THE UNIVERSITY OF CALGARY

Analysis of Drought Time Series

by

Lei He

A DISSERTATION

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

September, 2007

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a Dissertation entitled "Analysis of Drought Time Series" submitted by Lei He in partial fulfillment of the requirements for the degree of Master of Science.

Gemai Chen

Supervisor, Dr. Gemai Chen
Department of Mathematics and Statistics

Xuewen Lu

Dr. Xuewen Lu
Department of Mathematics and Statistics

Daniel V Gordon

Dr. Daniel V. Gordon
Department of Economics

Sept 13, 2007

Date

Abstract

Severe droughts of the twentieth century have had large impacts on economics, society, and the environment, especially in the Great Plains. Droughts of the 1930s displaced up to 20% of humans in central North America and droughts in 1988-1989 caused agricultural losses of $35 billion US dollars. Past drought variability should be investigated to gain an improved understanding needed for society to anticipate and plan for droughts of the future. However, drought is not a phenomenon that can be easily measured directly. In this thesis, we study a quantitative way to define droughts based on 2000 years diatom-inferred salinity time series from three lakes in the Canadian Prairies. Through two simulation studies, we show that the Lomb-Scargle periodogram can be used to investigate the periodicities of droughts even when the drought time series are unevenly spaced and value-truncated, and we carry out this investigation for Humboldt Lake in Saskatchewan, Chauvin Lake in Alberta and Nora Lake in Manitoba.

Acknowledgments

I have been fortunate in having studied and worked among people in the Mathematics and Statistics Department who are very kind. They are always generous in lending a hand when I am in need.

I would first like to express my gratitude to my supervisor, Dr. Gemai Chen. When I first met Dr. Chen, he told me: "The type of person one is will determine the type of research he does.". It impressed me very much and it guided me through my whole graduate study. He helped me develop an intuition in Statistics and learn methods to solve problems. It is impossible to express with words the amount of knowledge I received from him.

I am very grateful to Dr. John Collins. At my most difficult time, he gave me endless and warm-hearted encouragement to help me stand up and keep going forward. Without his encouragement, I could not have finished my thesis. I also want to express my thanks to Dr. Peter Ehlers, who whenever I ask him always showed great patience to help me with R programming skills. I would like to say thanks to Ms. Joanne Mellard. She helped me feel at home and secure.

During my graduate studies, I received substantial financial support from the Department of Mathematics and Statistics and the Alberta Heritage Scholarship. Thank you!

Finally, I want to say thanks to my parents and sister. They have given me endless care, support and encouragement throughout my whole life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the 1930s, there was an extreme large-scale drought in North America and it was called 'Dust Bowl' (Laird et al., 1996). It reduced agricultural output up to 40%, contributed to topsoil degradation, caused widespread farm abandonment and displaced up to 20% of the regional population (Maybank et al., 1995). In the summer of 1988, there was another 'big' drought across United States. This drought lasted for several years (Trenberth et al., 1988) and caused agricultural losses of $35 billion USD (Woodhouse and Overpeck, 1998), despite compensatory increases in commodity values.

Since severe droughts have large impacts on economies, society, and the environment, the following questions become important: Are droughts periodic? If yes, how can we find their periods? When will the next drought arrive? How long will the next drought last? It is hard to get accurate answers, but it is valuable to spend time and money on investigating the behavior of droughts in order to understand more about droughts.

Drought is a normal, recurrent feature of climate, although many erroneously consider it a rare and random event. It occurs in virtually all climatic zones, but its characteristics vary significantly from one region to another. Drought is a temporary aberration; it differs from aridity, which is restricted to low rainfall regions and is a permanent feature of climate. Drought originates from a deficiency of precipitation over an extended period of time, usually a season or more. This deficiency results in

a water shortage for some activity, group, or environmental sector. Drought should be considered relative to some long-term average condition of balance between precipitation and evapotranspiration (i.e., evaporation + transpiration) in a particular area, a condition often perceived as "normal". It is also related to the timing (i.e., principal season of occurrence, delays in the start of the rainy season, occurrence of rains in relation to principal crop growth stages) and the effectiveness (i.e., rainfall intensity, number of rainfall events) of the rains. Other climatic factors such as high temperature, high wind, and low relative humidity are often associated with drought in many regions of the world and can significantly aggravate its severity. Research by Donald A. Wilhite, director of the National Drought Mitigation Center, U.S., and Michael H. Glantz, the National Center for Atmospheric Research, U.S., in the early 1980s uncovered more than 150 published definitions of drought. The definitions reflect differences in regions, needs, and disciplinary approaches. Wilhite and Glantz categorized the definitions in terms of four basic approaches to measuring drought: meteorological, hydrological, agricultural, and socioeconomic. The first three approaches deal with ways to measure drought as a physical phenomenon. The last deals with drought in terms of supply and demand, tracking the effects of water shortfall as it ripples through socioeconomic systems.

All these tell us that it is difficult to find a universal way to define droughts, especially quantitatively. It therefore brings many difficulties to studying droughts quantitatively. Many researchers have tried to find a way or method to study droughts quantitatively and to catch the characteristics of this phenomenon; however, they all studied droughts indirectly. For example, most researchers studied drought through studying tree rings (Stockton and Meko, 1983; Stahle and Cleaveland, 1988; Gon-

zalez et al., 2003). Smakhtin and Hughes (2007) studied droughts by using rain fall data and developed a new software package for automated estimation, display and analysis of various drought indices - continuous functions of precipitation that allow quantitative assessment of meteorological drought events. They use five different indices to quantitatively describe drought based on the data from Southern Asia. Laird et al. (1996, 2003) and Yu and Ito (1999) studied drought in the northern Great Plains of the United States by using Diatom-inferred salinity data. Yu and Ito (1999) also studied the relationship between solar-oscillation periods and drought periods and showed that solar minima are in phase with drought periods in the northern Great Plains (using data from Rice Lake in North Dakota) and cold periods in Greenland.

Even though the above drought studies are quantitative, the researchers did not define drought directly through the data they used. The reason is that drought is not a phenomenon like price, index, sales, rainfall, water level in a reservoir etc., that we can find a way to measure directly. In fact, it is not entirely clear how to measure drought directly to generate a time series that can reflect the history of droughts. Also, similar quantitative studies of drought do not seem to have been done for the Canadian Prairies. Therefore, in this thesis, we continue the project led by Professor Peter Leavitt (University of Regina) and Professor Gemai Chen (University of Regina and University of Calgary), in which droughts were defined directly using diatom-inferred salinity data for the Canadian Prairies. We use 2000-year paleoclimatic records from saline lakes and define droughts as any event in which the detrended diatom-inferred salinity exceeded the value recorded for Moon Lake in 1988-1989 in standardized unit, the last regional drought; see Chapter 2 for more

details.

Our main interest in this thesis is to explore the periodic information hidden in the drought history. The common technique used to study periodicities is spectral analysis by calculating the periodogram from the Fourier transformation of the time series and searching for sharp peaks in the periodogram. If the time series is evenly spaced, the periodogram calculation is simplified and can be quickly evaluated with the fast Fourier transform (FFT) (Priestley, 1981).

However, our drought time series are unevenly spaced time series. In order for FFT to be employed, we must first perform linear interpolation on our data. Unfortunately, interpolation leads to an underestimation of high frequency components in a spectrum independent of the employed interpolation scheme (Schulz and Stattegger, 1997).

Lomb (1976) developed a method for unevenly spaced time series by using least-squares fitting of sine curves of various periods to the time series data set. Subsequently, Scargle (1982) extended Lomb's work by defining the *Lomb-Scargle periodogram* using periodogram analysis approach and proved that the periodogram analysis is exactly equivalent to least-squares fitting of sine curves to the data. Horne and Baliunas (1986) rediscussed the normalization of the periodogram given by Scargle (1982) and studied some properties of the Lomb-Scargle periodogram. Press and Rybicki (1989) proposed a practical mathematical formulation of the Lomb-Scargle periodogram.

Even though the Lomb-Scargle periodogram can be used for unevenly spaced time series, we are not sure if we can apply it directly to our drought time series. Our drought time series are a bit unlike the usual unevenly spaced time series in

that they are value-truncated, unevenly spaced time series which are explained in detail in Chapter 2. Therefore, in this thesis, through simulation study in different situations, we demonstrate that the Lomb-Scargle periodogram may be used for our drought time series under some conditions.

The plan of this thesis is as follows. In Chapter 2, we describe the method of data collection, data cleaning and a quantitative definition of drought. In Chapter 3, we provide the theoretical background of *spectral analysis of time series* and the Lomb-Scargle periodogram. In Chapter 4, we conduct simulation studies and apply the Lomb-Scargle periodogram to study the drought time series defined in Chapter 2. In Chapter 5, we summarize the thesis and discuss some possible further work.

Chapter 2

Data

2.1 Data Collection

In central Canada, water-shortage is the largest single source of crop insurance losses, yet drought risk assessments are based on fewer than 50 years of data. Unlike most researchers using tree rings or rain fall data, in this thesis we use 2000 years diatom-inferred salinity time series from saline lakes to quantitatively define and study the periodic behavior of droughts for the Canadian Prairies.

High-resolution paleoclimate records were derived from analysis of fossil diatoms in saline lakes using standard paleoecological techniques (Fritz 1996, Gasse et al. 1997). These algae are common and abundant members of the flora of inland saline lakes whose distribution is strongly related to lakewater salinity (Cumming et al. 1995, Gasse et al. 1997) and whose fossil species composition allows quantitative reconstruction of past lake-water salinity (Fritz 1996, Laird et al. 1998). In the Canadian Prairies, where potential evaporation exceeds precipitation, lake level and lakewater salinity are regulated mainly by temperature and the balance between precipitation and evaporation, the same factors that control drought occurrence (Skinremi et al. 1996). During warmer or drier periods, lake levels decline in closed basins and dissolved salts concentrate. Conversely, cool wet climates result in high lake levels and dilution of concentrated brines. As a result, strong osmotic stresses imposed by changes in lakewater salinity are the main determinant of diatom species

composition in saline lakes (Cumming et al. 1995, Fritz 1996, Gasse et al. 1997).

Candidate lakes for climatic reconstructions were identified from aerial photographs (c.1940-1997) that demonstrated lake-level change in response to known climatic events. Short cores of lake sediment were then obtained from about 35 lakes, sectioned in 1-cm intervals, and analyzed for preservation of fossil diatoms and evidence of historical changes in saline fossil taxa during the 20^{th} century. Chauvin Lake, Alberta ($52\,°41.41'N$, $110\,°06.02'W$), Humboldt Lake, Saskatchewan ($52\,°08.5'N$, $105\,°0.648'W$), and Nora Lake (Lake 100), Manitoba ($50\,°28.30'N$, $99\,°56.19'W$), satisfied these criteria and were selected for full analyses (see Figure 2.1).

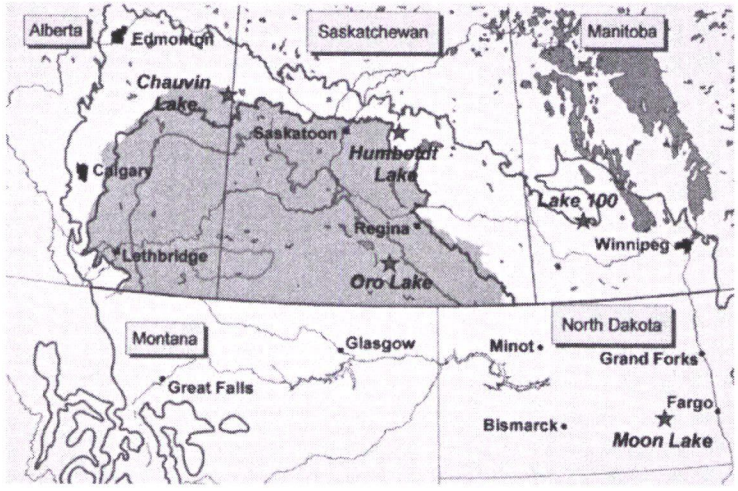

Figure 2.1: The map of sites of candidate lakes.

At each site, a 1.75-m piston core encompassing approximate 2000 years was obtained and sectioned continuously in 2.5 mm intervals (about 2.7 yr resolution). This time period encompasses the present climate system, as well as the most recent major climate regime shift (Laird et al. 2003). Fossil diatoms were isolated, identified to

Table 2.1: Relationship

	% Variance Explained	Study Interval	Significant Predictor Variables
Chauvin Lake	28.0	1938-1998	spring wheat production, cumulative departure from mean precipitation, total annual evaporation
Humboldt Lake	64.8	1938-1998	flax production, wheat production, annual cumulative evaporation, cumulative departure from mean precipitation, standardized precipitation index, Palmer drought severity index
Nora Lake	29.2	1965-1999	oats production, barley production, total annual precipitation

species and quantified in alternate samples using standard paleoecological techniques (Laird et al. 2003). Sediment ages were determined using radiometric analysis of ^{210}Pb(10-15 dates core^{-1}) and ^{14}C activities (4-6 AMS dates core^{-1}) (Laird et al. 2003).

Fossil records from the 20th century were calibrated against concomitant historical records to evaluate lake sensitivity to past droughts and congruence with documented crop failures. Canonical correspondence analysis indicated that 28-65% of the past variance in diatom species composition was explained by past changes in climate and crop production (Table 2.1 (Hall et al., 1999)).

In addition, diatom-inferred salinity was negatively correlated ($r = -0.45$ to -0.54, P-value. < 0.05) with the production of the major regional crop at Humboldt Lake, a site with the most extensive historical records (> 60 years). Further, comparison of fossil diatom profiles with those of δ^{15}N and pigments from algae (Rusak et al. 2004) demonstrated that while land-use practices influenced the nutrient chem-

Table 2.2: Basic descriptions for the salinity data from the three lakes

Lakes	Length	Max. Value	Min. Value	Time Range (years)
Humboldt Lake	355	24.717	0.621	2000
Chauvin Lake	381	28.642	0.305	2000
Nora Lake	312	16.827	0.135	2000

istry of Humboldt Lake after 1940, these changes did not affect the accuracy of climatic reconstructions. In all cases, salinity remained the principle environmental factor explaining diatom community variation during the 20^{th} century (Laird et al. 2003). Taken together, these historical comparisons show both that fossil diatoms accurately recorded changes in climate that influence production of economically important crops, and that variability in lake chemistry during the last 100 years arose in response to climatic variability, rather than changes in land-use and nutrient flux.

2.2 Preliminary Analysis

First, we have a look at the salinity data from Humboldt Lake, Chauvin Lake and Nora Lake (see Table 2.2 and Figure 2.2). From Figure 2.2 (a), we see that the mean μ_t is not constant over t. From the beginning to around year 400, μ_t is approximately 15; and after that, it seems to decrease slowly. This means that the Humboldt Lake salinity time series is not stationary in the mean. Similarly, from Figure 2.2 (b) and (c), we can conclude that the salinity time series from Chauvin Lake and Nora Lake are not stationary as well.

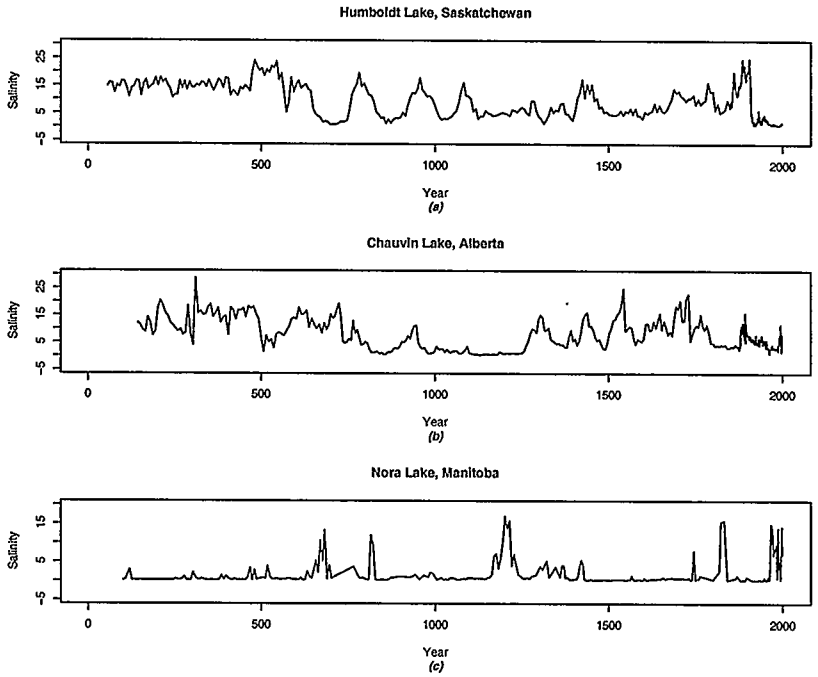

Figure 2.2: The salinity time series from the three lakes.

2.2.1 Detrend

Since there is a certain trend in μ_t in the three salinity time series, we need to eliminate it. A common method to remove a trend is to suppose that the time series follows the trend stationary model; that is, the process has stationary behavior around a trend,

$$x_t = \mu_t + y_t, \tag{2.1}$$

where x_t is our time series (observations), μ_t denotes the trend, and y_t is a stationary process. If we obtain a reasonable estimate of the trend component, say $\hat{\mu}_t$, then we

can work with the detrended series (residuals)

$$\hat{y}_t = x_t - \hat{\mu}_t. \tag{2.2}$$

By checking our time series above, we might try two different models for the trend component. First, we assume that a straight line is a reasonable model for the trend, i.e.,

$$\mu_t = \beta_1 + \beta_2 t.$$

Under this model, we estimate the trend, based on the salinity time series of Humboldt Lake as an example, by using the ordinary least squares (LS) method, and find

$$b_1 = 14.785$$

$$b_2 = -0.005$$

where b_1 and b_2 are the LS estimates for β_1 and β_2. Therefore, we have

$$\hat{\mu}_t = 14.785 - 0.005t$$

for Humboldt Lake.

To obtain the detrended time series y_t, we subtract $\hat{\mu}_t$ from the observations, x_t, that is,

$$\hat{y}_t = x_t - 14.785 + 0.005t,$$

and the plot of the detrended time series from Humboldt Lake is given in Figure 2.3 (a). We see that the trend disappears; namely, the mean of the series becomes constant. Similarly, the detrended salinity time series from Chauvin Lake and Nora Lake are plotted in Figure 2.3 (b) and (c), respectively.

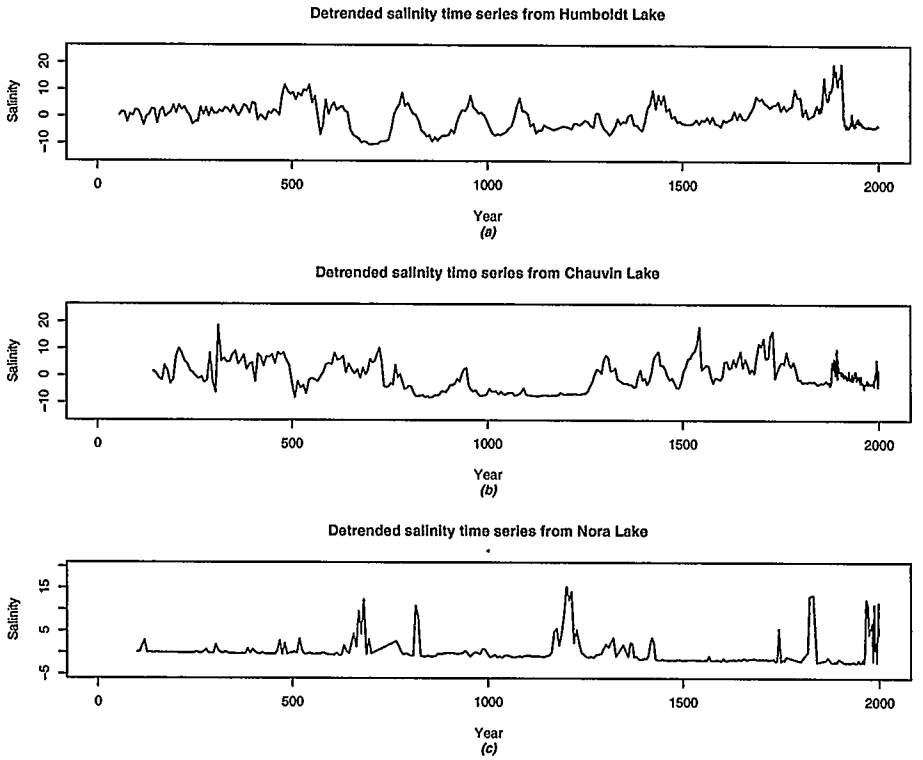

Figure 2.3: The detrended salinity time series of the three lakes using simple linear regression.

Second, we assume that a second order polynomial curve is a reasonable model for the trend, i.e.,

$$\mu_t = \beta_1 + \beta_2 t + \beta_3 t^2.$$

Under this model, we obtain the following LS estimates for the salinity time series

from Humboldt Lake,

$$b_1 = 17.57$$

$$b_2 = -0.012$$

$$b_3 = 3.168 \times 10^{-6}$$

where b_1, b_2 and b_3 are the LS estimates for β_1, β_2 and β_3. This leads to

$$\hat{\mu}_t = 17.57 - 0.012t + 3.168 \times 10^{-6}t^2$$

and the detrended time series

$$\hat{y}_t = x_t - 17.57 + 0.012t - 3.168 \times 10^{-6}t^2.$$

This detrended salinity time series is plotted in Figure 2.4 together with the detrended salinity time series using straight line regression.

Because b_3 is very close to zero, it seems that there is no need to employ the second order polynomial trend model. This can also be seen from Figure 2.4.

Similarly, we apply the same idea to the other two salinity time series from Chauvin Lake and Nora Lake. We find that we only need to use straight line as the trend model to detrend our salinity time series for all three lakes.

2.2.2 Defining drought

Moon Lake, North Dakota, U.S., was studied (Laird et al., 2003) prior to the three Canadian Prairies lakes discussed here. As a result of this study, it was suggested, using the drought in 1988 $\sim$ 1989 as reference, that 1.4 times the standard deviation of all positive detrended salinity values can be used as a threshold value to define droughts.

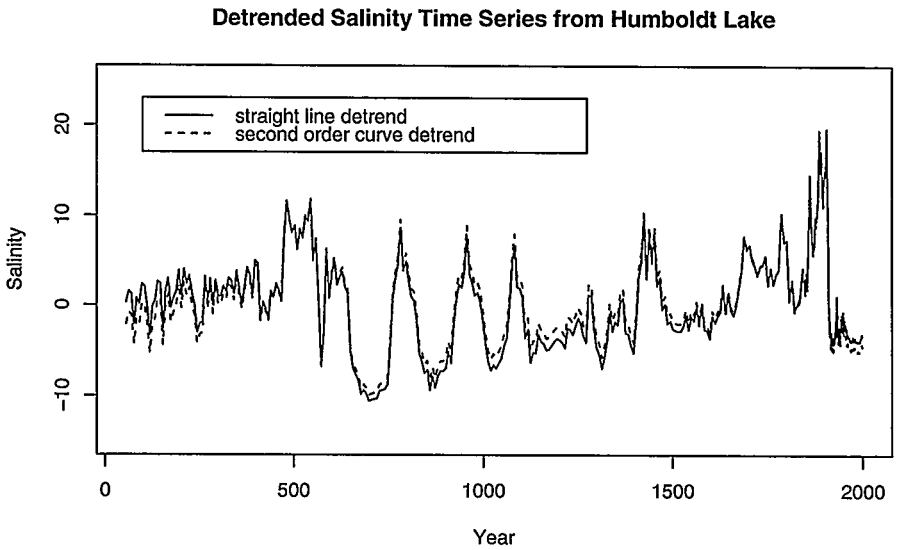

Figure 2.4: The detrended salinity time series of Humboldt Lake using two different trend models.

We illustrate below how to define droughts using Humbodlt Lake as an example. Similar results for Chauvin Lake and Nora Lake will be summarized in Table 2.3 and Figure 2.7.

For the detrended Humboldt Lake salinity time series, we find the sample standard deviation of the positive salinity values to be $s = 4.1485$. Then, we compute the threshold value as

$$\text{threshold} = 1.4 \times s = 5.8079.$$

Figure 2.5 can help us understand the definition of drought graphically. In Figure 2.5, the horizontal line represents the threshold and those salinity values that are greater than the threshold are droughts.

Table 2.3: Basic descriptive statistics and thresholds for the three lakes

Lakes	s	Threshold
Humboldt Lake	4.1485	5.8079
Chauvin Lake	3.3407	4.6769
Nora Lake	4.4364	6.2109

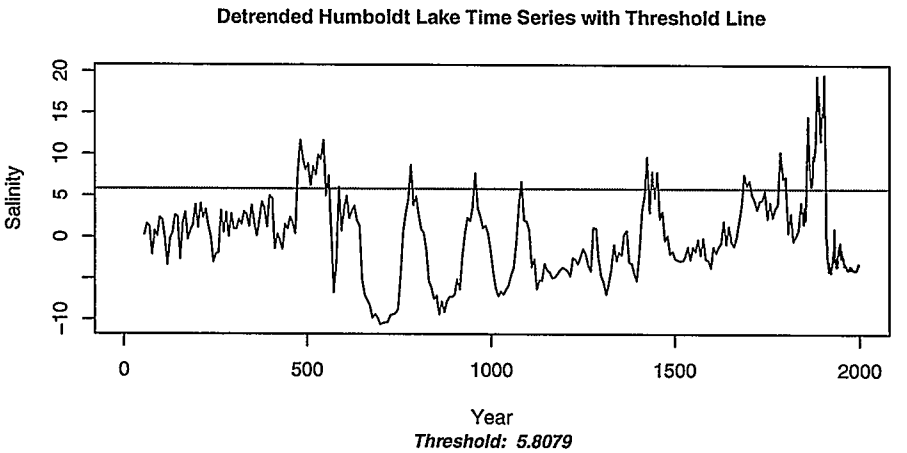

Figure 2.5: The detrended salinity time series of Humboldt Lake with threshold.

To emphasize the drought feature, all the salinity values that are less than or equal to the threshold value are set to the threshold value and we keep the original values unchanged if they are greater than the threshold value. We call the behavior this reset *value-truncation* and this reset time series *drought time series.* See Figure 2.6 for a plot.

The original time series is a salinity time series. It contains information other than that pertaining to droughts. For example, the very low salinity values may represent floods. If we study the periodicities of the whole original time series and

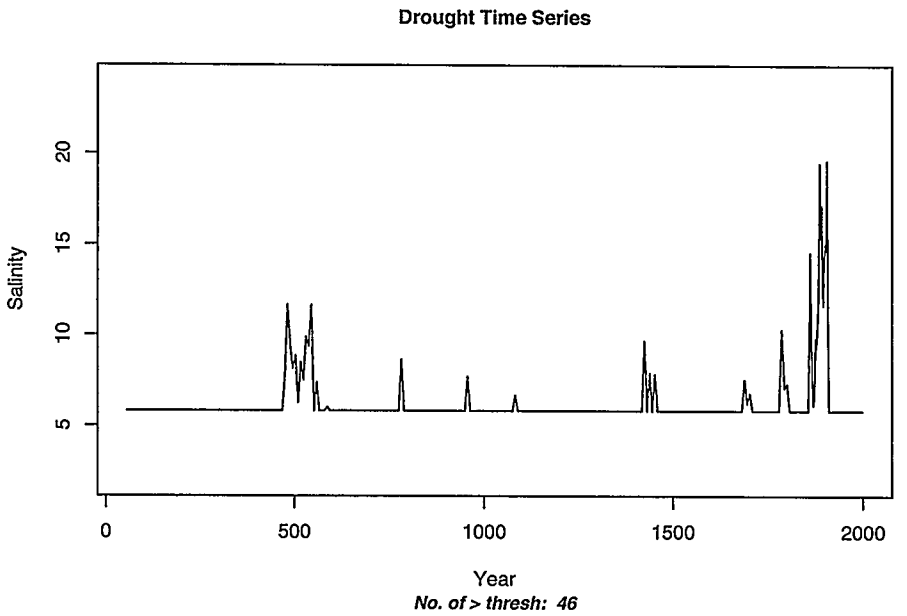

Figure 2.6: The drought time series from Humboldt Lake.

find some periods, it is hard to say which periods are for droughts and which are for floods. Therefore, we isolate the data above our drought threshold to quantitatively obtain drought time series and analyze the behavior of droughts by focusing on this drought time series.

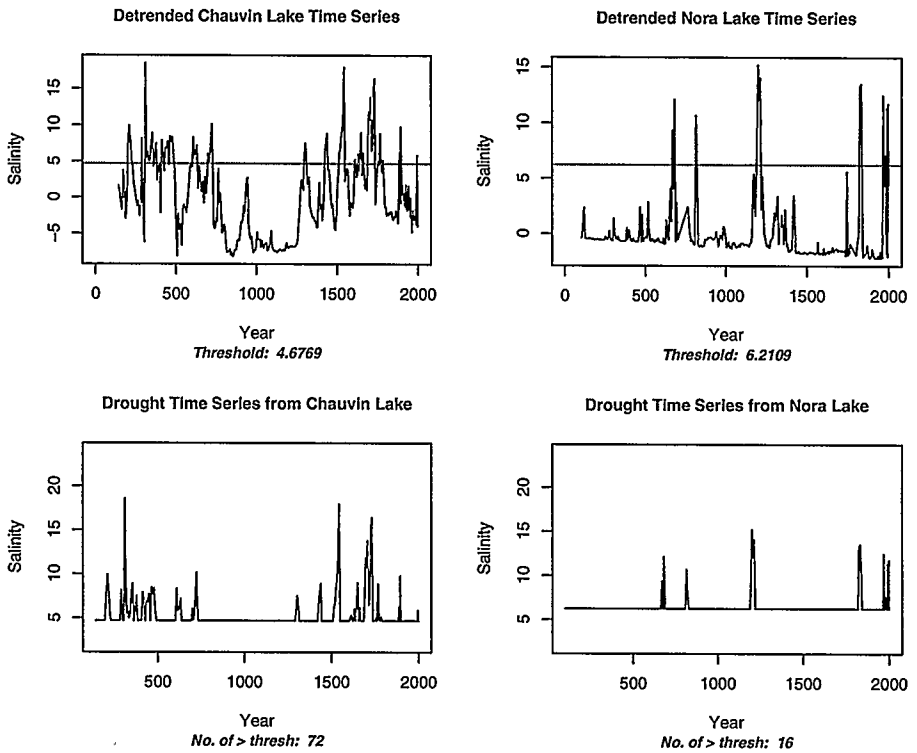

Figure 2.7: The detrended salinity and drought time series from Chauvin Lake and Nora Lake with threshold lines marked.

Chapter 3

Methodology

3.1 Introduction

There are two approaches to analyze time series: *time domain approach* and *frequency domain approach.* In time domain, a time series is described by the models based on prediction of the present as a regression on the past. In frequency domain, which has advantages for periodic processes, a time series is decomposed into different sine and cosine components with different frequencies. It is like a prism that splits light into its constituent colors and their strengths that are called the *spectrum* of the light. Shumway and Stoffer (2006) showed that any stationary stochastic process (time series) may be thought of, approximately, as a random superposition of sines and cosines oscillating at various frequencies.

Since we are interested in investigating the periodic behavior of drought, this thesis will focus on a specific frequency domain approach, namely, estimation of the spectrum by means of periodogram, which is often called *spectral analysis.* A fundamental objective of spectral analysis is to identify the dominant frequencies in a time series and to find an explanation of the phenomenon from which the measurements were taken. It may be done by searching for sharp peaks in the periodograms calculated from the Fourier transformation of the time series. If the time series is evenly spaced, the periodogram carries simple statistical behavior. Its calculation is simplified and can be quickly evaluated with the fast Fourier transform

(FFT) (Priestley, 1981).

However, palaeoclimatic data sets, such as the salinity time series we have, are normally unevenly spaced time series, which is also true for our drought time series. This is often a product of the non-linear relationship that commonly exists between depth and time, resulting in the transformation of a sampling regime that is equidistant in the depth domain into a non-uniformly spaced series in the time domain.

One way to solve this problem is to linearly interpolate the time series into an evenly spaced time array before applying the FFT. Unless performed carefully such an interpolation procedure can lead to aliasing of the signal (Schulz and Stattegger, 1997) resulting in the introduction of spurious components that may influence or even dominate the signal in the frequency domain.

Another way to solve this problem is to find an alternative method to use. Fortunately, there exists an alternative approach, which was first introduced in astrophysics called the *Lomb-Scargle preiodogram* that can overcome the problem caused by uneven spacing.

Astronomers could not always control viewing times, telescope availability and the position of an object in the sky—all of which is reminiscent of similar problems in using palaeoclimatic data sets. When studying variable stars in astronomy, Lomb (1976) sought a way to find periodicities in unevenly spaced time series by using least-squares fitting of sine waves of various periods to the data. Scargle (1982) extended Lomb's work by defining the *Lomb-Scargle periodogram* using periodogram analysis approach and proved that the periodogram analysis is exactly equivalent to least-squares fitting of sine curves to the data. Horne and Baliunas (1986) re-

discussed the normalization of the periodogram given by Scargle (1982) throught clarifying the proper definition of the variance that is used to normalize the Lomb-Scargle periodogram and studied some properties of the Lomb-Scargle periodogram as well. Press and Rybicki (1989) proposed a practical mathematical formulation of the Lomb-Scargle periodogram which was implemented in C (Press et al., 2002).

3.2 Spectral Analysis of Time Series

Spectral analysis primarily refers to the process of calculating and interpreting a spectrum for deciphering information from time series in the frequency domain. To carry out a spectral analysis, periodogram is the most commonly used quantity to detect periodic components of signals hidden in noise when the observed times are evenly or unevenly spaced. Periodogram has its population counterpart called the power spectrum, and its estimation is a main goal of spectral analysis.

From now on, all *time series* will be referred to as *stationary time series* except when otherwise stated. We use a set of real valued functions of time t

$$X_t = X(t)$$

to represent a phenomenon, and we use

$$\{x_{t_j} : j = 1, 2, \ldots, N\}$$

to represent a time series of size N which is arbitrarily sampled (evenly or unevenly) from $X(t)$. If the observed times are evenly spaced at interval h, it is customary to take $h = 1$, and $t_j = j$. So, we write evenly spaced time series of size N as

$$\{x_j : j = 1, 2, \ldots, N\}.$$

A cosine wave of time t is represented by

$$A\cos(2\pi\omega t),$$

where A is the amplitude and ω is the frequency. A sine wave is defined similarly.

We measure frequency, ω, by cycles per time unit and discuss the implications of certain frequencies in terms of the problem context. Of descriptive interest is the period T of a time series, defined as the number of points in a cycle, i.e.,

$$T = \frac{1}{\omega}.$$

We call the combination of a sine wave and a cosine wave a *harmonic function*, namely,

$$A\cos(2\pi\omega t) + B\sin(2\pi\omega t).$$

When we do spectral analysis on a time series $\{x_{t_j} : j = 1, 2, 3, \ldots, N\}$, we use the following model,

$$x_{t_j} = g_{t_j} + \varepsilon_{t_j} \text{ for } j = 1, 2, \ldots, \tag{3.1}$$

where g_{t_j} are signals and ε_{t_j} are random observational errors, which are often called *noise*. Hereafter, we assume that the signal will be taken to be strictly periodic and the errors at differnt times are independent; that is, ε_{t_i} is statistically independent of ε_{t_j} for $i \neq j$. We also assume that ε_{t_j} is normally distributed with zero mean and constant variance σ^2.

3.2.1 Orthogonal Transformation of Time Series

Spectral analysis can be viewed in terms of an *orthogonal transformation*. The basic idea behind this view is to decompose a time series into a number of components,

each one of which can be associated with a particular frequency. It is similar to vector decomposition in Calculus and Linear Algebra: an n-dimensional vector can be represented as a linear combination of n orthogonal vectors in n-dimensional space.

Let $\{f_i(t) : i = 1, 2, \ldots, N\}$ be a set of real-valued functions such that

$$\sum_{t=1}^{N} f_i(t) f_j(t) = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

We then analyze our time series with respect to the basis provided by $\{f_i(t)\}$ by computing the transformation coefficients a_i defined as

$$a_i = \sum_{t=1}^{N} x_t f_i(t), \ i = 1, 2, \ldots, N. \tag{3.3}$$

Given the transformation coefficients, we reconstruct our time series from the a_i using

$$x_t = \sum_{i=1}^{N} a_i f_i(t). \tag{3.4}$$

If the process $\{X_t\}$ represents some physical process such as a current or voltage, the total *energy* dissipated by the process in any time interval is equal to the sum of the amounts of energy dissipated by each component. Therefore, the key to a meaningful analysis is to pick a transformation such that the a_i have some physical interpretation so that the energy decomposition with respect to these a_i is relevant to our time series. There are several orthogonal transformations that can do so. The Fourier transformation, which provides the basis for spectral analysis, is such a transformation.

3.2.2 The Fourier Transform

The Fourier transform of a time series is an orthogonal transformation in which the a_i are 'stretched' and normalized versions of the trigonometric functions $\cos(t)$ and $\sin(t)$. Stretching is accomplished by introducing the notion of angular frequency $2\pi\omega$ to produce $\cos(2\pi\omega t)$ and $\sin(2\pi\omega t)$.

We note that there are the following relationships between cosine and sine functions,

$$\begin{cases} \int_{t_1}^{t_1+T} \cos^2(2\pi n\omega t)\mathrm{d}t = \int_{t_1}^{t_1+T} \sin^2(2\pi n\omega t)\mathrm{d}t = \dfrac{T}{2}, \\ \int_{t_1}^{t_1+T} \cos(2\pi m\omega t)\cos(2\pi n\omega t)\mathrm{d}t = \int_{t_1}^{t_1+T} \sin(2\pi m\omega t)\sin(2\pi n\omega t)\mathrm{d}t = 0, \quad m \neq n, \\ \int_{t_1}^{t_1+T} \sin(2\pi m\omega t)\cos(2\pi n\omega t)\mathrm{d}t = 0, \end{cases} \tag{3.5}$$

where $T = \dfrac{1}{\omega}$ is the period of the sine and cosine functions and m,n are integers.

The equations (3.5) illustrate that sine and cosine are orthogonal in the interval $(t_1, t_1 + T)$ for any t_1. Therefore, in this interval, a combination of $\cos(2\pi n\omega t)$ and $\sin(2\pi n\omega t)$ for all $n = 0, 1, 2, \ldots$ forms a set of orthogonal functions. When n goes to infinity, it is a *complete* orthogonal set.

Therefore, any real-valued periodic function $g_T(t)$, which is a deterministic function of t with period T, may be expressed as an infinite linear combination of sine and cosine functions,

$$g_T(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty} \left(a_n \cos(2\pi n\omega t) + b_n \sin(2\pi n\omega t)\right), \tag{3.6}$$

where $T = \dfrac{1}{\omega}$.

The expression on the right-hand side of (3.6) is called a *Fourier series*, and the constants $\{a_n\}$ and $\{b_n\}$ are called *Fourier coefficients*. We also call the right-hand

side of (3.6) the *sum of harmonics.*

3.2.3 Periodogram for evenly spaced time series

An evenly spaced time series $\{x_j : j = 1, 2, \ldots, N\}$ can be expressed by the Fourier series model as

$$x_j = \alpha_0 + \sum_{i=1}^{K} \Big(\alpha_i \cos(2\pi\omega_i j) + \beta_i \sin(2\pi\omega_i j) \Big) + \varepsilon_j, \tag{3.7}$$

where $\omega_i = \dfrac{i}{N}$ is the i^{th} harmonic of the fundamental frequency $\dfrac{1}{N}$, $K = \dfrac{N-1}{2}$ if N is odd, and $K = \dfrac{N}{2}$ if N is even. We treat (3.7) as a multiple regression model, and use the least-squares method to find the estimates of α_0, α_i and β_i as follows:

If $N = 2q + 1$ is odd, where q is an integer,

$$\begin{cases} a_0 &= \bar{x}_j, \\ a_i &= \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} x_j \cos(2\pi\omega_i j), \\ b_i &= \dfrac{2}{N} \displaystyle\sum_{j=1}^{N} x_j \sin(2\pi\omega_i j), \end{cases} \tag{3.8}$$

where $i = 1, 2, \ldots, q$.

Then, we define the *periodogram* as consisting of $q = (N-1)/2$ values

$$I(\omega_i) = \frac{N}{4} \left(a_i^2 + b_i^2 \right), \; i = 1, 2, \ldots, q. \tag{3.9}$$

If $N = 2q$ is even, where q is an integer, equations (3.8) and (3.9) will remain the

same only for $i = 1, 2, \ldots, (q-1)$, and for $i = q$,

$$\begin{aligned} a_q &= \frac{1}{N}\sum_{j=1}^{N}(-1)^j x_j, \\ b_q &= 0, \\ I(\omega_q) &= I(0.5) = N \cdot a_q^2. \end{aligned}$$

See Box et al. (1994) for details.

Once we have α_0, $\{\alpha_i\}$ and $\{\beta_i\}$, $i = 1, 2, \ldots, q$, estimated under model (3.7), we have a mathematical expression of the decomposition of the time series. It is still not easy to use it to detect the dominant frequencies. Periodogram, on the other hand, gives us an effective way. If the phenomenon $X(t)$ contains a sinusoidal component of frequency ω_r, then at and near ω_r, $a_r^2 + b_r^2$ makes a large contribution to $X(t)$. Hence the presence of a sinusoid is indicated by a large value of I near one value of ω, i.e., as a distinct narrow peak in the spectrum.

3.3 Periodogram for Unevenly Spaced Time Series

Here we introduce the Lomb-Scargle periodogram using the periodogram approach instead of the least-squares approach.

3.3.1 Lomb-Scargle periodogram

Given a time series $\{x_{t_j} : j = 1, 2, \ldots, N\}$, the discrete Fourier transform (DFT) is defined as

$$
\begin{aligned}
d(\omega) &= x_{t_1}e^{-2\pi i\omega t_1} + x_{t_2}e^{-2\pi i\omega t_2} + \cdots + x_{t_N}e^{-2\pi i\omega t_N} \\
&= \textstyle\sum_{j=1}^{N} x_{t_j}e^{-2\pi i\omega t_j} \\
&\quad \text{(By Euler formula } e^{i\theta} = \cos\theta + i\sin\theta) \\
&= \textstyle\sum_{j=1}^{N} x_{t_j}\Big(\cos(-2\pi\omega t_j) + i\sin(-2\pi\omega t_j)\Big) \\
&= \textstyle\sum_{j=1}^{N} x_{t_j}\Big(\cos(2\pi\omega t_j) - i\sin(2\pi\omega t_j)\Big).
\end{aligned}
\tag{3.10}
$$

The periodogram is then conventionally defined as

$$
\begin{aligned}
I(\omega) &= \frac{1}{N}|d(\omega)|^2 \\
&= \frac{1}{N}\Big|\sum_{j=1}^{N} x_{t_j}e^{-2\pi i\omega t_j}\Big|^2 \\
&= \frac{1}{N}\Big|\sum_{j=1}^{N} x_{t_j}\Big(\cos(2\pi\omega t_j) - i\sin(2\pi\omega t_j)\Big)\Big|^2 \\
&= \frac{1}{N}\Big|\sum_{j=1}^{N} x_{t_j}\cos(2\pi\omega t_j) - i\sum_{j=1}^{N} x_{t_j}\sin(2\pi\omega t_j)\Big|^2 \\
&= \frac{1}{N}\left(\left(\sum_{j=1}^{N} x_{t_j}\cos(2\pi\omega t_j)\right)^2 + \left(\sum_{j=1}^{N} x_{t_j}\sin(2\pi\omega t_j)\right)^2\right)
\end{aligned}
\tag{3.11}
$$

(Scargle, 1982).

$I(\omega)$ may be evaluated for any value of the frequency ω_j. For evenly spaced time series $\{x_j : j = 1, 2, \ldots, N\}$, equation (3.11) may be rewritten as

$$
\begin{aligned}
I(\omega) &= \frac{1}{N}|d(\omega)|^2 \\
&= \frac{1}{N}\Big|\sum_{j=1}^{N} x_j\Big(\cos(2\pi\omega j) - i\sin(2\pi\omega j)\Big)\Big|^2 \\
&= \frac{1}{N}\left(\left(\sum_{j=1}^{N} x_j\cos(2\pi\omega j)\right)^2 + \left(\sum_{j=1}^{N} x_j\sin(2\pi\omega j)\right)^2\right).
\end{aligned}
\tag{3.12}
$$

There are two problems with periodogram (3.11) and (3.12). First, $I(\omega)$ is very noisy even when the data are only slightly noisy, and the noise does not diminish

in amplitude with increasing sample size. Second, there is *spectral leakage*. This means that for a sinusoidal signal at a given frequency, ω_r, the spectrum in the periodogram not only appears at ω_r, but is also present in other frequencies. This problem is inherent to frequency analysis with a finite amount of data. *Aliasing* is one particular leakage which is a leakage of spectrum from high frequencies to *much lower* frequencies. Fortunately, anything from a slight to major unevenness in the sampling substantially reduces aliasing.

Due to these two problems, Scargle (1982) developped a slightly modified periodogram

$$I(\omega) = \frac{1}{2}\left\{ \frac{\left[\sum_{j=1}^{N} x_{t_j} \cos(2\pi\omega(t_j - \tau))\right]^2}{\sum_{j=1}^{N} \cos^2(2\pi\omega(t_j - \tau))} + \frac{\left[\sum_{j=1}^{N} x_{t_j} \sin(2\pi\omega(t_j - \tau))\right]^2}{\sum_{j=1}^{N} \sin^2(2\pi\omega(t_j - \tau))} \right\}, \quad (3.13)$$

where τ is defined through

$$\tan(4\pi\omega\tau) = \left(\sum_{j=1}^{N} \sin(4\pi\omega t_j)\right) \Big/ \left(\sum_{j=1}^{N} \cos(4\pi\omega t_j)\right). \quad (3.14)$$

Like periodogram (3.11), (3.13) reduces to equation (3.12) if the sampling spacing is even and has time-translation invariance. Beyond this, the Lomb-Scargle periodogram (3.13) has two other useful properties. First, it is equivalent to the least-squares fitting of sine waves. Second, and the most important, it has a simple statistical property that if the signal X_{t_j} is pure *Gaussian* noise, then $I(\omega)$ is exponentially distributed. This exponential distribution provides a convenient estimate of the probability that a given peak is a true signal, or it is the result of randomly distributed noise.

3.3.2 Distribution of $I(\omega)$

To study the distribution of $I(\omega)$, we rewrite (3.13) as

$$I(\omega) = \frac{1}{2}\Big(C^2(\omega) + S^2(\omega)\Big), \tag{3.15}$$

where

$$C(\omega) = A(\omega)\sum_{j=1}^{N} x_{t_j}\cos\left(2\pi\omega(t_j-\tau)\right), \tag{3.16}$$

$$S(\omega) = B(\omega)\sum_{j=1}^{N} x_{t_j}\sin\left(2\pi\omega(t_j-\tau)\right), \tag{3.17}$$

$$A(\omega) = \left(\sum_{j=1}^{N}\cos^2\left(2\pi\omega(t_j-\tau)\right)\right)^{-\frac{1}{2}}, \tag{3.18}$$

$$B(\omega) = \left(\sum_{j=1}^{N}\sin^2\left(2\pi\omega(t_j-\tau)\right)\right)^{-\frac{1}{2}}. \tag{3.19}$$

If $\{X_{t_j}\}$ are iid normally distributed with zero mean and constant variance σ^2, then $C(\omega)$ is a normal random variable as well since $C(\omega)$ is a linear combnation of independent and normally distributed random variables, and

$$\begin{aligned}
\mu_c &= \mathrm{E}\left(C(\omega)\right) = 0, \\
\sigma_c^2 &= \mathrm{Var}\left(C(\omega)\right) \\
&= \mathrm{Var}\left(A(\omega)\sum_{j=1}^{N} x_{t_j}\cos\left(2\pi\omega(t_j-\tau)\right)\right) \\
&= A(\omega)^2\sigma^2\sum_{j=1}^{N}\cos^2\left(2\pi\omega(t_j-\tau)\right) \\
&\quad \text{(Since } \{X_{t_j}\} \text{ are iid normally distributed with common variance } \sigma^2.\text{)} \\
&= \sigma^2 \text{ (by equation (3.18)).}
\end{aligned}$$

Similary, $S(\omega)$ is a normal random variable with

$$\begin{aligned}
\mu_s &= \mathrm{E}\left(S(\omega)\right) = 0, \\
\sigma_s^2 &= \mathrm{Var}\left(S(\omega)\right) \\
&= \mathrm{Var}\left(B(\omega)\sum_{j=1}^{N} x_{t_j}\sin\left(2\pi\omega(t_j-\tau)\right)\right) \\
&= B(\omega)^2\sigma^2\sum_{j=1}^{N}\sin^2\left(2\pi\omega(t_j-\tau)\right) \\
&\quad \text{(Since } \{X_{t_j}\} \text{ are iid normally distributed with common variance } \sigma^2.\text{)} \\
&= \sigma^2 \text{ (by equation (3.19)).}
\end{aligned}$$

Therefore, $I(\omega) = \frac{1}{2}\Big(C^2(\omega) + S^2(\omega)\Big)$ has the distribution function

$$\mathrm{P}\left(I(\omega) \le z\right) = \frac{1}{\sigma^2}\exp\left(\frac{-z}{\sigma^2}\right), \quad z > 0.$$

Thus, if we define the normalized periodogram $I_N(\omega)$ as $I_N(\omega) = I(\omega)/\sigma^2$, then $I_N(\omega)$ has the standard exponetial distribution with the density function $\mathrm{f}(z) = e^{-z}$.

3.3.3 Detecting periodic components

Our goal of using the Lomb-Scargle periodogram is to detect periodic components from a set of time series data. Therefore, we desire to find a power spectrum level z_0 so that we can claim the detection of a sinusoidal component when the calculated power spectrum exceeds this level, and we will be wrong with only a small probability, say α, correspondingly.

Suppose that the time series data are pure Gussian noise, then at any frequency ω_r, we have $\mathrm{P}(I_N(\omega_r) < z) = 1 - e^{-z}$. Let $\omega_1, \omega_2, \ldots, \omega_{N_{\mathrm{ind}}}$ be a set of frequencies in a Lomb-Scargle periodogram such that $I_N(\omega_r)$, $r = 1, 2, \ldots, N_{\mathrm{ind}}$, are independent,

then for any power spectrum level z_0, we have

$$\mathrm{P}\left(\max_{1\le r\le N_{\text{ind}}} I_N(\omega_r) \ge z_0\right) = 1-(1-e^{-z_0})^{N_{\text{ind}}}. \tag{3.20}$$

If we let α equal the probability in equation (3.20), we get

$$z_0 = -\ln(1-(1-\alpha)^{1/N_{\text{ind}}}). \tag{3.21}$$

We call z_0 the critical value at level α, and call α the significance level. In Scargle (1982), α is called the *false alarm probability.*

From (3.21), if a peak at frequency ω_r in a Lomb-Scargle periodogram is as high as or higher than z_0, we report a sinusoidal component at frequency ω_r with $(1-\alpha)\%$ confidence.

Horne and Baliunas (1986) performed extensive Monte Carlo simulations and gave a simple least squares formula to estimate the number of independent frequencies N_{ind} from the number of observations, N, in a time series:

$$N_{\text{ind}} \approx -6.362 + 1.193N + 0.00098N^2.$$

To illustrate the above procedure, let the periodic signal be $\cos(2\pi 0.005t)$ (Figure 3.1 (a)), which is contaminated by noises from $\mathrm{N}(0, 2^2)$ (Figure 3.1 (b)). We randomly take a sample of size 200 from the contaminated signal. It is an unevenly spaced time series (Figure 3.1 (c)). The Lomb-Scargle periodogram caculated from the 200 contaminated observations (Figure 3.2) shows that there is a peak at frequency 0.005 that is higher than the critical value at $\alpha = 0.05$ level of significance. That is, we detect the periodic signal with frequency $\omega = 0.005$ with 95% confidence.

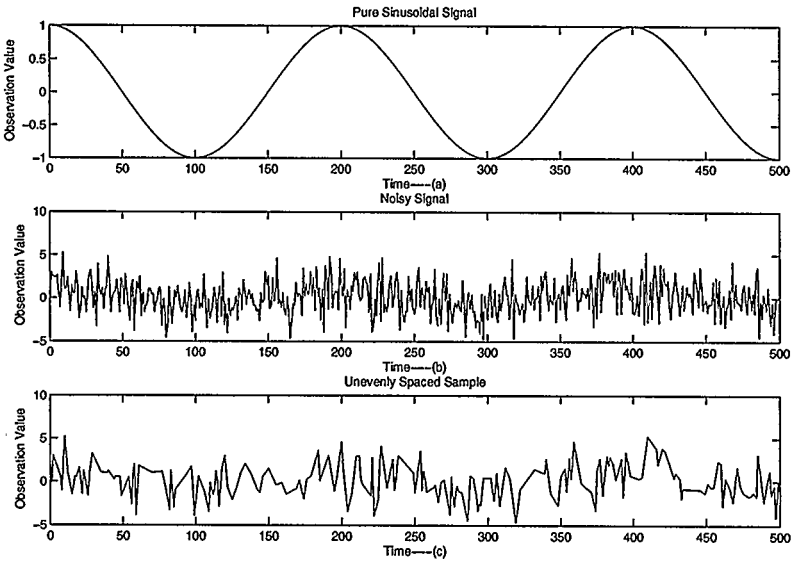

Figure 3.1: Illustration of the use of the Lomb-Scargle periodogram in detecting periodic component—data plots.

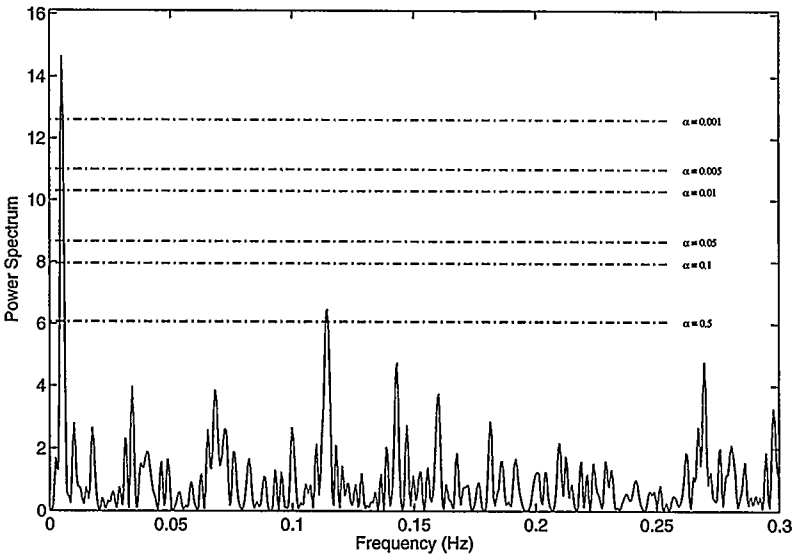

Figure 3.2: Illustration of the use of the Lomb-Scargle periodogram in detecting periodic component—periodogram.

Chapter 4

Analysis and Results

We plan to use the Lomb-Scargle periodogram to analyze the drought time series from the three lakes to see whether the droughts in the Canadian Prairies have any periodic pattern. As mentioned in Chapter 2, our drought time series are *unevenly spaced* and *value-truncated* time series. We know that the Lomb-Scargle periodogram works for stationary unevenly spaced time series. However, we are not sure whether it works for value-truncated time series as well.

In this chapter, we first simulate value-truncated time series in two different situations and check to see whether the Lomb-Scargle periodogram may be used to detect periodic signals. Then we apply the Lomb-Scargle periodogram to our real drought time series.

4.1 Simulation One

We simulate drought time series according to the following model:

$$X(t) = S_1(t) + S_2(t) + Noise(t),$$

where $S_1(t)$ and $S_2(t)$ are pure sinusoidal waves with frequencies ω_1 and ω_2, respectively, and t represents time. We use capital letters to represent the model and lower case letters to denote the specific simulated series.

4.1.1 Simulation set up

Sinusoidal components

Case 1:

$$X_1(t) = S_{11}(t) + S_{12}(t) + Noise(t), \tag{4.1}$$

where the sinusoidal components are

$$\begin{aligned} S_{11}(t) &= 3\cos(2\pi 0.005t), \\ S_{12}(t) &= 7\cos(2\pi 0.02t + \frac{\pi}{6}). \end{aligned}$$

Case 2:

$$X_2(t) = S_{21}(t) + S_{22}(t) + Noise(t), \tag{4.2}$$

where the sinusoidal components are

$$\begin{aligned} S_{21}(t) &= 3\cos(2\pi 0.005t), \\ S_{22}(t) &= 3\cos(2\pi 0.02t + \frac{\pi}{6}). \end{aligned}$$

The frequencies 0.005 and 0.02 used in Case 1 and Case 2 correspond to periods of 200 years and 50 years, respectively. Evenly spaced time series of length 2000 are first generated as the sum of the sinusoidal components, where $t \in \{1, 2, \ldots, 2000\}$.

Noise addition

Through the addition of noise it is possible to reduce the signal-to-noise ratio by increasing the variance of the noise. Here we assume the noise is normally distributed with mean 0 and variance σ^2_{noise}, i.e., Noise $\sim$ N(0, σ^2_{noise}). We simulate 2000 iid noise terms from N(0, σ^2_{noise}) and add them to the 2000 pure sinusoidal values already generated and denote the result by $\{x_t : t = 1, 2, \ldots, 2000\}$.

Table 4.1: Levels for factors σ_{noise}, n and TC

Factor	Levels in Case 1	Levels in Case 2
σ_{noise}	2, 3, 5, 7	1, 2, 3, 4
n	355, 300, 200	355, 300
TC	0.5, 1, 1.4, 2, 2.2	0.5, 1, 1.4, 2, 2.2, 2.4

Taking arbitrarily spaced samples

We randomly take a sample of size n, $\{x_{t_j} : j = 1, 2, \ldots, n\}$ from $\{x_t : t = 1, 2, \ldots, 2000\}$. Since this sample is a random sample, the spacing for $\{x_{t_j} : j = 1, 2, \ldots, n\}$ is uneven.

Value-truncation

We use different *threshold* values to truncate the unevenly spaced time series $\{x_{t_j} : j = 1, 2, \ldots, n\}$ to obtain unevenly spaced and value-truncated time series, $\{y_{t_j} : j = 1, 2, \ldots, n\}$. Since we use 1.4 times SD of the positive detrended values as the threshold value to define droughts, in our simulation we use different values (denoted by TC hereafter) to time SD to set up different threshold values.

Processing simulation

We consider σ_{noise}, n and TC as factors that may affect the period detection for the simulated unevenly spaced and value-truncated time series by using the Lomb-Scargle periodogram. The levels that each factor will take are listed in Table 4.1.

In each combination of σ_{noise}, n and TC, for example, $\sigma_{\text{noise}} = 3$, $n = 300$ and $TC = 1.4$, we run the above steps 500 times. For each run we compute the Lomb-Scargel periodogram for the simulated time series, and in the 500 runs we count the percentage of times the known built-in frequencies are detected. We call this

percentage the *successful detection rate* (SDR). For some combinations, we conduct 5000 runs to see the convergence trend of the successful detection rate.

Note here that when we say a built-in frequency ω is detected, we mean that the Lomb-Scargle periodogram contains a 95% confident peak at a frequency ω' such that ω' is in the interval $\omega \pm 0.05\omega$.

4.1.2 Simulation results

Case 1:

The simulation results for Case 1 are listed in Table 4.2 to Table 4.6. We report four different kinds of SDR, namely, SDR for S_{11}, SDR for S_{12}, SDR for both S_{11} and S_{12} and SDR for either S_{11} or S_{12}. Table 4.2 to Table 4.5 are based on 500 runs and Table 4.6 is based on 5000 runs. From these tables, we can clearly see that all three factors σ_{noise}, n and TC affect the SDR significantly. For example, if we decrease the sample size from 300 to 200, when we keep σ_{noise} and TC at low levels, say 2 and 0.5, respectively, the SDR for S_{11} will drop sharply from 90.8% to 54.2%. If we keep sample size n at high levels, say 355, and keep $TC = 1.4$, then when σ_{noise} increases from 3 to 5, the SDR for S_{11} drops from 81.6% to 19.8%. It is also clear to see that when TC increases SDR decreases.

From Table 4.6, we see that the SDRs based on 5000 runs are only slightly different from those based on 500 runs. The largest difference is less than 1%. This means that the results based on 500 runs are representative.

It is easier to express how strong the noise is relatively to the signal by using *Signal-to-Noise Ratio (SNR)*. In our case, the signal is the composite of the two sinusoidal components. The SNRs are listed in Table 4.7. The last line is in dB

Table 4.2: Successful Detection Rates in Case 1—Part 1

detect *signal 1*; 500 runs

	Standard Deviation of Noise (σ_{noise})			
SDR (%)	2	3	5	7
Sample Size (n) = 355				
$TC = 0.5$	98.4	89.4	43.6	15.6
$TC = 1.0$	99.2	87.8	31.8	8.8
$TC = 1.4$	99.4	81.6	19.8	5.0
$TC = 2.0$	98.0	56.2	6.0	0.4
$TC = 2.2$	92.8	39.6	3.4	0.2
Sample Size (n) = 300				
$TC = 0.5$	90.8	73.0	26.2	5.8
$TC = 1.0$	92.2	69.6	15.6	2.2
$TC = 1.4$	92.0	60.0	7.2	1.4
$TC = 2.0$	84.4	32.6	1.4	0.0
$TC = 2.2$	73.8	18.2	0.6	0.0
Sample Size (n) = 200				
$TC = 0.5$	54.2	34.8	8.6	2.8
$TC = 1.0$	57.8	30.4	4.8	0.8
$TC = 1.4$	57.0	21.8	1.8	0.2
$TC = 2.0$	37.2	6.2	0.0	0.0
$TC = 2.2$	24.2	2.2	0.0	0.0

Table 4.3: Successful Detection Rates in Case 1—Part 2

detect *signal 2*; *500* runs

	Standard Deviation of Noise (σ_{noise})			
SDR (%)	2	3	5	7
Sample Size (n) = 355				
$TC = 0.5$	100	100	100	100
$TC = 1.0$	100	100	100	100
$TC = 1.4$	100	100	100	99.8
$TC = 2.0$	100	100	99.8	76.6
$TC = 2.2$	100	100	90.4	50.6
Sample Size (n) = 300				
$TC = 0.5$	100	100	100	100
$TC = 1.0$	100	100	100	99.8
$TC = 1.4$	100	100	100	98.6
$TC = 2.0$	100	100	94.0	57.6
$TC = 2.2$	100	99.4	71.6	25.6
Sample Size (n) = 200				
$TC = 0.5$	100	100	100	98.8
$TC = 1.0$	100	100	99.6	90.0
$TC = 1.4$	100	100	97.8	66.8
$TC = 2.0$	99.6	94.6	41.8	10.2
$TC = 2.2$	96.6	68.4	17.0	1.8

Table 4.4: Successful Detection Rates in Case 1—Part 3

detect *both signals*; *500* runs

	Standard Deviation of Noise (σ_{noise})			
SDR (%)	2	3	5	7
Sample Size (n) = 355				
$TC = 0.5$	98.4	89.4	43.6	15.6
$TC = 1.0$	99.2	87.8	31.8	8.8
$TC = 1.4$	99.4	81.6	19.8	5.0
$TC = 2.0$	98.0	56.2	6.0	0.0
$TC = 2.2$	92.8	39.6	3.4	0.0
Sample Size (n) = 300				
$TC = 0.5$	90.8	73.0	26.2	5.8
$TC = 1.0$	92.2	69.6	15.6	2.2
$TC = 1.4$	92.0	60.0	7.2	1.4
$TC = 2.0$	84.4	32.6	1.2	0.0
$TC = 2.2$	73.8	18.2	0.6	0.0
Sample Size (n) = 200				
$TC = 0.5$	54.2	34.8	8.6	2.8
$TC = 1.0$	57.8	30.4	4.8	0.8
$TC = 1.4$	57.0	21.8	1.8	0.2
$TC = 2.0$	37.2	6.0	0.0	0.0
$TC = 2.2$	23.8	2.0	0.0	0.0

Table 4.5: Successful Detection Rates in Case 1—Part 4

detect *either signal*; *500* runs

	Standard Deviation of Noise (σ_{noise})			
SDR (%)	2	3	5	7
Sample Size (n) = 355				
$TC = 0.5$	100	100	100	100
$TC = 1.0$	100	100	100	100
$TC = 1.4$	100	100	100	99.8
$TC = 2.0$	100	100	99.8	77.0
$TC = 2.2$	100	100	90.4	50.8
Sample Size (n) = 300				
$TC = 0.5$	100	100	100	100
$TC = 1.0$	100	100	100	99.8
$TC = 1.4$	100	100	100	98.6
$TC = 2.0$	100	100	94.2	57.6
$TC = 2.2$	100	99.4	71.6	25.6
Sample Size (n) = 200				
$TC = 0.5$	100	100	100	98.8
$TC = 1.0$	100	100	99.6	90.0
$TC = 1.4$	100	100	97.8	66.8
$TC = 2.0$	99.6	94.8	41.8	10.2
$TC = 2.2$	97.0	68.6	17.0	1.8

Table 4.6: Successful Detection Rates in Case 1—Part 5

detect *signal 1*; *5000* runs

	(σ_{noise})		
SDR (%)	3	5	7
Sample Size (n) = 355			
$TC = 0.5$	88.68	44.54	15.64
$TC = 1.0$	86.82	32.38	8.52
$TC = 1.4$	81.10	20.52	4.24
$TC = 2.0$	56.16	5.22	0.60
$TC = 2.2$	40.40	2.44	0.28

detect *signal 2*; *5000* runs

Sample Size (n) = 355			
$TC = 0.5$	100	100	100
$TC = 1.0$	100	100	100
$TC = 1.4$	100	100	99.58
$TC = 2.0$	100	99.04	77.26
$TC = 2.2$	99.96	91.20	49.92

detect *both signals*; *5000* runs

Sample Size (n) = 355			
$TC = 0.5$	88.68	44.54	15.64
$TC = 1.0$	86.82	32.38	8.52
$TC = 1.4$	81.10	20.52	4.22
$TC = 2.0$	56.16	5.18	0.40
$TC = 2.2$	40.40	2.36	0.14

detect *either signal*; *5000* runs

Sample Size (n) = 355			
$TC = 0.5$	100	100	100
$TC = 1.0$	100	100	100
$TC = 1.4$	100	100	99.60
$TC = 2.0$	100	99.08	77.46
$TC = 2.2$	99.96	91.28	50.06

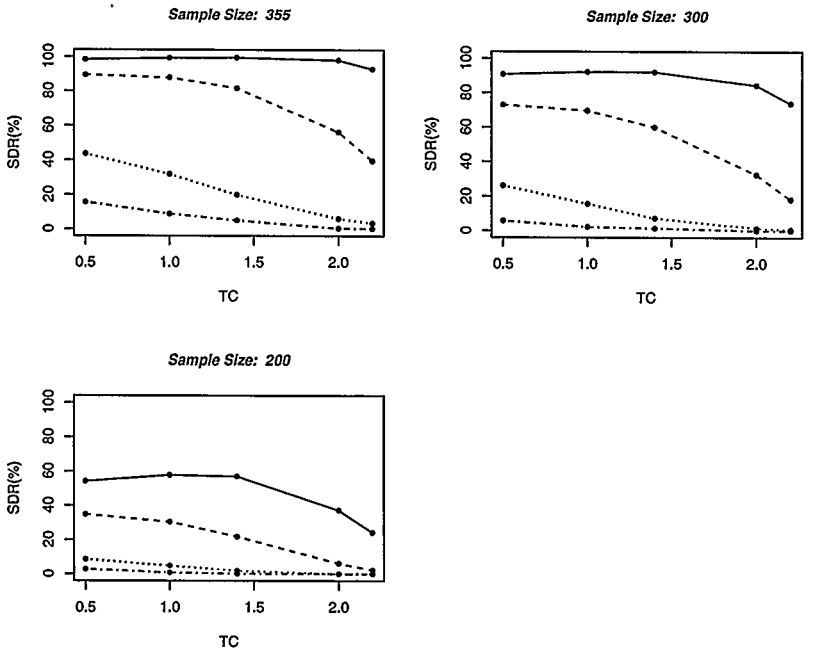

Figure 4.1: Successful detection rate for Signal 1 in Case 1.

unit, which is 10 times $\log_{10}$ of the line labeled *mean*. The larger the value is, the more power the signal contains compared with the noise. The negative value in dB means the power carried by the signal is less than the power carried by the noise, which means the data is very noisy. Using SNR, we display the successful detection rates for various combinations of σ_{noise}, n and TC in Figure 4.1 to Figure 4.4. In each figure, a 'solid line' represents 8.60dB SNR, a 'dashed line' represents 5.08dB, a 'dotted line' represents 0.64dB and a 'dotdash line' represents -2.28dB.

To summarize the results in Case 1, we conclude that for SNR no less than 5.08dB, $n = 355$ (the length of the drought time series from the three lakes) and

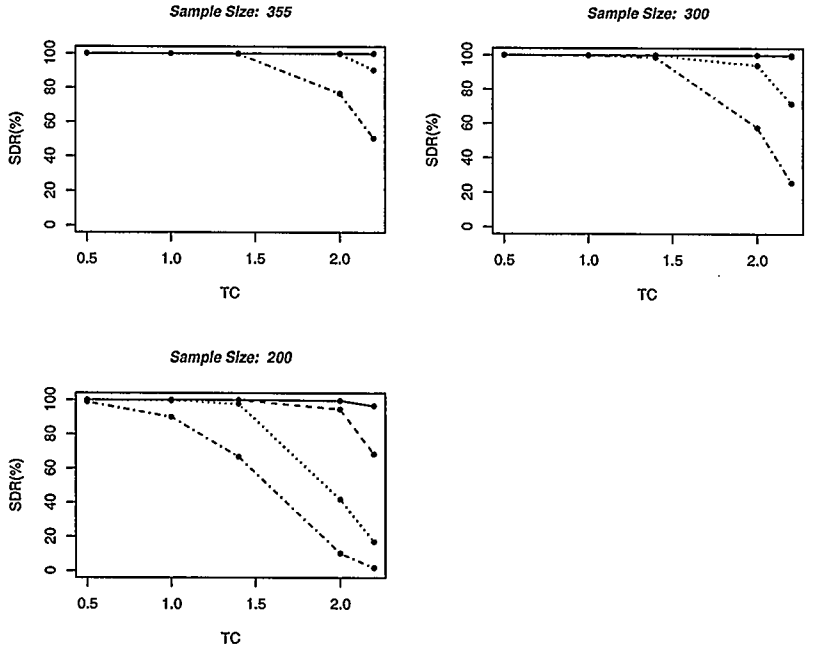

Figure 4.2: Successful detection rate for Signal 2 in Case 1.

threshold values no more than 1.4 times SD, the SDRs are above 80%.

Case 2:

From the simulation results in Case 1, we notice that the SDR for S_{11} and SDR for S_{12} are very different: it is harder to detect S_{11} than to detect S_{12}. This is of course related to the difference in amplitudes in S_{11} and S_{12} (S_{11} has an amplitude 3 and S_{12} has an amplitude 7), suggesting that amplitude is another factor affecting the SDR.

In Case 2, the amplitude of S_{21} and the amplitude of S_{22} are both set to 3. We focus on $n = 355$ and $n = 300$, and add one more value for TC, namely, $TC = 2.4$.

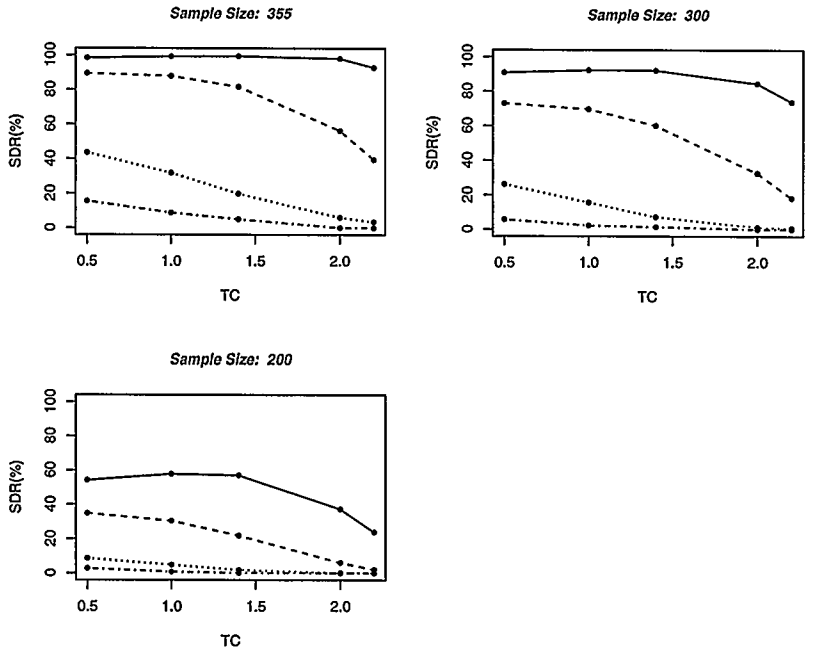

Figure 4.3: Successful detection rate for both Signal 1 and Signal 2 in Case 1.

The simulation results are listed in Table 4.8 and Table 4.9, and are displayed in Figure 4.5 and Figure 4.6. SNRs corresponding to the 4 values of σ_{noise} are listed in Table 4.10. In Figure 4.5 and Figure 4.6, a 'solid line' represents 9.54dB SNR, a 'dashed line' represents 3.52dB, a 'dotted line' represents 0dB and a 'dotdash line' represents -2.50dB.

From Table 4.8 and Table 4.9 and Figure 4.5 and Figure 4.6, we can see that the SDR for S_{21} is now quite similar to the SDR for S_{22}. Compared to the results in Case 1, the SDR for both S_{21} and S_{22} and for either S_{21} or S_{22} is higher than the SDR for both S_{11} and S_{12} and for either S_{11} or S_{12}, respectively. For SNR no more

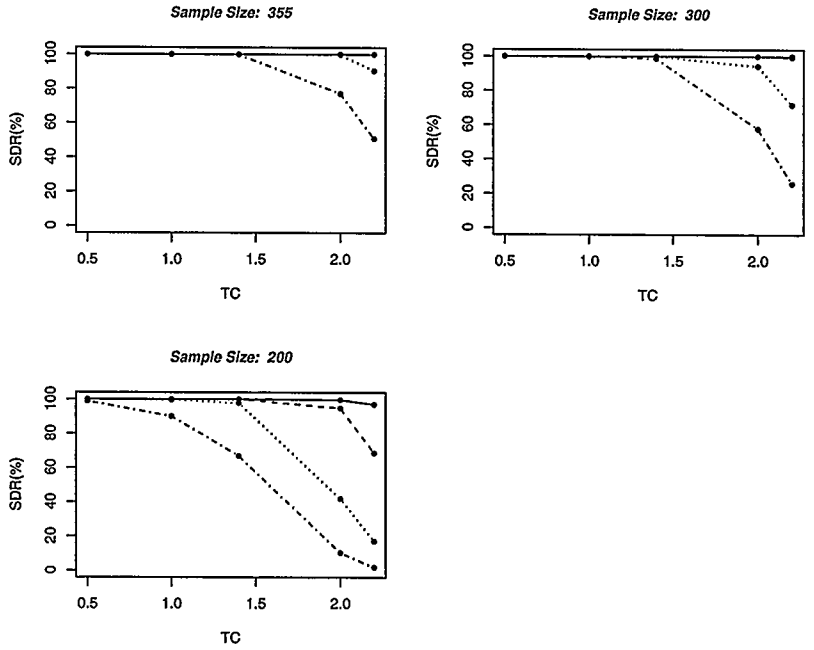

Figure 4.4: Successful detection rate for either Signal 1 or Signal 2 in Case 1.

than 0dB, $n = 355$ and threshold values no more than 1.4 times SD, the SDR is above 94% in Case 2.

4.2 Simulation Two

The simulation results in the last section show that the Lomb-Scargle periodogram has the potential to detect periodicity buried in unevenly spaced and value-truncated time series. In this section we want to set up a new simulation having the features of the salinity time series from the three lakes in the Canadian Prairies. By doing so, we move to a better position to analyze the real time series from the three lakes

Table 4.7: Signal-to-Noise Ratios in Case 1

for case 1

	Standard Deviation of Noise (σ_{noise})			
	2	3	5	7
Signal-to-Noise Ratio				
mean	7.2472	3.2210	1.1596	0.5916
95% C.I.	(7.2401, 7.2544)	(3.2178, 3.2242)	(1.1584, 1.1607)	(0.5910, 0.5922)
(dB)	8.60	5.08	0.64	-2.28

using the Lomb-Scargle periodogram.

In this new simulation, we first extract potential periodic components of droughts from salinity time series and treat those components as the sinusoidal components of the simulation model. Then we take those sinusoidal components away from the salinity time series and find an approximate representation for the series without the potential droughts information. Finally, we add the potential sinusoidal components of droughts back to the approximate representation to have a way to generate time series that look like the original salinity time series.

4.2.1 Extraction of salinity time series features

Denoting a detrended salinity time series by $\{x_{t_j} : j = 1, 2, \ldots, n\}$, we can represent $\{x_{t_j}\}$ using an infinite Fourier series:

$$\begin{aligned} x_{t_j} &= \sum_{i=1}^{\infty} \Big[a_i \cos(2\pi\omega_i t_j) + b_i \sin(2\pi\omega_i t_j) \Big] \\ &= \sum_{i=1}^{\infty} \Big[A_i \cos(2\pi\omega_i t_j + \phi_i) \Big], \quad j = 1, 2, \ldots, n, \end{aligned} \tag{4.3}$$

Table 4.8: Successful Detection Rates for signals in Case 2—Part 1

detect *signal 1*; *500* runs

	(σ_{noise})			
SDR (%)	1	2	3	4
Sample Size $(n) = 355$				
$TC = 0.5$	100	100	99.9	97.8
$TC = 1.0$	100	100	99.5	89.6
$TC = 1.4$	100	100	96.8	72.8
$TC = 2.0$	100	98.6	64.8	25.3
$TC = 2.2$	100	92.7	41.5	11.3
$TC = 2.4$	99.9	72.2	19.1	4.3
Sample Size $(n) = 300$				
$TC = 0.5$	99.9	100	99.8	90.7
$TC = 1.0$	100	100	97.4	76.8
$TC = 1.4$	100	99.9	89.3	53.5
$TC = 2.0$	100	91.1	40.4	10.0
$TC = 2.2$	99.9	74.6	18.1	3.6
$TC = 2.4$	99.0	43.3	5.6	1.3

detect *signal 2*; *500* runs

Sample Size $(n) = 355$				
$TC = 0.5$	100	100	100	97.8
$TC = 1.0$	100	100	99.8	91.6
$TC = 1.4$	100	100	97.8	73.0
$TC = 2.0$	100	98.0	61.3	23.9
$TC = 2.2$	100	88.3	36.8	10.3
$TC = 2.4$	99.4	58.4	14.4	2.6
Sample Size $(n) = 300$				
$TC = 0.5$	100	100	99.5	93.0
$TC = 1.0$	100	100	97.9	76.9
$TC = 1.4$	100	99.9	89.3	55.0
$TC = 2.0$	100	86.3	38.0	11.2
$TC = 2.2$	99.5	63.3	16.8	4.2
$TC = 2.4$	89.3	27.6	4.6	1.0

Table 4.9: Successful Detection Rates for signals in Case 2—Part 2

detect *both siganls*; *500* runs

	(σ_{noise})			
SDR (%)	1	2	3	4
Sample Size (n) = 355				
$TC = 0.5$	100	100	99.9	95.6
$TC = 1.0$	100	100	99.3	81.8
$TC = 1.4$	100	100	94.8	52.2
$TC = 2.0$	100	96.7	41.2	6.3
$TC = 2.2$	100	82.8	17.3	1.1
$TC = 2.4$	99.3	46.7	4.3	0.0
Sample Size (n) = 300				
$TC = 0.5$	99.9	100	99.3	84.0
$TC = 1.0$	100	100	95.3	57.9
$TC = 1.4$	100	99.8	79.2	28.3
$TC = 2.0$	100	79.2	16.0	0.8
$TC = 2.2$	99.5	50.4	4.2	0.2
$TC = 2.4$	88.9	15.9	0.2	0.0

detect *either signal*; *500* runs

Sample Size (n) = 355				
$TC = 0.5$	100	100	100	100
$TC = 1.0$	100	100	100	99.4
$TC = 1.4$	100	100	99.8	93.6
$TC = 2.0$	100	99.9	84.9	42.9
$TC = 2.2$	100	98.2	61.0	20.5
$TC = 2.4$	100	83.9	29.2	6.9
Sample Size (n) = 300				
$TC = 0.5$	100	100	100	99.7
$TC = 1.0$	100	100	100	95.8
$TC = 1.4$	100	100	99.4	80.2
$TC = 2.0$	100	98.2	62.4	20.4
$TC = 2.2$	99.9	87.5	30.7	7.6
$TC = 2.4$	99.4	55.0	10.0	2.2

Table 4.10: Signal-to-Noise Ratios in Case 2

for case 2

	Standard Deviation of Noise (σ_{noise})			
	1	2	3	4
Signal-to-Noise Ratio				
mean	8.9940	2.2485	1	0.5621
95% C.I.	(8.9835, 9.0046)	(2.2459, 2.2511)	(0.9982, 1.0005)	(0.5615, 0.5628)
(dB)	9.54	3.52	0.00	-2.50

where

$$\begin{cases} a_i &= c\Big[\sum_{j=1}^{n} x_{t_j} \cos(2\pi\omega_i t_j)\Big], \\ b_i &= c\Big[\sum_{j=1}^{n} x_{t_j} \sin(2\pi\omega_i t_j)\Big], \\ A_i &= \sqrt{a_i^2 + b_i^2}, \\ \phi_i &= \arctan\left(-\frac{b_i}{a_i}\right), \end{cases} \tag{4.4}$$

and c is a coefficient of the form, $\frac{d}{n}$, where $d > 0$ that ensures the Fourier series approximation to x_{t_j} is on the original scale.

We use the detrended salinity time series from the Humboldt Lake to illustrate how to extract drought periodicity features. In Figure 4.7, we plot the original detrended time series and three approximations to the original series using 1001, 1429 and 2001 terms in equation (4.3). We see from Figure 4.7 (d) that the approximation becomes very close to the original by using 2001 terms. We use x'_{t_j} to denote the approximation using 2001 terms, that is,

$$x'_{t_j} = \sum_{i=1}^{2001} \Big[A_i \cos(2\pi\omega_i t_j + \phi_i)\Big], \quad j = 1, 2, \ldots, 355. \tag{4.5}$$

Now for the fixed threshold 1.4 times SD of the positive detrended salinity values, we find 46 detrended values higher than the threshold. This is shown in Figure 4.8

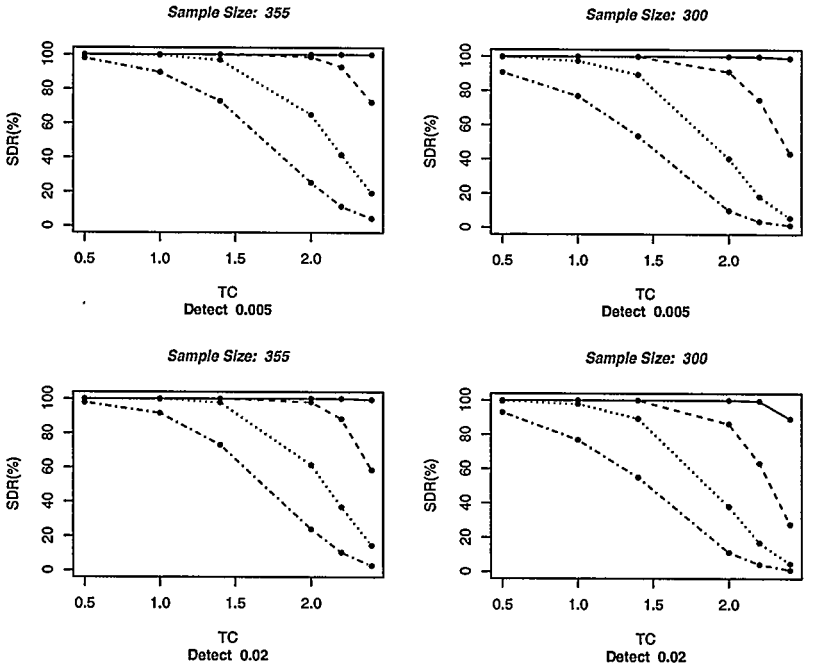

Figure 4.5: SDR in Case 2—Part 1.

by using dots in the bottom plot. These 46 values carry the information about the drought history in Humboldt Lake according to our drought definition. For each of the 46 values, we order the 2001 terms in its approximation from the largest to the smallest. We find that in 30 out of the 46 cases the frequencies 0.0045 and 0.0065 appear in the top two largest terms in the ordered approximation terms, and appear in the third and fourth place in some of the remaining cases. These two frequencies together with the corresponding amplitudes and phases are listed in Table 4.11, respectively. We consider the two corresponding periods 222 years and 154 years as the periodic drought features for Humboldt Lake.

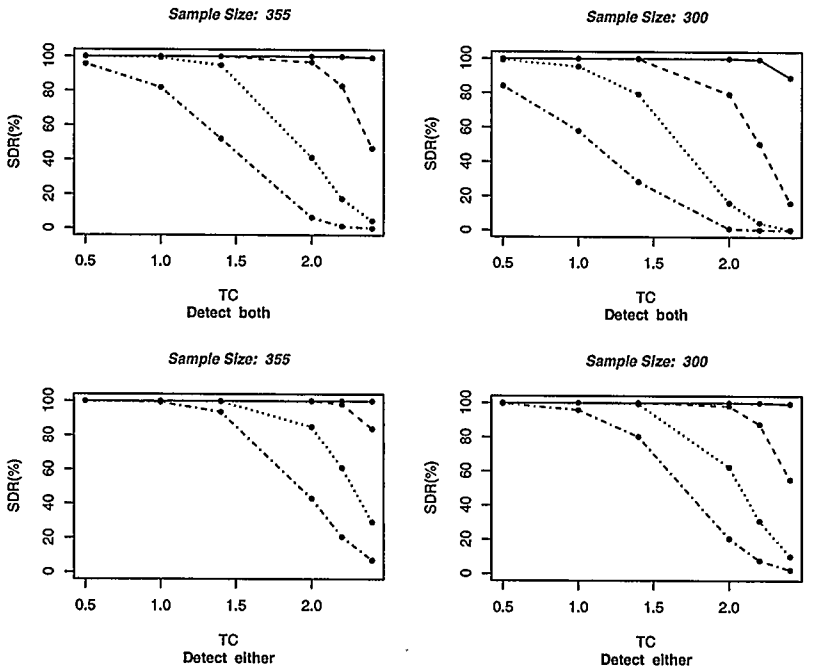

Figure 4.6: SDR in Case 2—Part 2.

Note that the above drought periodic feature extraction process cannot be treated as a formal way to find drought periodic features. However, we use this process to set up a way to generate time series that look like the real detrended salinity time series from the three lakes.

4.2.2 Simulation set up

Sinusoidal components of droughts

From Table 4.11, we use the sum of the two listed sinusoidal waves

$$0.2713\cos(2\pi 0.0045t_j + 0.2513) + 0.2949\cos(2\pi 0.0065t_j - 1.4447)$$

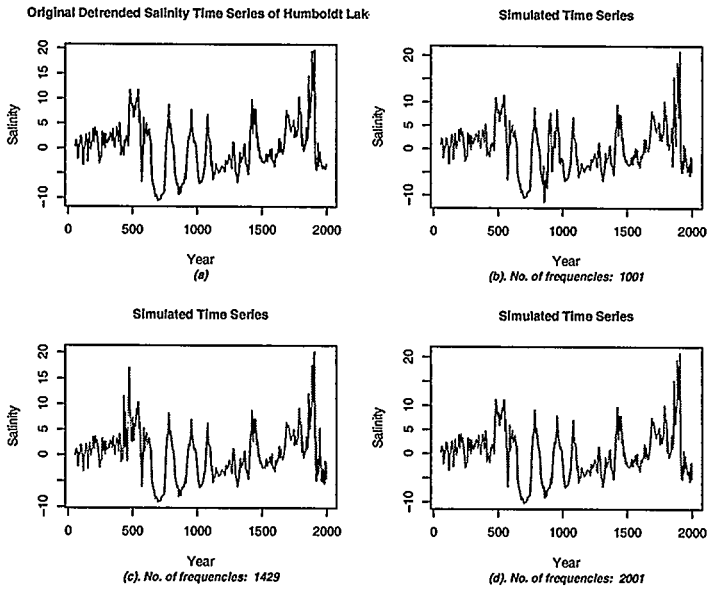

Figure 4.7: The plot of different approximations to represent the detrended salinity time series in Humboldt Lake.

to represent two periodic drought components.

Representing the salinity time series without the potential droughts

First, we take out the two terms in (4.5) that correspond to the two frequencies 0.0045 and 0.0065. To do this completely, we actually take away a band of frequencies from 0.003 to 0.008. This means that we take away the 11 terms from (4.5) that are associated with the frequencies from 0.003 to 0.008 and use the remaining 1990 terms as an approximation to the detrended salinity time series without the potential droughts denoted by $x''_{t_j}, \quad j = 1, 2, \ldots, 355$.

Next, we take the first 400 terms in x''_{t_j} to represent the major features of the detrended salinity time series without the potential droughts and denote this by

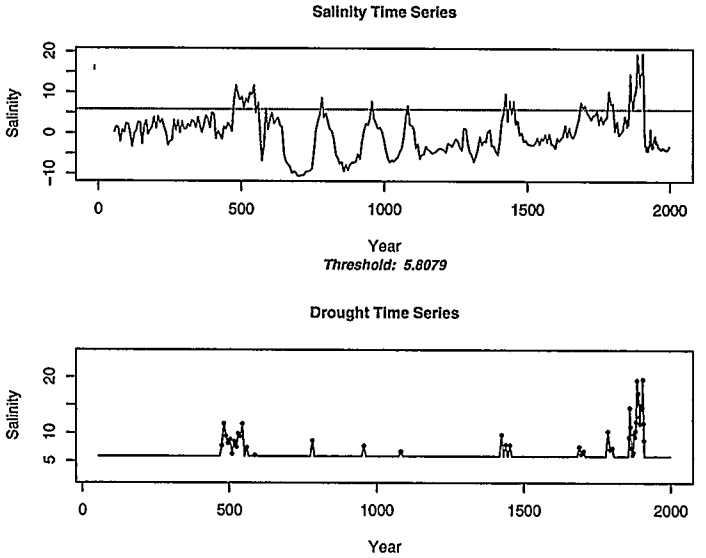

Figure 4.8: The plot of drought time series of Humboldt Lake and the plot of the 46 detrended salinity values greater than the threshold.

x'''_{t_j}, $j = 1, 2, \ldots, 355$. Plots of x''_{t_j} virsus t_j; x'''_{t_j} virsus t_j; $r_{t_j} = (x''_{t_j} - x'''_{t_j})$ virsus t_j and the histogram of r_{t_j} are displayed in Figure 4.9. We see from Figure 4.9 that x'''_{t_j} looks largely like x''_{t_j} and r_{t_j} can be described by N(0, σ^2), where

$$\sigma^2 = \frac{1}{355}\sum_{j=1}^{355}\left(r_{t_j} - \frac{1}{355}\sum_{k=1}^{355} r_{t_k}\right)^2 = 2.35^2.$$

Together, we use

$$z_{t_j} = 0.2713\cos(2\pi 0.0045 t_j + 0.2513) + 0.2949\cos(2\pi 0.0065 t_j - 1.4447) + x'''_{t_j} + e_{t_j} \tag{4.6}$$

to generate time series similar to the detrended salinity time series from the Humboldt Lake by simulating iid sequences e_{t_j}, where $e_{t_j} \sim \mathrm{N}(0, 2.35^2)$.

Table 4.11: Potential drought components for Humboldt Lake

Frequency (Hz)	Amplitude	Corresponding Phase
0.0045	0.2713	0.2513
0.0065	0.2949	-1.4447

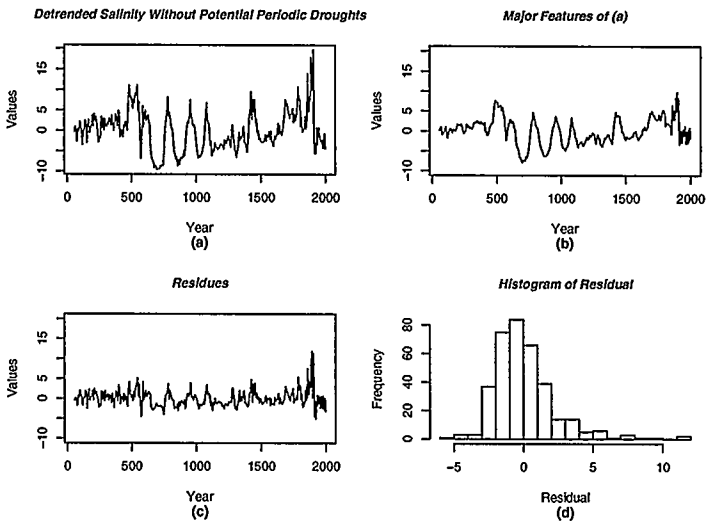

Figure 4.9: The plots related to Simulation Two.

Value-truncation

We use different *threshold* values to truncate the unevenly spaced time series $\{z_{t_j} : j = 1, 2, \ldots, 355\}$ to obtain the unevenly spaced and value-truncated time series $\{y_{t_j} : j = 1, 2, \ldots, 355\}$. We use three different values (denoted by TC hereafter) times SD of the positive values of $\{y_{t_j}\}$ to set up different threshold values.

Processing simulation

Since the t_j in (4.6) are from the real salinity time series and the standard deviation of the noise term is $\sigma = 2.35$, we only set up one factor TC to see how it affects the detection of the potential periodic components of droughts by using the Lomb-Scargle periodogram. The levels assigned to TC are 1, 1.4 and 1.8.

At each level of TC, we generate time series according to (4.6) either 500 times or 5000 times, and in each run we apply the Lomb-Scargle periodogram to detect the built-in periodic components of droughts. We count the percentage of times the built-in potential periodic components of droughts are detected out of 500 runs and 5000 runs separately. This percentage is called the successful detection rate (SDR) and four different kinds of SDR, namely, SDR for frequency 0.0045, SDR for frequency 0.0065, SDR for both 0.0045 and 0.0065, and SDR for either 0.0045 or 0.0065 are calculated and summarized in Table 4.12.

4.2.3 Simulation Results

The results at the top of Table 4.12 are based on 500 runs, the results in the middle are based on 5000 runs, and the differences between 500 runs and 5000 runs are listed at the bottom of Table 4.12. We see from the top of Table 4.12 that at $TC = 1.4$, the SDR for both 0.0045 and 0.0065 is 89.4%; the SDR for 0.0045 and the SDR for 0.0065 go up to 91.2% and 97.2%, respectively, and the SDR for either 0.0045 or 0.0065 is up to 99%. If we decrease TC to 1.0, then we see that the SDR for either 0.0045 or 0.0065 is 100% and the other three SDRs are at least 97%. Even when we increased TC to 1.8, the lowest SDR is still as high as 63%.

From the middle and the bottom of Table 4.12, we see that the SDRs based on

Table 4.12: Successful Detection Rates for simulated drought time series

500 runs

	Categories of Signal Detection			
	0.0045	0.0065	Both	Either
SDR (%)				
$TC = 1.0$	97.40	99.80	97.20	100
$TC = 1.4$	91.20	97.20	89.40	99.00
$TC = 1.8$	69.20	84.20	63.80	89.60

5000 runs

	Categories of Signal Detection			
	0.0045	0.0065	Both	Either
SDR (%)				
$TC = 1.0$	96.02	99.86	95.90	99.98
$TC = 1.4$	90.04	97.96	88.06	99.18
$TC = 1.8$	67.82	83.30	62.80	89.30

difference between 500 runs and 5000 runs

	Categories of Signal Detection			
	0.0045	0.0065	Both	Either
Difference (%)				
$TC = 1.0$	1.30	0.06	1.40	0.02
$TC = 1.4$	0.98	0.76	1.34	0.18
$TC = 1.8$	1.38	0.90	1.00	0.30

Note: 'Both' means both 0.0045 and 0.0065; 'Either' means either 0.0045 or 0.0065.

500 runs and the SDRs based on 5000 runs are very close with the biggest difference being 1.38%.

To summarize the results in Simulation Two, we conclude that for the salinity time series from Humboldt Lake, we have reasonable confidence to apply the Lomb-Scargle periodogram to our unevenly spaced and value-truncated drought time series to detect the periodic components of droughts. If we set the threshold value to 1.4 times the SD of positive values of the detrended salinity time series, the successful detection rate is above 89.4%.

4.3 Applications

Based on the results of Section 4.1 and Section 4.2, in this section, we use the Lomb-Scargle periodogram to explore the periodic behavior of droughts in the Canadian Prairies represented by the salinity time series from Humboldt Lake in Saskatchewan, Chauvin Lake in Alberta and Nora Lake in Manitoba.

4.3.1 Humboldt Lake

The periodogram plot from computing the Lomb-Scargle periodogram using 1.4 times SD threshold for the detrended salinity time series from Humboldt Lake is displayed in Figure 4.10. We see from Figure 4.10 that there are several high peaks at low frequencies, and there are another cluster of peaks around frequency 0.43. At 95% confidence level, the peaks at low frequencies are related to periodic components of droughts and the cluster of peaks centered around frequency 0.43 are related to random noise.

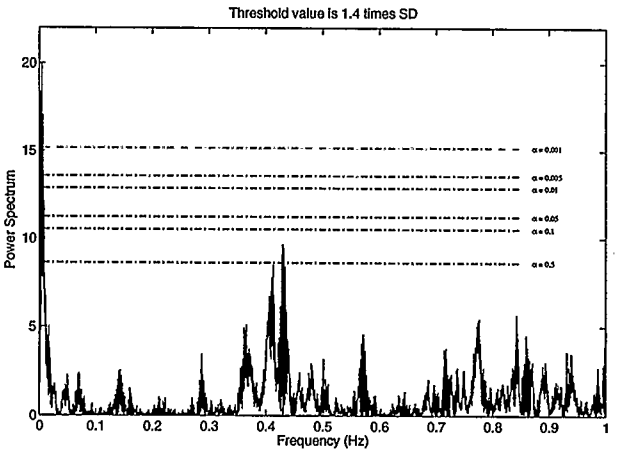

Figure 4.10: The Lomb-Scargle periodogram for the drought time series from Humboldt Lake.

Enlarging the low frequency peaks in Figure 4.10 and displaying these peaks in Figure 4.11, we see that the top three significant peaks correspond to frequencies 0.0022, 0.0044 and 0.0062, respectively, or in terms of periods, 454, 227 and 156 years, respectively. This means that droughts of a magnitude at least equal to that of the late 1980s may reoccur approximately every 450 years, every 227 years and every 156 yeas in Humboldt Lake area. These three periods are the main drought periods for Humboldt Lake, which are comparable to the finding of Yu and Ito (1999) based on a study of Rice Lake in North Dakota, U.S..

If we use 1.2 times SD or 1.6 times SD as the threshold, the findings are both similar to those of using 1.4 times SD threshold.

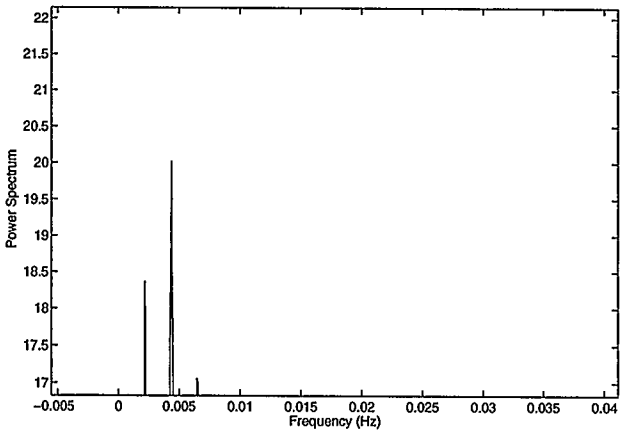

Figure 4.11: Top three powers of the Lomb-Scargle periodogram for the drought time series from Humboldt Lake.

4.3.2 Chauvin Lake

We take the same approach to analyze the detrended salinity time series from Chauvin Lake. The periodogram plot for 1.4 times SD threshold is given in Figure 4.12. After enlarging the low frequency peak in Figure 4.12 and displaying it in Figure 4.13, we see from Figure 4.12 and Figure 4.13 that, at 5% significance level, there are two significant peaks near frequencies 0.0008 and 0.85. This amounts to 1250 years and 1 year in terms of period. Looking back at Figure 2.7, the 1 year period seems to come from the truncation.

At 1.2 times SD threshold and 1.6 times SD threshold, the Lomb-Scargle periodograms are plotted in Figure 4.14 (top and bottom, respectively). We see from Figure 4.14 that the results are both similar to those of using 1.4 times SD threshold. However, if using 1.9 times SD threshold, we can only detect one significant peak at

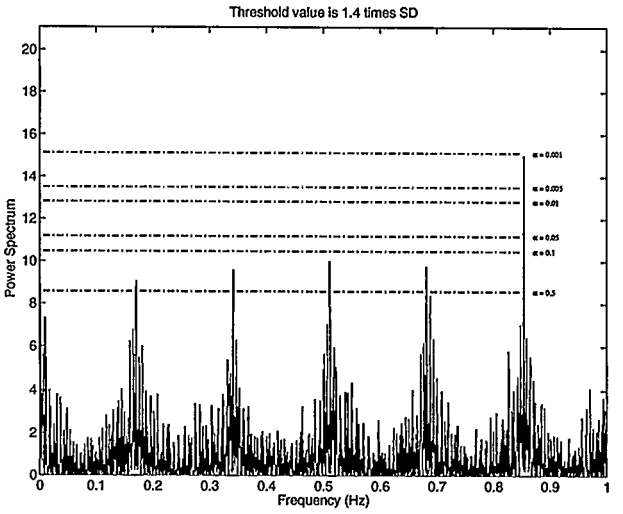

Figure 4.12: The Lomb-Scargle periodogram for the drought time series from Chauvin Lake.

frequency 0.0008.

4.3.3 Nora Lake

The story for Nora Lake is a little different. At 1.2, 1.4 and 1.6 times SD thresholds, no peaks are detected at 5% significant level; see Figure 4.15 for details. To understand why, we refer to Figure 2.7. From the bottom right corner of Figure 2.7 we see that after value-truncation, there is little drought information left. However, if we use 0.7 times SD threshold, a peak at frequency 0.0078 is detected at 5% significance level, which amounts to a period of 128 years. Other information is needed to see whether this period relates to any kind of droughts.

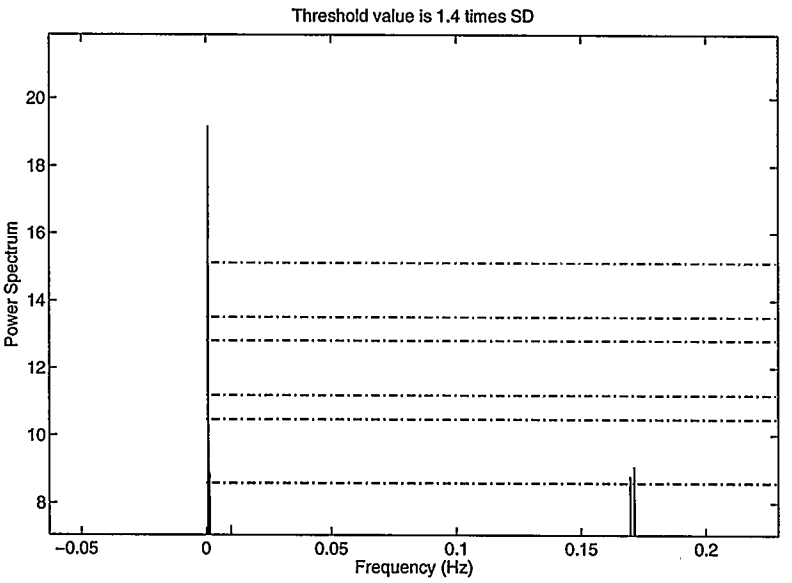

Figure 4.13: The enlargement of the Lomb-Scargle periodogram for the drought time series from Chauvin Lake at low frequencies.

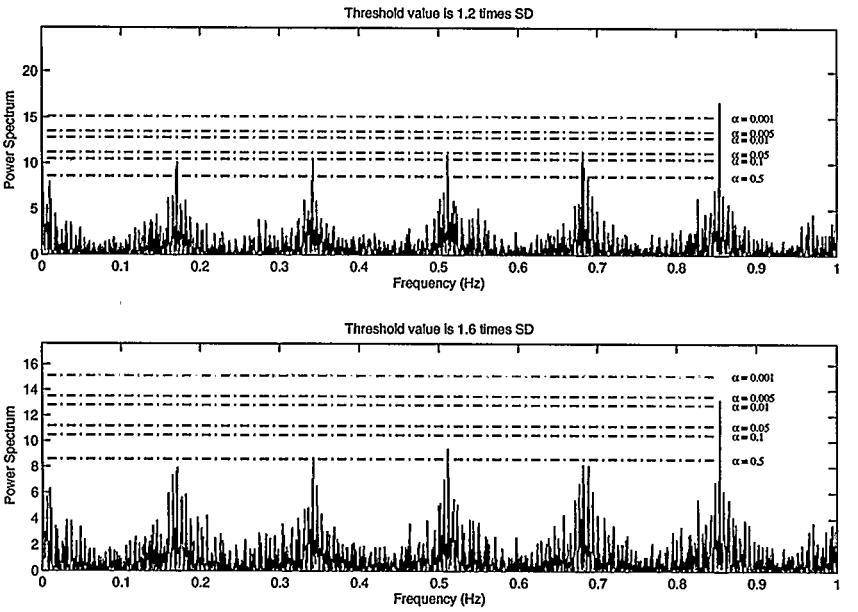

Figure 4.14: The Lomb-Scargle periodograms for the drought time series from Chauvin Lake using 1.2 and 1.6 times SD as threshold values.

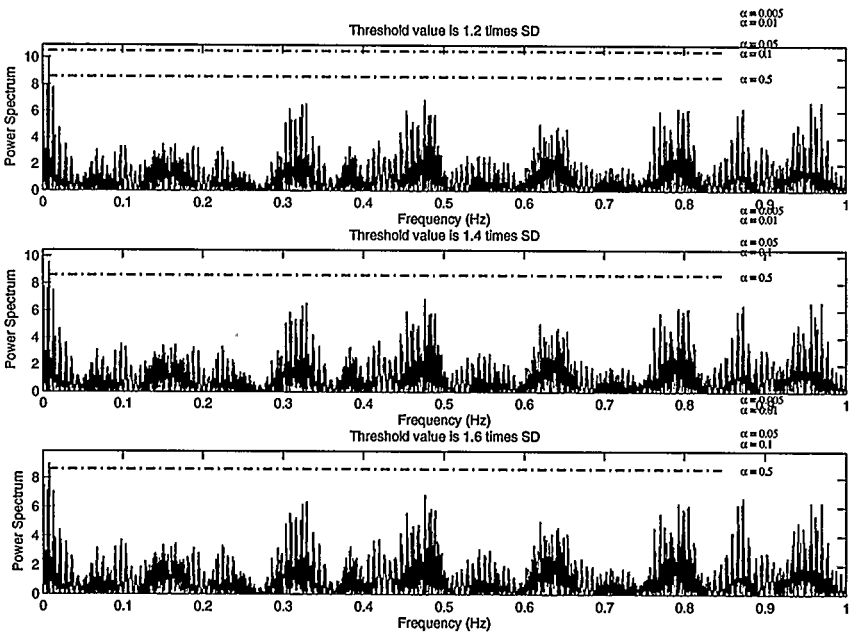

Figure 4.15: Lomb-Scargle periodograms for the drought time series from Nora Lake.

Chapter 5

Summary and Future Work

5.1 Summary

Motivated by the need to understand the droughts in the Canadian Prairies better, in this thesis we have investigated the periodic behaviors of droughts by using the 2000 years diatom-inferred salinity time series from Humboldt Lake in Saskatchewan, Chauvin Lake in Alberta and Nora Lake in Manitoba. In doing so, we have addressed the following important issues. First, we have taken a direct way to define droughts, using the 1988-1989 drought as a reference. This approach is intuitive as well as quantitative, but the resulting drought time series become value-truncated and unevenly spaced time series.

Second, to analyze value-truncated and unevenly spaced time series, we have adapted the Lomb-Scargle periodogram to our problem. Through simulation studies, we have confirmed that the Lomb-Scargle periodogram is applicable to study the drought periodicities based on the value-truncated and unevenly spaced drought time series.

Finally, we have analyzed the three drought time series from the three lakes using the Lomb-Scargle periodogram and found that for droughts of a magnitude as large as or larger than that of the drought in 1985-1986, Humboldt Lake seems to have had droughts with periods 454, 227 and 156 years, Chauvin Lake seems to have had droughts with a period of 1250 years, and Nora Lake seems to have had no periodic

drought history.

5.2 Future Work

Drought is a very complicated phenomenon. It may be affected by many factors that we cannot use a simple way to measure. Our simulation study did not consider dependence among the error terms, but in applications dependence is a fact that cannot be ignored. More simulation studies need to be conducted.

Recently, some researchers used wavelet analysis to reveal periodicity of a phenomenon. Tian et al. (2006) investigated late-Holocene drought and Brown et al. (2005) studied fire cycles in North American interior grasslands and their relation to prairie drought by using wavelet analysis. Spectral analysis can be viewed in terms of an orthogonal transformation. The Fourier transformation, on which the Lomb-Scargle periodogram is based, is such an orthogonal transformation that has led to some (partial) solutions to problems related to periodicity, and the Wavelet transformation is another one. Therefore, it would be interesting to apply wavelet technique to study our drought time series. In particular, it would be desirable to develop a wavelet version of the Lomb-Scargle periodogram.

On a more intuitive ground, since we mainly care about whether a salinity value is greater than a threshold value or not, we can transform the salinity time series to a binary series according to a threshold. For example, all the salinity values that are greater than a threshold value are reset to be "1" and the rest reset to "0". Figure 5.1 is a plot obtained by using the drought time series from Humboldt Lake and 1.4 times SD threshold value.

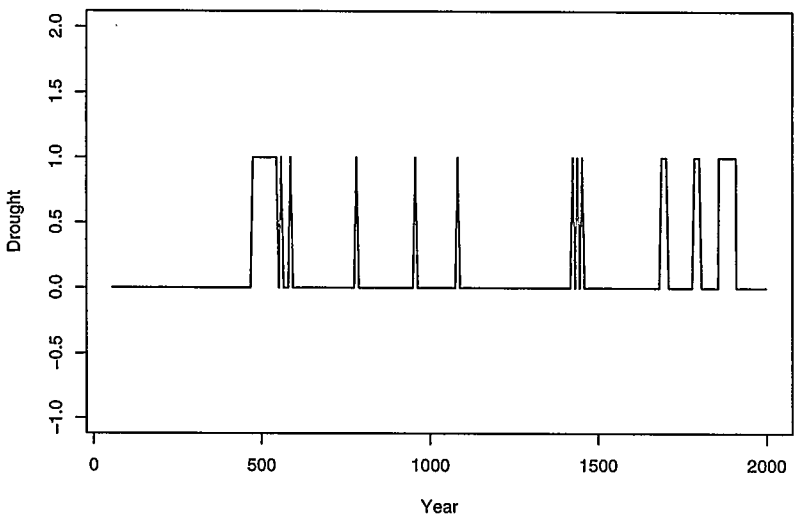

Figure 5.1: The binary version of the drought time series from Humboldt Lake.

Many researchers have investigated ways to predict the pattern of binary data sets by using non-linear and adaptive prediction techniques for digital communication signals. It would be interesting to try this approach to study drought periods.

Appendix A

The following are the codes for a MatLab program which computes the Lomb-Scargle periodogram for a given detrended time series (mydata) and user supplied threshold (ThreshCoeff time SD) and user supplied highest frequency examined (highestf).

```
function lspgrm(mydata, ThreshCoeff, highestf)
% This program is based on a Lomb-Scargle implementation in
% Press, Teukolsky, et al. Numerical Recipes, "Spectral
% Analysis of Unevenly Sampled Data." and is a modified version
% of Brett Shoelson's.

global dataid datamatrix effm ep fhi freqs hifac info jmax ...
       lines n nmax nout np ofac prob px py s sigfreqs ...
       sigpowers tmax tmin x y

inputdata = textread(mydata);
dataid=1;

if size(inputdata,2) ~= 2
error('Input data must be an n x 2 matrix of numbers.')
end

datamatrix = cut(inputdata,ThreshCoeff);
```

```
%Plot input data points
figure('numbertitle','off','name','Plots of data sets');
subplot(2,1,1); plot(inputdata(:,1),inputdata(:,2),'k-');
set(gca,'xlim',[min(real(inputdata(:,1))) max(real(...
inputdata(:,1)))],'ylim',[1.1*min(real(inputdata(:,2))) ...
1.1*max(real(inputdata(:,2)))]);
axis([min(inputdata(:,1)) max(inputdata(:,1)) ...
    (min(inputdata(:,2))-...
0.5*abs(min(inputdata(:,2))))...
    1.1*max(inputdata(:,2))]);
set(gca,'Fontsize',8);
xlabel('Year');ylabel('Salinity');
subplot(2,1,2); plot(datamatrix(:,1),datamatrix(:,2),'k-');
set(gca,'xlim',[min(real(datamatrix(:,1))) max(real(...
datamatrix(:,1)))],'ylim',[1.1*min(real(datamatrix(:,2))) ...
1.1*max(real(datamatrix(:,2)))]);
axis([min(datamatrix(:,1)) max(datamatrix(:,1)) ...
    (min(datamatrix(:,2))...
-0.5*abs(min(datamatrix(:,2))))...
    1.1*max(datamatrix(:,2))]);
set(gca,'Fontsize',8);
xlabel('Year');ylabel('Salinity');
lines=1;
```

```
freqs=[];sigfreqs=[];brkvals=[];freqsofint=[];sigpositions=[];
funcper=[];sigpowers=[];
np=0;

x=datamatrix(:,1);
y=datamatrix(:,2);
tmin=min(x);
tmax=max(x);
n=length(x);
fprintf('\n\nn = %i\n',n);
info{lines}=sprintf('n = %i',n);
lines=lines+1;
fprintf('\ntmin = %f, tmax = %f\n\n',tmin,tmax);
info{lines}=sprintf('tmin = %f, tmax = %f',tmin,tmax);
lines=lines+1;

period(highestf);

if ~isempty(freqs)
%CREATE SPECTRUM FIGURE
figure('name','Power Spectrum','NumberTitle','off');
plot(freqs(:,1),freqs(:,2),'color','k');
spectimage=gca;
```

```
axis([0 fhi 0 1.1*max(freqs(:,2))]);
set(gca,'Fontsize',14);
xlabel('Frequency (Hz)');ylabel('Power Spectrum');
    title(['Threshold value is ',num2str(ThreshCoeff),...
        ' times SD']);
    axes(spectimage);
end %if ~isempty(freqs)

if ~isempty(freqs)

%GENERATE TABLE OF PROBABILITIES
%Generating expytable.
expytable=exp(-freqs(:,2));
%Generating corresponding alpha values.
alphas=1-(1-expytable).^effm;
%Correcting for alpha = 0. (This ensures unique values
    %in "highly significant" regions.)
for i=1:length(alphas)
if alphas(i)==0
alphas(i)=rand/1e20;
end
    end
```

```
%CALCULATE GIVEN SIGNIFICANCE LEVELS FOR GRAPH
alph=[0.001 0.005 0.01 0.05 0.1 0.5];
lineat=log(1./(1-(1-alph).^(1/effm)));

%PROBABILITY LINES and LABELS
for i=1:length(alph)
line([freqs(1,1),0.85*fhi],[lineat(i),lineat(i)],...
        'color','black','linewidth',i/5,'linestyle','-.');
text(0.87*fhi,lineat(i),['a = ' num2str(alph(i))],...
        'fontsize',8,'fontname','symbol');
    end

%DETERMINE WHICH FREQUENCIES ARE SIGNIFICANT
fvals=find(alphas<=0.05)';
freqsofint=freqs(fvals,1)';
lenstring=length(freqsofint)-1;
for i=1:lenstring
%fprintf('Checking frequency %i of %i for
        %significance.\n',i,lenstring);
if freqsofint(i)>=freqsofint(i+1)
freqsofint=freqsofint(1:i);
end
end
alphasofint=alphas(fvals,1)';
```

```
%FIND POSITIONS OF BREAKS IN FVALS (FOR DATA CLUSTERING)
for i=1:length(fvals)-1
if fvals(i)~=fvals(i+1)-1
brkvals=[brkvals i];
end
end
brkvals=[brkvals length(fvals)];


fprintf...
    ('\n\nLAST STAGE... locating significant frequencies.\n\n');
%LOCATE SIGNIFICANT FREQUENCIES
for i=1:length(brkvals)
if i==1
minalph=min(alphasofint(1:brkvals(1)));
sigfreqs=[sigfreqs freqsofint(alphasofint==...
            minalph)];
else
minalph=min(alphasofint(brkvals(i-1)+...
            1:brkvals(i)));
sigfreqs=[sigfreqs freqsofint(alphasofint==...
            minalph)];
end %if i=1;
end %for i=1:...
```

```
%END LOCATE

    %TO WRITE THE SIGNIFICANT FREQUENCIES OUT OT A FILE
    %if ~isempty(sigfreqs)
    %xlswrite('sigfreqs',sigfreqs');
    %end

    %POWER AT SIGNIFICANT FREQUENCIES
for i=1:length(sigfreqs)
sigpositions=[sigpositions find(freqs(:,1)==...
        sigfreqs(i))];
end
sigpowers=freqs(sigpositions,2)';
if ~isempty(sigfreqs)
fprintf(['\n\nSignificant frequencies (in Hz) at ',...
            num2str(sigfreqs)]);
info{lines}=sprintf(['Significant frequencies at ',...
            num2str(sigfreqs)]);
lines=lines+1;
fprintf(['\nCorresponding powers:              ',...
            num2str(sigpowers)]);
info{lines}=sprintf(['Corresponding powers:    ',...
            num2str(sigpowers)]);
lines=lines+1;
```

```
else
fprintf('\nNo significant frequencies found.\n');
info{lines}=...
        sprintf('No significant frequencies found.');
lines=lines+1;
    end


end %if isempty(freqs)


fprintf('\n\nANALYSIS IS COMPLETE\n\n');


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%SUBFUNCTION
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function period(highestf)
global arg ave c cc cwtau effm ep expy fhi freqs hifac ...
       info ival jmax jval lines n nmax nout np ofac ...
       pnow prob px py pymax s ss sumc sumcy sums sumsh ...
       sumsy swtau tmax tmin variance wi wpi wpr wr ...
       wtau wtemp xmin xmax x xave xdif y yy


fhi = highestf;
```

```
hifac=fhi*2*(tmax-tmin)/n;
ofac = 4;
pause(0.1);
close(findobj('name','ofac'));

fprintf('\nhifac = %f',hifac);
info{lines}=sprintf('hifac = %f',hifac);
lines=lines+1;
fprintf('\nofac = %f',ofac);
info{lines}=sprintf('ofac = %f',ofac);
lines=lines+1;

np=ofac*hifac*n*0.5;
nout=floor(0.5*ofac*hifac*n);
ave=mean(y);
variance=var(y);
xmin=tmin;
xmax=tmax;
xdif=tmax-tmin;
xave=0.5*(xmax+xmin);
pymax=0;
pnow=1/(xdif*ofac);

fprintf('\nComputing %i JVALS.\n\n',n);
```

```
hbar = waitbar(0,'JVALS...');
for jval=1:n
waitbar(jval/n);
arg=2*pi*((x(jval)-xave)*pnow);
wpr(jval)=-2*sin(0.5*arg)^2;
wpi(jval)=sin(arg);
wr(jval)=cos(arg);
wi(jval)=wpi(jval);
end
close(hbar)

fprintf('\nIVAL: Computing %i values.\n\n',nout);

hbar = waitbar(0,'IVALS...');
for ival=1:nout
waitbar(ival/nout);
px(ival)=pnow;
sumsh=0;
sumc=0;
for jval=1:n
c=wr(jval);
s=wi(jval);
sumsh=sumsh+s*c;
```

```
sumc=sumc+(c-s)*(c+s);
end
wtau=0.5*atan2(2*sumsh,sumc);
swtau=sin(wtau);
cwtau=cos(wtau);
sums=0;
sumc=0;
sumsy=0;
sumcy=0;
for jval=1:n
s=wi(jval);
c=wr(jval);
ss=s*cwtau-c*swtau;
cc=c*cwtau+s*swtau;
sums=sums+ss^2;
sumc=sumc+cc^2;
yy=y(jval)-ave;
sumsy=sumsy+yy*ss;
sumcy=sumcy+yy*cc;
wtemp=wr(jval);
wr(jval)=(wr(jval)*wpr(jval)-wi(jval)*wpi(jval))...
        +wr(jval);
wi(jval)=(wi(jval)*wpr(jval)+wtemp*wpi(jval))...
        +wi(jval);
```

```
end
py(ival)=0.5*(sumcy^2/sumc+sumsy^2/sums)/variance;

%WRITE OUTPUT
freqs(ival,1)=px(ival);
freqs(ival,2)=py(ival);
pnow=pnow+1/(ofac*xdif);
end
close(hbar);

pymax=max(py);
jmax=find(py==pymax);

expy=exp(-pymax);

%effm is an estimate of the number of
%'independent' frequencies
effm=2*nout/ofac;

if ~isempty(effm) & effm~=0
prob=1-(1-expy)^effm;

fprintf('\npymax = %f',pymax);
info{lines}=sprintf('pymax = %f',pymax);
```

```
lines=lines+1;
fprintf('\nfmax = %f',px(jmax));
info{lines}=sprintf('fmax = %f',px(jmax));
lines=lines+1;
fprintf('\neffm = %f',effm);
info{lines}=sprintf('effm = %f',effm);
lines=lines+1;
fprintf('\nexpy = %f',expy);
info{lines}=sprintf('expy = %f',expy);
lines=lines+1;
fprintf('\nnout = %i',nout);
info{lines}=sprintf('nout = %i',nout);
lines=lines+1;
fprintf('\nalpha = %f',prob);
info{lines}=sprintf('alpha = %f',prob);
lines=lines+1;

else
fprintf('period.m: No frequencies to analyze.');
end %if ~isempty(effm) & effm~=0
return
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function [DroughtData] = cut(noisedata,coeff)
```

```
indicatorPosi = find(noisedata(:,2)>0);

posi = noisedata(indicatorPosi,2);

SdofPosi = std(posi);

thresh = coeff*SdofPosi;

thresh

indicatorThresh = find(noisedata(:,2)<thresh);

noisedata(indicatorThresh,2) = thresh;

DroughtData = noisedata;

return
```

Bibliography

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice-Hall Inc, 1994.

Brown, K. J., J. S. Clark, E. C. Grimm, J. J. Donovan, P. G. Mueller, B. C. S. Hansen, and I. Stefanova. Fire cycles in north american interior grasslands and their relation to prairie drought. *Proceedings of the National Academy of Sciences*, 102:8865–8870, 2005.

Cumming, B. F., S. E. Wilson, R. I. Hall, and J. P. Smol. *Diatoms from British Columbia (Canada) Lakes and their Relationship to Salinity, Nutrients and Other Limnological Variables.* J. Cramer, Stuttgart, 1995.

Fritz, S. C. Paleolimnological records of climatic change in north america. *Limnology and Oceanography*, 41:882–889, 1996.

Gasse, F., P. Barker, P. A. Gell, S. C. Fritz, and F. Chalie. Diatom-inferred salinity in palaeolakes: An indirect tracer of climate change. *Quaternary Science Reviews*, 16:541–563, 1997.

Gonzalez, J., J. B. Valdes, and F. Asce. Bivariate drought recurrence analysis using tree ring reconstructions. *Journal of Hydrologic Engineering*, 8:247–258, 2003.

Hall, R. I., P. R. Leavitt, R. Quinlan, A. S. Dixit, and J. P. Smol. Effects of agriculture, urbanization, and climate on water quality in the northern great plains. *Limnology and Oceanography*, 44:739–756, 1999.

Horne, James H. and Sallie L. Baliunas. A prescription for period analysis of unevenly sampled time series. *Astrophys. J. (USA)*, 302:757–763, 1986.

Laird, K. R., B. F. Cumming, S. Wunsam, J. A. Rusak, R. J. Oglesby, S. C. Fritz, and P. R. Leavitt. Lake sediments record large-scale shifts in moisture regimes across the northern prairies of north america during the past two millennia. *Proceedings of the National Academy of Sciences of the United States of America*, 100:2483–2488, 2003.

Laird, K. R., S. C. Fritz, and B. F. Cumming. A diatom-based reconstruction of drought intensity, duration, and frequency from moon lake, north dakota: a sub-decadal record of the last 2300 years. *Journal of paleolimnology*, 19:161–179, 1998.

Laird, K. R., S. C. Fritz, K. A. Maasch, and B. F. Cumming. Greater drought intensity and frequency before ad 1200 in the northern great plains, usa. *Nature*, 384:552–554, 1996.

Lomb, N. R. Least-squares frequency analysis of unequally spaced data. *Astrophys. Space Sci. (Netherlands)*, 39:447–462, 1976.

Maybank, J., B. Bonsal, K. Jones, R. Lawford, E. G. O'Brien, E. A. Ripley, and E. Wheaton. Drought as a natural disaster. *Atmosphere-ocean*, 33:195–222, 1995.

Press, W. H. and G. B. Rybicki. Fast algorithm for spectral analysis of unevenly sampled data. *Astrophys. J. (USA)*, 338:277–280, 1989.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery.

Numerical recipes in C++. Cambridge University Press, Cambridge, 2002. ISBN 0-521-75033-4. The art of scientific computing, Second edition, updated for C++.

Priestley, M. B. *Spectral analysis and time series. Vol. 1.* Academic Press Inc. [Harcourt Brace Jovanovich Publishers], London, 1981. ISBN 0-12-564901-0. Univariate series, Probability and Mathematical Statistics.

Rusak, J. A., P. R. Leavitt, S. McGowan, G. Chen, O. Olson, S. Wunsan, and B. F. Cumming. Millennial-scale relationships of diatom species richness and production in two prairie lakes. *Limnology and Oceanography*, 49:1290–1299, 2004.

Scargle, Jeffrey D. Studies in astronomical time series analysis. II. statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J. (USA)*, 263:835–853, 1982.

Schulz, M. and K. Stattegger. Spectrum: Spectral analysis of unevenly spaced paleoclimatic time series (STMA V40 1186). *Computers & Geosciences*, 23:929–945, 1997.

Shumway, Robert H. and David S. Stoffer. *Time series analysis and its applications.* Springer Texts in Statistics. Springer, New York, second edition, 2006. ISBN 978-0387-29317-2; 0-387-29317-5. With R examples.

Smakhtin, V. U. and D. A. Hughes. Automated estimation and analyses of meteorological drought characteristics from monthly rainfall data. *Environmental Modelling & Software*, 22:880–890, 2007.

Stahle, D. W. and M. K. Cleaveland. Texas drought history reconstructed and analyzed from 1698 to 1980. *Journal of Climate*, 1:59–74, 1988.

Stockton, C. W. and D. M. Meko. Drought recurrence in the great plains as reconstructed from long-term tree-ring records. *Journal of Climate and Applied Meteorology*, 22:17–29, 1983.

Tian, J., D. M. Nelson, and F. S. Hu. Possible linkages of late-holocene drought in the north american midcontinent to pacific decadal oscillation and solar activity. *Geophysical Research Letters*, 33:L23702, 2006.

Trenberth, Kevin E., Grant W. Branstator, and Phillip A. Arkin. Origins of the 1988 north american drought. *Science*, 242:1640–1645, 1988.

Woodhouse, C. A. and J. T. Overpeck. 2000 years of drought variability in the central united states. *Bulletin of the American Meteorological Society*, 79:2693–2714, 1998.

Yu, Z. and E. Ito. Possible solar forcing of century-scale drought frequency in the northern great plains. *Geology*, 27:263–266, 1999.