\left(\frac{t - t_{d_i}}{h_{dw}}\right)} \tag{7.3}$$

where n_{wd} is the number of transitions in the historical record from wet day to dry day, n_{dw} is the number of transitions in the historical record from dry day to wet day, n_d is the number of dry days in the historical record, n_w is the number of wet days in the historical record, K(.) is the kernel function (or weight function) and $h_{(.)}$ is a kernel bandwidth, t is the calendar day of interest and the $t_{(.)}$'s have the definitions described earlier. Note that the estimates on any calendar day t are obtained by using the information from days in the range $[t - h_{(.)}, t + h_{(.)}]$. Note that the definition of calendar dates is periodic, i.e., day 365 and day 1 are recognized as 1 day apart for a non-leap year. The contribution to the estimate of an event that lies within this range is determined by the kernel or weight function K(.),which is described below.

Since we have a discrete situation (i.e. each day being discrete), we use the discrete kernel developed by Rajagopalan and Lall [in press] as:

$$K(x) = \frac{3h}{(1-4h^2)}(1 - x^2) \quad \text{for } |x| \leq 1 \tag{7.4}$$

where $x = (t-t_{(.)})/h_{(.)}$, that measures how far an event $t_{(.)}$ that lies within a bandwidth $h_{(.)}$ of the day t, is from t; and $h_{(.)}$ is an integer.

The kernel in Equation (7.3) was derived from the consideration that the sum of all weights ascribed to events that lie within a bandwidth $h_{(.)}$ of t sum to 1, i.e., $\sum_{x=-1}^{1} K(x) = 1$; that the weights be symmetric on either side of t, i.e., $\sum_{x=-1}^{1} xK(x) = 0$; that each weight be positive; and that the resulting estimate of probability have minimum mean square error.

The estimators in Equations (7.2) and (7.3) are fully defined once the respective bandwidths are specified. We choose the bandwidth using the least squared cross validation (LSCV) procedure [Scott, 1992], where the bandwidth is chosen that minimizes a LSCV function, which is given as:

$$LSCV(h) = \frac{1}{n} \sum_{i=1}^{n} (1 - \hat{P}_{-t_i}(t_i))^2 \qquad (7.5)$$

where $\hat{P}_{-t_i}(t_i)$ is the estimate of the transition probability ($\hat{P}_{wd}$ or $\hat{P}_{dw}$) on day t_i dropping the information on day t_i, n is the number of observations (n_{dw} or n_{wd}). The observed probability of transition is taken to be 1 on the days on which transitions have occurred hence the 1 in the Equation (7.5). The bandwidth is searched from 1 to 182 (length of half year). Once the transition probabilities are estimated for each day in the historical record, the simulation of the precipitation occurrence for each day using the transition probability matrix of the previous day is possible.

Precipitation amount generation

Precipitation amounts for the wet days are generated from a kernel probability density estimated from all wet days that fall within a time interval or bandwidth centered on the calendar day of interest over all the years of historical record. This amounts to two steps: (1) choosing the time interval or bandwidth and (2) generating from the kernel-estimated PDF.

An appropriate bandwidth for localizing the estimate of the probability density of precipitation amount may be obtained by determining the bandwidth appropriate for estimating the probability that a day is wet. If the probability of daily precipitation is low, the precipitation data will be sparse, and the bandwidth needed for stabilizing the variance of the estimated probability distribution of precipitation will be large. Conversely, as the probability of daily precipitation is high, a large number of days with precipitation will occur and the bandwidth needed to localize the estimate can be smaller.

Consequently, we first consider the smoothing of the proportion of wet days ($p_t = n_t/NT$, n_t is the number of times calender day t was wet; NT is the total number of calendar

day t in the historical record) on each calendar day t = 1,2,..,366. These raw proportions are smoothed using the discrete kernel (DK) estimator of Rajagopalan and Lall [in press] which in this case is:

$$\hat{p}_t = \sum_{j=1}^{366} K(\frac{t-j}{h_p})\, p_j \qquad (7.6)$$

where K(.) is the discrete kernel as defined by Equation (7.3), and h_p is the bandwidth that we are interested in. The bandwidth h_p can be obtained using the LSCV procedure similar to Equation (7.5) as given by Rajagopalan and Lall [in press] as:

$$LSCV(h_p) = \sum_{t=1}^{366} (\hat{p}_t)^2 \; - \; 2 \sum_{t=1}^{366} \hat{p}_{-t}\, p_t \qquad (7.7)$$

where $\hat{p}_{-t}$ is the estimate of the calendar day t, by dropping the information on that day.

Once we estimate the time interval h_p, the next step is to pick the precipitation amounts on all the wet days that fall within the time interval h_p from the day of interest in all the years of the historical record. Let us say that the precipitation amounts so picked from the historical records are $y_1, y_2, .., y_{np}$ and $t_1, t_2, .., t_{np}$ are the corresponding calendar day index. The task now is to generate precipitation amount for the calendar day t, which is a wet day. This can be accomplished by fitting a conditional PDF f(y|t) (see Equation [7.10]) and then simulating from it. This step is carried out for each wet day that is simulated. Before describing the simulation procedure we introduce a kernel density estimator for continuous variables, which is given as:

$$\hat{f}(y) = \frac{1}{h_y\, np} \sum_{i=1}^{np} K_c(\frac{y - y_i}{h_y}) \qquad (7.8)$$

where $K_c(.)$ is a univariate, continuous kernel, and h_y is the bandwidth. Here we use the Epanechnikov kernel given by :

$$K_c(x) = 0.75(1. - x^2) \quad \text{for } |x| \leq 1 \qquad (7.9)$$
$$= 0. \quad \text{otherwise}$$

where $x = \frac{y - y_i}{h_y}$. For a detailed exposition of kernel density estimation for continuous variables and issues related to bandwidth selection, we refer the reader to Silverman [1986] and Scott [1992], and for kernel density estimation methods with specific application to precipitation modeling we refer to Lall et al., [1995] and Rajagopalan et al., [1995].

A logarithmic transform of the precipitation data prior to density estimation is often considered. Such a transformation is also attractive in the kernel density estimation context, since it can provide an automatic degree of adaptability of the bandwidth (in real space). This alleviates the need to choose variable bandwidths with heavily skewed data, and also alleviates problems that the kernel density estimation has with PDF estimates near the boundary (e.g., the origin) of the sample space. The resulting estimator works out as:

$$\hat{f}(y) = \frac{1}{np} \sum_{i=1}^{np} \frac{1}{h_{LY} y} K_c\left(\frac{\log(y) - \log(y_i)}{h_{LY}}\right) \qquad (7.10)$$

where h_{LY} is the bandwidth of the log transformed data. This is chosen using a recursive approach due to Sheather and Jones [1991] (SJ) to minimize the mean integrated square error (MISE) and recommended by Rajagopalan et al., [1995] typically for precipitation data.

The two-step procedure discussed above can be more formally considered through the conditional PDF $\hat{f}(y|t)$, defined using a product kernel representation as:

$$\hat{f}(y|t) = \frac{1}{y\, h_{LY}} \sum_{i=1}^{np} K_c\left(\frac{\log(y) - \log(y_i)}{h_{LY}}\right) K\left(\frac{t - t_i}{h_p}\right) \Big/ \sum_{i=1}^{np} K\left(\frac{t - t_i}{h_p}\right) \qquad (7.11)$$

Equation (7.11) shows that the conditional probability density of a rainfall amount y on calendar day t is obtained by considering a window of width h_p centered at t, weighting the precipitation amounts on wet days that fall within this window using the kernel K(.), and then forming a density estimate by further weighting these amounts with the kernel $K_c(.)$. Strictly speaking, the bandwidths h_p and h_{LY} should be chosen by optimizing a criterion relevant to the conditional density. The description of our procedure given earlier shows that we are essentially choosing these bandwidths independently. McLachlan [1992] discusses the simultaneous selection of bandwidths in each coordinate versus the use of the optimal univariate bandwidths in each direction. It is not clear that the additional effort of simultaneous selection of the two bandwidths is justified. Consequently, we choose the bandwidths h_{LY} and h_p by the methods described for the univariate case. Rajagopalan et al., [1995] show that bandwidths selected in this way are often satisfactory. For simulation from the kernel estimated PDF (such as Equation [7.11]) it is not necessary to explicitly estimate the density $\hat{f}(y|t)$. The estimation of the bandwidths h_{LY} and h_p and subsequent perturbation of the historical data is sufficient.

Simulation procedure

The simulation procedure from the NM model can be described in the following steps.

1. From the historical precipitation sequence evaluate the transition probabilities

($P_{wd}(t)$, $P_{ww}(t)$, $P_{dw}(t)$ and $P_{dd}(t)$) for each calendar day t using the estimators described earlier. Similarly evaluate the probability density function for precipitation amount on day t using the procedure described in the previous section.

2. Start the simulation with a wet or dry day (deciding by generating a uniform random number U in [0,1], if $U \leq 0.5$ then wet else dry).

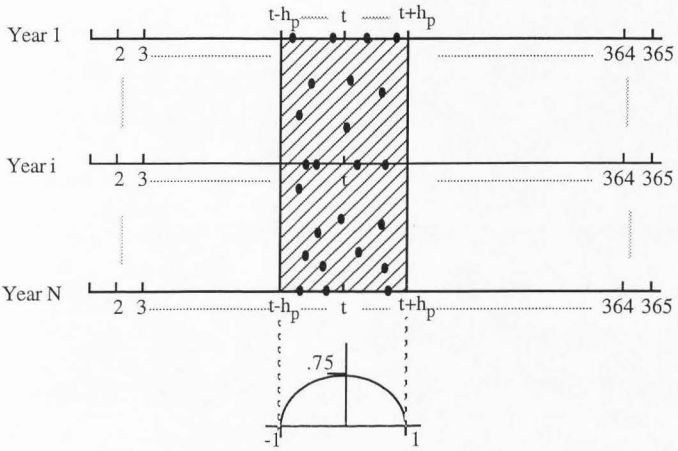
3. The precipitation state for the next day is simulated from the transition probability matrix for the current day (as estimated in step 1).

4. Precipitation amounts on wet days are generated following the process illustrated in Figure 7.2, which is described below:

(i) Pick all the wet day precipitation amounts (e.g., $y_1, y_2, .., y_{np}$) from all the years in the historical record that fall within the window h_p centered on the corresponding calender day of interest and also the corresponding calendar day indices $t_1, t_2, .., t_{np}$.

(ii) For the calendar day of interest, pick a historical wet day to perturb using the bandwidth h_p and the kernel K(x) to specify the resampling metric. Recall that the kernel function describes the weight given to each calendar day that lies within h_p of calendar day t, which depends on the "distance" between the two dates relative to the bandwidth h_p, and the kernel function given in Equation (7.4). Let the weights associated with each of np wet days that are thus identified be $wt_1, wt_2, .., wt_{np}$. Now generate a random integer j between 1 and np from a probability metric given by these weights.

(iv) The simulated precipitation amount is $y^* = \exp(\log(y_j) + Uh_{LY})$ where y_j is the precipitation on the historical day point picked to be perturbed. The random variate U is generated from the probability density corresponding to the kernel function $K_c(.)$. As mentioned earlier, we have used the Epanechnikov kernel in this study and simulation from this kernel is easily accomplished using the two-step procedure described in Silverman [1986].



t is the calendar day on which precipitation is required
h_p is the time interval around the calendar day t
1,...,N are the years in the historical record
Thick dots are the rainy days in the historical record
The kernel function shown at the bottom is used to weight the rainfall amounts on each of the
rainy day.

Figure 7.2. Precipitation amount generation process.

5. The process (steps 3 and 4) is repeated day by day until the desired length of record is generated.

Model Application

The model described was applied to daily rainfall data from Salt Lake City in Utah. Thirty years of daily weather data were available from the period 1961-1991. Salt Lake

City is at 40°46' N latitude, 111° 58' W longitude and at an elevation of 1288 m. Most of the precipitation comes in the form of winter snow. Rainfall occurs mainly in spring, with some in fall.

We shall first list some measures of performance that were used to compare the historical record and the model simulated record, and then outline the experimental design. The aim here is to capture the frequency structure of the events (i.e., the underlying PDF), which then amounts to the reproduction of all the statistics. By events we mean the wet spell lengths, dry spell lengths, and the wet day precipitation. The wet and dry spell lengths are defined as the successive wet or dry days. Clearly the wet spell lengths and dry spell lengths are defined through the set of integers greater than 1. We look at the model performance both at the seasonal scale and the annual scale. For the seasonal scale comparison we have the year divided into four seasons: winter or season 1 (Jan - Mar), spring or season 2 (Apr - Jun), summer or season 3 (Jul - Sep), and fall or season 4 (Oct - Dec).

Performance measures

1. Probability mass function of wet spell length, dry spell length, and probability density function of wet day precipitation in each season and annual.

2. Mean of wet spell length, dry spell length, and wet day precipitation in each season and annual.

3. Standard deviation of wet spell length, dry spell length, and wet day precipitation in each season and annual.

4. Length of longest wet spell and dry spell in each season and annual.

5. Maximum wet day precipitation in each season and annual.

6. Percentage of yearly precipitation in each season and annual.

7. Fraction of wet and dry days in each season annual.

Experiment design

Our purpose here is to test the utility of the NM model. The main steps involved in this are described below.

1. Thirty sets of synthetic records of 30 years each (i.e., the historical record length) are simulated using the NM model .

2. The statistics of interest are computed for each simulated record, for each season, and are compared to statistics of the historical record using boxplots. The PMFs of wet and dry spell lengths are estimated using the DK estimator of Rajagopalan and Lall (in press) (same as the estimator in Equation [7.6]) and the PDFs of the wet day precipitation is estimated using the estimator in Equation (7.10). The statistics listed in the previous section are computed for the simulated record and compared with those of the historical record.

Results

In this section we present comparative results of the NM model for the Salt Lake City data. The PDFs/PMFs of the simulated records are compared with those for the historical record using boxplots while other statistics are summarized in Tables 7.1, 7.2, and 7.3. A box in the boxplots (e.g., Figure 7.3) indicates the interquartile range of the statistic computed from thirty simulations, and the line in the middle of the box indicates the median simulated value. The solid lines correspond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics. The plots of the PDFs are truncated to show a common range across seasons and to highlight differences near the origin (mode).

Figure 7.3 shows the boxplots of kernel estimated PDFs of simulated data of wet day precipitation and the historical data. It can be seen that the historical PDFs are very well reproduced by the simulations in all the four seasons. The other statistics are also seen to be well reproduced by the model for all the seasons and also annual, as can be noticed from Table 7.1.

Boxplots of kernel estimated PMFs of simulated data of wet spell length are found to enclose the PMF of the historical data of wet spell length for all the four seasons in Figure 7.4 and for the annual in Figure 7.6. The other statistics are also preserved quite well by the simulations, as seen from Table 7.2. Good performance of the model in reproducing the dry spell statistics can be seen from Figures 7.5 and 7.7 and also from Table 7.3. The coefficient of skew, the coefficient of variation, the 25% quantile, and the 75% quantile were also preserved for all the three variables, but are not shown here. The extreme statistics (e.g.,. longest spell length or maximum wet day precipitation) exhibit a high degree of variability in the simulations (refer Tables 7.1, 7.2, and 7.3) and an asymmetric sampling distribution, as one would expect.

Note that none of the statistics that we have listed in the section under performance measures are explicitly or implicitly considered in the model. Hence the good reproduction of PDFs/PMFs of the three variables is quite heartening.

Summary and Conclusions

A nonhomogeneous Markov model for simulating daily precipitation is presented in this paper. The traditional Markov chain model is extended to consider the a smooth variation in the transition probabilities from day to day, thus attempting to capture the nonstationarity in the precipitation occurrence process. The 2x2 daily transition probability matrix is estimated nonparametrically. The primary intended use of the model is as a simulator that is faithful to the historical data sequence, obviating the need to divide the year

Table 7.1. Statistics of Wet Day Precipitation for Salt Lake City, UT, 1961-1991 from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

	Mean Wet Day PPT (inches)	Std. Dev. Wet Day PPT (inches)	Fraction of Yearly PPT	Maximum Wet Day PPT (inches)
Season 1				
25% quantile	0.16	0.19	0.23	1.26
Median	0.16	0.20	0.23	1.36
75% quantile	0.17	0.21	0.24	1.59
historical	0.15	0.17	0.21	0.92
Season 2				
25% quantile	0.19	0.24	0.26	1.74
Median	0.19	0.25	0.27	1.86
75% quantile	0.20	0.26	0.28	2.18
historical	0.20	0.24	0.28	1.62
Season 3				
25% quantile	0.18	0.27	0.24	1.94
Median	0.18	0.28	0.26	2.3
75% quantile	0.19	0.30	0.26	2.87
historical	0.18	0.29	0.26	2.28
Season 4				
25% quantile	0.16	0.19	0.24	1.37
Median	0.17	0.21	0.24	1.7
75% quantile	0.18	0.23	0.25	2.16
historical	0.17	0.19	0.25	1.23
Annual				
25% quantile	0.18	0.24		2.35
Median	0.18	0.25		2.55
75% quantile	0.19	0.25		3.45
historical	0.17	0.22		2.30

Table 7.2. Statistics of Wet Spell Length for Salt Lake City, UT, 1961-1991 from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

	Mean Wet Spell Length (days)	Std. Dev. Wet Spell (days)	Fraction Of Wet Days	Longest Wet Spell Length (days)
Season 1				
25% quantile	1.89	1.29	0.31	9
Median	1.92	1.37	0.32	10
75% quantile	1.99	1.43	0.33	11.8
historical	1.86	1.29	0.32	10
Season 2				
25% quantile	1.87	1.27	0.25	8
Median	1.91	1.34	0.25	9
75% quantile	1.95	1.41	0.26	10
historical	2.12	1.47	0.27	12
Season 3				
25% quantile	1.79	1.23	0.19	8
Median	1.86	1.29	0.20	9
75% quantile	1.91	1.37	0.20	10
historical	1.60	0.9	0.18	7
Season 4				
25% quantile	1.85	1.27	0.25	8
Median	1.87	1.32	0.26	9
75% quantile	1.92	1.38	0.27	10
historical	1.97	1.36	0.26	9
Annual				
25% quantile	1.88	1.32	0.26	10
Median	1.91	1.36	0.26	11
75% quantile	1.94	1.39	0.26	13
historical	1.91	1.31	0.26	12

Table 7.3. Statistics of Dry Spell Length for Salt Lake City, UT, 1961-1991 from Historical Precipitation Record and Averaged over 30 Simulated Precipitation Records

	Mean Dry Spell Length (days)	Std. Dev. Dry Spell (days)	Fraction of Dry Days	Longest Dry Spell Length (days)
Season 1				
25% quantile	3.8	3.5	0.67	23
Median	3.92	3.63	0.68	25
75% quantile	4.0	3.75	0.68	27
historical	3.91	3.64	0.68	30
Season 2				
25% quantile	5.21	5.64	0.74	39
Median	5.48	5.91	0.75	46
75% quantile	5.59	6.25	0.76	50
historical	5.5	5.41	0.73	28
Season 3				
25% quantile	6.82	7.12	0.79	44
Median	7.05	7.53	0.80	52
75% quantile	7.26	7.943	0.81	72
historical	6.87	6.92	0.82	55
Season 4				
25% quantile	4.91	5.47	0.73	38
Median	5.09	5.71	0.74	43
75% quantile	5.28	5.91	0.75	51
historical	5.21	5.38	0.74	31
Annual				
25% quantile	5.29	6.13	0.74	58
Median	5.41	6.32	0.74	70
75% quantile	5.54	6.67	0.74	86
historical	5.45	5.99	0.74	61

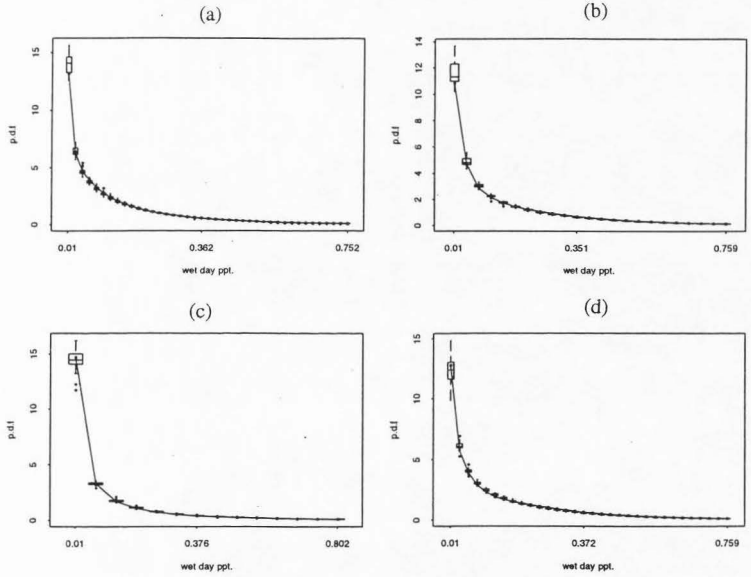


Figure 7.3. Boxplots of PDF of wet day precipitation of model simulated records along with the historical values for (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

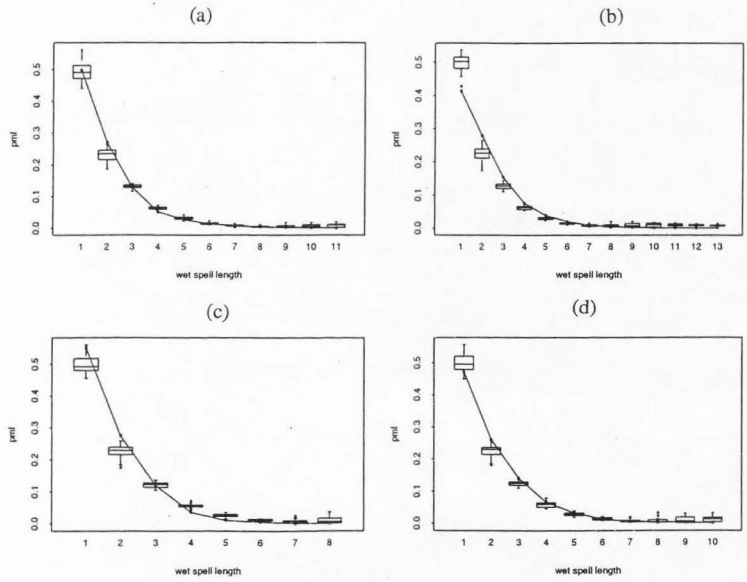


Figure 7.4. Boxplots of PMF of wet spell length of model simulated records along with the historical values for (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

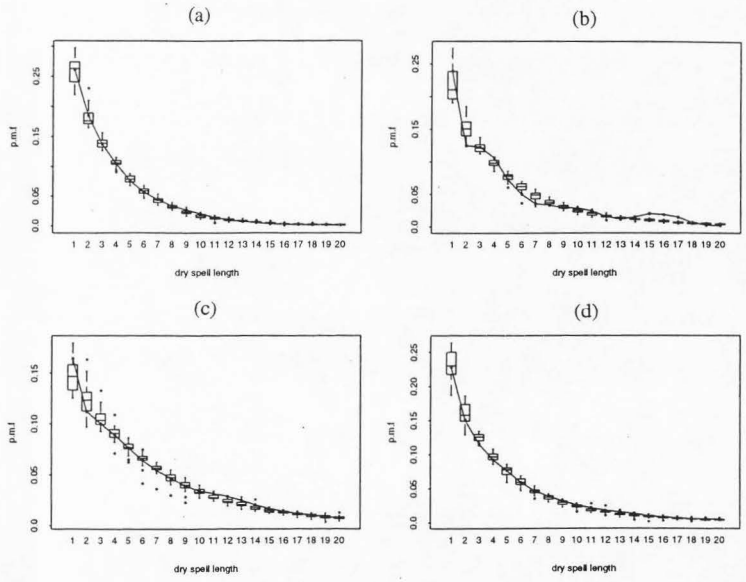


Figure 7.5. Boxplots of PMF of dry spell length of model simulated records along with the historical values (a) season 1, (b) season 2, (c) season 3, and (d) season 4.

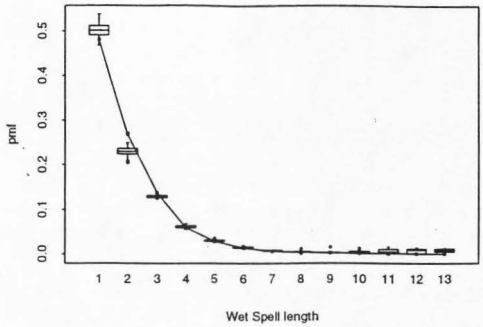


Figure 7.6. Boxplots of PMF of wet spell length over the whole year for model simulated records along with the historical values.

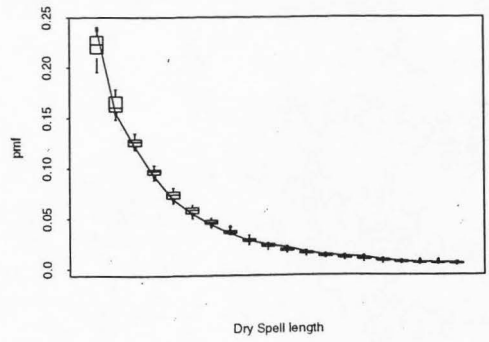


Figure 7.7. Boxplots of PMF of dry spell length over the whole year for model simulated records along with the historical values.

into seasons and subsequently fitting the Markov chain parameters separately for each season. Simulations from the model are shown to preserve the frequency structure (PDF/PMF) of the wet spell length, dry spell length, and wet day precipitation at both the seasonal and annual time scales.

In many cases, the Fourier series approach to addressing seasonal variation in Markov chain parameters may be just as effective. Recall that the Fourier series approach can be shown to be a subset of the kernel approach with a specific kernel choice. The kernel approach presented here is attractive because it is relatively parsimonious, locally adaptive, and extends quite naturally to localizing the probability density estimation for precipitation amount as well. Extensions to higher order chains or those with more states follow directly. One needs to define the appropriate events as was done here and go through the solution of the corresponding smoothing problem.

References

Chin, E.H., Modeling daily precipitation occurrence process with Markov Chain, *Water Resources Research*, 13, 949-956, 1977.

Eubank, R.L., *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, Inc., New York, 1988.

Feyerherm, A.M., and L.D. Bark, Statistical methods for persistent precipitation patterns, *Journal of Applied Meteorology*, 4, 320-328,1965.

Feyerherm, A.M., and L.D. Bark, Goodness of fit of a markov chain model for sequences of wet and dry days, *Journal of Applied Meteorology*, 6, 770-773, 1967.

Gabriel, K.R., and J. Neumann, A Markov chain model for daily rainfall occurrence at Tel Aviv, *Quarterly Journal of Royal Meteorological Society,* 88, 90-95, 1962

Guzman, A.G., and C.W. Torrez, Daily rainfall probabilities: conditional upon prior occurrence and amount of rain, *Journal of Climate and Applied Meteorology*, 24(10), 1009-1014 , 1985.

Haan, C.T., D.M. Allen, and J.O. Street, A Markov chain model of daily rainfall. *Water Resources Research*, 12(3), 443-449, 1976.

Hopkins, J.W., and P. Robillard, Some statistics of daily rainfall occurrence for the canadian prairie provinces, *Journal of Applied Meteorology*, 3, 600-602, 1964.

Lall, U., Nonparametric function estimation: Recent hydrologic applications, *US National Report, 1991-1994, International Union of Geodesy and Geophysics*, 1994.

Lall, U., B. Rajagopalan, and D.G. Tarboton, A Nonparametric Wet/Dry Spell model for resampling daily precipitation, Working Paper WP-95-HWR-UL/006, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York, 1992.

Rajagopalan, B., and U. Lall, A kernel estimator for discrete distributions, *Journal of Nonparametric Statistics*, (in press)

Rajagopalan, B., U. Lall, and D.G. Tarboton, Simulation of Daily Precipitation from A Nonparametric Renewal Model, Working Paper WP-95-HWR-UL/007, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Rajagopalan, B., and U. Lall, Seasonality of Precipitation along a meridian in the western U.S.,Working Paper WP-95-HWR-UL/006, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Roldan J., and D.A. Woolhiser, Stochastic daily precipitation models 1. A comparison of occurrence processes, *Water Resources Research*, 18(5), 1451-1459, 1982.

Scott, D.W., *Multivariate Density Estimation: Theory, Practice and Visualization*, Wiley series in probability and mathematical statistics, John Wiley, New York, 1992.

Sheather, S.J., and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society*, B. 53, 683-690, 1991.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York, 1986.

Smith, J.A., and H.A. Schreiber, Point processes of seasonal thunderstorm rainfall. 1. Distribution of rainfall events, *Water Resources Research* 10(3), 418-423, 1973

Srikanthan, R., and T.A. McMahon, Stochastic simulation of daily rainfall for australian stations. *Transactions of the ASAE*, 754-766, 1983.

Stern, R.D., and R. Coe, A model fitting analysis of rainfall data (with discussion), *Journal of Royal Statistical Society, Series, A.*, 147, 1-34, 1984.

Todorovic, P., and D.A. Woolhiser, Stochastic model of *n*-day precipitation, *Journal of Applied Meteorology*, 14(1), 17-24,1975.

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, *Journal of Applied Meteorology*, 18, 34-42,1979.

Woolhiser, D.A., E.W. Rovey, and P. Todorovic, Temporal and spatial variation of parameters for the distribution of n-day precipitation, in *Floods and Droughts, Proceedings of the Second International Symposium in Hydrology*, Fort Collins, 605-614, 1973.

CHAPTER VIII

MULTIVARIATE NONPARAMETRIC RESAMPLING SCHEME FOR GENERATION OF DAILY WEATHER VARIABLES[1]

Abstract

A nonparametric resampling technique for generating daily weather variables at a site is presented. The method samples the original data with replacement while smoothing the empirical conditional distribution function. The technique can be thought of as a smoothed conditional Bootstrap and is equivalent to simulation from a kernel density estimate of the multivariate conditional probability density function. This improves on the classical Bootstrap technique by generating values that have not occurred exactly in the original sample and by alleviating the reproduction of fine spurious details in the data. Precipitation is generated from the nonparametric wet/dry spell model as described in Lall et al., [1995]. A vector of other variables (solar radiation, maximum temperature, minimum temperature, average dew point temperature and average wind speed) is then simulated by conditioning on the vector of these variables on the preceding day and the precipitation amount on the day of interest. An application of the resampling scheme with 30 years of daily weather data at Salt Lake City, Utah, USA is provided.

Introduction

Daily weather variations influence agricultural and engineering management decisions. Crop yields and hydrological processes such as runoff and erosion are very sensitive to weather. Recognizing the inherent variability in climate, it is often necessary to assess management scenarios for a number of likely input sequences. Stochastic models are

[1]Coauthored by Rajagopalan Balaji, Upmanu Lall, David G. Tarboton and David S. Bowles.

consequently useful for simulating weather scenarios. Such models need to simulate sequences that are representative of the data. While there is a substantial literature for rainfall simulation and for other variables one at a time, only a few "multivariate" models have been developed.

In this chapter we develop and exemplify nonparametric procedures for resampling a vector of daily weather variables, such that selected lag 0 and lag 1 dependence characteristics are preserved. Dependence is defined in terms of joint or conditional probabilities, rather than correlation.

This work is an off-shoot of the ongoing Water Erosion Prediction Project (WEPP) of the United States Department of Agriculture (USDA). WEPP is a key model for soil and forest conservation studies. WEPP includes a climate generator (CLIGEN) and the work presented here intends to improve it. Hill slope erosion is driven largely by precipitation and a suite of other weather variables. Hence, the main objective is to generate weather sequences that will be used by WEPP to estimate hill slope erosion. In this study, we chose a set of five daily variables (solar radiation [SRAD], maximum temperature [TMX], minimum temperature [TMN], avg. wind speed [WSPD] and avg. dew point temperature [DPT] in addition to precipitation [P], that are of interest for erosion prediction. Most of these weather variables are sensitive to precipitation. Solar radiation, dew point temperature, maximum temperature, and minimum temperature are more likely to be below normal on rainy days than on dry days, while the wind speed may be above normal on rainy days than on dry days. Consequently precipitation is chosen as the driving variable of the models developed so far. Typically [see Jones et al., 1972; Nicks and Harp, 1980; Richardson, 1981], daily precipitation is generated independently and the other variables are generated by conditioning on precipitation events (i.e., whether a day is wet or dry).

Throughout this chapter we denote the historical time series of the five weather variables chosen above as $[\mathbf{z}]_{mkj}$ (m=1,..,NY, k = 1,..,366, j=1,..,NV), where NY is the number of years of record, and NV(=5) is the number of variables considered (SRAD, TMX, TMN, DPT, and WSPD). Further, we define $[\overline{\mathbf{Z}}]_{kj}$ and $[\mathbf{STD}]_{kj}$ as the corresponding mean and standard deviation vector for each calendar day k (k=1,..,366) of each variable j (j=1,..,5). The historical time series of the precipitation is denoted as $[P]_{mk}$.

We now discuss key attributes of some strategies for resampling or synthesizing vectors of these variables.

Resampling Approaches

Multivariate stochastic simulation of weather variables has not been studied as extensively as streamflow or precipitation. Two broad approaches that are possible are (1) parametric, and (2) nonparametric - Bootstrap (raw, conditional and smoothed).

Parametric

The parametric approach is the traditional method [see Jones et al., 1972; Bruhn et al., 1980; Nicks and Harp, 1980; Lane and Nearing, 1989; Richardson 1981] for stochastic daily weather simulations. Figure 8.1 summarizes the general structure of the parametric approaches. The general strategy is to generate precipitation independently and the other variables conditioned on the status of precipitation (i.e., rain or no rain on the day). The other variables are generated from either independent statistical distributions fitted separately to each of the variables for each of the two precipitation states (i.e., rain, no rain). Independently or jointly fitted auto regressive models of order 1 (AR-1) are sometimes used.

Usually the year is divided into periods (seasons) and moments (i.e., mean standard deviation and skew) are calculated for each variable for each period for each

precipitation state. The moments are used to fit statistical distributions or models. Dividing the year into various periods assumes homogeneity within each period and offers a

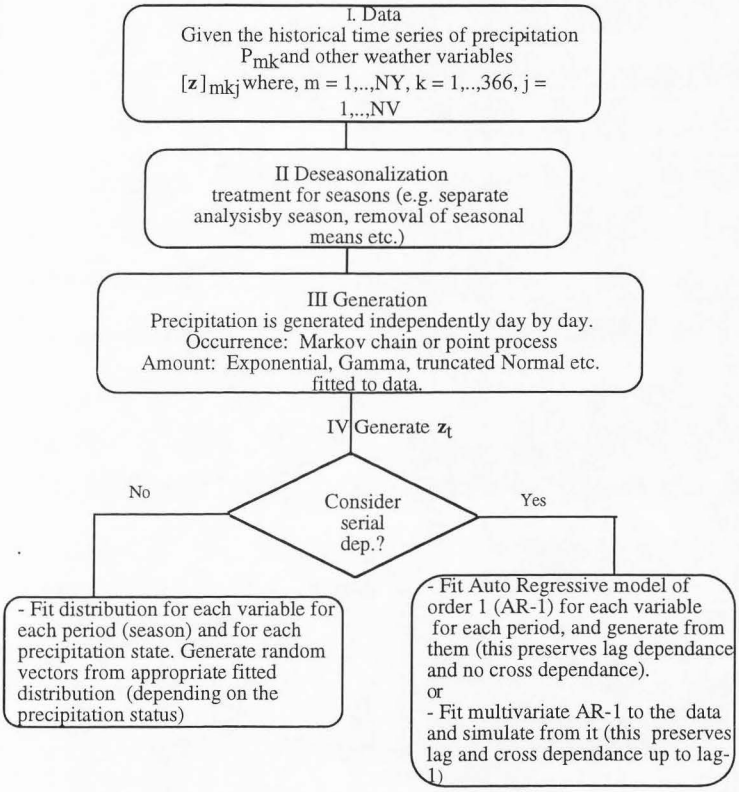


Figure 8.1. General structure of parametric approaches.

treatment of seasonality. Jones et al.[1972], Bruhn et al., [1980], Nicks and Harp [1980], and CLIGEN [Lane and Nearing, 1989] divide the year into 14-day and one month periods respectively in their works. Richardson [1981] adopted a method wherein the means and standard deviations of each periods and each precipitation state are smoothed using Fourier series. The smoothed daily values of the means and standard deviations are subsequently used for deseasonalization.

Daily precipitation is typically generated from a fitted first order Markov chain for precipitation occurrence and by sampling from the distribution (such as Gamma, Exponential, Truncated Normal, etc.) fitted for the daily precipitation amounts for each period.

One approach to generate the other variables is to fit distributions independently for each variable for each period and for each precipitation state. Here, the simulations are made under the assumption that each variable is independent and identically distributed (i.i.d). This approach and its variants are used by Jones et al. [1972], Bruhn et al. [1980], and CLIGEN [Lane and Nearing, 1989]. In CLIGEN each variable is assumed to be an independent Gaussian variable for each month, with parameters dependent on the precipitation state transition (e.g., wet to wet, dry to wet, etc.). This approach does not consider the dependence between the variables or the serial dependence for each variable. Only the dependence on the precipitation state or the precipitation transition is considered.

Serial dependence was incorporated by Nicks and Harp [1980], who fit auto regressive models of order one (AR-1) independently to each variable for each period. Consideration of dependence across variables is added by Richardson [1981] who used a multivariate auto regressive model of order one (MAR-1). When the cross dependence terms are neglected in MAR-1, it reduces to an AR-1 process. These AR models suffer from the drawback of assuming the data to be normally distributed. As a result only linear

dependence can be reproduced. In practice, changes in the weather variables relative to a change in precipitation or other weather variables are not proportional and the assumption of linearity is questionable. Transformation of the data to be multivariate normal may be difficult and may lead to biased statistics upon transforming back to the original space.

The parametric approaches discussed have three main drawbacks, which are (1) choice of a model (i.e., a statistical distribution or the order) is often subjective and rarely formally tested on a site by site basis, (2) reliance on an implicit Gaussian framework (e.g., AR or MAR), which preserves only linear dependance and is not appropriate for bounded variables, and (3) the fitted models have limited portability in the sense that procedures/distributions used at one site may not be best at other sites. The last point is important where an agency wishes to prescribe a uniform procedure over its domain.

Nonparametric

Nonparametric techniques do not require preselected distributions or models to be fit to data. The Bootstrap (or Raw Bootstrap) is a nonparametric technique introduced by Efron [1979]. It is often used for constructing a confidence region, attaching a standard error to an estimate, carrying out a test of a hypothesis, or estimating the sampling distribution of some statistic. Historical data are resampled with replacement. Since they are the same data, the simulations by construction have the same distributional properties as that of the historical data. Since each resampled observation is drawn independently, serial dependence is not preserved. Serial dependence can be accommodated by using the *block-resampling scheme* (a conditional bootstrap) developed by Kunsch [1989] and Liu and Singh [1988]. Here a block of k observations is resampled as opposed to a single observation in the Bootstrap. Serial dependence is preserved within, but not across a block. The block length k determines the order of the serial dependence that can be preserved.

A property of the Bootstrap technique is that the simulated samples will only have values that have occurred in the historical data and consequently the simulations are restricted to the historical set of values. Silverman [1986] points out that this behavior may reproduce spurious fine structure in the original data. This is not a desirable feature while applying the technique to simulation of daily weather variables, where we may wish to have simulated values that have not been observed in the historical data and may be also beyond the maximum/minimum of the observed data. This problem can be alleviated by using a "Smoothed Bootstrap".

In the Smoothed Bootstrap [Silverman, 1986], each observation y_i (i=1,..,n) is considered to be representative of a region (y_i-h , y_i+h) around it. The extent of this region h is called the bandwidth and is determined from the data. Intuitively, it is desirable to resample such that the maximum weight is given to the observation y_i and weights decrease when moving towards y_i-h or y_i+h. This is accomplished by having a weight function centered at each observation. The weight function is usually chosen to be a valid probability density function, such as the Gaussian (N(0,1)). The simulation proceeds by picking an observation y_i with replacement from $\{y_i,..,y_n\}$ and then generating a value from N(y_i, h) with h specified. Formally, the Smoothed Bootstrap is equivalent to resampling from a kernel density estimate.

In this paper, we develop a Smoothed Conditional Bootstrap that considers multivariate and serial dependence amongst the variables of interest. Hereafter, we refer to the scheme presented as the NP model. We first provide the motivation and main ideas of the model. The simulation algorithm is outlined next. The utility of the model is then illustrated through application to daily weather data at Salt Lake City, Utah, USA. In a related work Sharma et al.[1995] describe the application of the NP model to simulation of monthly streamflow.

Main Ideas of the NP Model

Our goal is to develop an approach that is driven directly by the observed data with reasonable assumptions, is easy to implement, is readily transferable from site to site and captures the relative frequencies of the data in a natural manner. We do this by defining the appropriate probability densities that we need to resample from and then discuss their estimation.

Overview of the NP model

A conceptual flow chart of the model is shown in Figure 8.2. The historical data of the other weather variables other than precipitation is standardized as $[\mathbf{x}]_{lkj}=([\mathbf{z}]_{lkj} - [\overline{\mathbf{Z}}]_k)/[\mathbf{STD}]_{kj}$, where l, k, and j are the same as defined earlier. This removes the seasonality present in each variable. Precipitation for day 't' (P_t) is generated from the wet/dry spell model as described in Lall et al.[1995] that is briefly summarized in later in this chapter. However, the user can generate daily precipitation from any other model that is considered appropriate.

In the NP model the year is divided into four periods or seasons (for the Salt Lake City example, these are season 1 (Jan-Mar), season 2 (Apr-Jun), season 3 (Jul-Sep), and season 4 (Oct-Dec)). Simulations for days in any particular period are made using the historical data of that period. Subsequently, the comparison between the simulations and the historical data are also made by season. One could choose different periods (e.g., monthly, weekly, etc.). We chose the above four periods so as to be consistent with the wet/dry spell model [Lall et al., 1995] for daily precipitation.

The aim of the model is to capture the day-to-day dependence present between the variables. The standardized vector of variables $\mathbf{x}_t$ for any day 't' is simulated from the multivariate conditional PDF $f(\mathbf{x}_t \mid \mathbf{V}_t)$. Here, $\mathbf{x}_t$ is a standardized vector

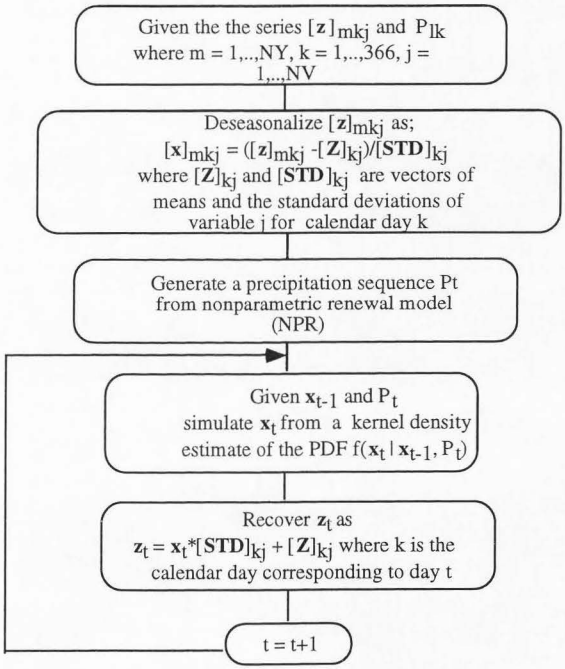


Figure 8.2. Overview of development of the NP model.

$[SRAD,TMX, TMN, WSPD, DPT]_t$ of length d(=5) that is to be generated for day t, P_t is the generated precipitation for day t from the wet/dry spell model, and $\mathbf{V}_t = [\mathbf{x}_{t-1}, P_t]$ is the conditioning vector of length d'(=6). The joint density is estimated in a space of dimension dg (=d+d').

The conditional density $f(\mathbf{x}_t \mid \mathbf{V}_t)$ is defined as:

$$f(\mathbf{x}_t \mid \mathbf{V}_t) = \frac{f(\mathbf{x}_t, \mathbf{V}_t)}{\int f(\mathbf{x}_t, \mathbf{V}_t) d\mathbf{x}_t} = \frac{f(\mathbf{x}_t, \mathbf{V}_t)}{f_V(\mathbf{V}_t)} \tag{8.1}$$

where $f_V(\mathbf{V}_t)$ is the marginal density of $\mathbf{V}_t$.

The standardized sequences $\mathbf{x}_t$ are then transformed to $\mathbf{z}_t = \mathbf{x}_t * [\mathbf{STD}]_k + [\overline{\mathbf{Z}}]_k$, where k is the calendar day associated with day t. Thus, the key idea here is the estimation of this conditional probability density function from the historical data using nonparametric density estimators (kernel estimators) and subsequently simulating or bootstrapping from it. The mechanism of kernel density estimation and the algorithm for simulation from a conditional PDF (as in Equation 8.1) using kernel density estimators is developed and outlined in later sections.

Precipitation model

The seasonal wet/dry spell model for daily precipitation described fully in Lall et al. [1995] has three random variables--wet spell length, L_w days; dry spell length, L_d days; and wet day precipitation amount, P inches. The periods (seasons) are as defined in the previous section. Variables wsp and dsp are defined through the set of integers between 1 and the season length, and P is defined as a continuous, positive random variable. A mixed set of discrete and continuous random variables is thus considered. The simplified version of the wet/dry spell model described in Lall et al. [1995] that considers successive wet

days' precipitation amount and successive wet and dry spell lengths to be independent is adopted in this study. Correlation statistics computed for the data sets analyzed supported these assumptions.

The PDFs of wet day precipitation amount f(P) and the probability mass functions (PMFs) of wet spell length $f(L_w)$ and dry spell length $f(L_d)$ are estimated for each season using kernel density estimators.

A dry spell is first generated using $f(L_d)$. Then a wet spell is generated using $f(L_w)$. Precipitation for each of the L_w wet days is then generated from f(P). The process is repeated with the generation of another dry spell. If a season boundary is crossed, the PDFs used for generation are switched to those for the new season. This procedure continues until a synthetic sequence of the desired length has been generated. The PDFs $f(L_w)$, $f(L_d)$ and f(P) are estimated using kernel density estimators detailed in Lall et al. [1995] and Rajagopalan et al., [1995] and are described below.

Kernel density estimation

The kernel density estimator generalizes the frequency histogram as an estimator of the PDF. While the histogram is capable of showing some features of the data, it has several drawbacks. It is difficult to manipulate analytically, it is not easy to visualize for multivariate situations, and it allows for no extrapolation beyond the data. The histogram is sensitive to the class width, as well as the origin of each class. Silverman [1986] illustrates these problems graphically. One can improve the histogram by centering rectangular boxes at each observation (to gain independence from choice of origin). A kernel density estimator, introduced by Rosenblatt [1956], is formed by centering a smooth kernel function at each observation.

An attractive feature of kernel estimators of the PDF is that they are local (use only a neighborhood around the point of estimate) and hence are not globally affected by outliers. Since they make weak prior assumptions of the underlying probability density function, they are data driven and robust and are portable across sites/data sets. For details on kernel density estimation, refer to Silverman [1986] and Scott [1992].

Univariate continuous variables. The kernel density estimator for a continuous variable (such as the wet day precipitation P) is defined as

$$\hat{f}(P) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{P-P_i}{h}\right) \tag{8.2}$$

where K(.) is a kernel function centered on the observation P_i, and can be any valid probability density function and h is a bandwidth. The bandwidth h controls the amount of smoothing of the data in the density estimate. Bandwidth h may be constant or variable, taking on different values at different locations. An estimator with constant bandwidth h (like in Equation 2) is called a fixed kernel estimator. Commonly used kernel functions are:

Gaussian Kernel	$K(t) = (2\pi)^{-1/2} e^{-t^2/2}$		(8.3a)
Epanechnikov Kernel	$K(t) = 0.75\ (1 - t^2)$	$\lvert t \rvert \leq 1$	(8.3b)
Bisquare Kernel	$K(t) = (15/16)\ (1 - t^2)^2$	$\lvert t \rvert \leq 1$	(8.3c)

The kernel function represents the weight given to the observation P_i based on distance between P and P_i. One can see from Equation (8. 2) that the kernel estimator is a convolution estimator that forms a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. This is illustrated in Figure 8.3. The kernel function, K(.), prescribes the relative weights, while h prescribes the range of

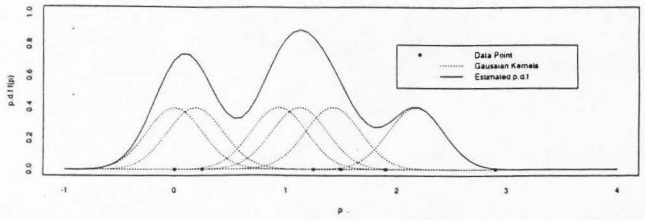


Figure 8.3. Example of kernel density estimation using 5 data points with Gaussian Kernel, h = 0.5

data values over which the average is computed. The PDF of wet day precipitation $\hat{f}(P)$ is obtained by applying a kernel density estimator to log transformed data. Note that most of the data of wet day precipitation is concentrated near the lower boundary (i.e., 0.). This is a problem for kernel density estimation methods since modifications to kernel density estimate are necessitated within a bandwidth of the boundary. The kernel centered at an observation that is within one bandwidth of the boundary extends past the boundary thereby leading to leakage of probability mass in the resulting density estimate (i.e., an increase in the bias of the estimate). This boundary problem can be avoided by applying the kernel density estimator to logarithmically transformed data. The resulting estimator is given as:

$$\hat{f}(P) = \frac{1}{n\,h\,P}\sum_{i=1}^{n} K\left(\frac{\log(P)-\log(P_i)}{h}\right) \qquad (8.4)$$

The Epanechnikov kernel is used and the bandwidth h is chosen for the log transformed data using the recursive approach of Sheather and Jones [1991] to minimize the mean integrated square error (MISE) of estimate of f(log(P)).

Silverman [1986] points out that in terms of mean square error of the estimated density, the kernel density estimator is more sensitive to the choice of the bandwidth than to that of the kernel, and the general practice is to choose a kernel and then seek an optimal estimate of the bandwidth h, under some criteria.

Univariate discrete variables. In this section, we present procedures for the estimation of the univariate probability mass functions for discrete variables (such as wet spell lengths w, dry spell lengths d). We recommend the discrete kernel (DK) estimator developed in Rajagopalan and Lall [in press]. The DK estimator for the PMF $\hat{f}(L)$, where L is either w or d, and n is the corresponding sample size, is given as:

$$\hat{f}(L) = \sum_{j=1}^{L_{max}} K_d\left(\frac{L-j}{h}\right) \tilde{\alpha}_j \qquad (8.5)$$

where $\tilde{\alpha}_j$ is the sample relative frequency (n_j/n) of spell length j, n_j is the number of spells of length j, L_{max} is the maximum observed spell length (note that $\sum_{j=1}^{L_{max}} \tilde{\alpha}_j = 1$), $K_d(.)$ is a discrete kernel function, and L, j, and h are positive integers. The kernel function $K_d(.)$ is given as:

$$K_d(t) = at_j^2 + b \quad \text{for } |t| \leq 1 \qquad (8.6)$$

The expressions for a and b for the interior of the domain, $L > h+1$ and the boundary region $L < h$, are developed in Rajagopalan and Lall [in press].

The bandwidth h is estimated by minimizing a least squares cross validation (LSCV) function given as:

$$LSCV(h) = \sum_{j=1}^{L_{max}} (\hat{f}(j))^2 - 2\sum_{j=1}^{L_{max}} \hat{f}_{-j}(j)\,\tilde{\alpha}_j \qquad (8.7)$$

where, $\hat{f}_{-j}(j)$ is the estimate of the PMFof spell length j, formed by dropping all the spells of length j from the data. This method has been shown by Hall and Titterington [1987] to automatically adapt the estimator to an extreme range of sparseness types. Monte Carlo results showing the effectiveness of the DK estimator with bandwidth selected by LSCV are presented in Rajagopalan and Lall [1995].

Multivariate continuous variables. By extending the idea of the kernel density estimator for univariate continuous variables, a kernel density estimate of the multivariate PDF of a vector **y** is defined as [Silverman, 1986]:

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} K(u) \qquad (8.8)$$

where $u = \dfrac{(\mathbf{y}-\mathbf{y}_i)^T S^{-1}(\mathbf{y}-\mathbf{y}_i)}{h^2}$, and K(u) is a multivariate Gaussian kernel function. $\mathbf{y} = [y_1, y_2, .., y_d]^T$ denotes the d dimensional random vector whose density is being estimated with $\mathbf{y}_i = [y_{1i}, y_{2i}, .., y_{di}]^T$ i = 1 to n the sample values of **y**, n is the number of sample vectors, h is a bandwidth and S is the sample covariance matrix. The Gaussian kernel function used is given as:

$$K(u) = \frac{1}{(2\pi)^{d/2} det(S)^{1/2}\, h^d} \exp(-u/2) \qquad (8.9)$$

Just as in the univariate case described in the earlier section, K(u) represents the weight given to an observation $\mathbf{y}_i$ that is based on distance between $\mathbf{y}$, and $\mathbf{y}_i$. The distance used here is the Euclidean distance modified to recognize the covariance of the $\mathbf{y}$. It can be seen that the estimator in Equation (8.8) is similar to the univariate estimator in Equation (8.2) since it is a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. Here too the kernel function K(.) prescribes the relative weights, h prescribes the range of data values over which the average is computed, and the covariance S provides the orientation. of the weight function.

Here, we chose the bandwidth h as the one that minimizes mean integrated square error in $\hat{f}(\mathbf{y})$ if the underlying distribution is assumed to be multivariate Gaussian. Silverman [1986] gives an appropriate h to use for a multivariate Gaussian PDF. using the Gaussian kernel as:

$$h = \{(4/(2d+1))^{1/(d+4)}\} n^{-1/(d+4)} \qquad (8.10)$$

Here n is the number of observations and d is the dimension. As the dimension d increases, h also increases. This happens because in higher dimensions large regions of high density may be completely devoid of observations in a sample of moderate size. The bandwidth in such a situation has to be bigger to cover these large regions.

The above choice of bandwidth is optimal for PDFs that are near Gaussian and is an adequate choice for many cases [Silverman, 1986]. Cross validation [see Sain et al., 1994] or plug-in methods [see Wand and Jones, 1994] could be used here to choose h as in the wet/dry spell model. However, this increases the computational burden substantially. Recall that the parametric approaches often assume a Gaussian distribution. In a Bayesian context, using this bandwidth can be thought of as developing a posterior kernel density

estimate with a Gaussian prior. The resulting tail behavior and degree of smoothing supplied will be consistent with an underlying Gaussian PDF, with some adaption to local features.

In the Bootstrap context we have a region that each observation $\mathbf{y}_i$ represents. The orientation and shape of the region are given by the scaling factor hS and the kernel function K(u). Resampling from the kernel density estimate entails picking a point $\mathbf{y}_i$ uniformly in $[\mathbf{y}_1,..,\mathbf{y}_n]$ and then simulating from the kernel K(u), i.e., $N(\mathbf{y}_i,h^2S)$. We extend this approach formally for simulation from a multivariate conditional PDF in the following section.

Kernel Density Estimation of Multivariate Conditional PDF

For the simulation of interest here an estimate of the conditional PDF $f(\mathbf{x}_t \mid \mathbf{V}_t = \mathbf{V}^*)$ is needed. The strategy used here is similar to the one used by Sharma et al. [1995] for streamflow simulation. Applying the estimator in Equation (8.8) to the conditional PDF in Equation (8.1) with sample vectors $\mathbf{x}_i=[\mathbf{x}_t, \mathbf{x}_{t-1},\mathbf{P}_t]_i$ denoted as $[\mathbf{x}_i,\mathbf{V}_i]$ we get:

$$f(\mathbf{x}_t \mid \mathbf{V}_t = \mathbf{V}^*) = \frac{1}{nh^d} \frac{1}{f_v(\mathbf{V}^*)} \sum_{i=1}^{n} \frac{1}{\det(S)^{1/2}} K\left(\frac{[\mathbf{x}_t-\mathbf{x}_i;\ (\mathbf{V}^*-\mathbf{V}_i)^T]S^{-1}\begin{bmatrix} \mathbf{x}_t-\mathbf{x}_i \\ \mathbf{V}^*-\mathbf{V}_i \end{bmatrix}}{h^2}\right) \qquad (8.11)$$

where S is the dg by dg covariance matrix of the vector $(\mathbf{x}_i,\mathbf{V}_i)$ estimated from historical data. Let the matrix S be partitioned as:

$$S = \begin{bmatrix} S_x & S_{xv}^T \\ S_{xv} & S_v \end{bmatrix} \qquad (8.12)$$

where S_X is the d by d covariance matrix of $\mathbf{x}$, S_V is the d' by d' covariance matrix of $\mathbf{V}$, and S_{XV} is the d by d' cross covariance between $\mathbf{x}$ and $\mathbf{V}$. Using the Gaussian kernel function (i.e., Equation 8.9) Equation (8.11) can be reduced to a weighted sum of Gaussian functions:

$$f(\mathbf{x}_t \mid \mathbf{V}_t = \mathbf{V}^*) = \sum_{i=1}^{n} w_i N(\mathbf{b}_i, \mathbf{c}_i) \tag{8.13}$$

where

$$w_i = w'_i / \sum_{i=1}^{n} w'_i \,,\; w'_i = \exp(-a_i/2);\; a_i = \frac{([\mathbf{V}^* - \mathbf{V}_i]^T [S_v]^{-1} [\mathbf{V}^* - \mathbf{V}_i])}{h^2}; \tag{8.14}$$

$$\mathbf{b}_i = \mathbf{x}_i + ([\mathbf{V}^* - \mathbf{V}_i]^T)[Sv]^{-1}[Sxv]\,;\quad \mathbf{c} = h^2 (S_X - S_{XV}{}^T S_V{}^{-1} S_{XV}) \tag{8.15}$$

Note that $\sum_{i=1}^{n} w_i = 1$

From Equation (8.13) we see that the conditional PDF reduces to a weighted sum of Gaussian functions. It can be thought of as a slice through a multivariate density function, estimated as a weighted sum of slices with the same orientation through the kernels placed on each observation. Simulation from the conditional PDF can be achieved by picking a point $\mathbf{x}_i$ with probability w_i, then sampling from $N(\mathbf{b}_i, \mathbf{c})$.

NP Simulation Algorithm

The simulation proceeds as:

1. Simulate precipitation for all the days of the year from the wet/dry spell model

2. Estimate the NP model parameters (i.e., bandwidth h and the covariance matrix S) from the data for each season.

3. At the start of each period of interest, initialize t=0, $\mathbf{x}_t$ = one of the historical observation randomly selected.

4. Generate $\mathbf{x}_t$ sequentially (day by day) from $f(\mathbf{x}_t \mid \mathbf{V}_t)$, where the conditioning vector $\mathbf{V}_t$ consists of the previous day's vector $\mathbf{x}_{t-1}$ and the current day's generated precipitation P_t
(i.e., $\mathbf{V}_t = [\mathbf{x}_{t-1}, P_t]$) as:

i. Estimate weights (w_i) associated with each data point ($\mathbf{x}_i$) (Equation 8.14)

ii. Resample an index i using w_i (i = 1,..,n) as probabilities.point $\mathbf{x}_i$ and $\mathbf{V}_t$ (Equation 8.15)

iv. Generate vector $\mathbf{x}_t = \mathbf{b}_i + \varepsilon$ where ε is from a multivariate normal distribution with mean **0** and variance **c** [see Devroye, 1986]

5. Recover $\mathbf{z}_t$ as $\mathbf{z}_t = \mathbf{x}_t * [\mathbf{STD}]_k + [\overline{\mathbf{X}}]_k$ where k is the calender day corresponding to day t.

6. At the start of a new simulation go to step 3.

Model Application

To demonstrate the utility of the resampling model for generation of daily weather variables, the model was applied to daily weather data from the station Salt Lake City in Utah. Thirty years of daily weather data were available from the period 1961-1991. Salt Lake City is at 40°46' N latitude, 111° 58' W longitude and at an elevation of 1288 m. Most of the precipitation comes in the form of winter snow. Rainfall occurs mainly in spring, with some in fall.

We shall first outline the experiment design and then some measures of performance used to judge the utility of the model.

Experiment design

Our purpose here is to test the utility of the NP generation scheme. The main steps involved in accomplishing this are

1. Daily precipitation is generated from the wet/dry spell model.

2. The other variables are generated following the simulation algorithm described in the previous section.

3. Twenty-five synthetic records of 30 years each (i.e., the historical record length) are simulated using the NP model.

4. The statistics of interest (described below) are computed for each simulated record, for each period, and are compared to statistics of the historical record using boxplots.

Performance measures

The following statistics were considered to be of interest in comparing the historical record and the NP simulated record of other weather variables.

Moments:

1. Mean of each variable for each season.

2. Standard deviation of each variable for each season.

3. Skew of each variable for each season.

4. Coefficient of variation of each variable for each season.

Relative Frequencies:

5. 25% quantile of each variable for each season.

6. 75% quantile of each variable for each season.

Dependence:

7. Cross correlation on any given day between the variables for each season.

8. Lag-1 daily cross correlation between the variables for each season.

9. Lag-1 daily correlation of each variable for each season.

Results

The statistics of interest calculated from the simulations are compared with those for the historical record using boxplots. A box in the boxplots (e.g., Figure 8.4) indicates the interquartile range of the statistic computed from twenty-five simulations, and the line in the middle of the box indicates the median simulated value. The solid lines correspond to the statistic of the historical record. The boxplots show the range of variation in the statistics from the simulations and also show the capability of the simulations to reproduce historical statistics.

Figures 8.4 through 8.7 show the boxplots of moments and relative frequency measures of Solar Radiation, Maximum Temperature, Minimum Temperature, and Average Dew Point Temperature, respectively. It can be seen that the historical values of mean, and the quantiles are well reproduced, while standard deviation, coefficient of skew, and coefficient of variation are not quite well reproduced. This is to be expected as the kernel methods inflate the variance by a factor equal to $(1+h^2)$ [see Silverman, 1986], which in turn effects the skew and the coefficient of variation. This inflation can be corrected through an appropriate scaling of the random terms during simulation [see Silverman, 1986]. However, it may be desirable to have to have a slight increase in the variance of the simulations as compared to that of the historical.

Illustrative statistics of wet spell lengths, dry spell lengths and wet day precipitation for the simulations from the wet/dry spell model are also estimated and are shown in Figures 8.8, 8.9, and 8.10, respectively. Figure 8.8 shows the boxplots of average wet

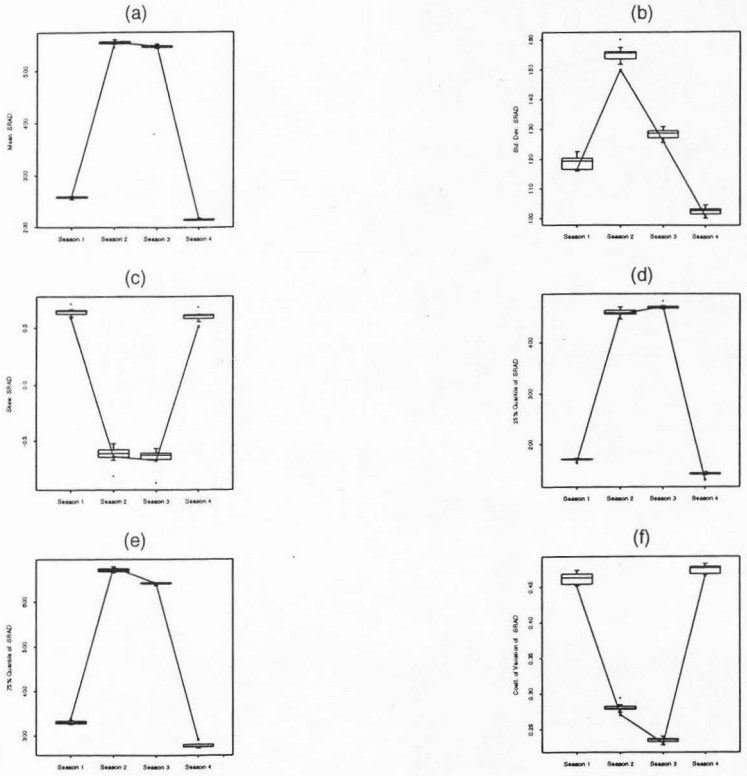


Figure 8.4. Boxplots of statistics of SRAD (a) mean SRAD, (b) standard deviation of SRAD, (c) skew of SRAD, (d) 25% quantile of SRAD, (e) 75% quantile of SRAD, and (e) coefficient of variation of SRAD for model simulations along with the historical values for the four seasons.

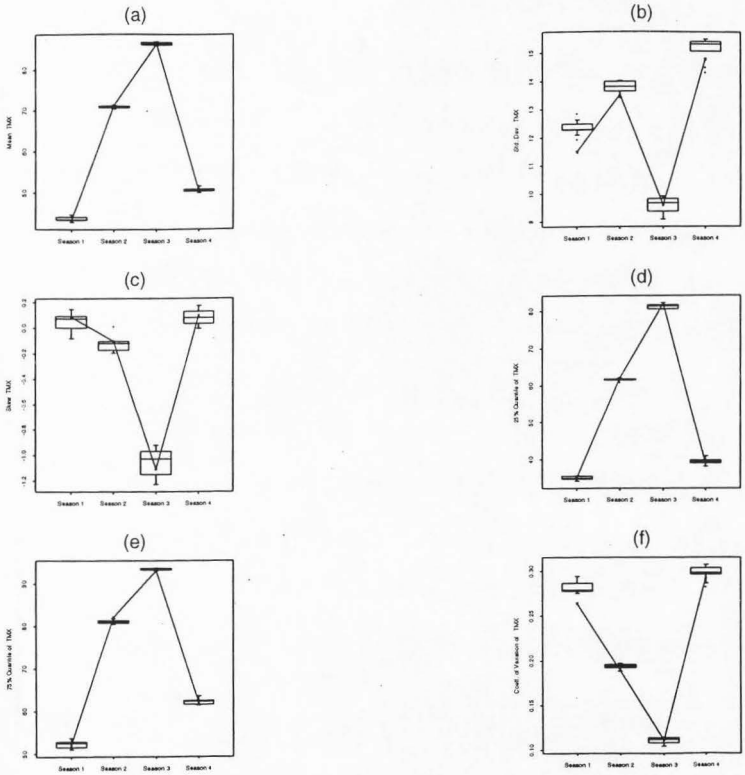


Figure 8.5. Boxplots of statistics of TMX (a) mean, (b) standard deviation, (c) skew, (d) 25% quantile, (e) 75% quantile, and (e) coefficient of variation of TMX for model simulations along with the historical values for the four seasons.

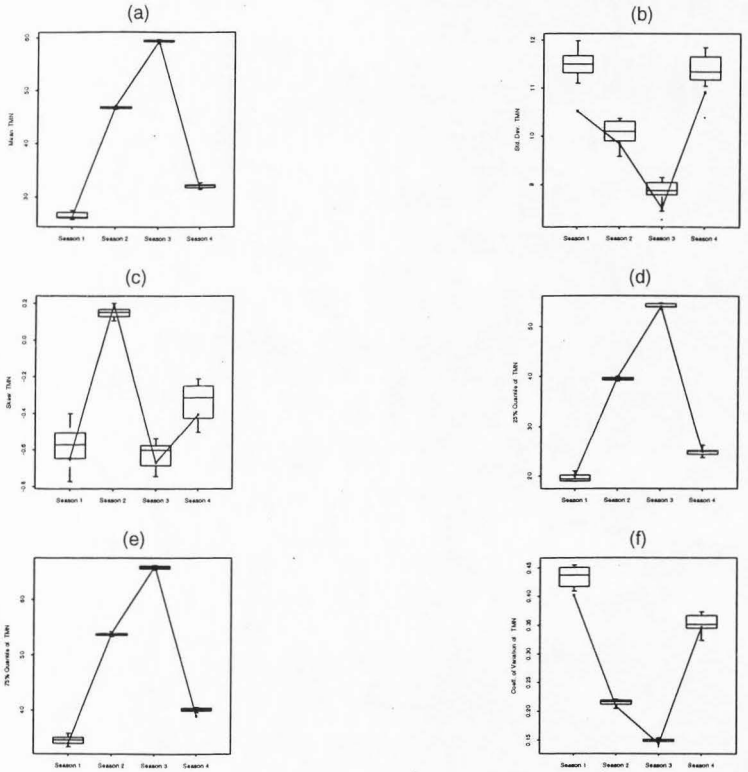


Figure 8.6. Boxplots of statistics of TMN (a) mean, (b) standard deviation, (c) skew, (d) 25% quantile, (e) 75% quantile, and (e) coefficient of variation for model simulations along with the historical values for the four seasons.

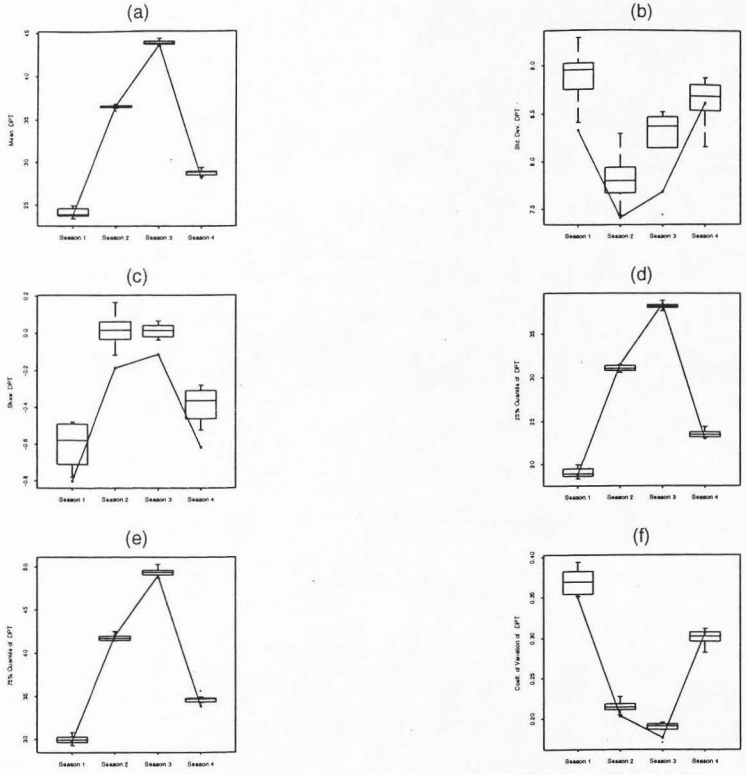


Figure 8.7. Boxplots of statistics of DPT (a) mean DPT, (b) standard deviation of DPT, (c) skew of DPT, (d) 25% quantile of DPT, (e) 75% quantile of DPT, and (e) coefficient of variation of DPT for model simulations along with the historical values for the four seasons.

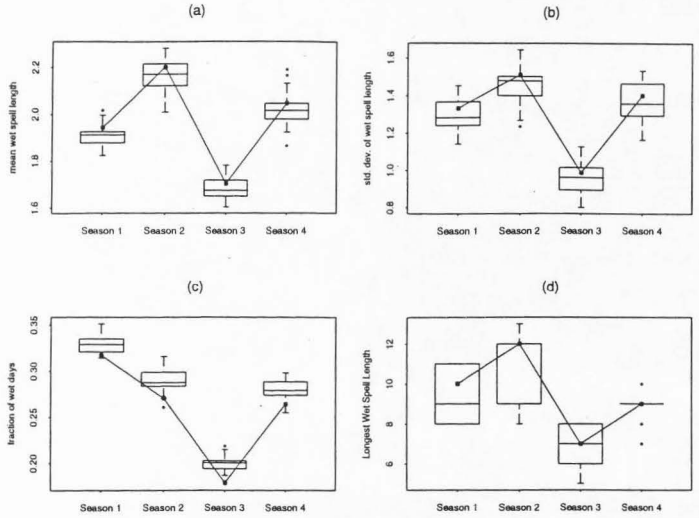


Figure 8.8. Boxplots of statistics of wet spell length (a) mean wet spell length, (b) standard deviation of wet spell length, (c) fraction of wet days, and (d) longest wet spell length for simulations from wet/dry spell model along with the historical values for the four seasons.

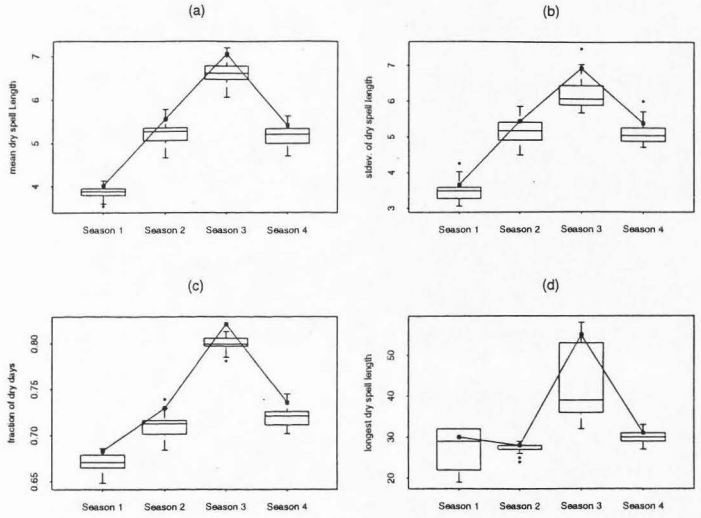


Figure 8.9. Boxplots of statistics of dry spell length (a) mean wet spell length, (b) standard deviation of wet spell length, (c) fraction of wet days, and (d) longest wet spell length for simulations from wet/dry spell model along with the historical values for the four seasons.

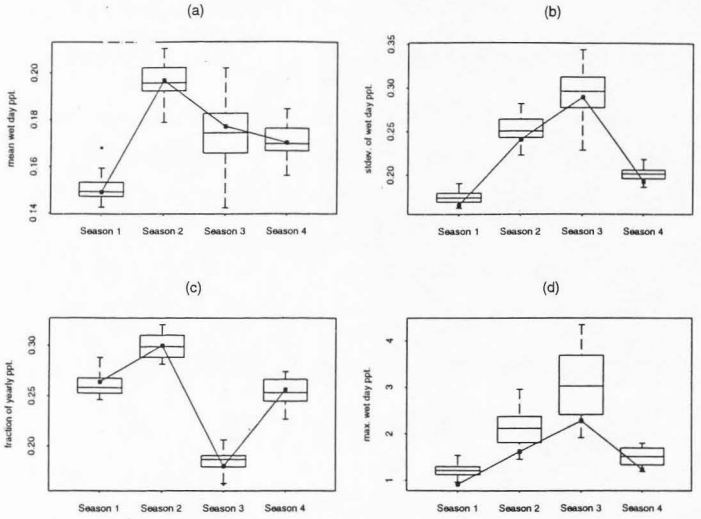


Figure 8.10. Boxplots of statistics of wet day precipitation (a) mean wet day precipitation, (b) standard deviation of wet day precipitation, (c) fraction of yearly wet day precipitation, and (d) maximum wet day precipitation for simulations from wet/dry spell model along with the historical values for the four seasons.

spell length, standard deviation of wet spell length, fraction of wet days, and length of longest wet spell length for each season. Figure 8.9 shows the boxplots of these statistics of the dry spell length. Figure 8.10 shows the boxplots of average wet day precipitation, standard deviation of wet day precipitation, percentage of yearly precipitation in each season. The boxplots in Figures 8.8, 8.9 and 8.10 show that the historical statistics are reproduced well by the simulations.

Figures 8.11 and 8.12 show the boxplots of the lag-0 cross correlation and lag-1 cross correlation between the variables. Figure 8.13 shows the lag-1 auto correlation of each variable for each of the four seasons. The correlations from the simulations and the historical correlations seem to be different in a number of cases. The correlations that are reproduced most poorly are the ones with precipitation. While the correlations of the variables with precipitation are very small as can be seen from these figures and in many cases seem insignificant.

One reason for this mismatch of the correlations is that the precipitation is supplied externally from the wet/dry spell model. As a result the covariance between $\mathbf{x}_{t-1}$ and P_t need not correspond to that of the historical covariance between them. This introduces a bias in the conditioning plane from which $\mathbf{x}_t$ is generated and results in a mismatch of the correlations. To verify this, we made twenty five simulations without conditioning on precipitation (i.e. simulated $\mathbf{x}_t$ from $f(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ where both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are of dimension 5). The correlations from this simulation are shown in Figure 8.14,8.15 and 8.16 respectively. It can be seen from these three figures that the correlations are well reproduced, which strongly suggests that the conditioning on the precipitation is the reason for mismatch of correlations in Figures 8.11, 8.12 and 8.13.

One way to get around this problem is to generate the precipitation also in the multivariate model, i.e. simulate $\mathbf{x}_t$ from $f(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ where both $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$ are of dimension 6. This should reproduce the correlation statistics. However, negative values

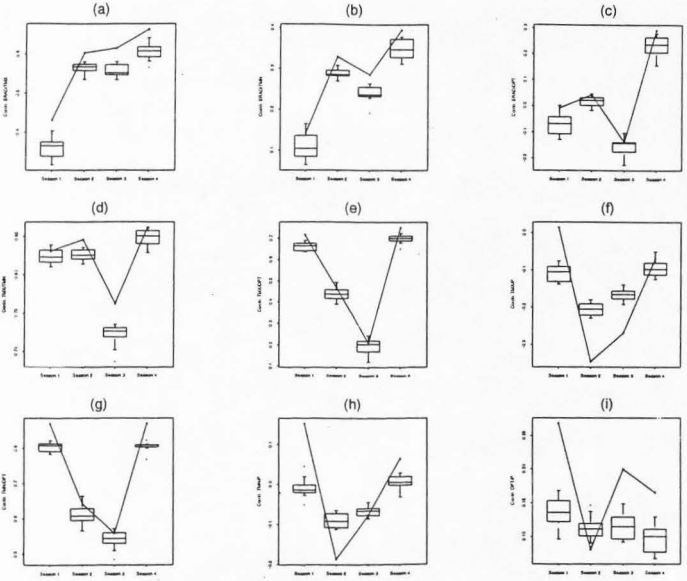


Figure 8.11. Boxplots of Lag-0 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, (f) TMX and P, (g) TMN and DPT, (h) TMN and P, and (i) DPT and P for model simulations along with the historical values for the four seasons.

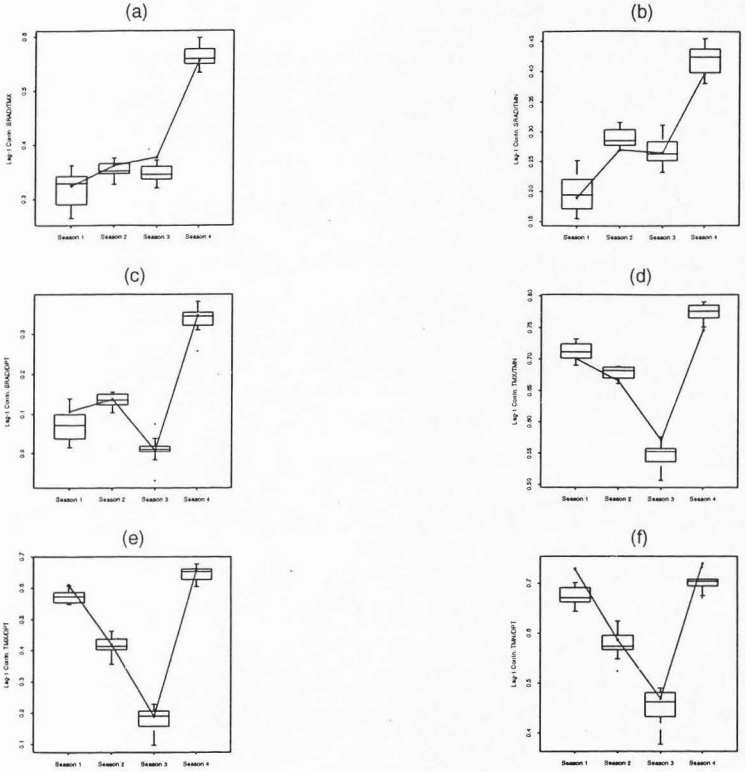


Figure 8.12. Boxplots of Lag-1 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations along with the historical values for the four seasons.

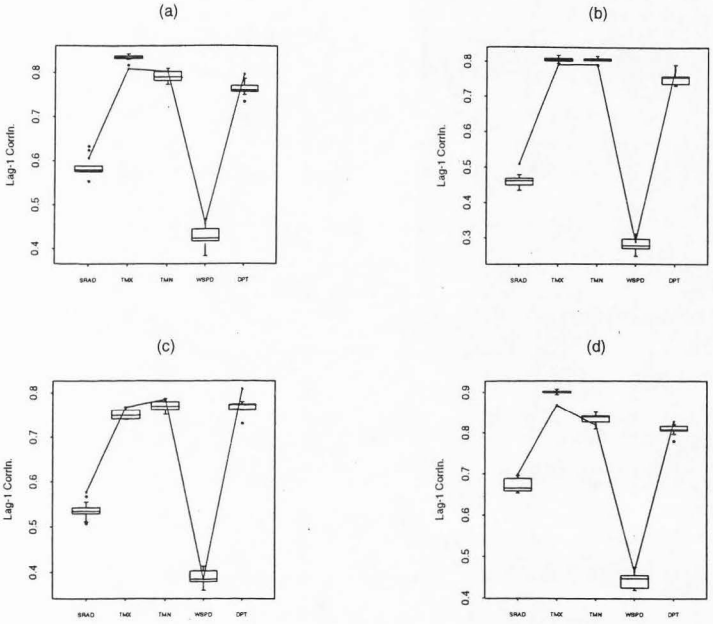


Figure 8.13. Boxplots of Lag-1 Auto Correlation of SRAD,TMX,TMN, WSPD, and DPT for (a) season 1, (b) season 2, (c) season 3, and (d) season 4 for model simulations along with the historical values.

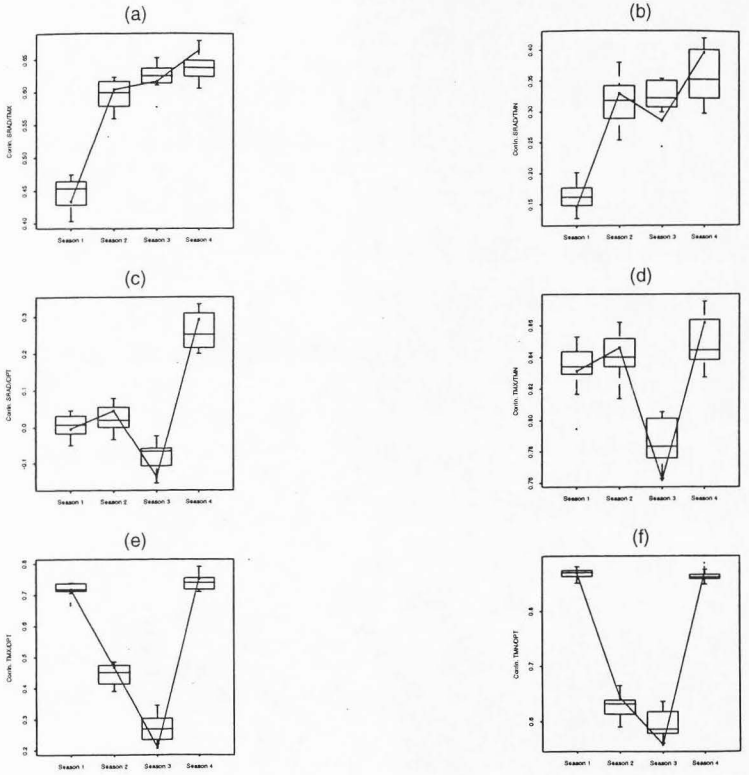


Figure 8.14. Boxplots of Lag-0 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations (without conditioning on precipitation) along with the historical values for the four seasons.

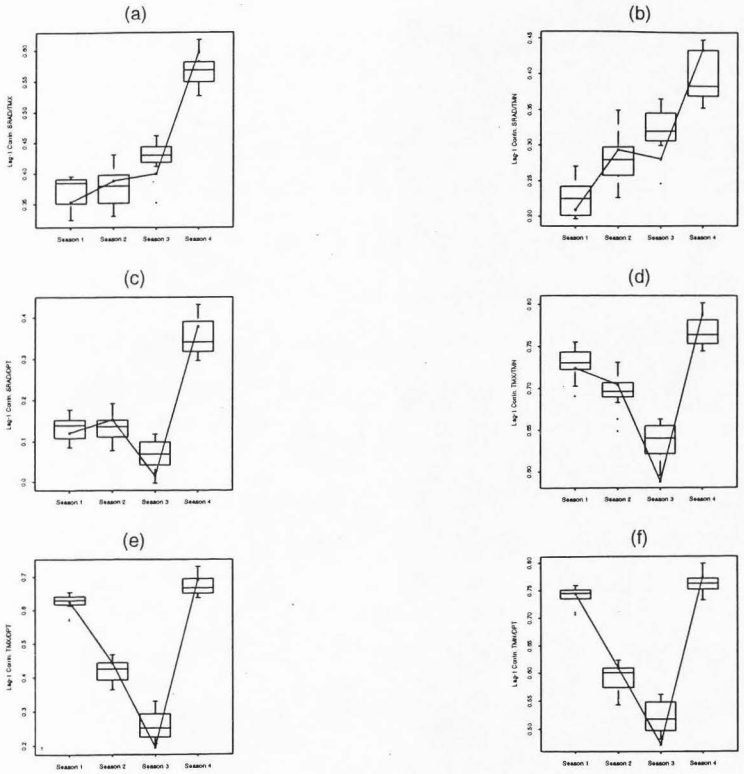


Figure 8.15. Boxplots of Lag-1 cross correlation between (a) SRAD and TMX, (b) SRAD and TMN, (c) SRAD and DPT, (d) TMX and TMN, (e) TMX and DPT, and (f) TMN and DPT for model simulations (without conditioning on precipitation) along with the historical values for the four seasons.

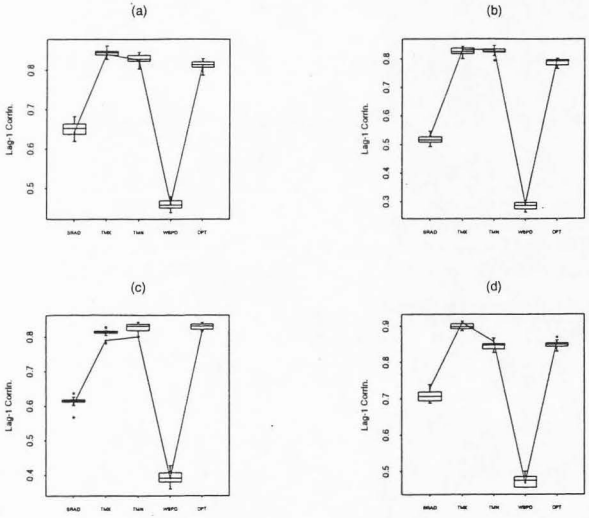


Figure 8.16. Boxplots of Lag-1 Auto Correlation of SRAD,TMX,TMN, WSPD, and DPT for (a) Season 1, (b) season 2, (c) season 3 and (d) season 4 for model simulations (without conditioning on precipitation) along with the historical values.

for precipitation may then be simulated. Since most of the precipitation is concentrated near 0., simulating precipitation also along with the other variables may lead to oversmoothing of the mode of the precipitation density.

Summary and Conclusions

A multivariate nonparametric model that aims at capturing dependence up to lag-1 was presented and illustrated. The simulations are made from the conditional PDF estimated from the data using kernel density estimators. The kernel estimators (being local average estimators) have the advantage of readily admitting arbitrary probability densities without requiring that they be hypothesized or formally identified. Broader dependence structures can be consequently considered. The need to choose/justify and fit the best PDF is side stepped.

The bandwidth is the key parameter in the NP model, because it determines the degree of smoothness that will be imparted to the PDF. The larger the bandwidth the smoother the PDF and vice versa. Choosing h automatically using cross validation [see Sain et al., 1994] or plug-in approaches [see Wand and Jones, 1994] from the data would be more appropriate than the choice used here. However, the additional variance in the choice of h induced by such an estimation process may detract from its use where the primary purpose is to resample the data. Bandwidth selection methods are undergoing continuous improvement. We expect to implement more formal selection procedures in due course. One could also use a local covariance matrix estimated at each data point using a few neighbors of that point (i.e., S_i instead of S in Equation 8.8). Sharma et al.[1995] use this method for streamflow simulation.

Another problem with simulations is the boundary effect. For the variables that are bounded (e.g., Solar Radiation and Precipitation), values that violate the bounds could be

generated. Typically these are censored to the bound. This may introduce a bias in the simulations. Procedures to better address this problem in univariate situations are described in Müller [1992] and Rajagopalan et al.[1995], but for multivariate situations effective methods are yet to be developed.

We chose to apply the NP model on a seasonal time scale, because the precipitation model that was used to drive the NP model is a seasonal model. However, we checked the results of the seasonal NP model at a monthly time scale, and found the performance to be similar (results are not presented here).

The NP model developed here underscores our growing conviction that nonparametric techniques have an important role to play in improving the synthesis of hydrologic time series. They can capture dependence structure present in the data, without imposing arbitrary distributional assumptions, and produce synthetic sequences that are statistically similar to the historic sequence. The idea of resampling the data with appropriate perturbation of each value while maintaining selected dependance characteristics (or data sequencing) is easy to accept as a practical matter.

References

Bruhn, J.A., W.E., Fry, and G.W., Fick, Simulation of daily weather data using theoretical probability distributions, *Journal of Applied Meteorology*, 19(9), 1029-1036, 1980.

Devroye, L., *Non-Uniform Random Variate Generation*. Springer-Verlag, New York, 1986.

Efron, B., Bootstrap methods: Another look at the Jackknife, *Annals of Statistics*. 7, 1-26, 1979.

Hall, P., and D.M. Titterington, On smoothing sparse multinomial data, *Australian Journal of Statistics*, 29(1), 19-37, 1987.

Jones, W., Rex, R.C., and D.E. Threadgill, A simulated environmental model of temperature, evaporation, rainfall, and soil moisture, *Transactions of the ASAE*, 366-372, 1972.

Kunsch, H.R., The Jackknife and the Bootstrap for General Stationary Observations, *Annals of Statistics*, 17, 1217-1241, 1989.

Lall, U., B. Rajagopalan, and D.G. Tarboton, A nonparametric wet/dry spell model for resampling daily precipitation, Working Paper WP-95-HWR-UL/006, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Lane, L.J., and M.A. Nearing, USDA - *Water Erosion Prediction Project: Hill slope Profile Model Documentation*, NSERL Report No.2, National Soil Erosion Research Laboratory, USDA-Agricultural Research Service, W. Lafayette, Indiana 47907, 1989.

Liu, R.Y., and K. Singh, *Using iid Bootstrap Inference for some Non-iid Models*, Preprint. Department of Statistics, Rutgers University, 1988.

Müller, H.G., Smooth optimum kernel estimators near endpoints, *Biometrika.*, 78(3), 521-530, 1992.

Nicks, A.D, and J.F. Harp, Stochastic generation of temperature and solar radiation data. *Journal of Hydrology*, 48, 1-7, 1980.

Rajagopalan, B., and U. Lall, A kernel estimator for discrete distributions, *Journal of Nonparametric Statistics*, (in press)

Rajagopalan, B., U. Lall, and D.G. Tarboton, Evaluation of kernel density estimation methods for daily precipitation resampling, Working Paper WP-95-HWR-UL/007, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Richardson, C.W., Stochastic simulation of daily precipitation, temperature and solar radiation. *Water Resources Research*, 17(1), 182-190, 1981.

Rosenblatt, M., Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*. 27, 832-837, 1956.

Sain, S.R., A.B. Keith, and D.W. Scott, Cross-validation of multivariate densities, *Journal of American Statistical Association.*, 89(427), 807-817, 1994.

Scott, D.W., *Multivariate Density Estimation*. John Wiley, New York, 1992.

Sharma, A, D.G. Tarboton, and U. Lall, Streaflow simulation: A nonparametric approach, Working Paper WP-95-HWR-UL/011, in Utah Water Research Laboratory, Utah State University, Logan, UT, 1995.

Sheather, S.J., and M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, B. 53, 683-690, 1991.

Silverman, B.W., *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, 1986.

Wand, M.P., and M.C. Jones, Multivariate Plug-In bandwidth selection, *Computational Statistics*, 9, 97-116, 1994.

CHAPTER IX
GENERAL SUMMARY

Results of the Study

This research developed nonparametric resampling procedures for simulation of daily precipitation and a suite of other weather variables. The procedures avoid prior assumptions as to the parametric forms of the underlying probability models. Consequently these procedures can be applied uniformly across regions/sites. A nonparametric seasonal wet/dry spell (NSS) model was developed for simulating the daily precipitation, and a multivariate nonparametric resampling scheme for simulating the daily values of other weather variables.

In the course of development of the wet/dry spell model, various nonparametric density estimation techniques for both discrete and continuous variables were compared, and a new discrete nonparametric estimator for estimation of discrete probabilities was developed.

Seasonal variation in precipitation was studied through the nonparametric estimation of the Poisson process rate, leading to a smooth representation of the occurrence process. Significant changes in seasonality were found in stations going from north to south along a meridional transect in the western U.S.

Recognizing that precipitation is one of the key variables that triggers several hydrologic processes, and also the increasing evidence that precipitation is strongly driven by large scale climatic fluctuations, an attempt to better understand the climatic fluctuations and their effect on precipitation patterns was also made in this study. Spectral analysis using the nonparametric multitaper method (MTM) was performed on monthly precipitation amounts at a a few stations along a meridional transect in the western U.S. revealed strong

signals in the 3-7 yr and 2-yr frequency band that were consistent with the atmospheric oscillations such as El Niño/Southern Oscillation (ENSO) and Quasi-biennial Osciallation (QBO). Significant spectral coherence was also found with the atmospheric indices e.g., Southern Oscillation Index (SOI) and Central Northern Pacific (CNP).

Results from the seasonality and the spectral analysis motivated us to seek a precipitation model that obviated the need to divide the year into seasons, which were found to vary from location to location. A nonhomogeneous Markov (NM) chain model that does not require the year to be divided into seasons was developed. The transition probabilities for each day were estimated using the discrete nonparametric estimator that was developed. The NM model is relatively parsimonious and locally adaptive. One objective of the study was to develop a multivariate resampling scheme for a suite of weather variables that would consider the day to day dependence and the precipitation status. A multivariate resampling procedure that simulates a vector of weather variables for any given day, conditioned on the vector of variables of the previous day and the precipitation status of the current day was developed. This approach is likened to a smoothed bootstrap, wherein the nature and amount of smoothness is provided by the multivariate kernel density estimators.

Precipitation Models

In the course of this research, two models for simulating daily precipitation were developed, the nonparametric seasonal wet/dry spell (NSS) model and the nonhomogeneous Markov (NM) model. A brief discussion regarding the nature and attributes of these two models is presented below.

NM model

The simplest traditional approach is the Markov chain model. Typically a first-

order, two-state (viz., a day is wet or dry) Markov chain is considered and the transition probabilities between the states are estimated from the data. The transition probabilities are assumed stationary over a chosen period (usually a a month) and hence this is one of the major drawbacks of this traditional approach. As a result they cannot reproduce long term persistence and clustering of events readily. Despite this the Markov chain model is attractive because of its largely nonparametric nature, ease of application and interpretability, relative parsimony, and well developed literature.

Nonparametric estimation methods readily offer to extend the traditional homogeneous Markov model to a nonhomogeneous situation. This admits first-order dependence parsimoniously. In the light of changing seasonality in precipitation, assumptions of homogeneity can be hard to justify; in such situations the NM that avoids the seasonality issue is better suited and is to be recommended.

Strong signals of low frequency variability have been seen in precipitation records at many sites. This indicates nonstationary behavior at the interannual time scales, contrary to the general assumption that precipitation process is stationary from year to year. The NM model can capture this interannual variability; however, inclusion of atmospheric indices (such as SOI, CNP, etc.) that quantify some of the know low frequency oscillations will improve the performance of the NM model. Attempts have been made to address this heterogeneous nature, by pursuing a hierarchical Markov chain model that considers "weather types" to describe the daily precipitation process, but the parameter estimation can be cumbersome. The NM model, on the other hand, can be modified easily in a parsimonious manner to accommodate this. Future work in this regard is needed.

NSS model

The main advantage of this representation is that it allows direct consideration of a composite precipitation event, rather than its discontinuous truncation into arbitrary daily

segments. This model can capture the clustering of events rather better than the traditional Markov chain. As the name suggests, a wet (dry) period is always followed by a dry (wet) period (i.e. no transition to the same state is possible).

The primary difficulties with this approach are (1) the inability to discriminate between rainfall events at short time scales, (2) the possible need for disaggregation of wet spell precipitation into daily or event precipitation, (3) justification of the independence between the dry and wet period lengths, and (4) effective reduction in the sample size by considering spells rather than days. The other objectionable aspect is the parametric specifications for probability distributions, and assumptions of independence of spells, especially in the light of heterogeneous nature of the data.

However, this structure is plausible and the NSS model developed in Chapter II addresses some of these difficulties.That model resamples precipitation traces under the assumption of stationarity within the season. If event characteristics are of interest, such as the wet and dry spells are of interest for planning crop/water management in arid regions, the NSS model is to be recommended. If the wet and dry spells are strongly correlated, the NSS model is highly recommended as it allows for conditional resampling. Also, the nature of the NSS model allows a rich structure for the wet and dry spell distribution.

However, the division of the year into fixed seasons is restrictive, especially in the event of significant changes in precipitation seasonality identified in Chapter V. This could be addressed by having a moving window, instead of the fixed season, and then estimating the probability density functions using the spells captured by this moving window, thereby capturing the nonstationarity. Data limitations detract from this moving window approach.

In closing, the two models developed here provide a very general framework for precipitation modeling, unlike the traditional approaches; however, future research to work around the problems mentioned in these two models is needed.

VITA

Rajagopalan Balaji

Research Interests: Statistical Climate modeling and its application to water related issues; nonparametric estimation of density and regression functions for time series analysis of climate data; nonlinear dynamical approach to climate modeling and forecasting.

Personal Data:

Born in Palayamkottai, (Tamil Nadu) India, December 4, 1967, son of Rajagopalan and Vagulavalli

Education:

B.Tech. in Civil Engineering, from Regional Engineering College, Kurukshetra, India, in 1989.
M.Tech. in Quality Reliability and Operations Research, from Indian Statistical Institute, Calcutta, in 1991.
PhD in Civil and Environmental Engineering with an emphasis in Stochastic Hydrology, from Utah State University, Logan, Utah, in 1995.

Experience:

Graduate Research Assistant (October 1991 - March 1995), Utah Water Research Laboratory,Utah State University, Logan, Utah.
Project training (April - July, 1991) on *Statistical Design of Experiments*, Alembic Chemicals Ltd., Baroda, India.
Practical training (June, July 1988) on *Structural Design*, Engineers India Ltd., Cochin, India.

Selected Publications and Conference Presentations:

Rajagopalan, B and U. Lall, A Kernel Estimator for Discrete Distributions, *Journal of Nonparametric Statistics* (in press), 1995.
Kshirsagar, M.M, B. Rajagopalan, B.and U. Lall, Optimal Parameter Estimation for Muskingum Routing with Ungauged Lateral Inflow, *Journal of Hydrology* (in press), 1995.
Rajagopalan, B., U. Lall, and D G. Tarboton, A Nonparametric Renewal Model for Modeling Daily Precipitation, published in the Proceedings of the *International Conference in Stochastic and Statistical Methods in Hydrology and Environmental Engg.*, Univ. of Waterloo, Waterloo, Canada, Time series anaslysis and forecasting, Ed by K. Hipel, Kluwer, 1994.
Rajagopalan, B., U. Lall, D.G. Tarboton and D. S. Bowles, Multivariate Nonparametric Simulation of Weather Variables, presented at *E.G.S, XIX General Assembly*, Grenoble, France, April 25-29, 1994.
Rajagopalan, B., and D. G. Tarboton, Understanding Complexity in the structure of rainfall, *Fractals*, 1(3), 606-628, 1993.
Rajagopalan, B., A.K. Sikka, D.S. Bowles, D.S., and Limaye, A.S., Spatial estimation techniques for precipitation analysis - application to a region in India, presented at *Intl. Conf. on Hydrology and Water Res.*,New Delhi, India, Dec. 20-22, 1993.