Spring 1999

Dynamic analysis of unevenly sampled data with applications to statistical process control

Laura Ann McSweeney
University of New Hampshire, Durham

DYNAMIC ANALYSIS OF UNEVENLY SAMPLED DATA WITH APPLICATIONS TO STATISTICAL PROCESS CONTROL

BY

Laura A. McSweeney

B.S., Bridgewater State College (1993)
M.S., University of New Hampshire (1996)

DISSERTATION

Submitted to the University of New Hampshire
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

in

Mathematics

May 1999

UMI Number: 9926028

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

This dissertation has been examined and approved.

Kevin M. Short

Director, Kevin M. Short
Professor of Mathematics

Debajyoti Sinha

Debajyoti Sinha
Professor of Mathematics

Ernst Linder

Ernst Linder
Professor of Mathematics

Philip J Ramsey

Phil Ramsey
Faculty in Residence, Mathematics

John Geddes

John Geddes
Faculty in Residence, Mathematics

4/20/99

Date

Dedication

This work is dedicated with love to my family
and to the memory of Uncle Mike.

Acknowledgments

There are many people who influenced the outcome of this work. While their level of involvement varied from direct input to subtle guidance, I would like to express my gratitude for their support.

First, I would like to thank my advisor, Kevin M. Short, who is the ideal advisor. He is an inspiring mentor, who showed by example how to be an excellent teacher and superior researcher. I appreciate how he always went out of his way to advise and assist me. He was always available to critique a presentation, proof an article, answer a question or suggest an interesting research problem. Thanks to his preparations and encouragement during my employment search, and the *many* presentations I had to give, I am no longer apprehensive about giving talks! This is just one small way in which he has helped me to mature mathematically and professionally.

I would also like to thank Debajyoti Sinha, Ernst Linder, Phil Ramsey and John Geddes; the other members of my Dissertation Committee. I appreciate their insightful comments on how to improve this work. More importantly though, I want thank them for sharing with me their knowledge and enthusiasm about applied mathematics and statistics.

Finally, I would like to thank my emotional support group: my parents Dave and Aldona McSweeney, my brother David, my sister-in-law Lisa and all my friends. Their encouragement, love and prayers have made the past six years memorable.

God bless you all!

TABLE OF CONTENTS

List of Tables

List of Figures

ABSTRACT

DYNAMIC ANALYSIS OF UNEVENLY SAMPLED DATA WITH APPLICATIONS TO STATISTICAL PROCESS CONTROL

by

Laura A. McSweeney
University of New Hampshire, May, 1999

Dynamic analysis involves describing how a process changes over time. Applications of this type of analysis can be implemented in industrial settings in order to control manufacturing processes and recognize when they have changed significantly. The primary focus of this work is to construct methods to detect the onset of periodic behavior in a process which is being monitored using a scheme where data is sampled unevenly.

Techniques that can be used to identify statistically significant periodic structure using the periodogram will be reviewed and developed. The statistical properties of the periodogram for unevenly sampled data will be calculated. These statistics reveal that standard methods applied to randomly sampled data give incorrect results, especially for small sample sizes. These standard tests are not designed specifically for data collected at random times. Monte Carlo methods are used to adjust the critical values used for testing the significance of spectral peaks. The effectiveness of the tests for determining periodic behavior are compared using the standard critical values and the adjusted values. The adapted test is then extended into a control chart which will signal when periodic behavior enters into an irregularly sampled process.

The new methods are applied to an industrial example from a silicon wafer coating process. The data was collected irregularly and the underlying dynamics of the process were investigated. Interesting periodic behavior was uncovered in the analysis.

When data has complicated oscillatory behavior, methods of nonlinear dynamic analysis can be used to make predictions. A new toroidal reconstruction technique is developed for data that appears to be driven predominantly by two or three frequencies. Comparisons between the new method and a standard time delay reconstruction utilizing nonlinear dynamic forecasting methods are made using simulated and real-world data collected from a vibrating warehouse air duct.

Chapter 1

INTRODUCTION

Statistical process control can be used to monitor many industrial processes. Common statistical process control techniques, like control charting, generally assume that the data, collected over time, is random and normally distributed. This may not be the case, since the data may have deterministic components which need to be modeled and removed before the analysis of the random error component can be done. The goal of modeling the deterministic part of the data is to find a model which not only fits the data, but also gives information regarding the underlying dynamics of the system. By learning how the process changes over time, factors which influence production might be identified. New information can then be incorporated to improve the process. Also, by combining dynamical modeling with statistical process control it should be possible to predict and detect process changes. A variety of methods are available to model the observations depending on what characteristics are present. Typically, statistical methods are used to remove any trends or correlation in the time series. However, data may have components which are periodic or chaotic in nature and statistical techniques may not be able to model these phenomena.

Let $\{X_j\}_{j=1}^{N}$ represent the evenly sampled time series where the value X_j is measured at time t_j. A general assumption for time series modeling is that the data is *weakly stationary*. Therefore, the mean and variance of the data remains constant over time and the autocorrelation structure depends only on the lag. A plot of the data as a function of time may indicate that the process is nonstationary. For example, the plot may reveal trends or cyclic behavior in the data. These effects would need to be removed before implementing standard stochastic models that require stationarity. Global trends in the time series data are usually removed first in order to make the mean constant. Methods such as simple linear or polynomial regression can be implemented to remove these trends. For example, if the overall mean of the data seems to increase linearly over time, as in Figure 1-1a, then the trend can be modeled by

$$X_j = \alpha + \beta t_j + \epsilon_j$$

where α and β are constants and ϵ_j is a random error term with zero mean. In this case the overall mean as a function of time is given by the quantity $\alpha + \beta t$. The parameters α and β can be estimated from the data using the method of least squares. The linear trend for the example is shown in Figure 1-1a. The residuals, ϵ_j, show how the data varies around the trend line. Figure 1-1c illustrates the residuals for the example. Notice that the residuals seems to fluctuate about zero indicating that the trend was removed. Clearly, if the trend appears to be nonlinear, a higher order polynomial can be fit to the data to remove the global trend.

If the data appears to have cyclic behavior, an analysis in the frequency domain can be conducted. Common spectral techniques like the discrete Fourier transform or the periodogram can be used to determine what discrete frequencies dominate the cyclic behavior. The oscillatory behavior can be modeled using least-squares to estimate a sinusoidal model. These specific periodic influences can be removed by subtracting the original data from the predicted values of the sinusoidal model. The result, called the residuals, can then be examined for randomness. Figure 1-2a shows an example of data that oscillates with a period of 5 samples. Least squares methods can be used on the observed data to estimate the constants μ, α and β in the sinusoidal model

$$X_j = \mu + \alpha \cos(2\pi t_j/5) + \beta \sin(2\pi t_j/5) + \epsilon_j.$$

The estimated cycle for this example is shown in Figure 1-2b. The residuals remaining after subtracting the predictions from the estimated model are shown in Figure 1-2c. It appears that the mean has been stabilized by removing the cyclic behavior.

Once the time series appears to be stationary, statistical process control methods can be applied to the residuals if they appear to be stochastically random and normally distributed. If the residuals still demonstrate autocorrelated behavior, statistical time series methods can be used to model the variability. For example, spectral techniques may not be able to remove short term autocorrelations and that would make the residuals look almost periodic.

The most common class of time series models are the autoregressive integrated moving average (ARIMA) processes [8]. These models attempt to predict the current status using the previously observed points. The different ARIMA models are identified by examining the autocorrelation and partial autocorrelation structure of the residuals. The simplest of these models is the autoregressive model of order p, denoted AR(p), where the current data point is dependent on the previous p data points. In particular,

$$X_j = \mu + \sum_{i=1}^{p} \alpha_i X_{j-i} + Z_j$$

where X_{j-i} is the measurement collected at time t_{j-i} and Z_j is an observation from a random process with zero mean and fixed variance. A regression model can be used to estimated the coefficients $\{\alpha_i\}_{i=1}^{p}$. These autoregressive models are used when the autocorrelation function of $\{X_j\}_{j=1}^{N}$ dies off at lag p and the partial autocorrelation function decays exponentially.

The ARIMA class has been extended to model seasonal processes (SARIMA)[14] and long term processes (ARFIMA)[38]. However, these models, by construction, require evenly sampled data. Therefore, these models do not extend naturally to unevenly sampled data.

It may also be possible to model the data using methods designed for chaotic data. While there is no universally accepted definition of chaos, there are agreed upon features of chaotic data. Chaotic data comes from a deterministic system that exhibits long-term aperiodic behavior [70]. Often, data from a chaotic system looks stochastically random due to the high degree of complexity of the system. These systems are aperiodic and have varying amplitude structures. However, since chaotic systems are deterministic, they often have an attractor, or an underlying geometric structure, that can be reconstructed from the data. The nice geometrical properties of the attractor make it possible to make short-term predictions. Long-term predictions are unlikely since chaotic systems have sensitive dependence on initial conditions. Therefore, points that are close together at a particular instant will separate exponentially over time and their trajectories will no longer be correlated. The residual errors produced after making nonlinear dynamic forecasting to make short-term predictions can be examined for randomness. Therefore, it is possible to follow the suggestion of Haykin and Li [34] who state that it is possible to "construct a nonlinear predictive model for a deterministic characterization" and then apply statistical tests "to the prediction error produced by the model".

Unfortunately, most time series modeling and control charting techniques are developed specifically for evenly sampled data. Statistical methods like ARIMA models do not extend naturally to irregularly sampled data. Although it may be possible to extend these ideas to unevenly sampled data, this work will focus mainly on the detection of sinusoids in unequally sampled data. Methods to identify statistically significant periodic behavior in unevenly sampled data using the classical periodogram will be considered. The concept of the periodogram will be introduced in Chapter 2. The next chapter will focus on reviewing and developing methods to detect periodic behavior in data. Various sampling schemes will be discussed. Chapter 4 will extend the statistical tests into control charts to monitor and detect when periodic behavior enters a process. The analysis of an industrial data set will be presented in Chapter 5 using the new methods designed for randomly sampled data. Chapter 6 will briefly describe techniques used to uncover the attractor of nonlinear data as well as to introduce a new toroidal reconstruction technique. Comparisons of two reconstruction methods will be made using nonlinear forecasting on another industrial data set of air handler vibrations. The final chapter will discuss future investigations and extensions.

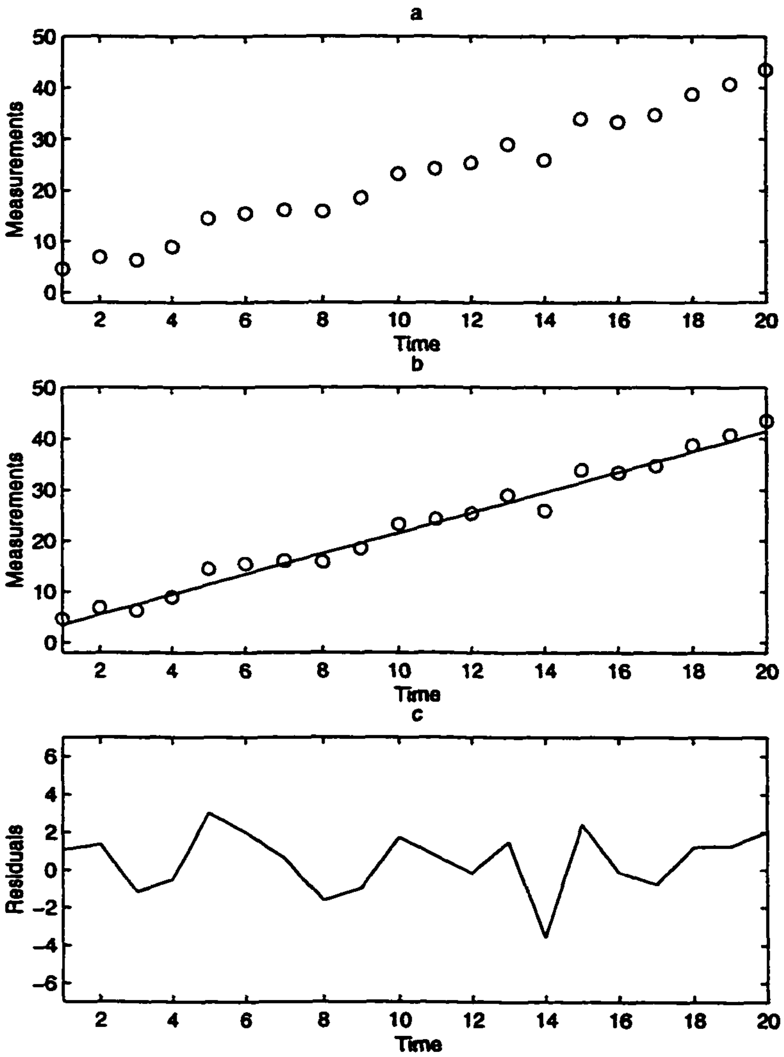


Figure 1-1 The removal of a linear trend

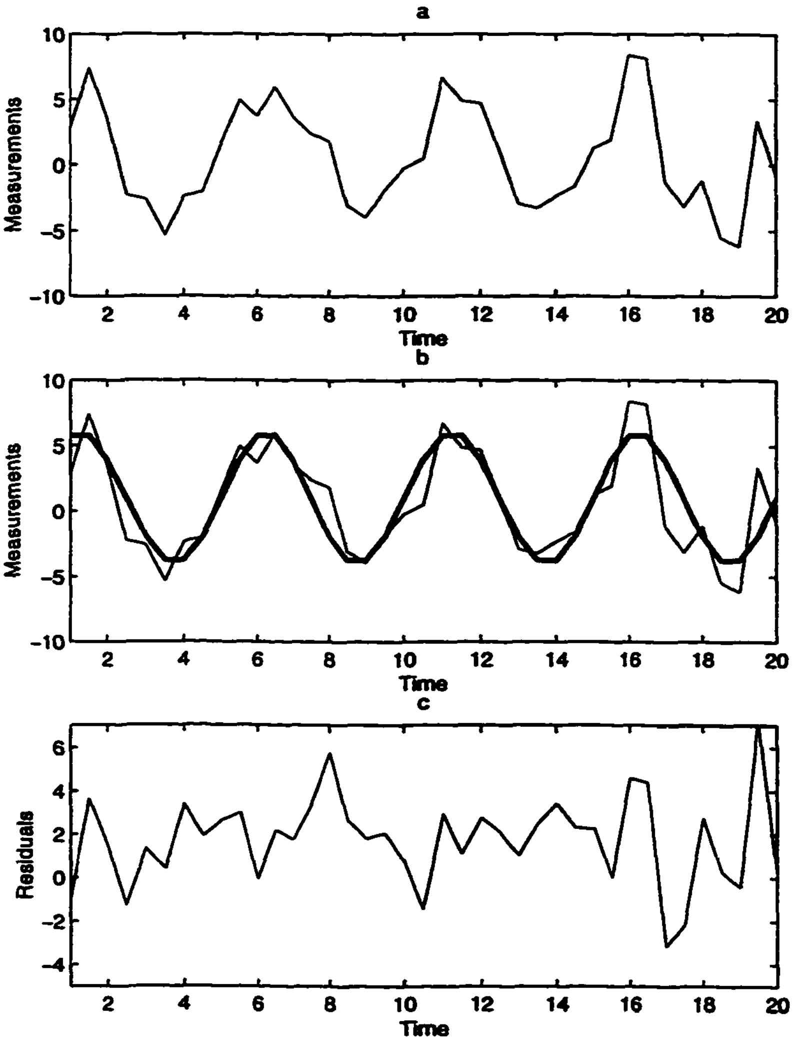


Figure 1-2 The removal of a cyclic component

Chapter 2

PERIODOGRAM ANALYSIS

Many research areas are concerned with the detection of periodic behavior in data. Information regarding cyclic behavior may provide insight about the underlying dynamics of the process being studied. For instance, astronomers may be interested in the cyclic behavior of solar flares, sunspot activity and star intensities. Industries may be concerned about instrument failure or the production of defective products and may wish to determine if these events are periodic in nature.

One common method used to detect cyclic (sinusoidal) behavior is the periodogram, introduced by Schuster [63]. The periodogram, which is quite easy to use, provides a method of searching for underlying periodic behavior. The standard definition of the periodogram will be introduced in section 2.1. In section 2.2, a geometric viewpoint of the periodogram will be discussed to provide a background as to how the periodogram works. These ideas will be demonstrated with simple sinusoid waves and then generalized to more complicated data. The MATLAB codes developed to analyze the data are discussed in the last section.

2.1 Definition

The analysis of time series data using spectral methods begins with the assumption that the data can be modeled by a linear combination of sine and cosine waves. For a given time series, the observations can be represented as $X_1, X_2, \ldots, X_j, \ldots, X_N$ where the jth sample is taken at time t_j and N is the total number of samples. In this section it will be assumed that the data is evenly sampled. When data is sampled regularly, the times can be represented as $t_j = j\Delta$ where Δ is the fixed sampling interval. The goal may be to determine if the process from which the data was collected has an underlying sinusoidal model of the form

$$X_j = \mu + a\cos(\omega t_j) + b\sin(\omega t_j) + \epsilon_j$$

for a particular, fixed (angular) frequency ω. Here ϵ_j represents a random noise term. The parameters μ, a, b are unknown but for a fixed ω they can be estimated from the data using least-squares estimations. If the parameters a and b are determined to be relatively large, one could conclude that there is a periodic cycle with frequency ω in the model.

Most likely, however, the frequency ω is unknown. This creates a more general question of whether the data exhibits any cyclic behavior. In this situation, we would want to check for periodic behavior over a range of ω values. However, the sampling rate, sampling scheme and the length of the data restrict the periodic behavior that can be detected. Consider data collected regularly with sampling interval Δ. In order to detect a sine or cosine wave, at least two samples need to be taken per cycle. Therefore, the highest frequency from which alias-free information can be obtained is $f_c = 1/(2\Delta)$. This frequency is called the Nyquist frequency [14]. The lowest frequency that could possibly be determined is the frequency which oscillates once in $N\Delta$ time units, the length of the time series. Thus, the lowest frequency that can be detected is $f_o = 1/(N\Delta)$.

It is important to note that if data is unevenly sampled, it is possible to detect frequencies higher than the Nyquist frequency without aliasing effects [54]. The lowest independent frequency that should be considered in unevenly sampled data is $1/(t_{max} - t_{min})$, the time span of the data. This frequency corresponds to the cycle that oscillates once through the course of the data.

In evenly sampled data, a systematic check for periodic behavior can be done using (angular) frequencies of the form

$$\omega_k = 2\pi\left(\frac{2f_c k}{N}\right) = \frac{2\pi k}{N\Delta} \quad \text{for} \quad k = 1, 2, \ldots, M = \lfloor N/2 \rfloor,$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. These frequencies, which shall be refered to as *Fourier basis frequencies*, fall between the lowest frequency that can be detected and the Nyquist frequency. Additionally, these frequencies form an orthogonal basis when the times are evenly sampled. Thus, it is possible to represent the finite set of observed data with a finite discrete Fourier series representation. In particular,

$$X_j = \frac{a_0}{2} + \sum_{k=1}^{M} \left[a_k \cos(\omega_k t_j) + b_k \sin(\omega_k t_j)\right]$$

where ω_k is the kth basis element [26]. The coefficients, a_0, a_k and b_k for $k = 1, 2, \ldots, M$ can be determined using least-squares estimation on the model

$$X_j = \mu + a_k \cos(\omega_k t_j) + b_k \sin(\omega_k t_j) + \epsilon_j.$$

Alternately, the same estimates can be derived using the orthogonality properties of the basis elements which are [14]

$$\sum_{j=1}^{N} \cos(\omega_k t_j) = \sum_{j=1}^{N} \sin(\omega_k t_j) = 0$$

$$\sum_{j=1}^{N} \cos(\omega_k t_j)\cos(\omega_l t_j) = \sum_{j=1}^{N} \sin(\omega_k t_j)\sin(\omega_l t_j) = \begin{cases} 0 & \text{if } k \neq l \\ N & \text{if } k = l = N/2 \\ N/2 & \text{if } k = l \neq N/2 \end{cases}$$

and

$$\sum_{j=1}^{N} \cos(\omega_k t_j)\sin(\omega_k t_j) = 0.$$

These two approaches yield the estimates of the amplitudes

$$\hat{a}_0 = \frac{\sum_{k=1}^{N} X_k}{N}$$

$$\hat{a}_k = \frac{2}{N}\sum_{j=1}^{N} X_j \cos(\omega_k t_j) \qquad k = 1, 2, \ldots, M$$

and

$$\hat{b}_k = \frac{2}{N}\sum_{j=1}^{N} X_j \sin(\omega_k t_j) \qquad k = 1, 2, \ldots, M.$$

The ultimate goal of *periodogram analysis* of the data is to determine which frequencies, if any, account for the variability in the data. The contribution of the kth frequency component to the model is given by

$$a_k \cos(\omega_k t) + b_k \sin(\omega_k t) = R_k \cos(\omega_k + \phi_k)$$

where the amplitude and phase are, respectively,

$$R_k = \sqrt{a_k^2 + b_k^2}$$

and

$$\phi_k = \tan^{-1}\left(\frac{-b_k}{a_k}\right).$$

The *periodogram estimate* for the kth frequency component is defined to be

$$\begin{aligned}
I(\omega_k) &= \frac{N}{2}\hat{R}_k^2 \\
&= \frac{N}{2}\left(\hat{a}_k^2 + \hat{b}_k^2\right) \\
&= \frac{N}{2}\left[\left(\frac{2}{N}\sum_{j=1}^{N} X_j \cos(\omega_k t_j)\right)^2 + \left(\frac{2}{N}\sum_{j=1}^{N} X_j \sin(\omega_k t_j)\right)^2\right] \\
&= \frac{2}{N}\left[\left(\sum_{j=1}^{N} X_j \cos(\omega_k t_j)\right)^2 + \left(\sum_{j=1}^{N} X_j \sin(\omega_k t_j)\right)^2\right].
\end{aligned}$$

The estimate for a particular ω_k measures the strength of the sinusoidal signal with frequency ω_k. The periodogram estimate is proportional to the amplitude R_k and, therefore, it measures the relative size of the amplitudes of the corresponding sine and cosine waves. For this reason, the periodogram estimate is also referred to as the power or intensity. Obviously, if either a_k or b_k is significantly large, the value R_k would be large and one could conclude that there is a cycle corresponding to ω_k in the observed data. This definition is also equivalent to the sum of squares associated with the kth component [8] which measures the contribution of the kth sinusoidal model to the total variation in the data. If the periodogram estimate for a particular ω_k is significantly large, then there is a signal associated to ω_k in the data. Statistical tests to determine the presence of a large estimate will be discussed and developed in the next chapter.

It should be noted that there are various definitions for the periodogram estimate [4, 7, 8, 14, 54, 80, 81]; however, the definitions are the same up to a scalar multiple. For the rest of this paper, unless otherwise stated, the definition of the *classical periodogram* will be

$$\begin{aligned} P_X(\omega) &= \frac{1}{N}\left[\left(\sum_{j=1}^{N} X_j \cos(\omega t_j)\right)^2 + \left(\sum_{j=1}^{N} X_j \sin(\omega t_j)\right)^2\right] \\ &= \frac{1}{N}\left|\sum_{j=1}^{N} X_j e^{i\omega t_j}\right|^2 . \end{aligned} \qquad (1)$$

The equivalent complex form, equation (1), can be obtained by expanding the complex exponential using Euler's formula and regrouping the complex conjugate pairs.

The periodogram for a given data set is simply a function of ω. Therefore, the power $P_X(\omega)$ can be plotted against ω or against the corresponding period $\tau = 2\pi/\omega$. For example, consider data collected from the sinusoidal model with a period of 4 time units where

$$X_j = 2 + 3\cos(2\pi t_j/4) + \epsilon_j$$

for $j = 1, 2, \ldots, 40$ and the ϵ_j represent identically and independently distributed standard normal errors. The corresponding periodogram as a function of the period is shown in Figure 2-1. This figure illustrates how the periodogram detects potential periodic behavior since the large peak in the periodogram at period 4 indicates that the amplitude of the sine or cosine associated with that period is relatively large. In this contrived situation, the peak is easily distinguishable. Unfortunately, with real world data, the peak values may not be as obvious. This issue will be addressed in the next chapter.

One final comment about the periodogram is its relationship to the discrete Fourier transform which is commonly used for the spectral analysis of discrete data. The discrete

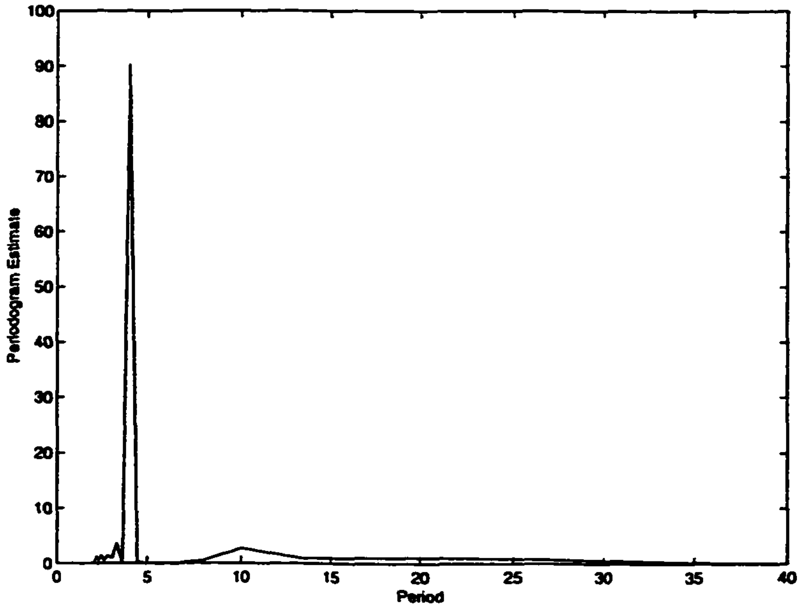


Figure 2-1 Example of a periodogram

Fourier transform, for a particular time series $\{X_j\}_{j=1}^{N}$ and frequency ω, is defined to be

$$\begin{aligned} F_X(\omega) &= \sum_{j=1}^{N} [X_j \cos(\omega t_j) + iX_j \sin(\omega t_j)] \\ &= \sum_{j=1}^{N} X_j e^{-i\omega t_j}. \end{aligned}$$

The periodogram is simply proportional to the squared magnitude of the discrete Fourier transform as shown by the relation

$$P_X(\omega) = \frac{1}{N} |F_X(\omega)|^2.$$

2.2 Geometric Interpretation

Although the standard definition of the periodogram looks notationally intensive, it is straightforward to use. Given a set of observations, the periodogram is easy to code, especially using the complex form of the definition. Examples of the code written for MATLAB are provided in Appendix B. While the periodogram is easy to use, it may not be apparent why a large peak would appear if there is an associated cycle in the data. Fortunately, this concept can be explained geometrically using the complex form of the periodogram found in equation (1).

The equation for $P_X(\omega)$ that was just derived has a nice geometric interpretation by considering the observations about the complex unit circle. Suppose we wish to determine if there is a period $\tau = 2\pi/\omega$ in the observed data. Each observation can be converted to polar coordinates with the radius equal to the observed measurement and the angle determined by the time. In particular, time is essentially wrapped around the unit circle in such a way that one orbit around the unit circle corresponds to the period of interest, τ. Thus an observation (t_j, X_j) can be viewed as the polar coordinate (r_j, θ_j) where $r_j = X_j$ and $\theta_j = \omega t_j = 2\pi t_j/\tau$. An associated vector can be drawn from the origin to this point in the complex plane. The *polar plot* will refer to the plot that results when all theses vectors are plotted. With this viewpoint, the periodogram estimate for a fixed τ and, hence, ω value is equivalent to the squared magnitude of the resultant sum of these vectors scaled down by a factor N, the number of data points. If the data contains a periodic component of frequency ω, the vectors plotted in the complex plane will exhibit a skewed orientation and the magnitude of the resultant will be large, whereas when a frequency is absent in the data, the vectors tend to cancel and produce a small resultant length. Consequently, when the data is collected from a random process. the periodogram estimate would be nearly zero. In contrast, when the data contains a signal of period τ in the data, the vector directions are not centered about the origin and a large periodogram estimate results.

Examples representing different sampling situations will be used to demonstrate how this method works. First, consider data collected evenly using a sampling rate $\Delta = .25$ days from a sine wave with an 8 day period as in the model

$$X(t_j) = 10\sin(2\pi t_j/8)$$

where $j = 1, 2, \ldots, 1000$ and $t_j = j\Delta = .25j$ days. Assuming the true model is unknown, let us check the data for a five and eight day cycle using the method described above for $\tau = 5$ and $\tau = 8$ days. Table 2.1 lists the measurements and the times of the first five observations from this process. The angle transformations for $\tau = 5$ and $\tau = 8$ days are also listed. To determine if the data exhibits a 5 day cycle, the vectors corresponding to the polar coordinates $(X_j, 2\pi t_j/5)$ would be plotted for $j = 1, \ldots, 1000$. The periodogram estimate is equal to the squared magnitude of the resultant of these vectors, divided by 1000. Similarly, the presence of an 8 day cycle can be checked by first plotting the polar coordinates of the form $(X_j, 2\pi t_j/8)$.

The periodogram estimate for $\tau = 5$ is 3.4747 and the polar plot is shown in Figure 2-2 where the endpoint of each vector is represented by the symbol 'o'. The grid lines and numerical labels on the plot indicate the radius and angle scale. It is not surprising that the estimate is small since the vectors are symmetric about both axes resulting in the cancellation

Measurement (Radius)	Time	Theta ($\tau = 5$)	Theta ($\tau = 8$)
1.9509	.25	$\pi/10$	$\pi/16$
3.8268	.5	$\pi/5$	$\pi/8$
5.5557	.75	$3\pi/10$	$3\pi/16$
7.0711	1	$2\pi/5$	$\pi/4$
8.3147	1.25	$\pi/2$	$5\pi/16$

Table 2.1 Polar Coordinate Transformation

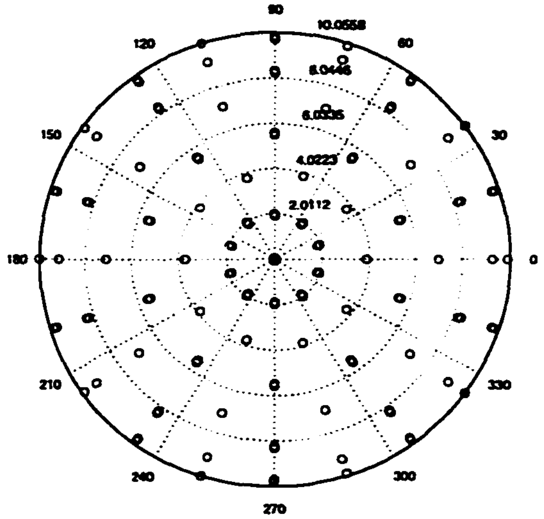


Figure 2-2 Polar plot for $\tau = 5$

of both components of the vectors. One would conclude that since the periodogram estimate is small, the data does not exhibit period-5 behavior. However, the polar plot for period 8, in Figure 2-3, has a periodogram estimate of 25048, which is relatively large indicating that the data contains an 8 day cycle.

The previous example demonstrates how the periodogram can detect periodic behavior, but unfortunately real world data is usually not so "clean." Data may have measurement error, multiple periodic components or be unevenly sampled. The geometric approach can be used to hypothesize and test what effect these factors have on the periodogram's ability

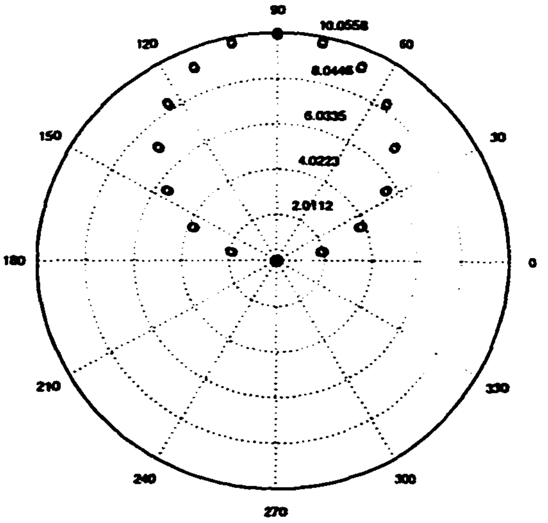


Figure 2-3 Polar plot for $\tau = 8$

to detect periodic behavior.

Consider data which is taken from the model

$$X(t_j) = 10\sin(2\pi t_j/8) + \epsilon_j$$

where the times are evenly sampled as in the previous example and the ϵ_j are random error terms with standard deviation equal to 3. Since the times have not changed, the angles for the noisy data are the same as the pure sinusoid example. The only difference from the first example is that the lengths of the vectors vary slightly from the lengths that occur with "clean" data. The difference in the length of the jth vector is governed by the error term, ϵ_j. Figures 2-4 and 2-5 show the polar plots for $\tau = 5$ and $\tau = 8$ days, respectively. Notice that even though the data points do not overlap, they still fall along the same rays as in Figure 2-2 and Figure 2-3. The periodogram estimates for $\tau = 5$ was 40.9678 and

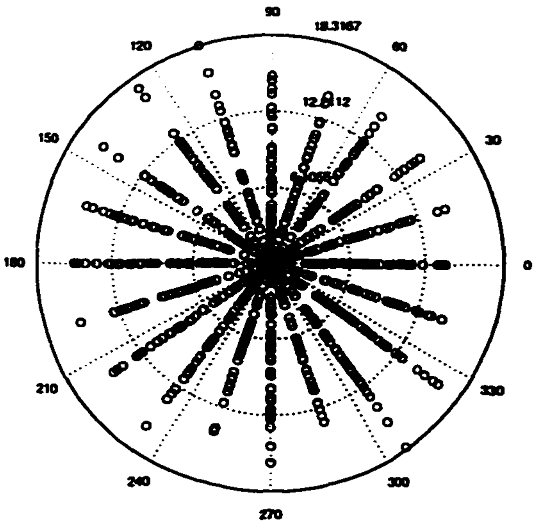


Figure 2-4 Polar plot for $\tau = 5$

the estimate for $\tau = 8$ was 26036. While the periodogram estimates are much larger than in the "clean" data, it is apparent that the periodogram estimate associated with $\tau = 8$ is significantly larger than the other estimate indicating that the data contains an eight day cycle.

The addition of multiple periodic components to the underlying model does not affect the geometric interpretation. Although the vectors scatter in the plane more, periodic behavior of period τ is still detected when the polar plot for τ indicates a pattern with a significantly large periodogram estimate. For example, consider data collected regularly from the following model, which will be called Model 1:

$$X_j = 10\sin(2\pi t_j/8) + 8\cos(2\pi t_j/\pi) + \epsilon_j.$$

For $\tau = 5, \pi$ and 8 days the periodogram estimates were 15.7937, 16276 and 24007. The

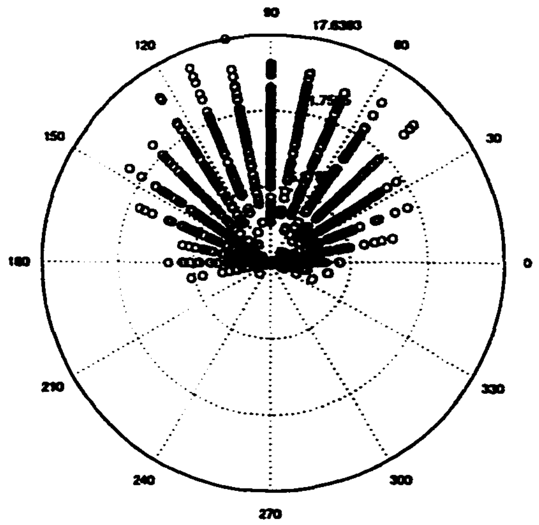


Figure 2-5 Polar plot for $\tau = 8$

data seems to exhibit a π and 8 day cycle, since the estimates for those periods are relatively large compared to the value for $\tau = 5$.

One question that arises is whether the periodogram will signal multiples of the fundamental period. For instance, if there is a cycle with period π in the data, will the periodogram incorrectly detect a 2π cycle? The polar plot with $\tau = 2\pi$ takes twice as as long to wrap time around the unit circle as it does for $\tau = \pi$. Therefore, the pattern formed in the π cycle will occur on the top and bottom half plane.

Figure 2-6 shows the polar plot for $\tau = 2\pi$ using the data collected from Model 1. Since the vectors form a symmetric pattern along both axes, the overall periodogram estimate is 12.052. Therefore, one would not conclude that there was a cycle of period 2π in the data. For comparison the analysis is repeated using data that was taken from the model

$$X_j = 10\sin(2\pi t_j/8) + 8\cos(2\pi t_j/\pi) + 6\sin(2\pi t_j/(2\pi)) + \epsilon_j.$$

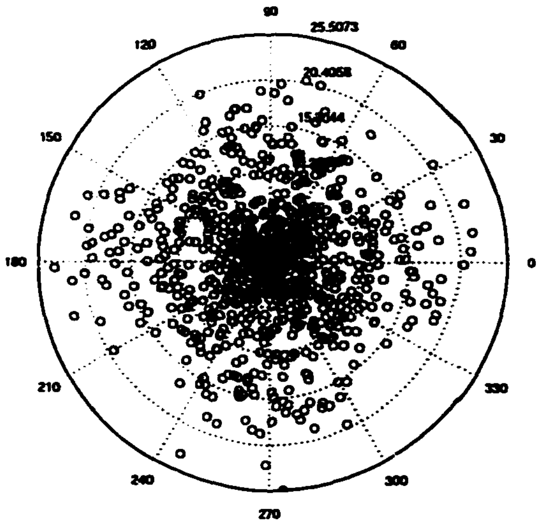


Figure 2-6 Polar plot for $\tau = 2\pi$

Here the data has both a π and 2π day cycle. In this case, the periodogram estimate for $\tau = 2\pi$ would be large (9141.5) since the polar plot has an asymmetric pattern, as shown in Figure 2-7.

These examples demonstrate that patterns in the polar plot seem to occur when τ is a fundamental period or a multiple of the fundamental period of a cycle in the data. However, a large periodogram estimate usually only arises in the situation where τ is a fundamental period.

Since the periodogram definition is flexible enough to handle unevenly sampled times, the geometric method can be used to detect cyclic behavior in unevenly sampled data. Such data often occurs in observational studies when the time of the samples can not always be regulated. For instance, measurements of the intensity of a particular star may be hindered by cloud cover. In these situations, the angles of the vectors can take on any value in $[0, 2\pi)$. To examine the effect of the sampling scheme on the periodogram, consider 1000 samples

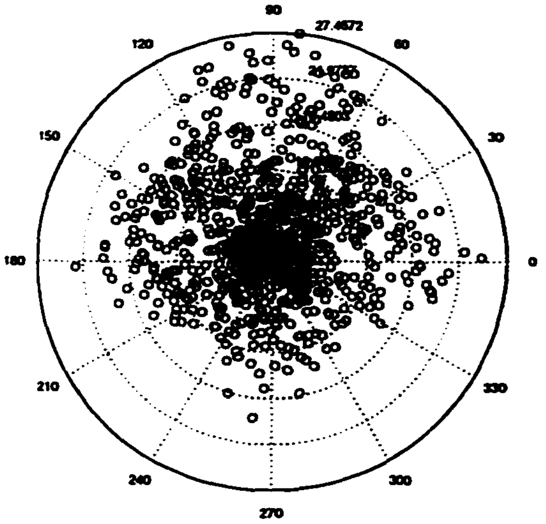


Figure 2-7 Polar plot for $\tau = 2\pi$

from Model 1 taken at random times. Here we let t_j correspond to the time in days of the jth sample and the ϵ_j are still assumed to be random noise. Figure 2-8 shows the polar plot for $\tau = 5$ days, a period not represented in the data. Since the times are unevenly sampled, the vectors no longer fall along the 10 rays as they did in the evenly sampled case. Figure 2-9 shows the polar plot of the 8 day cycle. Although the patterns in the polar plots change dramatically from the evenly sampled cases, periodic behavior is still indicated by a pattern in the plot and a significantly large periodogram estimate. Notice that the periodogram estimate for $\tau = 5$ days (91.8197) is still small relative to the periodogram estimate of the 8 day cycle (26128).

Until now the periodogram has only been evaluated at certain ω values. Since the underlying frequencies are often unknown, a more systematic check for cyclic behavior is desirable. Although the periodogram can be evaluated for any set of ω values, the traditional periodogram for evenly sampled data is calculated for each of the Fourier basis frequencies.

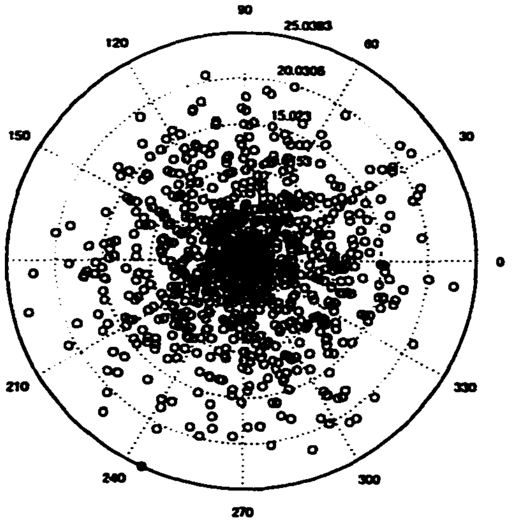


Figure 2-8 Polar plot for $\tau = 5$

However, for unevenly sampled data, a natural orthogonal frequency basis is no longer apparent [44, 54, 61]. In this situation, the choice of frequencies is somewhat arbitrary. The geometric interpretation of the periodogram can help practitioners to refine their choice of a frequency basis.

For instance, consider 1000 observations collected randomly from a noisy process with the underlying model

$$X_j = 5\cos(2\pi t_j/6) + 4\sin(2\pi t_j/2.5) + \epsilon_j.$$

The periodogram of the data evaluated for periods chosen arbitrarily to range from 2 to 21 days with a step size of 1 day, is shown in Figure 2-10. Notice that the periodogram correctly identifies the 6 day cycle, but does not indicate the 2.5 day period. The geometric approach can be used in conjunction with the standard application of the periodogram to

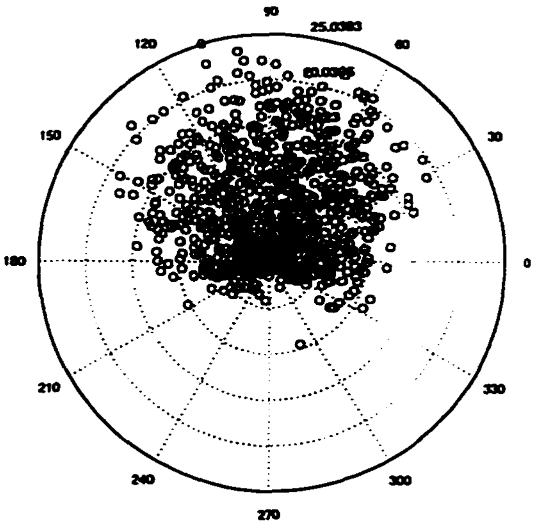


Figure 2-9 Polar plot for $\tau = 8$

help researchers decide if they need to refine their frequency basis. The polar plots for each period tested are shown in Figures 2-11 through 2-15. Notice that there are very distinct, non-random patterns for $\tau = 5, 6, 10, 12, 15$ and 18, yet only one of them has a large periodogram estimate, as shown in Table 2.2. That large periodogram estimate corresponds to the fundamental period of 6. Therefore, patterns at multiples of that period are expected. However, the patterns at 5, 10 and 15 indicate that we may have missed important information. If the frequency basis was refined slightly, by using τ increments of .5 instead of 1, the period of 2.5 would be detected.

This procedure can be quickly and easily incorporated into a movie. For the previous example, the movie would have frames corresponding to the plots in Figures 2-11 through 2-15. This visual environment makes it easy to spot patterns and potential subharmonics that the standard use of the periodogram may miss. Patterns in the polar plots will usually appear at multiples of periods where the periodogram estimate is large. Patterns at other

Period (in days)	Periodogram Estimate
5	78.89
6	5532
10	18.01
12	116.63
15	5.49
18	2.10
20	7.80

Table 2.2 Patterns in the Polar Plot and the Periodogram Estimates

locations indicate that other cyclic behavior may be present in the data. The basis of periods used to calculate the periodogram can then be refined to examine this.

2.3 The Mathematics Behind the MATLAB Code

The three main MATLAB scripts used for the construction of this chapter were polarplot.m, periodogram.m and makemovie.m. These programs can be found in Appendices A, B and C. The polarplot.m routine was used to generate the polar plots. The mathematics behind the program relies on the geometric interpretation, where for a predefined value of τ, the observations are converted into vectors and the squared magnitude of the resultant of all these vectors is calculated. This program requires the input of the observation times and measurements as well as the period τ that we are investigating. In each program, the observed data is centered about zero by subtracting the mean. Then the data is converted into polar coordinates. The radius is simply the measured value and the angle is defined by the operation $2\pi t/\tau$. This transformation can be done in one line due to MATLAB's matrix-vector syntax which eliminates unnecessary loops. The data can then be plotted using MATLAB's built-in polar.m program. The second part of the program calculates the length

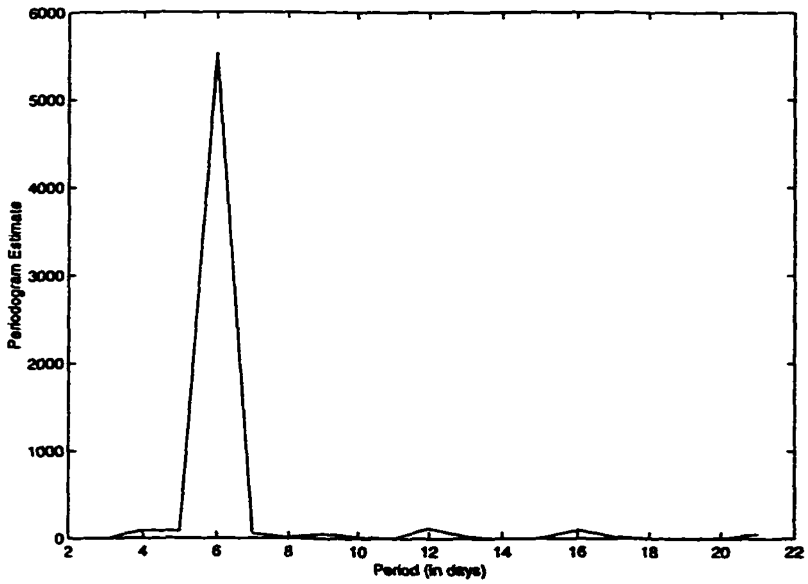


Figure 2-10 Periodogram for unevenly sampled data with a 6 and 2.5 day period

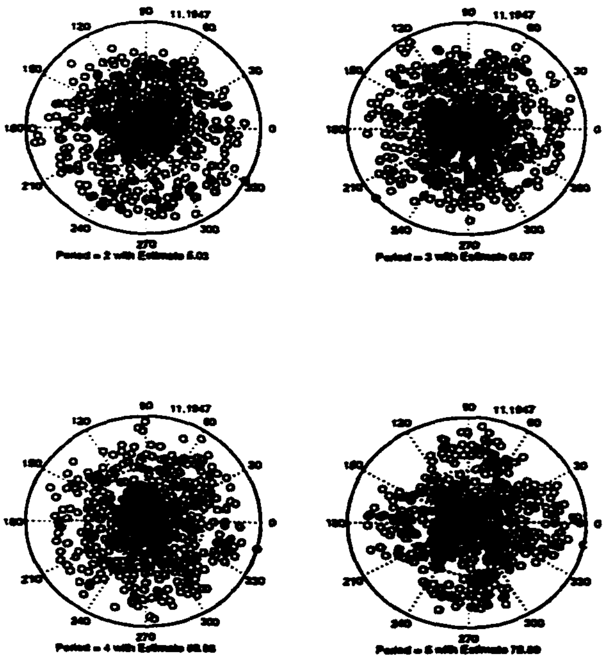


Figure 2-11 Polar plots for $\tau = 2, \ldots, 5$

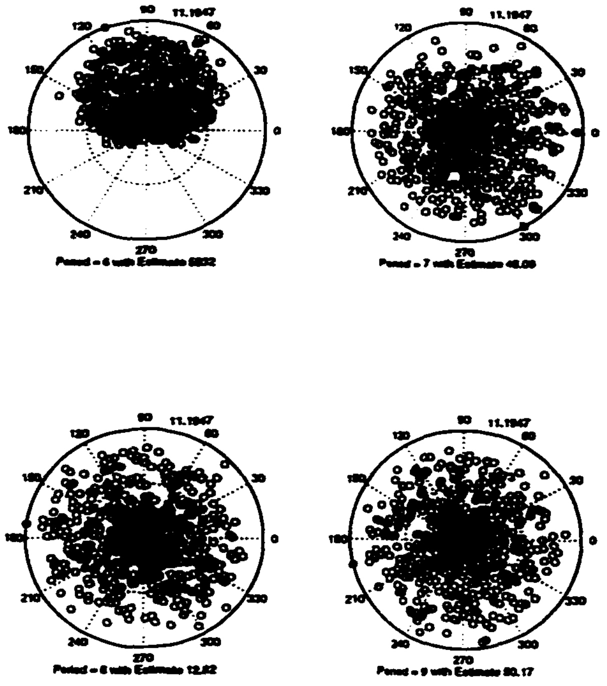


Figure 2-12 Polar plots for $\tau = 6, \ldots, 9$

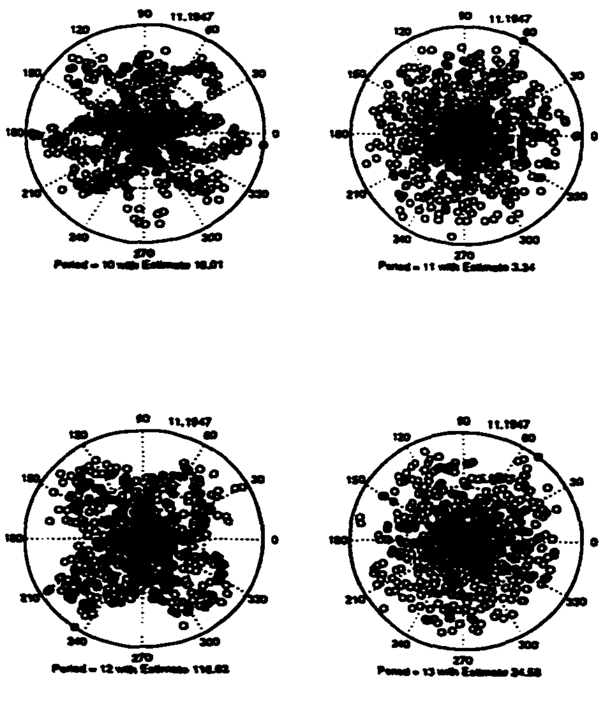


Figure 2-13 Polar plots for $\tau = 10, \ldots, 13$

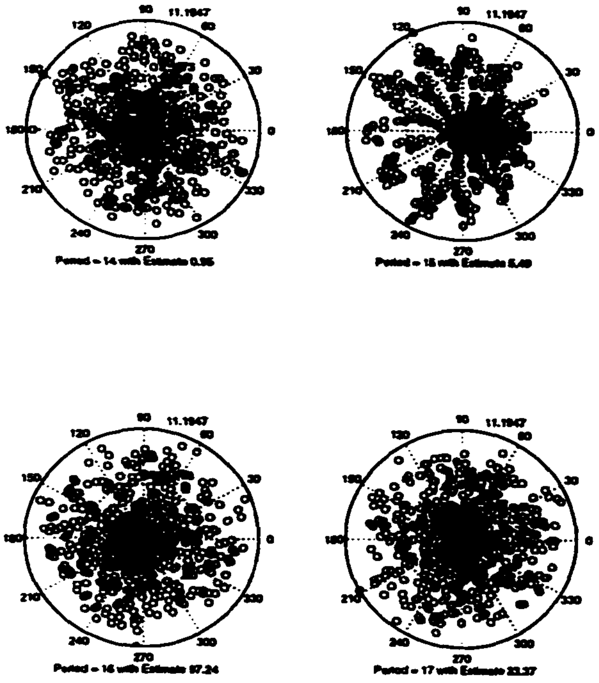


Figure 2-14 Polar plots for $\tau = 14, \ldots, 17$

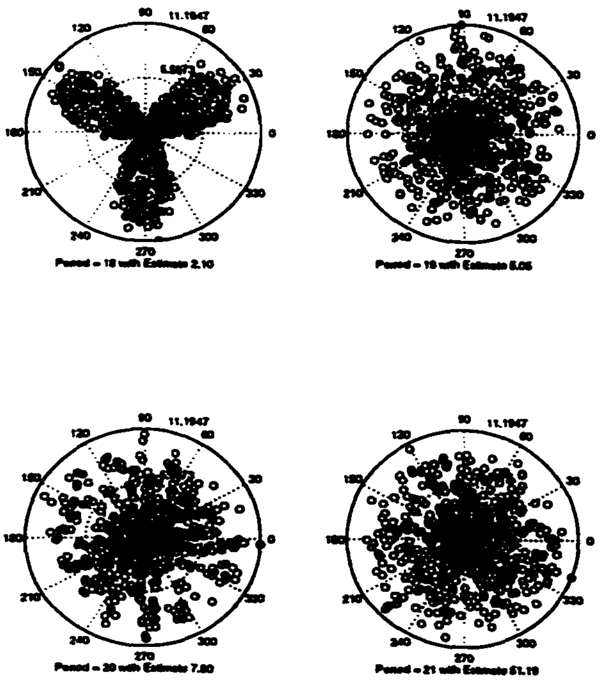


Figure 2-15 Polar plots for $\tau = 18, \ldots, 21$

of the resultant of the vectors. This is done simply by converting the polar coordinates into rectangular coordinates using the built-in function pol2rect.m and calculating the magnitude of the resultant. The periodogram estimate is found by squaring and normalizing by the number of data points.

The periodgram is also easy to code, very efficient to run and can be used for both evenly sampled and unevenly sampled data. The program periodogram.m takes advantage of MATLAB's matrix manipulations and complex data types. This allows the simultaneous calculation of the periodogram estimates for a large number of periods, in a matter of seconds. The observation times and measurements are inputted along with the periods where cyclic behavior is to be tested. The observations are centered and the times are then converted into angles as before. However, matrix multiplication allows us to convert the times into angles for all of the periods in one step. Multiplying the angle matrix by the measurements yields a column vector with the resultants corresponding to each period. The function abs.m calculates the length of the resultant for each period. Finally, the periodogram estimate is calculated by squaring the magnitudes of the resultant and normalizing by the number of samples.

To illustrate all the mathematics compressed into this eight-lined program, a simple example will be demonstrated. For the sake of notation vectors will be represented with arrows and matrices will be denoted by bold capital letters. Let the observation vector $\vec{x} = (x_1, x_2, x_3, x_4, x_5)$ and the time vector $\vec{t} = (t_1, t_2, t_3, t_4, t_5)$ be 1 × 5 row vectors. We assume that $\vec{x}$ has mean 0. Suppose we wish to examine the data for cyclic behavior of periods $\tau = \tau_1, \tau_2$ and τ_3. Following the algebraic manipulations and notation outlined in the periodogram program, the periods are first transformed into angular frequencies.

$$\vec{w} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \begin{bmatrix} 2\pi/\tau_1 \\ 2\pi/\tau_2 \\ 2\pi/\tau_3 \end{bmatrix}.$$

By multiplying $\vec{w}$ and $\vec{t}$ we create the matrix containing the angles defined by the times, for each period. Notice that each row contains the angles corresponding to a specific period.

$$\mathbf{W} = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} [t_1 \; t_2 \; t_3 \; t_4 \; t_5] = \begin{bmatrix} w_1t_1 & w_1t_2 & \cdots & w_1t_5 \\ w_2t_1 & w_2t_2 & \cdots & w_2t_5 \\ w_3t_1 & w_3t_2 & \cdots & w_3t_5 \end{bmatrix}.$$

Then the matrix $i\mathbf{W}$ is exponentiated. The matrix $\mathbf{E}$ contains all the complex angles that result when the times for each period are transformed.

$$\mathbf{E} = e^{i\mathbf{W}} = \begin{bmatrix} e^{iw_1t_1} & e^{iw_1t_2} & \ldots & e^{iw_1t_5} \\ e^{iw_2t_1} & e^{iw_2t_2} & \ldots & e^{iw_2t_5} \\ e^{iw_3t_1} & e^{iw_3t_2} & \ldots & e^{iw_3t_5} \end{bmatrix}.$$

The resultants associated with each period are calculated by multiplying matrix $\mathbf{E}$ and the transpose of the observation vector, $\vec{x}$.

$$\mathbf{R} = \mathbf{E}\vec{x}^T$$

$$= \begin{bmatrix} e^{iw_1t_1} & e^{iw_1t_2} & \dots & e^{iw_1t_5} \\ e^{iw_2t_1} & e^{iw_2t_2} & \dots & e^{iw_2t_5} \\ e^{iw_3t_1} & e^{iw_3t_2} & \dots & e^{iw_3t_5} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

$$= \begin{bmatrix} x_1e^{iw_1t_1} + x_2e^{iw_1t_2} + \dots + x_5e^{iw_1t_5} \\ x_1e^{iw_2t_1} + x_2e^{iw_2t_2} + \dots + x_5e^{iw_2t_5} \\ x_1e^{iw_3t_1} + x_2e^{iw_3t_2} + \dots + x_5e^{iw_3t_5} \end{bmatrix}.$$

Finally, the magnitude of the each resultant is calculated, squared and then scaled down by the number of samples. The plot of the periodogram is constructed by plotting P vs. $\vec{\tau}$ where

$$\begin{aligned} \vec{P} &= \frac{|\mathbf{R}|^2}{N} \\ &= \begin{bmatrix} \left|x_1e^{iw_1t_1} + x_2e^{iw_1t_2} + \dots + x_5e^{iw_1t_5}\right|^2/N \\ \left|x_1e^{iw_2t_1} + x_2e^{iw_2t_2} + \dots + x_5e^{iw_2t_5}\right|^2/N \\ \left|x_1e^{iw_3t_1} + x_2e^{iw_3t_2} + \dots + x_5e^{iw_3t_5}\right|^2/N \end{bmatrix} \end{aligned}$$

and

$$\vec{\tau} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}.$$

The output from this program is used by the script makemovie.m to create a visual display of the polar plots. In this format, patterns in the polar plot are very distinctive. Each frame of the movie corresponds to a polar plot at a specific period. The output of makemovie.m can then be viewed using the MATLAB program, movie.m. The individual frames from such a movie are shown in Figures 2-11 through 2-15.

Chapter 3

TESTING FOR PERIODIC BEHAVIOR

The periodogram is a method commonly used to detect periodic behavior in time series data. Cyclic behavior attributed to a certain frequency is determined to be present in a data set if there is a significantly large periodogram estimate at the frequency. Statistical analyses can be used to determine if an estimate is significantly large or not.

The focus of this chapter will be to review and to derive statistical tests which will determine if a peak is due to the presence of a periodic signal. Since the statistical distributions will depend on the sampling scheme the chapter will be divided into three sections which address evenly sampled data, unevenly sampled data and randomly sampled data. The first two sections will discuss methods previously developed for evenly and unevenly sampled data, in addition to some new applications of nonparametric and parametric tests. The final section will introduce new research in which data collected at random times is analyzed for periodic behavior.

3.1 Evenly Sampled Data

It is commonly assumed that data can be collected using a fixed sampling rate. Preferences for this type of data collection stem from a variety of reasons. It may be convenient and inexpensive to collect data regularly, especially with fast recording devices and computer equipment. In addition there are some nice properties that result from even sampling. For example, the discussion beginning on page 10 describes the existence of a set of orthogonal frequency basis elements for evenly sampled time series data. Therefore, it is possible to represent the data as a linear combination of sines and cosines of these basis elements. Also, the orthogonality property is equivalent to the independence of the sine and cosine terms. This property will be used to derive the statistical distribution of the periodogram in the following section.

3.1.1 Distribution of the Periodogram Estimate

Given an evenly sampled time series $\{X_j\}_{j=1}^{N}$ that exhibits cyclic behavior, it is often useful to determine how well the model

$$X_j = \mu + a\cos(\omega t_j) + b\sin(\omega t_j) + \epsilon_j$$

describes the behavior of the data for a particular ω value. This question can be answered by carrying out a hypothesis test which examines the validity of the null hypothesis

$$H_0 : a = b = 0$$

versus the alternative hypothesis

$$H_A : a \neq 0 \quad \text{or} \quad b \neq 0.$$

This test is also equivalent to periodogram analysis where one checks the significance of the periodogram estimate $P_X(\omega)$ [14]. In order to determine this the statistical distribution of the periodogram estimate needs to be calculated. It is commonly assumed that the errors are randomly and normally distributed with mean 0 and variance σ^2. The null hypothesis states that no sinusoids are present in the data. Under this assumption and the assumption that the times are fixed constants, the X_j's are normally distributed with mean μ and variance σ^2. Likewise, the terms $X_j \cos(\omega t_j)$ and $X_j \sin(\omega t_j)$ in the periodogram definition are normally distributed since the cosine and sine expressions are treated as constants. The terms $\sum_{j=1}^{N} X_j \cos(\omega t_j)$ and $\sum_{j=1}^{N} X_j \sin(\omega t_j)$ are also normally distributed by the linear property of normal variables. The squares of these two sum terms will have a chi-square distribution and since the sine and cosine functions are orthogonal, these two squared terms are independent. These facts are used to determine that the distribution of $2P_X(\omega)/\sigma^2$ is chi-squared with 2 degrees of freedom [26]. Let $V = 2Px(\omega)/\sigma^2$. Since V has a chi-squared distribution with 2 degrees of freedom, the probability density function is defined to be

$$f_V(v) = \frac{1}{2}e^{-v/2}.$$

This distribution is equivalent to an exponential distribution with mean parameter 2.

Using a straightforward transformation it can be shown that the normalized periodogram estimate, defined as $P_N(\omega) = P_X(\omega)/\sigma^2$, is exponentially distributed with mean parameter 1. To prove this fact, let $U = P_N(\omega)$. Hence, $U = V/2$ and

$$\begin{aligned}
f_U(u)\Delta u &\approx \Pr(u < U < u + \Delta u) \\
&= \Pr\left(u < \frac{V}{2} < u + \Delta u\right) \\
&= \Pr(2u < V < 2(u + \Delta u)) \\
&= 2(u + \Delta u) f_V(2u).
\end{aligned}$$

Dividing through by Δu and taking the limit as $\Delta u \to 0$ gives

$$\begin{aligned} f_U(u) &= f_V(2u)\frac{d}{du}2u \\ &= 2f_V(2u) \\ &= e^{-u}. \end{aligned}$$

One final transformation is needed to determine the distribution of the classical periodogram. If $Y = P_X(\omega)$, then $Y = \sigma^2 U$. The probability density function of Y can be found by the calculation

$$\begin{aligned} f_Y(y)\Delta y &\approx \Pr(y < Y < y + \Delta y) \\ &= \Pr(y < \sigma^2 U < y + \Delta y) \\ &= \Pr\left(\frac{y}{\sigma^2} < U < \frac{y+\Delta y}{\sigma^2}\right) \\ &= \left(\frac{y+\Delta y}{\sigma^2}\right) f_U\left(\frac{y}{\sigma^2}\right). \end{aligned}$$

In the limit,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma^2} f_U\left(\frac{y}{\sigma^2}\right) \\ &= \frac{1}{\sigma^2} e^{y/\sigma^2}. \end{aligned}$$

Consequently, the periodogram estimate $P_X(\omega)$ is exponentially distributed with mean parameter σ^2.

These facts allow the observed periodogram estimate to be tested to determine if it is significantly large. Therefore, under the null hypothesis, the probability of observing a particular estimate, z, is given by

$$\Pr\left[P_X(\omega) > z\right] = e^{-z/\sigma^2}.$$

Given a predetermined significance level, the presence or absence of periodic behavior can be judged. If the probability calculation yields a significantly small value, then the probability of having a peak as large as the observed estimate is small and the null hypothesis would be rejected. The data would then be considered to have cyclic behavior associated with the frequency ω

3.1.2 Testing Multiple Frequencies

While the exponential distribution allows one to check for periodic behavior for a particular frequency ω, it is often the case that the frequency is unknown. It is customary to check for periodic behavior at a range of frequencies. For evenly sampled data, cyclic behavior is often examined at the Fourier basis frequencies, as mentioned on page 10. Since these form an orthogonal basis, the periodogram estimates at these frequencies are independent. Therefore, consider the normalized periodogram estimates, $P_N(\omega) = P_X(\omega)/\sigma^2$, evaluated at each of the $M = \lfloor N/2 \rfloor$ Fourier basis frequencies and let z be the height of the largest peak. The probability that the spectral peak for any given frequency is less than z is $1 - e^{-z}$. Since the periodgram estimates at these different frequencies are independent, the probability that all M frequencies have a spectral peak less than z is $[1 - e^{-z}]^M$. Thus, the probability that at least one frequency has a corresponding peak greater than z is $1 - [1 - e^{-z}]^M$. This significance level provides a method to test whether a periodic signal is present in the data. In particular, if one observes a maximum peak at ω_0 for which the probability of observing such a height is smaller than a previously chosen level of significance, one can conclude that there is a significant period associated with ω_0 in the data.

Another test was developed by Fisher [21] which uses the statistic

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\frac{1}{M}\sum_{k=1}^{M} P_X(\omega_k)}$$

or equivalently, the statistic

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\sum_{k=1}^{M} P_X(\omega_k)}.$$

These test statistics measure the relative size of the largest peak of the periodogram. Therefore, if the maximum value of the periodogram estimate greatly exceeds what is expected, the test statistic will be large. In this case, the null hypothesis would be rejected and one could conclude that the data exhibits periodic behavior. Fisher derived a closed form distribution for the test statistic T based on the exponential distribution when the number of samples, N, is odd. The probability for T is given by the expression

$$\Pr[T > Mg] = \sum_{j=1}^{k} (-1)^{j-1} \left(\begin{array}{c} M \\ j \end{array} \right) (1 - jg)^{M-1},$$

where $g > 0$, $k = \lfloor 1/g \rfloor$ and M is the number of periodogram estimates calculated. Fuller [26, page 284] constructed a table listing the upper percentage points of Fisher's statistic for various significance levels and sample sizes. By comparing the test statistic of the observed data against the upper confidence bound, the presence or absence of periodic behavior in the observed data can be examined. This test will be refered to as *Fisher's test.*

3.1.3 Extensions of Standard Tests

While Fisher's test is very simple and powerful if the true periodic component is a Fourier basis frequency, it suffers when the true period corresponds to a frequency in between two Fourier basis frequencies. Spurrier and Thombs [69] suggest a test which can handle this situation by using a refined basis of frequencies and adapting Fisher's test accordingly. The new model

$$X_j = \mu + A_m \cos(\omega_m t_j) + B_m sin(\omega_m t_j) + \epsilon_j \quad \text{for} \quad j = 1, \ldots, N$$

has unknown parameters μ, A_m, B_m and m with $\omega_m = 2\pi/m$. Again, ϵ_j is assumed to be from a random normal distribution. Using a refined frequency basis, the authors consider the null hypothesis that $A_m = B_m = 0$ for all $m \in (1, 2]$. However, for ease of computation, they approximate the values by only considering values of $m \in \{1.01, 1.02, \ldots, 1.99\}$. The test statistic is found by using the multiple regression model

$$\mathbf{X} = \mathbf{Y}[\mu \;\; A_m \;\; B_m]'$$

for a particular m where

$$\mathbf{X} = (X_1, X_2, \ldots, X_N)'$$

is the vector of observed values and $\mathbf{Y}$ is the design matrix with the jth row

$$[1 \quad \cos(\omega_m t_j) \quad \sin(\omega_m t_j)]\,.$$

The test statistic is defined to be

$$T' = \frac{\max_{m\in[1.01,1.02,\ldots,1.99]} S(m|\mu)}{\sum_{j=1}^{N}(X_j - \bar{X})^2}$$

where

$$\bar{X} = \frac{1}{N}\sum_{j=1}^{N} X_j$$

is the mean of the observed values, $(\mathbf{Y'_m Y_m})^-$ is the generalized inverse and

$$S(m|\mu) = \mathbf{X'Y_m(Y'_m Y_m)^- Y'_m X} - N\bar{X}^2$$

represents the amount of variance explained by adding the terms associated with the frequency ω_m to the null model. Since the distribution of the test statistic is difficult to define analytically, critical values for this test statistic were generated using Monte Carlo simulations. Notice that the denominator estimates the total variance of the observed data. Therefore if T' exceeds the critical value, then one of the sinusoidal terms explains a large portion of the variability in the data. By construction, this test is capable of detecting frequencies between the Fourier basis frequencies, which is an improvement over Fisher's test; however, the test is not as powerful as Fisher's test when the true frequency corresponds to a Fourier basis frequency.

Tatum [73] makes an adjustment to the previous two tests by estimating the variance using previously observed data. Since it is usually assumed that the process is free of periodic behavior prior to testing, an estimate of the underlying variance can be found using data collected earlier. In otherwords, the denominator in T and T' can be replaced with the sample variance estimate from a previous section of data. The estimate of the variance used in the denominator is assumed to be free of any variability associated with cyclic behavior and therefore makes the test more reliable when determining if the maximum peak is significantly large. For example, let

$$s_R^2 = \frac{1}{R-1} \sum_{k=1}^{R} (X_k - \bar{X}_R)^2$$

be the sample variance and

$$\bar{X}_R = \frac{1}{R} \sum_{k=1}^{R} X_k$$

be the mean of the first R observations. Clearly, these calculations can be generalized for any previously collected block of R observations. The new statistics proposed by Tatum would be defined as

$$T_m = \frac{\max_k \{P_X(\omega_k)\}}{s_R^2} \quad \text{for} \quad k = 1, \ldots, M = \lfloor N/2 \rfloor$$

and

$$T_m' = \frac{\max_{n \in [1.01, 1.02, \ldots, 1.99]} S(n|\mu)}{s_R^2}$$

where T_m is the modified Fisher statistic and T_m' is adjusted from the statistic introduced by Spurrier and Thombs. This simple modification has improved the power of the tests as compared to Fisher's test and the Spurrier and Thombs test. The adapted Fisher's test has the exact distribution [20, 73]

$$\Pr[T_m \leq z] = \sum_{j=0}^{M} \begin{pmatrix} M \\ j \end{pmatrix} (-1)^j \left[\frac{jz}{R-1} + 1\right]^{-(R-1)/2} \quad \text{for} \quad z > 0$$

where $M = \lfloor N/2 \rfloor$ and R is the number of previously observed points used to estimate the variance. Although this modified test shows improvement over Fisher's test, it still is not very powerful at detecting frequencies between the Fourier basis frequencies. The adapted T'_m test indicates improvement over T' in that it retains the ability to detect non-Fourier frequencies and with increased power.

Another issue with testing a range of frequencies is the presence of more than one periodic component. The presence of additional periodic components hinders the detection abilities of Fisher's test. In this situation, the periodogram will have more than one large estimate which results in an inflated mean or sum of the periodogram estimate. The over-estimation of the denominator causes the estimate of Fisher's test statistic to be too small. Therefore, Fisher's test becomes less sensitive to the introduction of periodic behavior. Simple modifications of Fisher's test have been proposed by Bølviken [7] to alleviate some of these problems. He considers Fisher's statistic

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\sum_{k=1}^{M} P_X(\omega_k)}.$$

This statistic can be interpreted as the ratio between the largest peak and the total error variance, σ^2, under the null hypothesis of no periodic behavior. The presence of multiple periodic cycles yields several large periodogram estimates which tend to inflate the estimate of σ^2 affecting the sensitivity of the Fisher's test. Bølviken suggests replacing the denominator with a more "robust" estimate. He discusses generalizing the denominator by defining

$$\hat{\sigma}^2 = \sum_{k=1}^{M} b_k P_X(\omega_{(k)})$$

where $b_k \geq 0$ for $k = 1, \ldots, M$ are predetermined weights and the $P_X(\omega_{(k)})$'s are the ordered periodogram estimates with $P_X(\omega_{(1)}) \leq P_X(\omega_{(2)}) \leq \ldots \leq P_X(\omega_{(M)})$. A trimmed mean can be used to get a more accurate estimate of the noise variance. In this case, the weights can be defined as

$$b_k = \begin{cases} 1 & \text{for} \quad k \leq M - a \\ 0 & \text{for} \quad k > M - a \end{cases}$$

where a equals the number of peaks to ignore in the calculation of the mean. Thus, the test statistic as a function of a is

$$T_a = \frac{\max_k \{P_X(\omega_k)\}}{\sum_{k=1}^{M-a} P_X(\omega_{(k)})}.$$

The parameter a should not be chosen to be too low. For long time series, it is suggested that the number of estimates to trim should be approximately three to four times the suspected number of cycles present in the data. Notice that when $a = 0$, T_a is equal to Fisher's test statistic.

An alternate way to define the weights would be to winsorize the periodogram estimates in the calculation of the denominator. With this approach the a largest peaks are replaced by the value of the moderately sized peak $P_X(\omega_{(M-a)})$, where a is subjectively chosen. In particular,

$$b_k = \begin{cases} 1 & \text{for} \quad k < M - a \\ a+1 & \text{for} \quad k = M - a \\ 0 & \text{for} \quad k > M - a \end{cases} \cdot$$

This gives the test statistic

$$T_w = \frac{\max_k \{P_X(\omega_k)\}}{\left[\sum_{k=1}^{M-a-1} P_X(\omega_{(k)})\right] + (a+1)P_X(\omega_{(M-a)})}.$$

The distribution of the statistics T_a and T_w have been derived in closed form [6] and tables of the critical values for the T_a statistic have been published [7]. These statistical tests were shown to improve the ability to detect periodic behavior when more than one frequency is present in the data as compared to Fisher's test.

3.2 Unevenly Sampled Data

When analyzing evenly sampled data there are standard spectral methods to detect periodic behavior and statistical tests to determine how confident we are about the behavior. However, there may be situations where data is not collected regularly. The collection of measurements may be hindered by factors intrinsic to the data collection process or by outside influences. Data may be almost evenly sampled except that some samples are missing. The data could also be collected in clusters with a relatively large time span between the groups of data. Alternately, data may be sampled randomly. For example, astronomical time series data often has periodic behavior, but data collection is often collected irregularly. The sampling is usually limited to nights and only under certain conditions. Cloud cover and unavailability of equipment may prohibit evenly collected samples. In industrial settings, missing data could result when production lines are shut down for system maintenance, during power outages or during holiday breaks.

One early method dealing with this problem was to ignore the unequally spaced data and use standard periodogram methods designed for evenly sampled data [78]. Interpolating the data to make it evenly sampled is another option [54, 77]. While in certain cases these methods may suffice, it may be more beneficial to analyze the observed data directly [9, 16, 37, 44, 61, 76]. The subsequent sections in this chapter will address methods designed to be applied to the unevenly sampled observations.

3.2.1 Missing Data

Cipra [16] discusses two situations in which data is collected with a fixed sampling rate, however some observations are missing. The first case covers regularly missing data, or where missing data occurs periodically. Under the null hypothesis, the observed data $\{X_j\}_{j=1}^{N}$, is assumed to be from a normal distribution with zero mean and variance σ^2. Although data is missing it can be represented as evenly sampled if we declare a new random variable

$$Y_j = g(t_j)X_j \quad \text{for} \quad j = 1, 2, \ldots, N$$

where

$$g(t_j) = \begin{cases} 1 & \text{if } X_j \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

Since the data is missing regularly, the function $g(t)$ is periodic with period c. Let a denote the number of observations taken in one cycle. Thus a equals the number of t_j's with $g(t_j) = 1$ and $j \in \{1, 2, \ldots, c\}$. Under the null hypothesis that the data does not have any underlying cyclic behavior, the distribution of $Y_j \cos(\omega_k t_j)$ and the corresponding sine term have a normal distribution. Here the frequency basis is defined by

$$\omega_k = \frac{2\pi k}{N} \quad \text{for} \quad k = 1, \ldots, r$$

where

$$r = \lfloor \frac{(N/c) - 1}{2} \rfloor.$$

To prevent aliasing, the frequency basis is restricted so that the largest frequency that can be observed is π/c. Using the same argument as the evenly sampled case, the distribution of the periodogram is again exponential but with mean parameter $a\sigma^2/c$. Therefore, Cipra indicates that Fisher's test can be used on the new series $\{Y_j\}_{j=1}^{N}$ using critical values found from the distribution derived by Fisher.

The function of $g(t)$ can be redefined and applied to handle more general cases of almost evenly sampled data. For instance, $g(t)$ need not be periodic. The binary function would simply indicate whether an observation was collected at the particular time. However, $g(t)$ must satisfy the condition that the number of missing observations in a sample of N data points is on the order of $N^{\frac{1}{2}}$. For this case, the periodogram estimate, at any arbitrary frequency, approaches an exponential distribution as N tends to infinity. For large N, Fisher's test can be used to approximate the significance of the largest periodogram estimate.

The other scenario considered by Cipra is when evenly sampled data is missing randomly.

Here observations are missing according to a Bernoulli probability model. At each time period, the probability of taking a measurement is $p \in (0,1)$. The collected data, $\{X_j\}_{j=1}^{N}$ can be represented with the new random variable

$$Y_j = Z_j X_j$$

where

$$Z_j = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

is a Bernoulli variable independent of the observed measurement. The periodogram, as a function of $\{Y_j\}$, converges in distribution to an exponential distribution with mean parameter $p\sigma^2$. Again, for large N, Cipra suggests using Fisher's test using the transformed data and approximating the critical values with the ones already established.

3.2.2 Lomb Periodogram

Although methods for almost evenly sampled data have been addressed, people have been interested in defining a generalized periodogram that can be defined for any sampling scheme and yet still preserves the nice statistical distribution that exists for the evenly sampled case. Part of the difficulty is the ambiguity of the orthogonality conditions of the Fourier basis frequencies when the data is unevenly sampled.

In developing the alternative form of the periodogram for unevenly sampled data, Lomb [44] addressed the lack of a natural orthonormal basis by selecting a set of frequencies and then using least-squares fitting of sinusoids to the data to determine the power associated with each frequency, thus giving an estimate of the periodogram. The resulting form of the periodogram is

$$P_X(\omega) = \frac{1}{2}\left\{ \frac{\left[\sum\limits_{j=1}^{N} X_j \cos(\omega t_j)\right]^2}{\sum\limits_{j=1}^{N} \cos^2(\omega t_j)} + \frac{\left[\sum\limits_{j=1}^{N} X_j \sin(\omega t_j)\right]^2}{\sum\limits_{j=1}^{N} \sin^2(\omega t_j)} \right\},$$

where we have omitted the time shift that Lomb introduced to allow for time invariance, since it will not enter the discussion here. The Lomb periodogram has been carefully studied by Scargle [61] and he was able to show a connection between the original work of Lomb and the derivation of a generalized periodogram that would ostensibly preserve the statistical properties of the periodogram of evenly sampled data.

In deriving the statistical properties of the Lomb periodogram for random normal data, $\{X_j\}_{j=1}^{N}$, with mean zero and variance σ^2, Scargle starts with the definition of a generalized periodogram,

$$P_X(\omega) = \frac{A^2}{2N}\left(\sum_{j=1}^{N} X_j \cos(\omega t_j)\right)^2 + \frac{B^2}{2N}\left(\sum_{j=1}^{N} X_j \sin(\omega t_j)\right)^2$$

and then considers the mean and variance of the quantity $C(\omega) = A\sum_{j=1}^{N} X_j \cos(\omega t_j)$. To do so, he treats the cosine terms for a fixed set of t_j as constant coefficients and considers the sum as a sum of independent normally distributed random variables to derive the results that the mean $\mathrm{E}[C(\omega)] = 0$. Therefore, the variance is given by

$$\sigma_c^2 = \mathrm{E}\left[C^2(\omega)\right] = A^2 \sum_{j=1}^{N}\sum_{k=1}^{N} \mathrm{E}\left[X_j X_k\right] \cos\omega t_j \cos\omega t_k$$

$$= A^2\sigma_0^2 \sum_{j=1}^{N} \cos^2 \omega t_j.$$

This result is dependent on the particular sampling times t_j and the particular frequency ω. A similar result holds for the term $S(\omega) = B\sum_{j=1}^{N} X_j \sin(\omega t_j)$. Horne and Baliunas [37] show that when the Lomb periodogram is normalized by dividing by the variance of the observed data, the expected distribution of peaks is exponential with mean 1. This result, which is consistent with the evenly sampled case, now provides the basis for testing whether a peak is significant or not. If a periodogram estimate exceeds the critical value determined by the exponential distribution, there is evidence to support the presence of a cycle.

3.2.3 Classical Periodogram

Another technique which can be used to detect periodic signals in unevenly sampled time series is the classical periodogram. As mentioned in Chapter 2, the definition of the periodogram is flexible enough to handle unevenly sampled data, however, the frequency basis is no longer clearly defined. We also need to consider whether a periodogram estimate is significantly larger than zero, or equivalently if there is significant periodic structure in the data. One way of determining whether a peak is significant is to generate upper confidence bounds using parametric or nonparametric techniques.

Since the frequency basis that we choose may not be orthogonal due to the unevenly sampled times, the statistical tests developed previously may not be valid. An alternate method of testing for significant periodic behavior assuming that the observations are random and Gaussian can be done using *Monte Carlo methods* [56]. For each sample time, a random sample can be independently chosen from a Gaussian distribution with the same mean and variance. For each set of N samples from a random normal distribution, the periodogram can be evaluated and the highest peak can be obtained. Many simulations can be run to get a large sample of periodogram estimates. Percentiles can be calculated to form upper confidence bounds. The largest peak of the periodogram of the observed data can then be compared to the upper confidence bounds. If the peak of the observed spectrum exceeds the confidence bound, then a significant cycle is associated with the frequency. This method can also provide upper percentage bounds for the Lomb periodogram.

When the underlying distribution of the data is unknown, we suggest a nonparametric test based on *permutation resampling* to test whether the data is random or if the time ordering of the data is significant because of the underlying dynamics of the process [9, 62, 74, 75]. This method generates a new time series from the observed data by randomly assigning each sample time to an observed measurement without replacement. Since the true distribution of the observed data may be unknown, this surrogate data can be used as another realization from the process. By construction, this new realization has the same mean and variance as the observed data. Also by shuffling the data any temporal correlation in the data is destroyed and the effects of any time bias in the data are reduced. The term time bias refers to patterns in the sampling scheme which may cause spurious peaks in the periodogram. The next step would be to calculate the classical periodogram for this new data set for a particular frequency, ω. This resampling process would be repeated many times. After a sufficient number of trials, one can calculate the 95th percentile, or whatever level of significance is desired. The number of iterations should be large enough that the confidence bounds stabilize. If the periodogram estimate for frequency ω exceeds the upper confidence bound, the data exhibits a significant signal with frequency ω.

To demonstrate how these test work, the Lomb periodogram with confidence bounds found using the Monte Carlo method was compared to the permutation resampling scheme for the classical definition of the periodogram. For each method, 1000 simulations were used to construct the upper confidence bounds. The first example comes from a simulated data set with samples collected from a uniform distribution. The time values were generated to simulate 10 random hourly measurements taken every Tuesday and Thursday for a year. At a 99.9% upper confidence bound, the Lomb periodogram in Figure 3-1a incorrectly identifies this data set as periodic, with a period of 7 days. This period is attributed to the time bias introduced by the sampling scheme, and not the behavior of the underlying process. On the other hand, the permutation resampling technique did not find any significant cyclic behavior, as shown in Figure 3-1b where the 99.9% confidence bound is denoted by 'o' and the periodogram estimates are denoted by '*'.

Another example uses times which are randomly chosen from 0 to 1000 days and the measurements were taken from the sinusoidal process

$$X(t_j) = .5\cos(2\pi t_j/5) + \epsilon_j$$

for $j = 1, \ldots, 1000$ where ϵ_j are standard normal noise. This process is assumed to have a five day cycle. Since both periodogram methods indicated the 5 day cycle, its effect was modeled using least squares regression. Residuals were obtained by subtracting the predictions using the least squares model from the original data. The analysis was repeated on the residuals to see if any remaining periodic structure was identified. The plot in Figure 3-2a shows the Lomb periodogram using the residual series after extracting the 5 day cycle. At a 95%

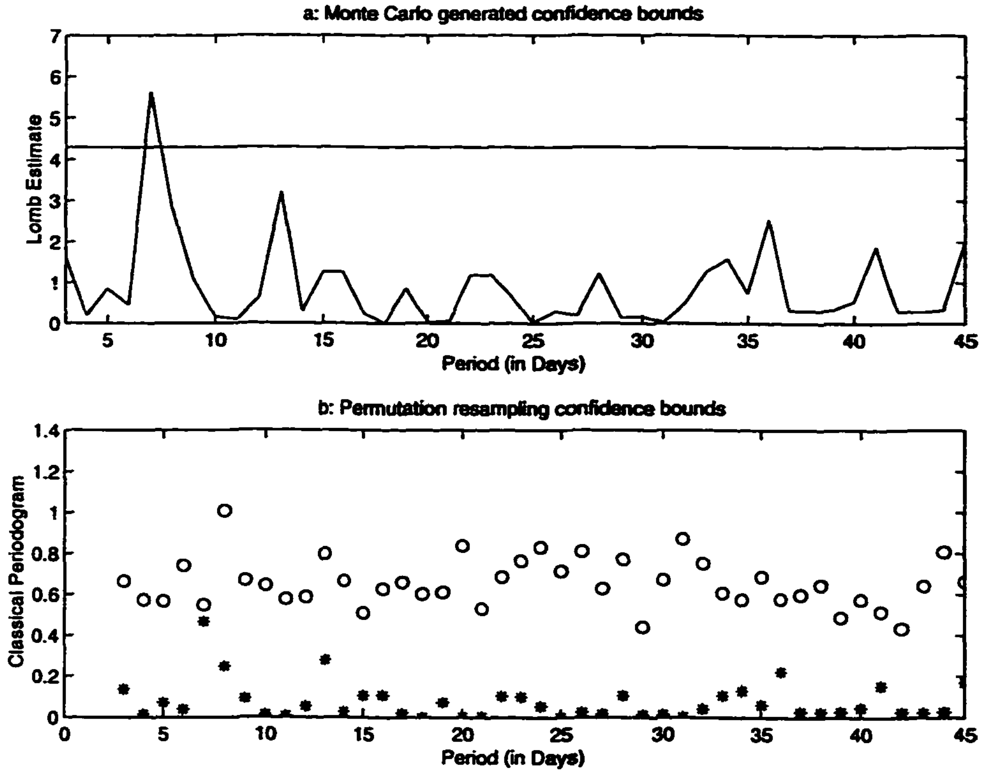


Figure 3-1 Testing for periodic behavior using different methods, example 1

level of significance the Lomb periodogram identifies an additional and significant period of 17 days which is not present in the actual process. The permutation resampling method indicates no additional periodic behavior as shown in Figure 3-2b. Thus in both examples, the permutation resampling method correctly identified the underlying dynamics regardless of the time bias introduced by the sampling procedure or the underlying distribution of the data. Consequently, in some cases the permutation resampling method appears to be more robust to noise and anomalies in the irregular sampling procedure than the Monte Carlo method. In these two simulated examples, the normalized Lomb periodogram, with the confidence bounds determined using Monte Carlo methods, found cycles which were not

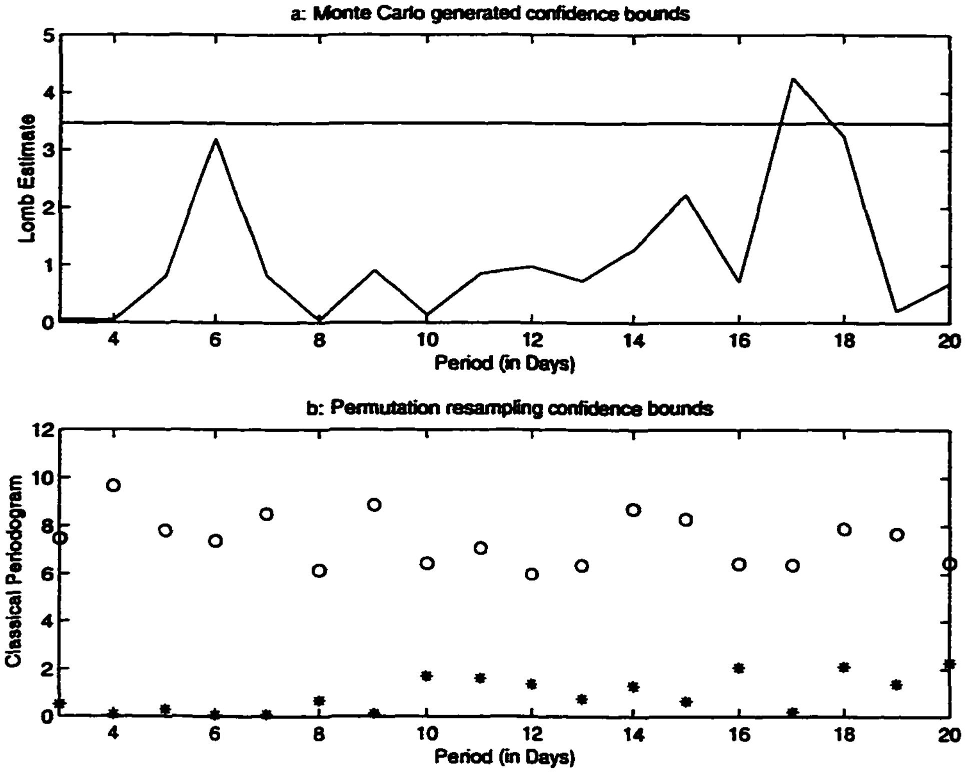


Figure 3-2 Testing for periodic behavior using different methods, example 2

actually present in the data while the resampling scheme detected only the true underlying dynamics.

Although the permutation resampling method can minimize the effects of time bias in certain situations, it too can sometimes detect false peaks. For instance, consider measurements shown in Figure 3-3a taken from the process

$$X(t_j) = .5\sin(2\pi t_j/120)$$

with a 120 hour (or 5 day) cycle. The t_j represent the sampling times when measurements are taken at 10 random hours every Monday through Friday for one year. Both the Lomb

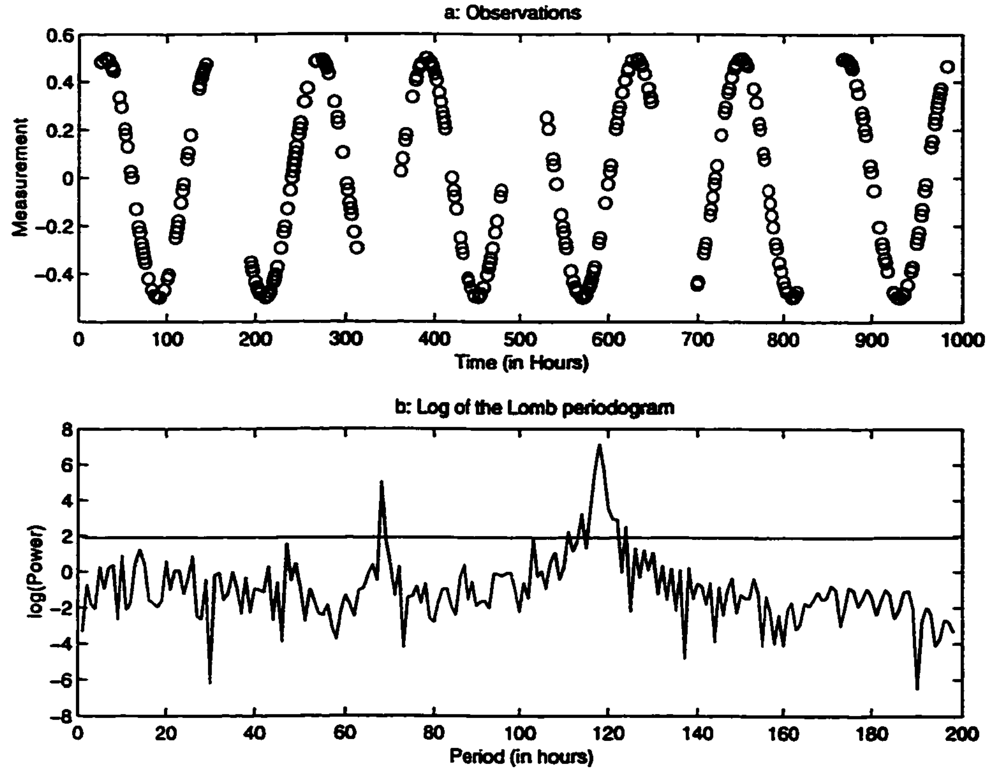


Figure 3-3 Both tests incorrectly identify a 70 hour cycle

periodogram and classical periodogram incorrectly detect a significant cycle of 70 hours, shown in Figure 3-3b. Clearly, further work needs to be done to determine why this time bias occurs and how it can be corrected. Since the goal is to find the hidden dynamics of the process, one must proceed with caution and examine the results using all tools available.

3.3 Randomly Sampled Data

In some processes it may not be feasible to collect data using a fixed sampling rate. When samples are collected at random intervals, the times should be considered to be realizations of a random variable. When we adopt this viewpoint, the distribution of the periodogram must be recalculated since the cosine and sine terms can no longer be considered constant coefficients. The added variability changes the distribution of those terms so that they are no longer normally distributed and thus the distribution properties of the periodogram are affected. Therefore, the misapplication of extending the traditional results to random data can result in the detection of spurious peaks.

For example, when the Lomb periodogram is applied to small samples of data collected at random times, the spectral peaks are typically not exponentially distributed. To support this claim, simulations were conducted using Monte Carlo methods. For each trial, a sample of 50000 spectral estimates were generated for a particular frequency ω and random time sequence. Both these factors were changed in each trial. The Kolmogorov-Smirnov goodness-of-fit test was also used to test the hypothesis that the distribution of the spectral peaks was exponentially distributed with mean parameter $\sigma^2 = 4$. Thus, the variance of the peaks should be $\sigma^4 = 16$. Table 3.1 shows the estimates of the mean and variance of some periodogram power estimates, along with the results of the Kolmogorov-Smirnov test. Large p-values indicate that the two distributions are statistically identical and suggest that the periodogram estimate is exponentially distributed. Therefore, for $N = 5$ and 10, the null hypothesis would be rejected and for $N = 25$ and 50, the results are mixed. Since the exponential distribution for the periodogram estimates may not hold for small samples of data we will focus on deriving statistics specifically for randomly sampled data for the classical periodogram.

N	Trial	Mean Estimate	Variance Estimate	p-value
5	1	3.957	18.661	0
	2	3.973	16.898	0
	3	3.996	17.342	0
10	1	3.999	19.826	0
	2	4.012	16.872	0.0051
	3	4.001	17.719	0
25	1	4.001	16.368	0.0745
	2	3.990	16.827	.0005
	3	4.007	16.465	0.1867
	4	4.004	16.524	0.2209
50	1	4.031	16.203	0.0557
	2	4.005	16.394	0.0817
	3	4.023	16.454	0.5555

Table 3.1 Estimates of Mean and Variance of Lomb Estimates for Random Samples

3.3.1 Statistical Distribution

One consequence of the assumption that the $\cos \omega t_j$ terms can be treated as constant coefficients is that the variance results must be interpreted in the sense that the recorded measurements were taken as one realization of data from an ensemble of possible data sets, but the ensemble must be taken at exactly the same times. Consequently, it is not possible to take a long data set and break it up into pieces to assemble an ensemble, as one might do when studying an ergodic process. If, instead, we consider the sampling times to be random variables, then the $\cos \omega t_j$ terms must be treated as random variables. Properties of the classical periodogram with this new interpretation will be considered in this section.

Consider the observed times $\{t_j\}_{j=1}^{N}$ as realizations of a random variable T which is sampled in such a way that the distribution of angles $W = \omega T$ for a fixed ω is uniformly distributed on $(-\pi, \pi)$. Under this assumption all angles are equally likely and the probability density function (pdf) of W is $f_W(w) = \frac{1}{2\pi}$ for $-\pi < w < \pi$. The method of transforming random variables can be used to find the pdf of $Y = \cos W$. In particular,

$$\begin{aligned} f_Y(y)\Delta y &\approx \Pr(y < Y < y + \Delta y) \\ &= 2\Pr(\cos^{-1}(y+\Delta y) < W < \cos^{-1}(y)) \\ &= 2f_W(\cos^{-1}(y))[\cos^{-1}(y) - \cos^{-1}(y+\Delta y)]. \end{aligned}$$

Dividing through by Δy and taking the limit as $\Delta y \to 0$ gives

$$\begin{aligned} f_Y(y) &= 2f_W(\cos^{-1}(y)) \times \frac{d}{dy}\left[-\cos^{-1}(y)\right] \\ &= 2\left(\frac{1}{2\pi}\right)\left(\frac{1}{\sqrt{1-y^2}}\right) \\ &= \frac{1}{\pi\sqrt{1-y^2}} \quad \text{for} \quad -1 < y < 1 \end{aligned}$$

and zero otherwise. This is a well-defined density function since $\int_{-\infty}^{\infty} f_Y(y)dy = 1$ and f_Y is non-negative and piecewise continuous.

Similarly, the distribution of

$$V = \sin(\omega T) = \sin(W) = \cos\left(W + \frac{\pi}{2}\right)$$

can be found using the method described above. Thus,

$$\begin{aligned} f_V(v)\Delta v &\approx \Pr(v < V < v + \Delta v) \\ &= 2\Pr\left[\cos^{-1}(v+\Delta v) - \frac{\pi}{2} < W < \cos^{-1}(v) - \frac{\pi}{2}\right] \\ &= 2f_W\left(\cos^{-1}(v) - \frac{\pi}{2}\right) \times \left\{\left[\cos^{-1}(v) - \frac{\pi}{2}\right] - \left[\cos^{-1}(v+\Delta v) - \frac{\pi}{2}\right]\right\} \end{aligned}$$

$$= 2\left(\frac{1}{2\pi}\right)\left\{\cos^{-1}(v) - \cos^{-1}(v+\Delta v)\right\}.$$

The pdf of V is obtained by dividing both sides by Δv and taking the limit as $\Delta v \to 0$

$$f_V(v) = \frac{1}{\pi\sqrt{1-v^2}} \quad \text{for} \quad -1 < v < 1.$$

Not surprisingly, $\sin \omega T$ and $\cos \omega T$ have the same distribution. Without loss of generality, the statistical properties of the periodogram will be determined by considering the distribution of $[X_j \cos(\omega t_j)]^2$ and extending the results to the sine term.

First, the cumulative distribution function (cdf) for the random variable

$$Z = X\cos(\omega T) = XY$$

is derived assuming that the observed data, $\{X_j\}_{j=1}^{N}$, is sampled from a random normal noise process X with zero mean and standard deviation σ. Additionally, it is assumed that the times and measurements are independent. Now, finding the probability that $Z < \alpha$ is equivalent to determining the probability that X and Y have a product less than α where $-1 < Y < 1$ and $-\infty < X < \infty$. For a fixed α this would require integrating the joint density of X and Y over the region where $XY < \alpha$. Geometrically this would correspond to integrating over the region D_Z bounded by the hyperbola $xy = \alpha$, as shaded in Figure 3-4. We will examine the cdf, F_Z in two cases. The cdf will be defined first for positive α values.

$$\begin{aligned} F_Z(\alpha) &= \Pr(Z < \alpha) \\ &= \Pr(x, y \in D_Z) \end{aligned}$$

$$
\begin{aligned}
&= \int_{D_Z}\int f_{X,Y}(x,y)dxdy \\
&= \int_{D_Z}\int f_X(x)f_Y(y)dxdy \\
&= 2\int_0^1\int_{-\infty}^{\frac{\alpha}{y}} \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-x^2}{2\sigma^2}}\frac{1}{\pi\sqrt{1-y^2}}dxdy \\
&= 2\int_0^1\left[\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-x^2}{2\sigma^2}}dx+\int_0^{\frac{\alpha}{y}} \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-x^2}{2\sigma^2}}dx\right]\frac{1}{\pi\sqrt{1-y^2}}dy \\
&= \int_0^1\left[1+\operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\sigma y}\right)\right]\frac{1}{\pi\sqrt{1-y^2}}dy,
\end{aligned}
$$

where

$$
\operatorname{erf}(\phi) = \frac{2}{\sqrt{\pi}}\int_0^{\phi} e^{-t^2}dt
$$

is the error function [2].

Similarly, for $\alpha < 0$

$$
\begin{aligned}
F_Z(\alpha) &= \Pr(Z<\alpha) \\
&= \Pr(x,y \in D_Z) \\
&= \int_{D_Z}\int f_{X,Y}(x,y)dxdy \\
&= \int_{D_Z}\int f_X(x)f_Y(y)dxdy \\
&= 2\int_0^1\int_{-\infty}^{\frac{\alpha}{y}} \frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-x^2}{2\sigma^2}}\frac{1}{\pi\sqrt{1-y^2}}dxdy \\
&= \int_0^1 \operatorname{erfc}\left(\frac{-\alpha}{\sqrt{2}\sigma y}\right)\frac{1}{\pi\sqrt{1-y^2}}dy \\
&= \int_0^1\left[1-\operatorname{erf}\left(\frac{-\alpha}{\sqrt{2}\sigma y}\right)\right]\frac{1}{\pi\sqrt{1-y^2}}dy
\end{aligned}
$$

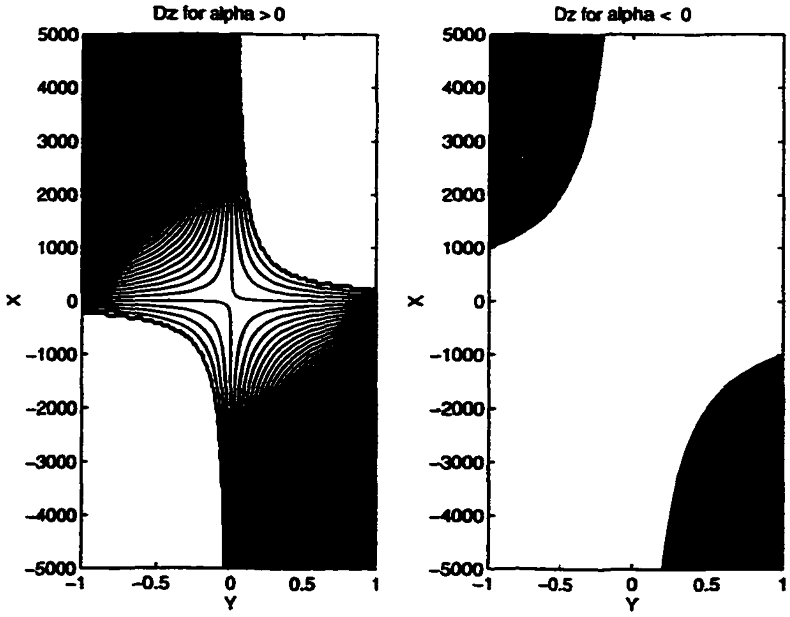


Figure 3-4 Integration region, D_Z, for $\alpha > 0$ and $\alpha < 0$

$$= \int_0^1 \left[1 + \text{erf}\left(\frac{\alpha}{\sqrt{2}\sigma y}\right)\right] \frac{1}{\pi\sqrt{1-y^2}} dy$$

where $\text{erfc}(\phi) = 1 - \text{erf}(\phi)$.

Thus, for all α, the cdf of Z is defined to be

$$F_Z(\alpha) = \int_0^1 \left[1 + \text{erf}\left(\frac{\alpha}{\sqrt{2}\sigma y}\right)\right] \frac{1}{\pi\sqrt{1-y^2}} dy.$$

Using the property that $\lim_{z\to\pm\infty} \text{erf}(z) = \pm 1$, the cdf is well defined since $\lim_{z\to-\infty} F_Z(z) = 0$ and $\lim_{z\to\infty} F_Z(z) = 1$.

The density function of Z, f_Z, can be found by taking the partial derivative of the cdf with respect to α as follows

$$\begin{aligned} f_Z(\alpha) &= \frac{\partial}{\partial \alpha} F_z(\alpha) \\ &= \frac{\partial}{\partial \alpha} \int_0^1 \left[1 + \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\sigma y} \right) \right] \frac{1}{\pi\sqrt{1-y^2}} dy. \end{aligned}$$

Since the integrand satisfies certain continuity requirements [58], the partial derivative and integral can be switched to obtain

$$\begin{aligned} f_Z(\alpha) &= \int_0^1 \frac{\partial}{\partial \alpha} \left[1 + \operatorname{erf}\left(\frac{\alpha}{\sqrt{2}\sigma y} \right) \right] \frac{1}{\pi\sqrt{1-y^2}} dy \\ &= \int_0^1 \frac{2}{\sqrt{\pi}} e^{\frac{-\alpha^2}{2\sigma^2 y^2}} \frac{1}{\sqrt{2}\sigma y} \frac{1}{\pi\sqrt{1-y^2}} dy \\ &= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^1 e^{\frac{-\alpha^2}{2\sigma^2 y^2}} \frac{1}{y\sqrt{1-y^2}} dy. \end{aligned}$$

The expected value of Z is defined as

$$E[Z] = \int_{-\infty}^{\infty} \alpha f_Z(\alpha) d\alpha = 0$$

since the integrand is an odd function. Another way to determine this fact is to use the assumption that X and Y are independent and X is normally distributed with zero mean. It follows that

$$E[Z] = E[XY] = E[X]E[Y] = 0. \qquad (1)$$

The variance of Z can be found using the formulation $\mathrm{Var}[Z] = E[Z^2] - E[Z]^2 = E[Z^2]$.

Now consider the random variable $S = Z^2$. We proceed to derive the density function and expected value of S which equals the variance of Z. The pdf of S can be found by transforming the random variable Z. In particular,

$$\begin{aligned}
f_S(s)\Delta s &\approx \Pr(s < S < s + \Delta s) \\
&= 2\Pr(\sqrt{s} < Z < \sqrt{s + \Delta s}) \\
&= 2f_Z(\sqrt{s})[\sqrt{s + \Delta s} - \sqrt{s}].
\end{aligned}$$

Finally, by dividing both sides by Δs and taking the limit as $\Delta s \to 0$, we have

$$\begin{aligned}
f_S(s) &= 2f_Z(\sqrt{s}) \times \frac{d}{ds}\sqrt{s} \\
&= 2f_Z(\sqrt{s})\left(\frac{1}{2\sqrt{s}}\right) \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^1 e^{\frac{-s}{2\sigma^2 y^2}} \frac{1}{\sqrt{s}} \frac{dy}{y\sqrt{1 - y^2}}
\end{aligned}$$

for nonnegative s.

Hence,

$$\begin{aligned}
E[S] &= \int_0^\infty s f_S(s) ds \\
&= \int_0^\infty s \left[\frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^1 e^{\frac{-s}{2\sigma^2 y^2}} \frac{1}{\sqrt{s}} \frac{dy}{y\sqrt{1 - y^2}}\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^\infty \left[\int_0^1 e^{\frac{-s}{2\sigma^2 y^2}} \sqrt{s} \frac{dy}{y\sqrt{1 - y^2}}\right] ds.
\end{aligned}$$

By substituting $u = \frac{1}{y}$ and $dy = \frac{-du}{u^2}$ and making the appropriate changes we obtain

$$\begin{aligned}
E[S] &= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_0^\infty \left[\int_1^\infty e^{\frac{-su^2}{2\sigma^2}}\sqrt{s}\frac{u}{u^2\sqrt{1-\frac{1}{u^2}}}du\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_0^\infty \left[\int_1^\infty e^{\frac{-su^2}{2\sigma^2}}\sqrt{s}\frac{du}{\sqrt{u^2-1}}\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_1^\infty \left[\int_0^\infty e^{\frac{-su^2}{2\sigma^2}}\sqrt{s}ds\right] \frac{du}{\sqrt{u^2-1}}.
\end{aligned}$$

where Fubini's Theorem [59] allowed us to change the order of integration. Then setting $t = \sqrt{s}$ and substituting $ds = 2tdt$ we have

$$\begin{aligned}
E[S] &= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_1^\infty \left[\int_0^\infty e^{\frac{-t^2u^2}{2\sigma^2}}2t^2dt\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{2\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_1^\infty \left[\int_0^\infty e^{\frac{-t^2u^2}{2\sigma^2}}t^2dt\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{2\sqrt{2}}{\pi\sigma\sqrt{\pi}}\int_1^\infty \left[\frac{\sigma^3\sqrt{\pi}}{\sqrt{2}u^3}\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{2\sigma^2}{\pi}\int_1^\infty \frac{du}{u^3\sqrt{u^2-1}} \\
&= \left(\frac{2\sigma^2}{\pi}\right)\left(\frac{\pi}{4}\right) \\
&= \frac{\sigma^2}{2}. \qquad (2)
\end{aligned}$$

Therefore, the random variable $S = [X\cos(\omega T)]^2$ has an expected value of $\sigma^2/2$ where σ^2 is the variance of the random normal noise process. The same result holds for $[X\sin(\omega T)]^2$. These facts will be used to derive the expected value of the periodogram. In order to simplify notation, let $G_j = X_j\cos(\omega T_j)$ and $H_j = X_j\sin(\omega T_j)$ for $j = 1, \ldots, N$ where the X_j are independently and identically distributed normal random variables with zero mean and variance σ^2 and the T_j are independent random variables as described at the beginning of this section. It is also assumed that the X_i and T_j are independent for all $i, j = 1, \ldots N$ and $i \neq j$. The calculations use the properties that for all j, $E[G_j] = E[H_j] = 0$ and $E[G_j^2] = E[H_j^2] = \sigma^2/2$.

First, consider the summand in the periodogram definition involving just the cosine terms. By expanding the square of the sum and using the linear properties of the expectation operator and the property that the expectation of a product of independent variables is the product of the expectations we have

$$\begin{aligned} E\left[\left(\sum_{j=1}^{N} X_j \cos(\omega T_j)\right)^2\right] &= E\left[\left(\sum_{j=1}^{N} G_j\right)^2\right] \\ &= E\left[\sum_{i=1}^{N} G_j^2 + \sum_{i=1}^{N} \sum_{j=i+1}^{N} 2G_iG_j\right] \\ &= \sum_{i=1}^{N} E\left[G_j^2\right] + \sum_{i=1}^{N} \sum_{j=i+1}^{N} 2E[G_i]E[G_j] \\ &= \sum_{i=1}^{N} \frac{\sigma^2}{2} + 0 \\ &= \frac{N\sigma^2}{2}. \end{aligned} \tag{3}$$

This result will also hold for the sine term.

Thus the expected value of the classical periodogram for data sampled randomly is given by

$$E\left[P_X(\omega)\right] = E\left[\frac{1}{N}\left\{\left(\sum_{j=1}^{N} G_j\right)^2 + \left(\sum_{j=1}^{N} H_j\right)^2\right\}\right]$$

$$
\begin{aligned}
&= \frac{1}{N}\left(\sum_{j=1}^{N} E\left[\left(\sum_{j=1}^{N} G_j\right)^2\right] + \sum_{j=1}^{N} E\left[\left(\sum_{j=1}^{N} H_j\right)^2\right]\right) \\
&= \frac{1}{N}\left(\sum_{j=1}^{N}\frac{\sigma^2}{2} + \sum_{j=1}^{N}\frac{\sigma^2}{2}\right) \\
&= \frac{1}{N}\left(\frac{N\sigma^2}{2} + \frac{N\sigma^2}{2}\right) \\
&= \sigma^2. \qquad (4)
\end{aligned}
$$

The variance of the periodogram can be found using similar techniques. The variance calculation requires the computation of $\mathrm{Var}[G_j^2]$, $\mathrm{Var}[H_j^2]$, $\mathrm{Var}[2G_iG_j]$, $\mathrm{Var}[2H_iH_j]$, and $\mathrm{Cov}[G_j^2, H_j^2]$ for $i, j = 1, \ldots, N$ and $i \neq j$. Other covariance terms are involved in the analysis; however, all the terms not listed have covariance equal to zero and are not included to simplify the equations. We first present the individual calculations of the terms listed above and then proceed to calculate the variance of the periodogram.

First, we calculate

$$
\begin{aligned}
\mathrm{Var}[G_j^2] &= E[G_j^4] - E[G_j^2]^2 \\
&= E[S^2] - E[S]^2 \\
&= \frac{9\sigma^4}{8} - \frac{\sigma^4}{4} \\
&= \frac{7\sigma^4}{8} \qquad (5)
\end{aligned}
$$

where $S = [X\cos(\omega T)]^2$ has been previously defined and has mean equal to $\sigma^2/2$. The calculation of the expected value of S^2 is similar to the calculation of $E[S]$. Therefore,

$$\begin{aligned}
E[S^2] &= \int_0^\infty s^2 f_S(s)ds \\
&= \int_0^\infty s^2 \left[\frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^1 e^{\frac{-s}{2\sigma^2 y^2}} \frac{1}{\sqrt{s}} \frac{dy}{y\sqrt{1-y^2}}\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^\infty \left[\int_0^1 e^{\frac{-s}{2\sigma^2 y^2}} s^{\frac{3}{2}} \frac{dy}{y\sqrt{1-y^2}}\right] ds.
\end{aligned}$$

By substituting $u = \frac{1}{y}$ and $dy = \frac{-du}{u^2}$ and making the appropriate changes we obtain

$$\begin{aligned}
E[S^2] &= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^\infty \left[\int_1^\infty e^{\frac{-su^2}{2\sigma^2}} s^{\frac{3}{2}} \frac{u}{u^2\sqrt{1-\frac{1}{u^2}}} du\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_0^\infty \left[\int_1^\infty e^{\frac{-su^2}{2\sigma^2}} s^{\frac{3}{2}} \frac{du}{\sqrt{u^2-1}}\right] ds \\
&= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_1^\infty \left[\int_0^\infty e^{\frac{-su^2}{2\sigma^2}} s^{\frac{3}{2}} ds\right] \frac{du}{\sqrt{u^2-1}}
\end{aligned}$$

where again Fubini's Theorem [59] allowed us to switch the order of integration. Then setting $t = \sqrt{s}$ and substituting $ds = 2tdt$ we have

$$\begin{aligned}
E[S^2] &= \frac{\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_1^\infty \left[\int_0^\infty e^{\frac{-t^2u^2}{2\sigma^2}} 2t^4 dt\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{2\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_1^\infty \left[\int_0^\infty e^{\frac{-t^2u^2}{2\sigma^2}} t^4 dt\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{2\sqrt{2}}{\pi\sigma\sqrt{\pi}} \int_1^\infty \left[\frac{3\sigma^5\sqrt{\pi}}{\sqrt{2}u^5}\right] \frac{du}{\sqrt{u^2-1}} \\
&= \frac{6\sigma^4}{\pi} \int_1^\infty \frac{du}{u^5\sqrt{u^2-1}} \\
&= \left(\frac{6\sigma^4}{\pi}\right)\left(\frac{6\pi}{32}\right) \\
&= \frac{9\sigma^4}{8}.
\end{aligned}$$

Since G_j and H_j have the same distribution, it follows that $\mathrm{Var}[H_j^2] = 7\sigma^4/8$.

Next we show that $\mathrm{Var}[2G_iG_j] = \mathrm{Var}[2H_iH_j] = \sigma^4$.

$$\begin{aligned}
\mathrm{Var}[2G_iG_j] &= 4\mathrm{Var}[G_iG_j] \\
&= 4\left(E[G_i^2G_j^2] - E[G_iG_j]^2\right) \\
&= 4\left(E[G_i^2]E[G_j^2] - (E[G_i]E[G_j])^2\right) \\
&= 4\left[\frac{\sigma^2}{2}\left(\frac{\sigma^2}{2}\right) - 0^2\right] \\
&= \sigma^4 \qquad\qquad (6)
\end{aligned}$$

due to the independence and normality assumptions.

The last computation uses the fact that a random normal variable with zero mean has a fourth moment equal to $3\sigma^4$ [51, page 117]. Another fact is that for Y uniformly distributed on $(-\pi, \pi)$ we have

$$\begin{aligned}
E[\cos^2(Y)\sin^2(Y)] &= \int_{-\pi}^{\pi} \cos^2(y)\sin^2(y) f_Y(y)dy \\
&= \int_{-\pi}^{\pi} \cos^2(y)\sin^2(y)\left(\frac{1}{2\pi}\right)dy \\
&= \frac{1}{2\pi}\left[\frac{-\sin(4y)}{32} + \frac{y}{8}\right]_{-\pi}^{\pi} \\
&= \frac{1}{2\pi}\left[\frac{\pi}{8} + \frac{\pi}{8}\right] \\
&= \frac{1}{8}.
\end{aligned}$$

Thus, since X_j is normally distributed with zero mean and ωT_j is uniformly distributed on

$(-\pi, \pi)$,

$$\begin{aligned}
\mathrm{Cov}[G_j^2, H_j^2] &= E\left[\left(G_j^2 - \frac{\sigma^2}{2}\right)\left(H_j^2 - \frac{\sigma^2}{2}\right)\right] \\
&= E[G_j^2 H_j^2] - \frac{\sigma^2}{2}E[G_j^2] - \frac{\sigma^2}{2}E[H_j^2] + \frac{\sigma^4}{4} \\
&= E[X_j^4 \cos^2(\omega T_j)\sin^2(\omega T_j)] - 2\left(\frac{\sigma^4}{4}\right) + \frac{\sigma^4}{4} \\
&= E[X_j^4]E[\cos^2(\omega T_j)\sin^2(\omega T_j)] - \frac{\sigma^4}{4} \\
&= 3\sigma^4\left(\frac{1}{8}\right) - \frac{\sigma^4}{4} \\
&= \frac{\sigma^4}{8}. \qquad (7)
\end{aligned}$$

Results (5)-(7) are used to calculate the theoretical variance of the periodogram. In particular,

$$\begin{aligned}
\mathrm{Var}\,[P_X(\omega)] &= \mathrm{Var}\left[\frac{1}{N}\left(\left(\sum_{j=1}^{N} G_j\right)^2 + \left(\sum_{j=1}^{N} H_j\right)^2\right)\right] \\
&= \frac{1}{N^2}\mathrm{Var}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} G_i G_j + \sum_{i=1}^{N}\sum_{j=1}^{N} H_i H_j\right] \\
&= \frac{1}{N^2}\left(\sum_{j=1}^{N}\mathrm{Var}\left[G_j^2\right] + \sum_{j=1}^{N}\mathrm{Var}\left[H_j^2\right] + \sum_{j=1}^{N} 2\mathrm{Cov}\left[G_j^2, H_j^2\right]\right. \\
&\quad \left. + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathrm{Var}\,[2G_iG_j] + \sum_{i=1}^{N}\sum_{j=i+1}^{N}\mathrm{Var}\,[2H_iH_j]\right) \\
&= \frac{1}{N^2}\left(2\sum_{j=1}^{N}\frac{7\sigma^4}{8} + \sum_{j=1}^{N}\frac{\sigma^4}{4} + 2\sum_{j=1}^{N}\sum_{j=i+1}^{N}\sigma^4\right) \\
&= \frac{1}{N^2}\left[\frac{14N\sigma^4}{8} + \frac{N\sigma^4}{4} + \frac{2N(N-1)\sigma^4}{2}\right] \\
&= \frac{1}{N^2}\left[\left(N^2 + N\right)\sigma^4\right] \\
&= \left(1 + \frac{1}{N}\right)\sigma^4. \qquad (8)
\end{aligned}$$

As mentioned previously, other covariance terms which arise in the variance calculation were not expressed above. These terms include Cov $[G_i^2, G_j^2]$, Cov $[H_i^2, H_j^2]$, Cov $[G_i^2, 2G_iG_j]$, Cov $[2H_iH_j, 2H_kH_l]$, Cov $[G_i^2, 2H_iH_j]$, Cov $[G_i^2, 2G_jG_k]$, Cov $[2H_iH_j, 2H_iH_k]$, Cov $[G_i^2, H_j^2]$, Cov $[G_i^2, 2H_jH_k]$, Cov $[H_i^2, G_jG_k]$, Cov $[2G_iG_j, 2G_kG_l]$, Cov $[H_i^2, 2G_iG_j]$, Cov $[H_i^2, H_iH_j]$, Cov $[2G_iG_j, 2H_iH_j]$, Cov $[2G_iG_j, 2H_iH_k]$, Cov $[2G_iG_j, 2H_kH_l]$, Cov $[2G_iG_j, 2G_iG_k]$ and finally Cov $[H_i^2, H_jH_k]$ for $i, j, k, l = 1, \ldots, N$ unequal. Some examples of the computation of the theoretical covariance between these cross-terms are described below and take advantage of the independence assumptions. The techniques can be extended to show that all of the other covariance terms are zero. For example, for $i \neq j$

$$\begin{aligned}
\mathrm{Cov}[G_i^2, G_j^2] &= E\left[\left(G_i^2 - \frac{\sigma^2}{2}\right)\left(G_j^2 - \frac{\sigma^2}{2}\right)\right] \\
&= E[G_i^2 G_j^2] - \frac{\sigma^2}{2}E[G_i^2] - \frac{\sigma^2}{2}E[G_j^2] + \frac{\sigma^4}{4} \\
&= E[G_i^2]E[G_j^2] - 2\left(\frac{\sigma^2}{2}\right)\left(\frac{\sigma^2}{2}\right) + \frac{\sigma^4}{4} \\
&= \frac{\sigma^2}{2}\left(\frac{\sigma^2}{2}\right) - \frac{\sigma^4}{4} \\
&= 0.
\end{aligned}$$

Another example uses the fact that for a random variable Y, uniformly distributed on $(-\pi, \pi)$,

$$\begin{aligned}
E[\cos(Y)\sin(Y)] &= \int_{-\pi}^{\pi} \cos(y)\sin(y)\left(\frac{1}{2\pi}\right) dy \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi} \cos(y)\sin(y) dy \\
&= \frac{1}{2\pi}\left[\frac{sin^2 y}{2}\right]_{-\pi}^{\pi}
\end{aligned}$$

$$= 0.$$

Since ωT_j is uniformly distributed on the appropriate range this result can be used to calculate

$$\begin{aligned}
\mathrm{Cov}[2G_iG_j, 2H_iH_j] &= E[(2G_iG_j - 0)(2H_iH_j - 0)] \\
&= E[4G_iH_iG_jH_j] \\
&= E\left[4X_i^2X_j^2 \cos(\omega T_i)\sin(\omega T_i) cos(\omega T_j)\sin(\omega T_j)\right] \\
&= 4E[X_i^2]E[X_j^2]E[\cos(\omega T_i)\sin(\omega T_i)]E[\cos(\omega T_j)\sin(\omega T_j)] \\
&= 4(\sigma^2)(\sigma^2)(0)(0) \\
&= 0.
\end{aligned}$$

3.3.2 Verification of Results

In this section the results of various simulations which support the theoretical calculations derived in the previous section will be presented. In particular, the results of equations (1) - (4) and (8) are examined to see if these hold for different N, ω and σ values. The methods take advantage of the fact that ensembles that do not have fixed sampling times.

First equations (1) - (3) are examined using simulated data. For each simulation, 50000 ensembles were created, each consisting of 1000 random times and measurements from a

$\omega = 2\pi/3, \sigma = 1$	Trial 1	Trial 2	Theoretical
E[G]	$-3.8e^{-5}$	$4.9e^{-5}$	0
V[G]	.5000	.5000	.5
E[$C^2(\omega)$]	500.8	502.7312	500
E[H]	$-1.6e^{-4}$	$1.1e^{-4}$	0
V[H]	.5001	.5000	.5
E[$S^2(\omega)$]	502.3	500.9249	500

Table 3.2 Simulation of Intermediate Calculations I

random normal distribution. The following terms were calculated

$$\begin{aligned}
(i)\quad & E[G] && \equiv E\left[X\cos(\omega T)\right] \\
(ii)\quad & V[G] && \equiv \mathrm{Var}\left[X\cos(\omega T)\right] \\
(iii)\quad & E\left[C^2(\omega)\right] && \equiv E\left[\left(\sum_{j=1}^{N} X_j\cos(\omega T_j)\right)^2\right]
\end{aligned}$$

as well as the sine counterparts which will be denoted as $E[H], V[H]$ and $E[S^2(\omega)]$. These simulations reveal statistical properties consistent with the analysis presented earlier. In particular, for a time series with N observations from a noise process with standard deviation σ, the theoretical values should be

$$\begin{aligned}
E[G] &= E[H] &&= 0 \\
V[G] &= V[H] &&= \tfrac{\sigma^2}{2} \\
E[C^2(\omega)] &= E[S^2(\omega)] &&= \tfrac{N\sigma^2}{2}.
\end{aligned}$$

The first example shows the estimates from two typical simulations with parameters $\sigma = 1$ and $\omega = 2\pi/3$. The results are summarized with the expected theoretical values in Table 3.2. Likewise, Table 3.3 shows the outcomes of two additional simulations where the parameters were defined to be $\sigma = 2$ and $\omega = 2\pi/5$.

$\omega = 2\pi/5, \sigma = 2$	Trial 1	Trial 2	Theoretical
E[G]	$7.2e^{-5}$	$-5.9e^{-6}$	0
V[G]	2.000	2.0010	2
E[$C^2(\omega)$]	1999.2	2002.5	2000
E[H]	$-1.8e^{-4}$	$-5.0e^{-5}$	0
V[H]	2.001	2.0002	2
E[$S^2(\omega)$]	2000.1	1990.3	2000

Table 3.3 Simulation of Intermediate Calculations II

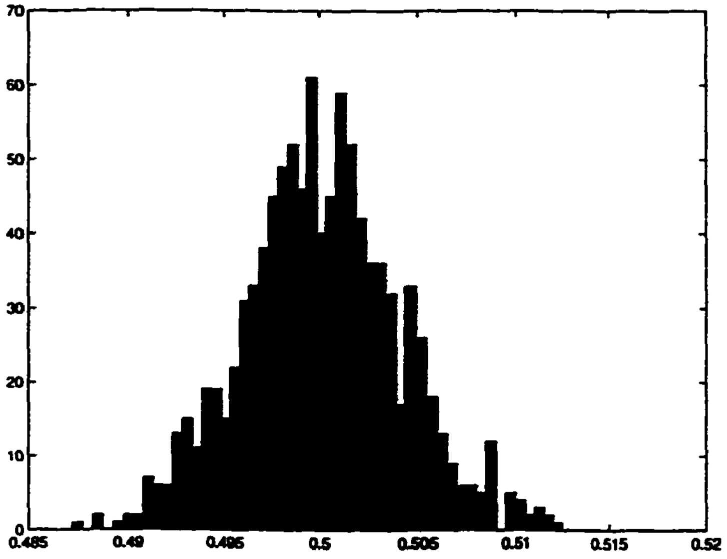


Figure 3-5 Histogram demonstrating that Var[G]/σ= 1/2

These two examples indicate that our method is sound. However, we wanted to verify that the results were independent of the frequency and the variance parameter. Thus, the estimates of Var[G] for 50000 simulations were generated where each trial had randomly chosen frequency and σ value. A histogram of Var[G]/σ for all the trials is shown in Figure 3-5. From our derivation, the quotient should be 1/2 and the histogram indicates that this is plausible.

Next, the mean and variance of the periodogram were examined. Simulations, summa-

N	σ	Trial	Theoretical Mean	Mean	Theoretical Variance	Variance	σ^4
5	1	1	1	1.0001	1.2	1.1999	1
		2		1.0002		1.2096	
10	2.5	1	6.25	6.2470	42.96875	42.9849	39.0625
		2		6.2665		42.9661	
40	2	1	4	3.9911	16.4	16.3890	16
		2		3.9971		16.3670	
100	3	1	9	9.0026	81.81	81.9285	81
		2		8.9917		81.7107	
1000	1	1	1	1.0016	1.001	1.0060	1
		2		1.0016		.9999	

Table 3.4 Simulated Mean and Variance of the Periodogram of Randomly Sampled Data

rized in Table 3.4, were conducted to verify the results of equations (4) and (8). The mean and variance at the periodogram estimates were calculated for 500000 different time series of length N, for $N = 5, 10, 40, 100$ and 1000. Different ω values were used in each trial to ensure that the calculations were not frequency dependent. The mean and variance of the simulations are listed along with the theoretical values according to our derivation. Notice that for small sample sizes, the variance of the periodogram is larger than σ^4, the variance expected for an exponential distribution with mean parameter σ^2.

3.3.3 Determination of Critical Values

The results developed show a different viewpoint on how to calculate statistical properties of the periodogram when the sampling times are chosen randomly. The results differ slightly from the approaches taken previously [37, 44, 61] since the times are not treated as constants. Hence the variance of

$$\sum_{j=1}^{N} X_j \cos(\omega t_j)$$

is not a function of the sampling times or the frequencies. However, the variance traditionally derived for this expression, namely

$$\sigma^2 \sum_{j=1}^{N} \cos(\omega t_j),$$

does approach our result as $\Delta t \to 0$. By definition,

$$\int_0^{2\pi} \cos^2(\omega t) dt = \lim_{\Delta t \to 0} \sum_j \cos^2(\omega t) \Delta t = \pi.$$

However, for small Δt,

$$\Delta t \approx \frac{2\pi}{N}$$

where N is the number of samples collected. So,

$$\pi \approx \sum_j \cos^2(\omega t) \Delta t \approx \frac{2\pi}{N} \sum_j \cos^2(\omega t).$$

Thus, in the limit,

$$\sum_j \cos^2(\omega t) = \frac{N}{2}$$

and therefore

$$\text{Var}\left[\sum_{j=1}^{N} X_j \cos(\omega t_j)\right] = \frac{N\sigma^2}{2}$$

which agrees with our findings.

The distribution of the periodogram for evenly sampled data is an exponential distribution with mean σ^2. Therefore, the variance of the periodogram is σ^4. The results of the

N = 10	p-value = 0
N = 50	p-value = .0282
N = 100	p-value = .1565
N = 1000	p-value = .3032

Table 3.5 Results of Kolmogorov-Smirnov Test

distribution properties of the periodogram for randomly sampled data presented here has the same mean; however, the variance differs slightly from the previous work by the factor of $[1 + \frac{1}{N}]$. In the simulations, especially for small sample sizes, the variance of the periodogram is larger than the variance traditionally proposed. Obviously, as N increases, the variance does approach σ^4. Also, the distribution of the periodogram seems to approach an exponential distribution as the sample size increases, as shown in Figure 3-6. This figure compares the distribution of the classical periodogram to the exponential distribution with mean σ^2 when $N = 10, 100$ and 1000 and $\sigma = 1$. As N increases, the quantile plot function of S-plus approaches the line $y = x$ indicating that the distribution approaches an exponential distribution. Thus, for small N the distribution is exponential-like; however, it has a heavier tail. This is not surprising since the variance of the periodogram for small sample sizes is slightly larger than what is expected for a true exponential distribution. In fact, Splus' Kolmogorov-Smirnov goodness-of-fit test, summarized in Table 3.5 indicates that the distribution of the classical periodogram with $N = 10$ is significantly different from an exponential distribution with mean 1 (p-value = 0). However for values of $N \geq 100$, the two distributions appear to be statistically equivalent.

Although the distribution of randomly sampled data, as defined above, approaches an exponential distribution, it is inappropriate to use the critical values derived for Fisher's test.

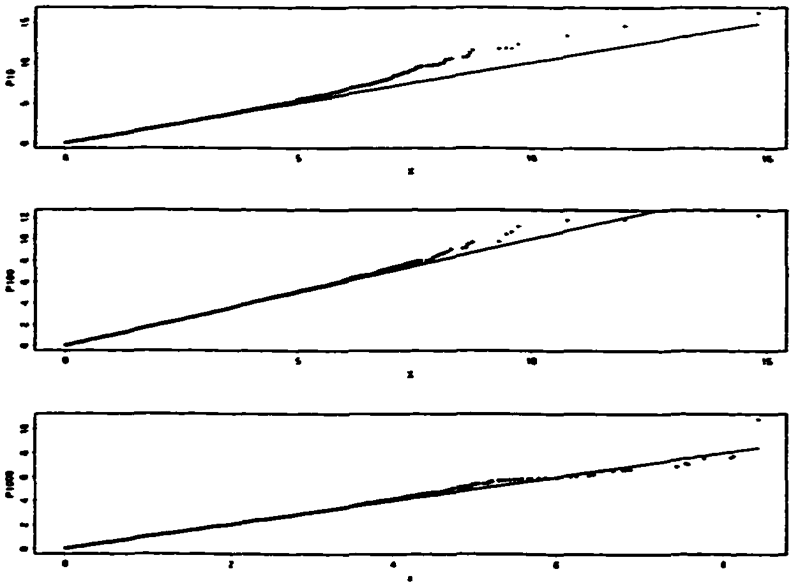


Figure 3-6 Quantile-Quantile Plot

The critical values derived for Fisher's test assume that the periodogram is exponentially distributed. Especially for smaller samples, this is not the case. The critical values for Fisher's test are greater than what occur in Monte Carlo simulations on normally distributed data sampled at random times. To demonstrate this, several Monte Carlo trials were carried out for $N = 11, 21, 31, 51$ and 101. Each trial began with the creation of a random time sequence which was fixed for the rest of the trial. Then, data was selected from a normal distribution and the periodogram of the simulated data was calculated at each of the Fourier basis frequencies. Fisher's test statistic (page 38),

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\frac{1}{M}\sum_{k=1}^{M} P_X(\omega_k)},$$

was calculated and recorded. This simulation was repeated 300000 times, where for each

No. of Periodogram Estimates	N	Fisher's Critical Value	Monte Carlo Critical Value
5	11	3.419	3.213
10	21	4.450	4.113
15	31	5.019	4.659
25	51	5.701	5.439
50	101	6.567	6.363

Table 3.6 Upper Critical Value (5% level of significance)

iteration new data was generated. The 95th percentile of the test statistics from the 300000 simulations is recorded as the Monte Carlo critical value and it was consistently lower than the critical value derived for evenly sampled data. Table 3.6 summarizes the results of a Monte Carlo trial for 5 different sample sizes, each with a different random sampling scheme. For comparison, Fisher's critical value for evenly sampled data is also listed. Notice that for all N, Fisher's critical value is larger than the value produced by the Monte Carlo method. In fact for $N = 10$, the value 3.419 actually corresponds to approximately the 97th percentile in the Monte Carlo trial and for $N = 101$ the value 6.567 corresponds to approximately the 96th percentile, instead of the 95th percentile.

Although this difference may not seem significant, it could be staggering for industrial companies which are monitoring their process using small samples sizes. Since Fisher's cutoff points are larger, periodic behavior may go undetected. This could be cost prohibitive for industries. Therefore, in order to test the null hypothesis that no periodic signal is contained in a collection of randomly sampled data we suggest finding critical values through Monte Carlo simulations.

3.3.4 The Performance of the Statistical Tests

In this section we will explore how various tests perform when data is randomly sampled. All critical values were found by Monte Carlo methods, using 30000 independent realizations for each sampling scheme and a 5% level of significance. These studies examine the relative *power* of Fisher's test statistic to detect sinusoidal behavior under certain conditions using the Monte Carlo generated confidence bounds. The power of a statistical test is estimated by the proportion of trials that correctly reject the null hypothesis because the test statistic exceeds the critical value.

First we examine how the power of Fisher's test is affected when the underlying sinusoidal model has a fixed period but the amplitude is varied. Consider a random sampling scheme of $N = 51$ times and observations generated from the model

$$X_j = B\cos(2\pi t_j/5.1) + \epsilon_j \quad \text{for} \quad j = 1, 2, \ldots, N$$

where the ϵ_j are independently and identically distributed standard normal errors and B ranges from 0 to 2.5 in increments of .25. For each amplitude, 10000 independent tests were conducted. The estimates of the power as a function of amplitude for two different random sampling schemes are shown in Figure 3-7. The test with modified critical values seems to perform well at detecting cyclic behavior at a Fourier frequency. In addition, the power of the test using Fisher's critical value is represented in Figure 3-7 by the dashed line. Since the Fisher's cutoff point is larger, it tends to have less power than the test using the Monte Carlo generated bounds.

In the next study, the ability of Fisher's test to detect non-Fourier frequencies was

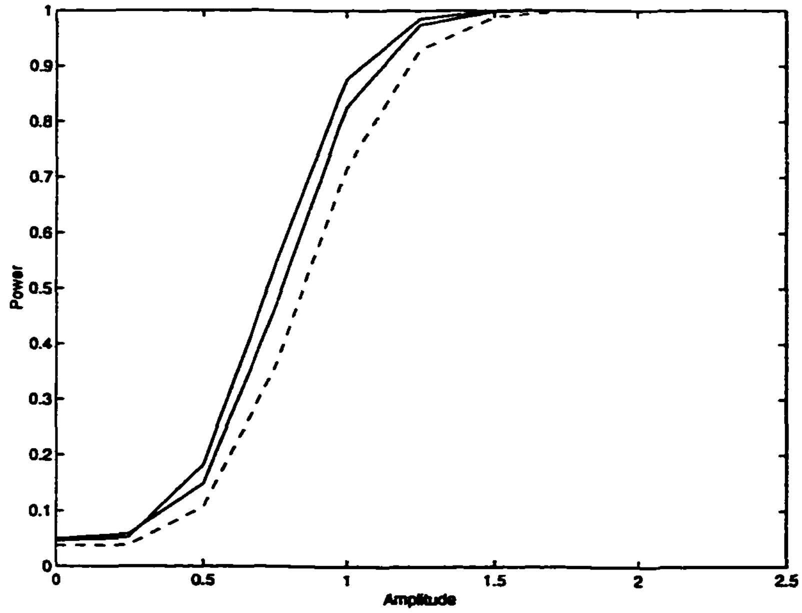


Figure 3-7 Estimated power of Fisher's test as a function of amplitude

tested. For this trial, the amplitude is fixed while the frequency is varied between three Fourier frequencies. The model used was

$$X_j = 1.5\cos(2\pi t_j f) + \epsilon_j$$

for $j = 1, 2, \ldots, N$ where $N = 51$ and $\{\epsilon_j\}_{j=1}^{N}$ are again assumed to be random, standard normal errors. The values of f were selected to run between the Fourier frequencies 10/51 and 12/51 with a step size of .1/51. Again, for each set of random sampling times, Monte Carlo simulations were used to estimate the critical values. Although Fisher's test detected the signal for frequencies near the three Fourier basis frequencies, represented with a Δ in the plot, the test does not perform well at detecting frequencies midway between Fourier basis frequencies, as shown in Figure 3-8.

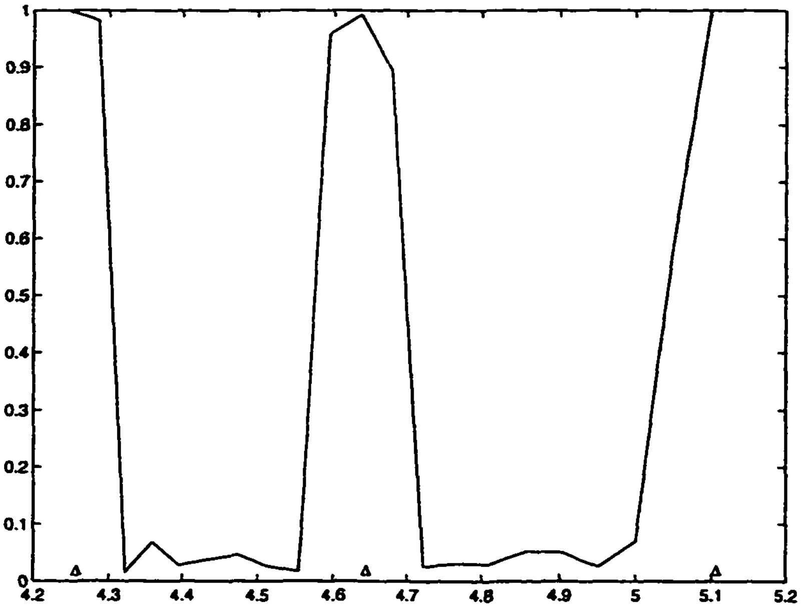


Figure 3-8 Estimated power of Fisher's test as a function of frequency

The final batch of simulations examine how the modified tests fare when multiple periodic components are present. The power of the tests were calculated when two and three periodic components were present in the underlying model. The first model considered was

$$X_j = B\cos(2\pi t_j/5.1) + B\cos(2\pi t_j/3) + \epsilon_j,$$

for $j = 1, 2, \ldots, N$ where $N = 51$ and B is varied from 0 to 2.5 with a step size of .25. The $\{\epsilon_j\}_{j=1}^{N}$ are defined as before. For this example the periods 3 and 5.1 correspond to two Fourier basis frequencies with the same amplitude. The power of Fisher's test, using the upper bound generated from the Monte Carlo method, was calculated using 10000 independent simulations. Two independent trials were conducted where data was collected at random from the process above. The results of these trials are shown in Figure 3-9. For comparison, two estimates of the power of Fisher's test applied to data that has only one periodic component are shown with a dotted line. Notice that the test loses power when an additional cycle is introduced to the data. The test performs even worse when a third periodic component is added. To illustrate this, an additional summand corresponding to the Fourier frequency 3/51 was included in the model above. The resulting power is diminished greatly, as shown in Figure 3-10.

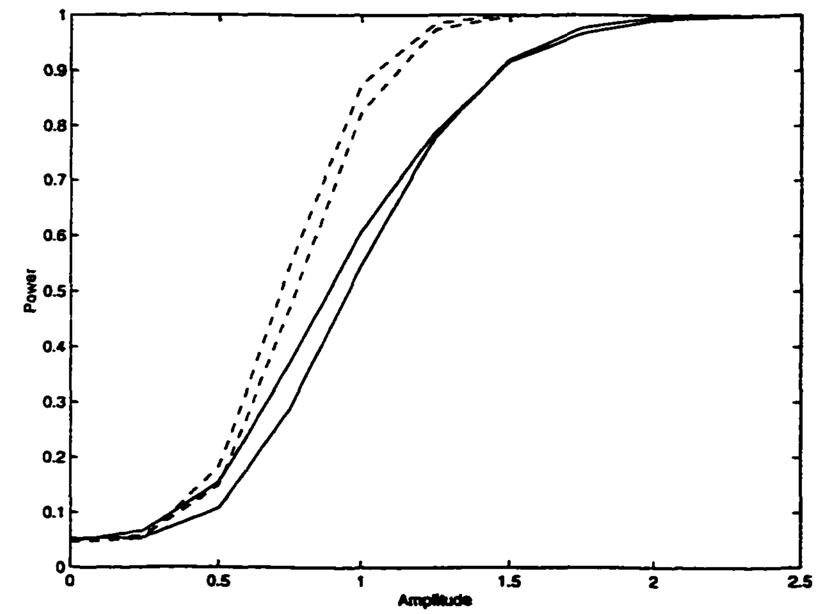


Figure 3-9 Estimated power of Fisher's test with two periodic components

Since Fisher's test performs poorly when multiple periods are involved, a simple modification proposed by Bølviken and discussed on page 43, can be made to improve the test. In this case, the Fisher test statistic

$$T = \frac{\max_k \{P(\omega_k)\}}{\frac{1}{M}\sum_{k=1}^{M} P(\omega_k)}$$

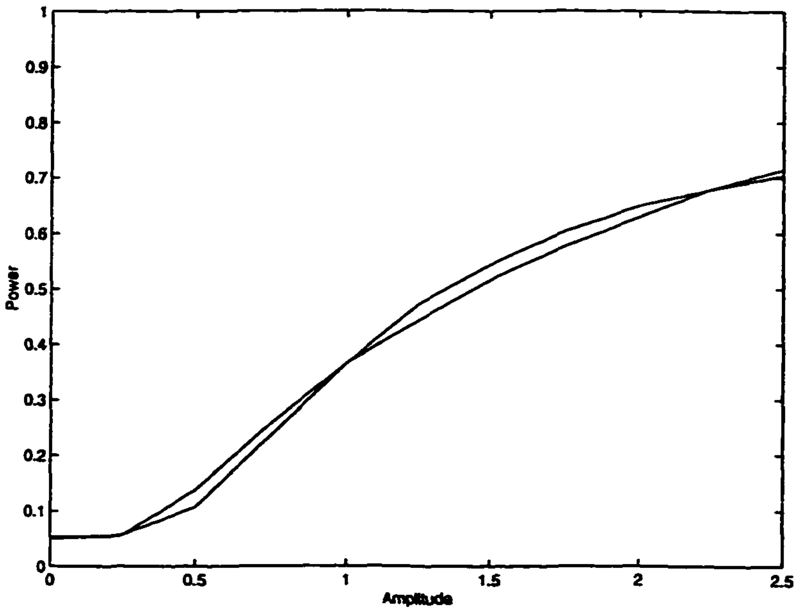


Figure 3-10 Estimated power of Fisher's test with 3 periodic components

has an inflated denominator which makes the ratio smaller. This ratio results in a test statistic that is less likely to exceed the critical value causing the test to become less accurate in terms of the significance level. The simple change of ignoring large peaks in the calculation of the mean makes the test more robust. Following the method proposed by Bølviken and applying it to the unevenly sampled data, the denominator is replaced by

$$\frac{1}{M-a}\sum_{k=1}^{M-a} P(\omega_{(k)})$$

where $P(\omega_{(k)})$ are the ordered peaks with $P(\omega_{(1)}) \leq P(\omega_{(2)}) \ldots \leq P(\omega_{(M)})$ and the value of a was chosen to be 3. This value for a was chosen to correspond to the three cycles known to be in the data. Critical values for the test are again formed using 30000 Monte Carlo simulations. The results, shown in Figure 3-11, show a dramatic increase in the power of

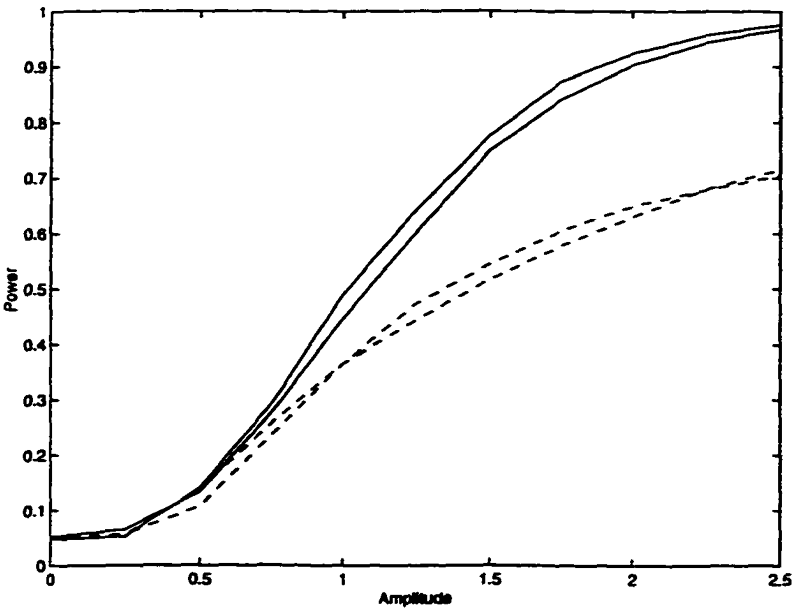


Figure 3-11 Estimated power of Bølviken's test

the test, as compared to the power of the standard application of Fisher's test without the trimming method, shown with the dotted line.

Not surprisingly, these tests on irregularly sampled data perform similarly to the tests applied for evenly sampled data. The main difference is the critical values used are generated from Monte Carlo simulations instead of the bounds derived for evenly sampled data. Without this adjustment the standard critical values would be too high and periodic behavior may go undetected. Therefore, the misapplication of using the inflated bounds to detect significantly large peak values results in a loss of power of Fisher's test. However, the use of Monte Carlo methods to modify the critical bounds improves the signal detection capabilites of Fisher's test.

Chapter 4

SPECTRAL CONTROL CHARTS AND OTHER MONITORING DEVICES

Statistical process control is used in industrial settings to improve productivity and the quality of goods. Industries use statistical techniques to monitor their processes to make sure that they are running efficiently. Samples are collected over time and are analyzed to detect deviations from the desired process. Walter Shewhart [65] introduced the concept of a *control chart* to monitor an industrial process.

At a given time, samples are collected from the process and an average measure is calculated and plotted on the control chart. The samples collected over time will undoubtedly have some natural variability or common cause variation. If the process is *in control* at a given time, meaning there is only common cause variation, the plotted measurement will fall within predetermined control limits about the target value. However, when unusual sources of variability are present in the process, the averages will plot outside the control limits, thus indicating a possible change in the process. The sources of variation, called assignable causes, might be attributed to factors like defective materials or equipment. The process is then considered *out of control* because it has changed significantly. When this happens it is important to determine whether deviations are produced by changes in the process mean, the process variance, a new random noise source or by the introduction of a non-random component to the production process. This chapter will review the premise of a control chart as well as address methods designed to detect the onset of periodic (sinusoidal) behavior in a system.

4.1 Control Chart Basics

A control chart is a graphical method of representing measurements of some quality characteristic over time. The measurements that are being monitored are plotted as a function of the sample number or as a function of time. To assist in determining the behavior of the process, the average value expected when the process is in control is represented by a horizontal line in the middle of the chart. If the process is in control, the measurements collected will vary randomly about this line. Two other horizontal lines are typically used to indicate the upper and lower control limits, denoted by UCL and LCL. Under normal conditions, most of the measurements will fall between these two bounds. In this situation, the process is deemed in control and no adjustments are made to the process. When an assignable cause is present the points tend to fall outside the control limits since the variability associated with an assignable cause is typically much larger than the common cause variability. Also, distinct patterns within the control limits imply there is an assignable cause affecting the process since the variation in the data no longer behaves randomly. Therefore, the process would be deemed out of control and an investigation for the assignable cause would be carried out. The process could then be properly adjusted to bring it back in control.

Part of the appeal of control charts is that they can be used on-line. As data is collected the control chart can be updated with the information provided by the newest sample. Given the current speed of computers the calculations can be done almost instantaneously which makes it ideal to use in industrial settings. An example of a control chart is shown in

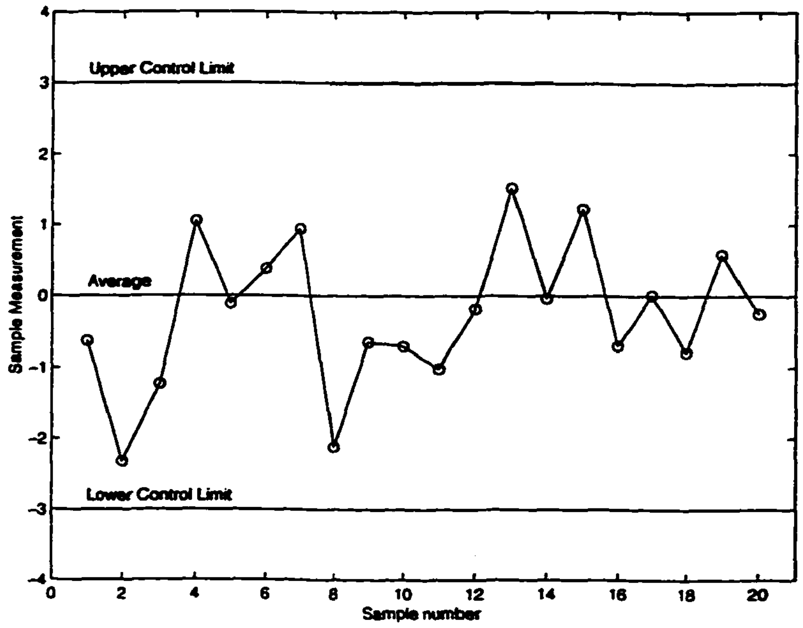


Figure 4-1 Example of a control chart

Figure 4-1. Notice that the process would be considered in control based on these 20 points since none of the points fall outside of the control limits and the points do not follow an obvious pattern.

If the characteristic which is being monitored is represented by the sample statistic X with mean μ_x and standard deviation σ_x then the parameters of the control chart are defined by [46]

$$\begin{aligned} \text{UCL} &= \mu_x + K\sigma_x \\ \text{Average} &= \mu_x \\ \text{LCL} &= \mu_x - K\sigma_x \end{aligned}$$

where K is determined by the desired level of significance and the distribution of the variable X. For example, consider the process of filling cereal boxes where the machines are calibrated to fill each box with $\mu = 16$ oz. of cereal and the known standard deviation is $\sigma = .4$ oz.

To monitor the process, $n = 4$ random samples are collected and weighed every hour. If X represents the mean of the samples taken every hour, then, by the central limit theorem, X is normally distributed with mean $\mu_x = 16$ oz. and with standard deviation $\sigma_x = \sigma/\sqrt{n} = .4/\sqrt{4} = .2$. For a significance level of $\alpha = .05$ the UCL and LCL would be

$$\text{UCL} = \mu_x + Z_{\alpha/2}\sigma_x = 16 + 1.96(.2) = 16.392$$

and

$$\text{LCL} = \mu_x - Z_{\alpha/2}\sigma_x = 16 - 1.96(.2) = 15.608$$

where $Z_{\alpha/2}$ is such that the $\Pr\left[Z > Z_{\alpha/2}\right] = \alpha/2$ for a standard normal variable Z. The process is considered to be in control whenever the average of the sample randomly falls in the interval $(15.608, 16.392)$.

A control chart is in some ways equivalent to conducting a hypothesis test each time a sample is collected [46]. Generally speaking, control charts test the null hypothesis that the process is in statistical control at a particular time. If an observation falls outside the control limits then the null hypothesis would be rejected and the process would be considered out of statistical control. Mathematically, the test examines whether the true mean of the process equals a value μ_0. If the test is rejected, it is assumed that the true process mean is $\mu_1 \neq \mu_0$ and therefore the process has changed. For instance, in the cereal box example, defining the control limits would be the same as conducting the hypothesis test

$$H_0 : \mu_x = 16 \quad \text{versus} \quad H_A : \mu_x \neq 16$$

for a significance level of $\alpha = .05$. This hypothesis test would be conducted each time a sample is gathered. Therefore, the null hypothesis would be rejected when the mean of the observations taken at a particular time fall outside the confidence bounds.

There are many reasons why the null hypothesis could be rejected [4] and control charts have been designed to detect the different behaviors. The control chart described in the cereal box example is called a Shewhart chart and it is very good at detecting a large, sudden shift in the process mean. The use of different supplies or an incorrect setting could cause the process mean to jump. A change in the underlying variability of the process could also indicate out of control behavior. Other assignable causes could be a trend or slow change in the mean resulting from the wearing out of machinery parts. This could be indicated by patterns in the control chart. Another reason for patterns in the control chart could be the introduction of a nonrandom component entering into the process. The next section will discuss statistical process control methods designed to detect periodic behavior which can enter and significantly alter a process.

4.2 Evenly Sampled Data

Factors which are periodic such as a motor cycling on and off, fluctuations in temperature, humidity or pressure or the shift rotations of employees could affect the output of a manufacturing process. Since these influences can result in an out of control process, it is important to be able to detect when specific periodic factors are present in a process. As discussed in Section 3.1.2, Fisher developed a statistical test to detect significant periodic behavior for evenly sampled data based on the relative size of the dominant peak of the periodogram. Since then many approaches have been proposed to improve the signal detection capability of Fisher's test on evenly sampled data when non-Fourier frequencies or multiple frequencies are involved [7, 33, 68, 69, 73, 80, 81]. Using these modified tests, control charts for the detection of periodic behavior were designed to supplement traditional mean and variance control charts.

The first control chart developed for detecting periodic behavior was proposed by Beneke, Leemis, Schlegel and Foote [4]. This control chart is based on Fisher's test statistic, defined on page 38, and monitors the ratio of the largest value of the periodogram to the average of the periodogram estimates. The value of the test statistic

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\frac{1}{M}\sum_{k=1}^{M} P_X(\omega_k)}$$

is calculated and compared against the upper control limit which is defined to be the critical value of Fisher's test statistic for a predetermined significance level. The process would be considered out of control due to the introduction of periodic behavior whenever T exceeds the critical value. Figure 4-2 illustrates the basic idea of the spectral control chart. In this example the process being monitored goes out of control at the eighteenth sample.

The spectral control chart proposed by Beneke et al. requires N evenly sampled data points which will be used to calculate the periodogram. The sample size N and sampling rate also determine the Fourier basis frequencies that will be tested. Choosing the sample size may depend on previous knowledge about the process and information about what frequency structures are of interest. Once the frequency basis has been chosen, the spectral control chart will use the most recent point in addition to the previous $N-1$ samples. Let $X_1, X_2, \ldots, X_N$ represent the first N samples collected using a fixed sampling rate. Here

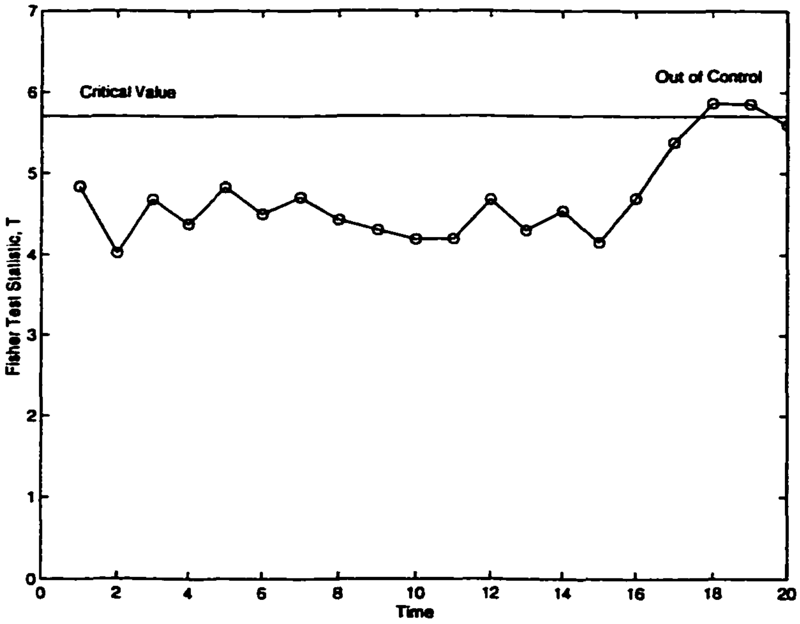


Figure 4-2 Example of a spectral control chart

the data value X_k may denote a single observation taken at time t_k or the average of a batch of measurements taken at t_k. It is also assumed that the data values are sampled from a random normal process. The value which is plotted on the control chart is

$$T = \frac{\max_k \{P_X(\omega_k)\}}{\frac{1}{M}\sum_{k=1}^{M} P_X(\omega_k)} \quad \text{for} \quad k = 1, 2, \ldots, M = \lfloor N/2 \rfloor$$

where $P_X(\omega_k)$ is the periodogram estimate for the frequency $\omega_k = 2\pi k/N$ using the N most recent data points $X_1, \ldots, X_N$. The value T would be plotted on the control chart and compared against the upper control limit. When the next observation X_{N+1} is taken, the statistic T will be recalculated using the data points $X_2, \ldots, X_{N+1}$. This procedure is repeated as more data is collected until an out of control point is signaled. A point which is out of control would indicate that the data has a periodic component and is not longer random. The frequency of the cyclic behavior could be determined by locating the frequency associated with the largest periodogram estimate.

Since this control chart is based on Fisher's test statistic it shares the same pitfalls as Fisher's test. While the control chart is effective at determining when periodic behavior corresponding to a Fourier basis frequency enters the process, it performs poorly when multiple periodic cycles are present or if the cycle corresponds to a frequency midway between two Fourier basis frequencies. Spurrier and Thombs [69] proposed a control chart which would signal out of control behavior for a wider range of frequencies. Their control chart is similar to the previous control chart except that the test statistic T', defined on page 40, is monitored. By construction, this control chart is better at detecting periodic behavior for a wider range of frequencies. However, the capabilities of the control chart deteriorate when multiple periodicities are present. Lastly, another control chart was proposed by Tatum [73] based on the modified statistical tests on page 41. These charts show significant improvement over their counterparts previously discussed.

4.3 Randomly Sampled Data

All these methods, however, assume that the data is collected regularly in time. In practice though it may not be feasible to collect evenly spaced samples. Beneke et al. [4, page 66] state "If the observations are at irregular time intervals ... the procedures of this article cannot be used directly" in reference to their control chart based on Fisher's test statistic. In fact the previous chapter demonstates that direct application of Fisher's test on unevenly sampled data is incorrect since the distribution of the periodogram for small N is not exponential. Therefore, the critical values used in the spectral control chart of evenly sampled data are not appropriate for randomly sampled data. Instead, the critical values need to be adjusted and can be estimated by Monte Carlo simulations.

A control chart which monitors Fisher's test statistic can be used to detect periodic behavior in data that is sampled randomly. Its setup will be similar to the control chart for evenly sampled data proposed by Beneke et al. except the value of Fisher's statistic is measured against the critical value found through Monte Carlo simulations. The process is considered to be free of cyclic behavior whenever the statistic falls below the critical value.

To evaluate the effectiveness of the control chart, the average run length (ARL) will be calculated. The ARL is the number of observations expected before an out of control point is signaled. Clearly, the ARL is a function of the critical value as well as the number of samples N used in the calculation of the statistic. The ARL is estimated using Monte Carlo simulations to determine how quickly periodic behavior is detected.

First the critical value of the periodogram estimate for standard normal data sampled at $N = 51$ random times was computed numerically using Monte Carlo simulations. A significance level of .0027 was chosen since it corresponds to a 3σ limit under the assumption of normality and this value is typically used in control charts. For this sample size, the critical value of Fisher's test statistic for the unevenly sampled data was found to be 7.2732 using the Monte Carlo method. In this simulation, 51 random sampling times were selected. For each sampling time, a data value was independently drawn from a standard normal distribution. Fisher's test statistic was calculated using the generated data and recorded. This procedure was repeated 300000 times and the critical value of 7.2732 corresponds to the 99.73rd percentile of the trials. This process was repeated for a different random sampling scheme and a similar critical value was found. The corresponding critical value found directly from the distribution of Fisher's test statistic for 51 evenly sampled data points is 7.914. Since some practitioners may incorrectly apply Fisher's test directly to unevenly collected data, the ARL between the control charts using these two critical values will be compared.

To estimate the ARL for the spectral control chart, $N = 51$ random sampling times were selected. The simulation began with the generation of 51 independent samples from a standard normal distribution. Fisher's test was carried out on these N points by calculating the periodogram estimate for each of the Fourier basis frequencies. The Fisher test statistic was then compared against the upper critical value. For each trial, Fisher's test was carried out using both the standard, evenly sampled critical value, 7.914, and the Monte Carlo determined critical value, 7.2732. If the chart did not signal, then an observation at a random time value was generated from a sinusoid model

$$X_j = A\cos(2\pi t_j/5.1) + \epsilon_j$$

where each ϵ_j is selected independently from a standard normal distribution and A represents the amplitude. The value A will be changed for different trials. Since a new observation was collected, the test statistic was computed using data values corresponding to $j = 2, 3, \ldots, 52$ and the current state of the process is determined. Another sample is selected at random whenever the process is found to be in control. The run length is equal to the number of times the test statistic is calculated before the process is determined to be out of control. The simulation was carried out 1000 times and the ARL is defined to be the average of these run lengths.

A summary of the ARL estimates from two such simulations for each amplitude $A =$

Method	Critical Value	Trial	$A = 0$	$A = 1$	$A = 2$	$A = 4$
Standard Fisher's Test	7.914	1	3496.0	80.1	33.6	24.3
		2	3761.6	87.6	33.5	24.3
Modified Fisher's Test	7.2732	1	1338.8	62.9	30.2	21.8
		2	1387.1	60.3	30.4	22.2

Table 4.1 Average Run Lengths for Randomly Sampled Data

Trial	$A = 0$	$A = 1$	$A = 2$	$A = 4$
1	1704.9	52.1790	29.9270	21.9500
2	1655.1	50.7640	29.6840	21.8700

Table 4.2 Average Run Lengths for Evenly Sampled Data

$0, 1, 2$ and 4 is found in Table 4.1. Not surprisingly the standard approach has higher average run lengths which is a factor of the inflated critical value. The modified method detects periodic behavior much more quickly than the standard method when the amplitudes are smaller. For comparison, the ARLs for evenly sampled data are listed in Table 4.2. The ARLs for the modified test are very similar to those obtained for evenly sampled data. Therefore, even though the exact statistical properties of Fisher's test are lost when data is collected irregularly, using critical values based on Monte Carlo simulations allow the construction of a control chart which performs about the same as the evenly sampled control chart based on exact distribution theory.

4.4 Other Monitoring Devices

The control chart methods described above help to determine whether or not periodic behavior has entered a process which is assumed to be random. However, this situation may not be applicable when the process being studied is assumed to have periodic behavior. In this case, there are several questions which may be of interest. First, it may be useful to determine if periodic behavior in a system persists or is a temporary effect that cycles in and out of the process. The evolutionary spectrum, discussed in the following section, is a tool which can be used to monitor cyclic behavior. Although the evolutionary spectrum is designed to be used on evenly sampled data, we have extended its use to unevenly sampled data.

Another piece of information that might be useful is an estimate of the underlying noise variance. Since the process is assumed to be periodic in nature, the sample variance will be inflated due to the oscillations of the data. By creating ensembles from the observed data in a moving window of data, the periodogram can be used to estimate the variability of the random error process. Preliminary results using evenly sampled data will be presented in Section 4.4.2.

4.4.1 Evolutionary Spectrum

The evolutionary spectrum is used to monitor the periodogram as it evolves over time. The periodogram is calculated within a time window which is then moved through the data. Either overlapping or nonoverlapping windows can be used. Contour and surface plots suggest how the periodogram estimates vary over time. Statistical information can be used to allow only statistically significant behavior to be represented on the plots. In other words, only periodogram estimates that exceed a particular critical value would be plotted.

For example, consider the data collected regularly from a sine model with a time-varying amplitude A_j

$$X_j = A_j \sin(2\pi t_j/5.1) + \epsilon_j$$

where $t_j = 1, 2, \ldots, 1000$ and $\{\epsilon_j\}_{j=1}^{N}$ are standard normal errors. We will demonstrate how the evolutionary spectrum works using examples involving two different amplitude functions. In the first example the amplitude function will be almost constant. The amplitude of the sine term will drift slightly from a mean value of 2, as shown in Figure 4-3a. In the second example the amplitude function will be treated as periodic since the amplitude of the sine term will be determined by the function

$$A_j = \cos^2(2\pi t_j/500),$$

shown in Figure 4-3b.

In both examples the evolutionary spectrum was calculated using nonoverlapping windows of 51 samples. Only values which exceed the critical value determined by a 5% level of significance are represented on the plot. A surface plot showing how the spectrum of the first example evolves over time is shown in Figure 4-4a. Recall that the amplitude of the sine wave with period 5.1 is held almost constant. A contour plot generated from the surface plot, shown in Figure 4-4b, indicates that there is persistent periodic behavior corresponding to the cycle of 5.1 time units. This behavior is represented by contours which are drawn close together located at the period 5.1. The contours and the peaks in the surface plots remain for the duration of the time series indicating the cyclic behavior in the process does not change significantly.

Compare these results to the second example in which the sine model

$$X_j = A_j \sin(2\pi t_j/5.1) + \epsilon_j$$

has a periodic amplitude

$$A_j = \cos^2(2\pi t_j/500).$$

Clearly, the process determined by the sine wave has an intrinsic cycle of 5.1 days, but the amplitude of that cycle varies over time. The periodogram plot of the data, Figure 4-5, correctly identifies that there is a 5.1 day cycle, however, this gives no information about the amplitude. Figure 4-6a shows the evolutionary spectrum of the periodogram values which exceed the 95th percentile using a moving window of 51 data points. A corresponding contour plot is shown in Figure 4-6b. Notice that the dark values in the contour plot indicate a large value of the periodogram. It is clear from both the contour and surface plots that the data has a 5.1 day cycle, but that cycle comes in and out of the process.

A final example demonstrates how the evolutionary spectrum can detect both of these behaviors simultaneously. Consider data collected regularly from the process

$$X_j = 4\sin(2\pi t_j/5.1) + A_j \sin(2\pi t_j/3) + \epsilon_j$$

where

$$A_j = \begin{cases} 3 & \text{for} \quad j = 601, 602, \ldots 800 \\ 0 & \text{otherwise} \end{cases}$$

and $\{\epsilon_j\}_{j=1}^{N}$ are normally distributed with zero mean and standard deviation equal to 2. The evolutionary spectrum, in Figure 4-6, clearly shows that the data has a constant period 5 cycle, but the period 3 cycle is a temporary effect which affects only points 600-800, approximately.

While typically used on evenly sampled data, as illustrated, we have adapted the concept for unevenly sampled data. The only additional constraint is the number of observed data in each window needs to be sufficiently large in order to get good estimates of the periodogram. Otherwise, one can skip that particular window since reliable estimates can not be found or a larger window size can be chosen. The evolutionary spectrum program, which can handle evenly or unevenly sampled data, can be found in Appendix D.

The evolutionary spectrum is another tool which can be used to supplement the periodogram. While the periodogram can detect periodic behavior, it does not provide information about the nature of the cyclic behavior. The evolutionary spectrum provides some knowledge about whether cyclic behavior is changing temporally or if the amplitudes are relatively constant. This additional information can give practitioners insight into the underlying dynamics in the process of interest.

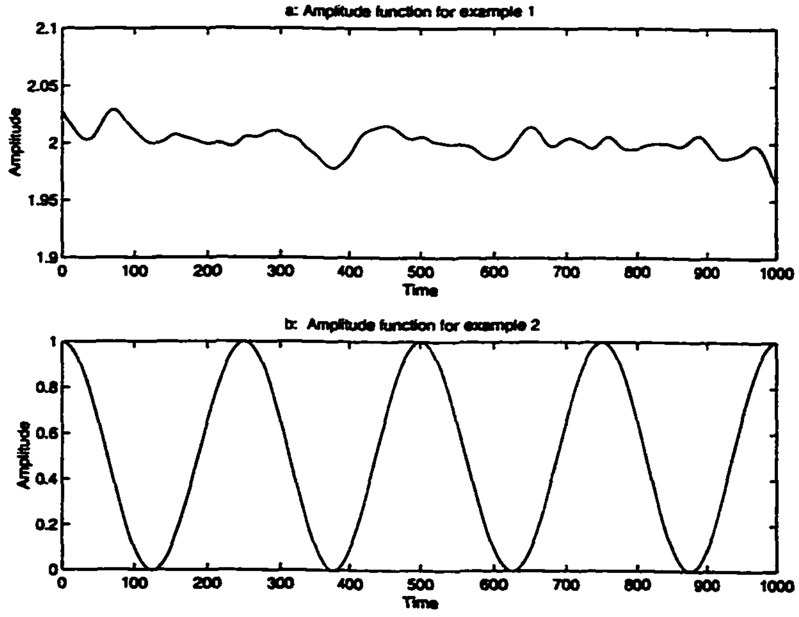


Figure 4-3 Time-varying amplitude

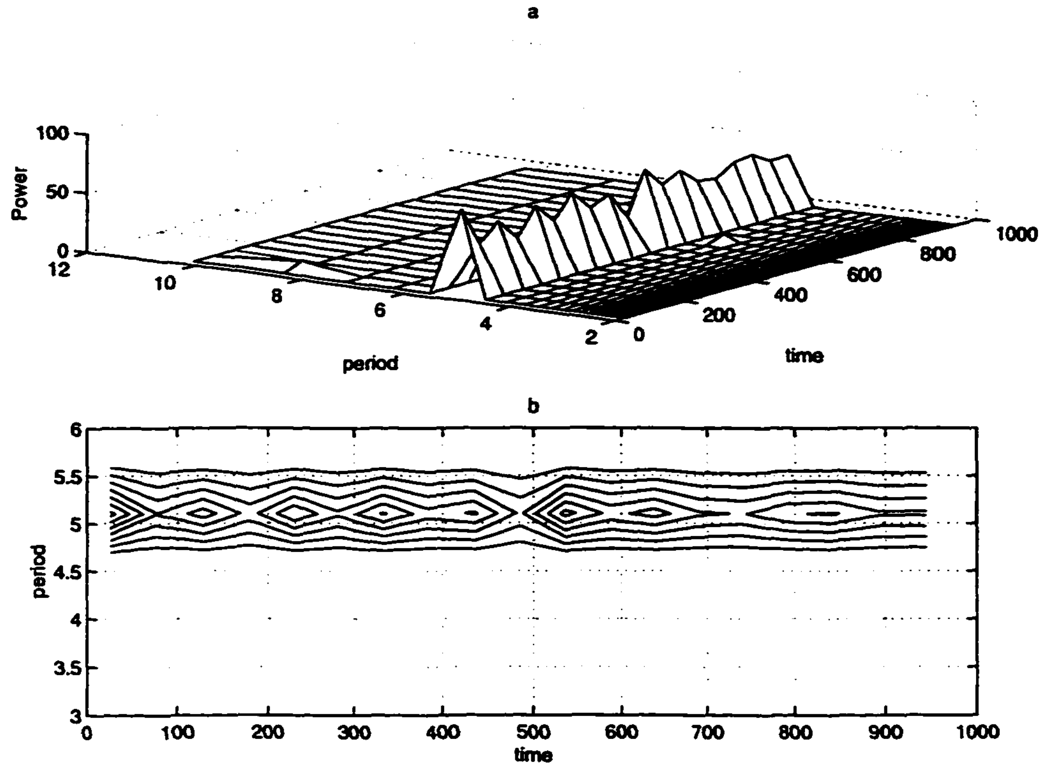


Figure 4-4 Evolutionary spectrum of data with time-varying amplitude

4.4.2 Estimating the Variance of the Noise Process

Another useful by-product of the periodogram is that it can be used to generate variance estimates for the underlying random process, even if the process is periodic. When the data has sinusoidal behavior it follows the model

$$X_j = \mu + \sum_k A_k \cos(\omega_k t_j) + \sum_k B_k \sin(\omega_k t_j) + \epsilon_j$$

where each ϵ_j is assumed to be an independent sample from a normal distribution with mean 0 and variance σ^2. Since the variability about the sinusoids are reflected by the error

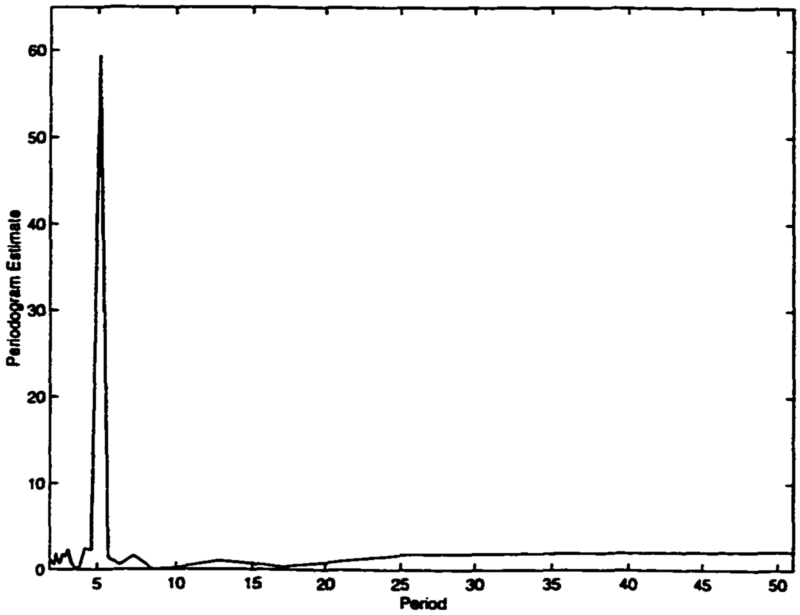


Figure 4-5 Periodogram of data with time-varying amplitude

term, it is usually important to estimate σ^2. The usual sample variance estimate

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_j - \bar{X})^2$$

with $\bar{X}$ equal to the mean of the observed data, cannot be used to estimate σ^2 since the estimate includes the variability associated with the oscillations of the sinusoids. The estimate $\hat{\sigma}^2$ overestimates the noise variance.

As seen with the evolutionary spectrum, the process may also have sinusoids which influence the process for brief periods of time and then disappear. Simply filtering the dominant frequencies out of the data may not appropriately model the time varying amplitudes that are present in the process. The residuals may not accurately represent the true errors.

An alternate method of generating variance estimates for the underlying random pro-

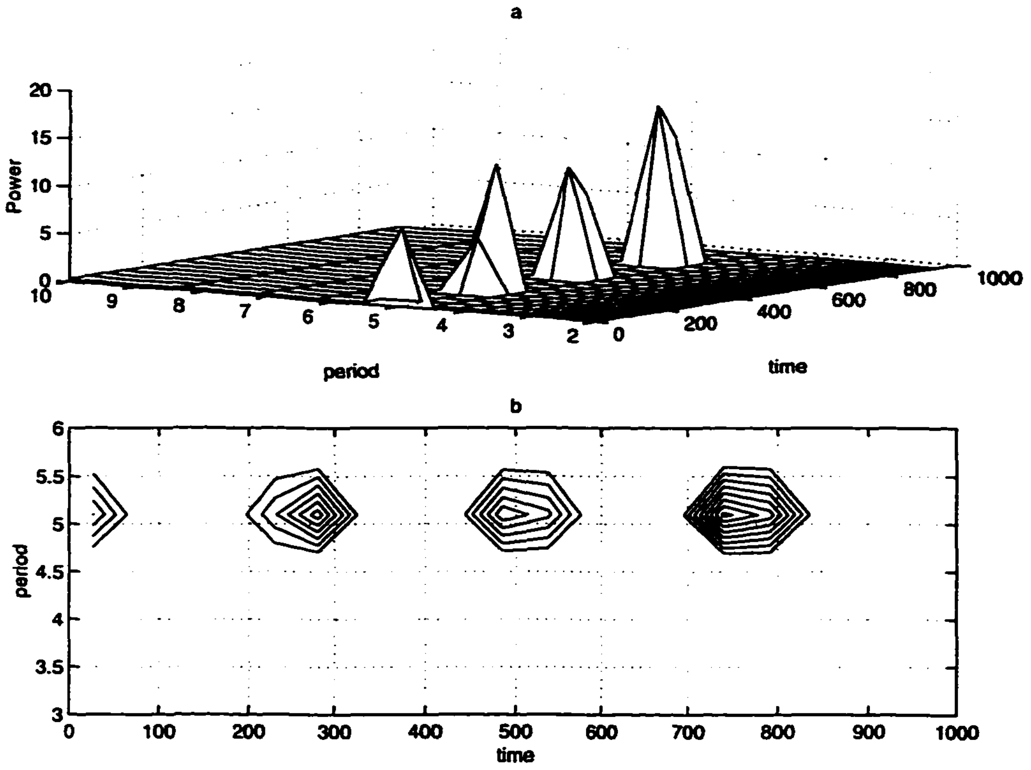


Figure 4-6 Evolutionary spectrum of data with cyclic amplitude

cess is with the periodogram. As shown in Chapter 2, the periodogram is exponentially distributed with mean parameter σ^2 under the null hypothesis that the data is collected regularly from a normal process with variance σ^2. Each periodogram estimate has an expected value of σ^2. Therefore, the variance parameter can be estimated by averaging the periodogram estimates across frequencies. A better estimate can be found by taking a long span of data windowed into an ensemble and evaluating the periodogram estimates for each ensemble. This way the variance estimate can be found by averaging the periodogram estimates across ensembles and frequencies.

Unfortunately, the data may be cyclic in nature and the periodogram estimates corre-

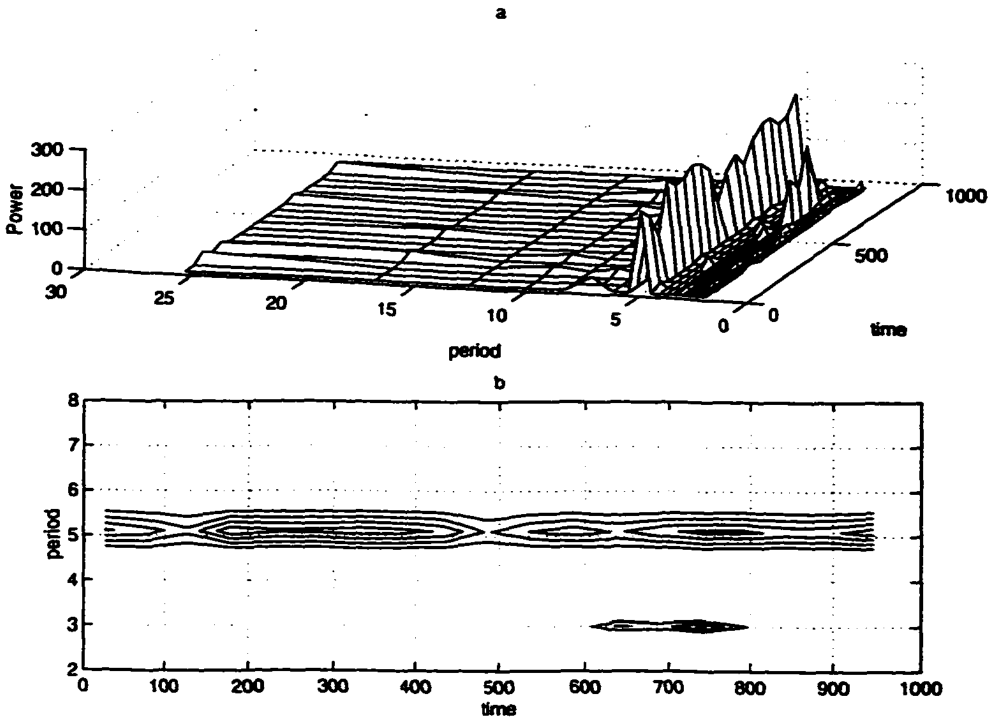


Figure 4-7 Evolutionary spectrum of data with persistent and temporary cyclic behavior

sponding to the frequencies present in the data will be large. These large values will therefore inflate the variance estimates if this procedure is used. To remove the influences of these large estimates, the null distribution of the periodogram estimates can be used to determine which values of the spectral power are significantly large. These values can then be ignored in the calculation of the average of the estimates across ensembles and frequencies. In particular, for a random variable X which is exponentially distributed with mean parameter σ^2, the probabilities of exceeding values larger than multiples of the mean are

$$\Pr[X > 3\sigma^2] = 0.0498$$

$$\Pr[X > 4\sigma^2] = 0.0183.$$

So, periodogram estimates which exceed $C\sigma^2$ for $C = 3,4$ or some suitable number in between, are assumed to be the effect of periodic behavior. These significantly large values will be ignored in the averaging process within the window. This trimming method should eliminate large estimates which are attributed to a periodic signal when calculating an estimate of the underlying noise variance.

For example, consider again the model

$$X_j = 4\sin(2\pi t_j/5.1) + A_j \sin(2\pi t_j/3) + \epsilon_j$$

where

$$A_j = \begin{cases} 3 & \text{for} \quad j = 601, 602, \ldots 800 \\ 0 & \text{otherwise} \end{cases}$$

and ϵ_j has zero mean and variance equal to 4. The estimated variance from the observed data is 12.7571. Clearly, this is not a good estimate of the underlying noise variance, which is 4. An estimate of the underlying noise variance at a given time can be found using the most recent collection of 255 data points. Five ensembles of length 51 were created from this data window, which can clearly be updated as new data is collected. The periodogram is calculated in each of the five ensembles and only those spectral estimates that are below the value $C\sigma^2$ are averaged to estimate the noise variance at the current sampling time.

Figure 4-8 compares the variance estimates using three different techniques. For comparison, the true noise variance is represented by the horizontal line at 4. One estimate was generated using the sample variance of the data in the window. Another estimate was the average of all the periodogram estimates in the five ensembles. These two estimates, the upper two plots in Figure 4-8, are similar and tend to estimate the total variance of the data and not the underlying noise variance. Since the data is periodic, it is not surprising these two estimates are not very good. The final estimate was the average of only those periodogram estimates that fell below the cutoff point $4\sigma^2 = 16$ and this gives a better estimate of the underlying noise variance. In fact, all of the estimates are within .8571 of the true variance.

Now since σ^2 is what we are trying to estimate, an estimate of the cutoff value $C\sigma^2$ can come from historical data or the variance estimate of the data. For example in the previous example the cutoff value was chosen to be four times the true variance. However, if that were unknown or could not be estimated from previously collected data, the cutoff value $3\hat{\sigma}^2 = 3(12.7571) = 38.2713$ could have been used. Figure 4-9 shows the results using this bound. The estimates using the trimmed method are still better than the estimates from the other two methods.

Caution must be used since a bad initial estimate of the variance can still lead to an overestimation of the noise variance. In this case, the cutoff value is too high and the influence of the periodicities still enter the calculation. The choice of C is therefore important. If the initial estimate of σ^2 is severely overestimated a lower value of C can be used to reduce the cutoff level and diminish the effect of periodicities. In the previous example, the value of C was reduced to 3 to account for some of the inflation of the initial estimate of the variance.

Another issue is the choice of a window size and number of ensembles to create. If the window size is too large, the variance estimation will be made using data that may be uncorrelated with the present data point. The number of ensembles to use is also somewhat arbitrary. Since the averaging process tends to converge when more data is available, a large

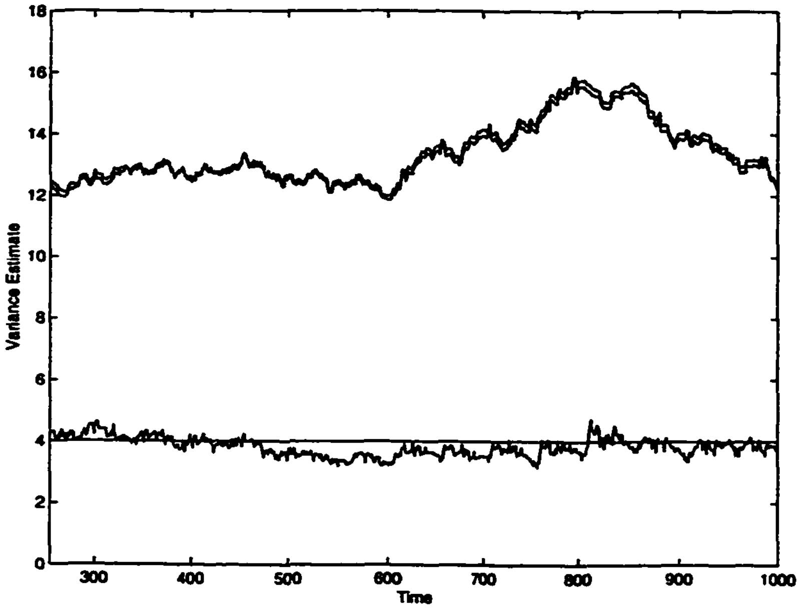


Figure 4-8 Monitoring the underlying noise variance with known initial variance

number of subwindows would be preferable, but this needs to be balanced with the width of the overall window.

These preliminary results indicate that this method of estimating the variance has some promising characteristics. The method should extend to unevenly sampled data, however, the distribution of the periodogram is no longer exponential. An appropriate upper bound can be obtained from the Monte Carlo simulations. In addition, window and ensembles need to be defined so that there are enough data points in each subwindow to construct the periodogram. These problems lay the foundation for future research.

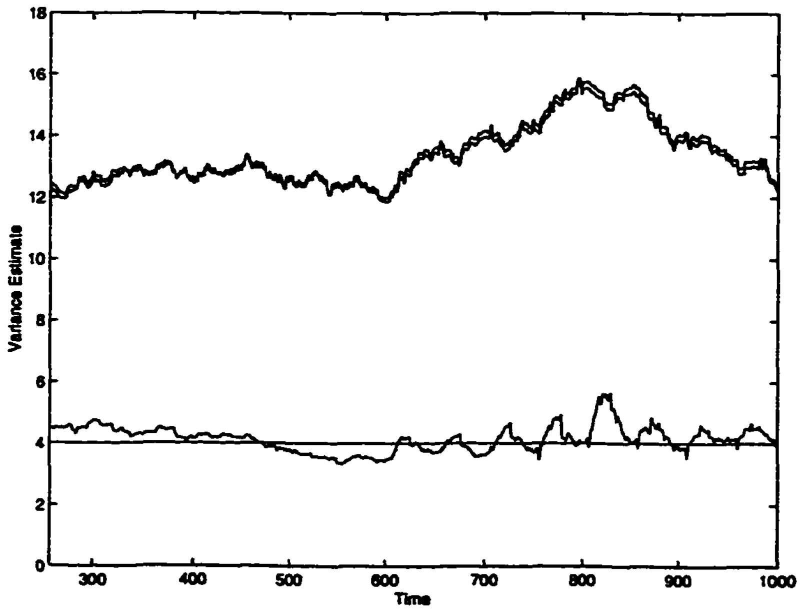


Figure 4-9 Monitoring the underlying noise variance using initial variance estimate

Chapter 5

ANALYSIS OF SILICON WAFER COATING PROCESS

5.1 Methods

Statistical process control techniques, like control charting, require randomly and normally distributed data. Frequently though, observed data is not random and normally distributed and is assumed to be the sum of deterministic and random error processes. In order to utilize statistical process control, these two components must be identified and separated. Modeling techniques are used to account for the determinism while control charts can be used to monitor the variation of the random process. Many procedures have been developed for modeling regularly sampled data. Popular time series models include the autoregressive (AR), moving average (MA) and autoregressive integrated moving average (ARIMA) processes. These models assume a fixed sampling rate and determine a model using the correlation structure of the time series. In addition, spectral methods can be used to model periodic behavior in data. Typical applications of periodogram analysis usually assume even sampling. However, data collected by industries sometimes cannot be sampled regularly due to constraints in the production and data collection schemes. In these cases, it is not appropriate to use these common time series models.

In addition to unevenly sampled measurements, multivariate measurements can be collected. These observations could be simultaneous measurements taken from different, possibly related processes. Alternatively, the term multivariate could refer to the individual dimensions of multidimensional data collected from a single process. Although each dimension of a multivariate time series contributes information regarding the underlying dynamics of the process, it may be hard to find a parsimonious, multivariate model. Section 5.1.1 will review how principal component analysis can be used to collapse multivariate time series data into a lower dimension while still retaining the interesting dynamics of the system. The dynamics in the lower dimensional time series can then be analyzed with spectral methods developed specifically for unevenly spaced data. These methods, introduced in previous chapters, identify significant periodic behavior in the data. The significant periodic signals that represent the deterministic part of the process can then be removed from the time series so that the remaining variability can be examined. Additionally, reasons for the cyclic behavior can be investigated. Section 5.1.2 reviews how to extract the periodic signal for the data using least-squares. Once the cyclic behavior has been removed, the remaining structure can be analyzed for normality and randomness. If the residuals, or errors, are normally distributed and random, statistical control methods can be implemented on the residuals. These methods are applied to an industrial data set in Section 5.2.

5.1.1 Principal Component Analysis

Multivariate time series, although typically easy to collect, may be hard to analyze. One reason is that there may be too many time series relative to the number of observations. This situation yields an underdetermined system which makes it difficult to fit a model to the data. Also, the individual series are typically correlated which can lead to an ill-conditioned model. In this case, any model found could be unstable. Lastly, it may be hard to interpret and generalize results of analyses done to individual series. Principal component analysis (PCA) [22, 55, 64] seeks to remedy these problems by finding linear combinations of the multivariate time series which capture the structure in the system and reduce dimensionality. Geometrically, PCA rotates the observations to a new coordinate system. The transformed coordinates have dimensions which are orthogonal and hence are uncorrelated. For example, in a situation with multivariate time series of dimension 2, the multivariate plot of the data may fall in an ellipse. If so, PCA would transform the x- and y-axes to coincide with the major and minor axes of the ellipse, thereby removing the linear correlation from the data.

Consider the p-dimensional multivariate time series where $Y_j = [Y_{j1} \quad Y_{j2} \quad \ldots \quad Y_{jN}]'$ represents the jth univariate time series for $j = 1, \ldots, p$. It is assumed that each time series has zero mean. Also, since PCA is not scale invariant, the individual time series should be standardized if the scales of the time series differ. Thus, let $X_j = (Y_j - \bar{Y}_j)/s_{Y_j}$ be the standardized series where $\bar{Y}_j$ and s_{Y_j} are the sample mean and sample standard deviation of Y_j, again for $j = 1, 2, \ldots, p$. Define the $p \times N$ design matrix

$$\mathbf{X} = \begin{bmatrix} X_1' \\ X_2' \\ \vdots \\ X_p' \end{bmatrix}$$

$$= \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1N} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2N} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ X_{p1} & X_{p2} & X_{p3} & \cdots & X_{pN} \end{bmatrix}$$

where X_{ji} corresponds to the ith sample of the jth time series. Notice that the ith column contains the ith sample for all of the p time series.

Consider multiplying $\mathbf{X}$ by an orthonormal rotation matrix $\mathbf{A}$. This transformation results in new coordinates defined by the relation $\mathbf{Z} = \mathbf{AX}$. The primary goal of PCA is to find a rotation matrix $\mathbf{A}$ which will yield row vectors which are uncorrelated. Geometrically, this can be viewed as finding the matrix $\mathbf{A}$ which rotates the axes of the cluster of points of the observed data to align with the principal axes. Hence, the matrix $\mathbf{Z}$ is required to have a sample covariance matrix

$$\mathbf{S_Z} = \begin{bmatrix} s^2_{Z_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & s^2_{Z_2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & s^2_{Z_p} \end{bmatrix}$$

where $s^2_{Z_k}$ is the sample variance of Z'_k, the kth row of $\mathbf{Z}$. But, since $\mathbf{Z} = \mathbf{AX}$, it follows that $\mathbf{S_Z} = \mathbf{ASA'}$ where $\mathbf{S}$ is the sample covariance matrix of the design matrix $\mathbf{X}$. Therefore, the spectral decomposition theorem [22] can be invoked to determine the form of $\mathbf{A}$. The matrix $\mathbf{A}$ must contain the (normalized) eigenvectors of $\mathbf{S}$. In addition, the resulting diagonal entries of matrix $\mathbf{S_Z}$ are the associated eigenvalues of $\mathbf{S}$. Thus,

$$\mathbf{A} = \begin{bmatrix} V_1' \\ V_2' \\ \vdots \\ V_N' \end{bmatrix}$$

and $s^2_{Z_k} = \lambda_k$ where V_k and λ_k are the kth eigenvector and eigenvalue of $\mathbf{S}$.

The p principal components are defined to be $Z_k = V_k'\mathbf{X}$ for $k = 1, \ldots, p$. The components represent the transformed variables in the new coordinate system. The variance of Z_k can be estimated by λ_k. Also, $\text{Var}(Z_1) \geq \text{Var}(Z_2) \geq \cdots \geq \text{Var}(Z_p)$ since the eigenvalues are ordered from largest to smallest. Using this property, it is possible to define the proportion of the total variance that the principal components explain. In particular, the proportion of variance explained by the first k principal components is equal to

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

It is this feature which assists in determining if dimension reduction is possible. For instance, if the correlation between the time series is high, the first few principal components will probably explain most of the variability. The reason for this is that the first few eigenvalues would be large. The remaining eigenvalues would be relatively small and the associated principal component may not contribute much information. Consider the geometric example where the data vectors lie in a flattened ellipsoid. Most of the information in the data, occurs a plane. In this situation, only the first two principal components would be needed to capture the structure of the data. The dimension could then be reduced from three dimensions to two.

Several criteria for selecting the dimension in a principal component analysis are outlined by Rencher [55]. Ideally, one hopes that the dimension will be reduced by ignoring some of the principal components. A general guideline is to choose the principal components which together account for at least 80% of the total variation. Another procedure specifies that principal components should be ignored if the associated eigenvalue is less than the average of the eigenvalues. An additional method is a graphical approach. A cutoff point between "large" and "small" eigenvalues can usually be found by plotting λ_i versus i. Only the principal components associated with "large" eigenvalues would be retained. It is important to note that the principal components which are not used in further analysis may carry important information. One must weigh the benefits of dimension reduction versus the potential loss of information. However, in most cases the first few principal components contain most of the information in the original data and can be used in further analyses instead of the original data. Thus, information from correlated series can be combined into a lower dimensional time series.

5.1.2 Extraction of Periodic Signals

Chapter 3 explored ways to statistically test for periodic signals in unevenly sampled data. Once we detect signals in the time series which are statistically significant, we will want to remove the periodic structure. The removal of the discrete periodic components of the time series is necessary to reveal the stochastic or random behavior of the process. Specifically, if $p_1, p_2, \ldots, p_k$ are the periods which are significant, the method of least squares can be used on

$$X_j = \sum_{i=1}^{k} A_i \cos(2\pi t_j/p_i) + \sum_{i=1}^{k} B_i \sin(2\pi t_j/p_i) + \epsilon_j \quad \text{for} \quad j = 1, \ldots, N$$

to find the amplitude estimates $\hat{A}_i$ and $\hat{B}_i$ for $i = 1, 2, \ldots, k$. Once the amplitudes have been estimated, the residual time series can be formed and analyzed. In particular, the residuals or errors,

$$R_j = X_j - \left[\sum_{i=1}^{k} \hat{A}_i \cos(2\pi t_j/p_i) + \sum_{i=1}^{k} \hat{B}_i \sin(2\pi t_j/p_i) \right] \quad \text{for} \quad j = 1, \ldots, N$$

can be evaluated for normality and randomness. If the residuals satisfy these conditions, a control chart can be created to monitor the process.

5.2 Industrial Application

The techniques described above were used to analyze two processes in which the thickness of photoresist coating on silicon wafers is monitored. The two processes are similar, except that different thicknesses of coating are applied. These two data sets will be called Resist A and Resist B. The data, consisting of the width of photoresist coating, was collected simultaneously from both processes over one year. The time lags between measurements range from one minute to 32 hours, so clearly the time series is not evenly sampled. Also, there were several batch measurements where multiple measurements were recorded at the same time. These multiple measurements were averaged to get one measurement per sample time. Plots of the two processes are shown in Figures 5-1a and b. The large span of missing data (approximately t=2.2 × 10^7 to 2.5 × 10^7 seconds) corresponds to corrupted data collected from Oct. 3, 1997 to Oct. 20, 1997. The other stretch of missing values starts at approximately 3.1×10^7, the Christmas-New Year holiday break. Despite these gaps, the raw data shows periodic behavior. This cyclic behavior needs to be removed before one can implement statistical process control. It also appears that the Resist B process was adjusted around $t=1.5 \times 10^7$ seconds (June 27, 1997). In order to account for this adjustment, the data after June 26, 1997 was shifted up by the mean. Overall, the Resist A and Resist B time series appear to have similar behavior, which is not surprising considering the two processes are almost identical. For this reason, multivariate techniques will be used instead of doing two separate univariate analyses. The time series values were paired to create two-dimensional vectors. Then singular value decomposition was applied to separate the principal components. The first principal component, shown in Figure 5-1c, explains 85.6% of the total variance and seems to capture the periodic behavior exhibited in both time series. Therefore, instead of working with the two original time series individually, the analysis will concentrate on the first principal component only.

Since the data was unevenly sampled and exhibited strong periodic behavior, the normalized Lomb periodogram was used to find dominant frequencies of the first principal component. The spectral estimate, shown in Figure 5-2, indicates that there is a strong period of 1.198×10^7 seconds (approximately 139 days). This verifies that the data seems to be dominated by a slowly oscillating cycle. Figures 5-4a and b show the signal and the residuals after removing this periodic component.

Table 5.1 shows additional significant periods indicated by the normalized Lomb periodogram and the percent variance each contributed. The extracted signal using all the periods listed in Table 1 does seem to follow the data, as shown in Figure 5-4c. However, the residuals, in Figure 5-5, may have additional periodic structure remaining. Since the

Period (Approx. Days)	% Variance Explained
139	17.76
8	6.17
52	3.40
75	2.53
25	1.27
20	1.22
3	1.11
26	.71
12	.75

Table 5.1 Significant Periods Found Using Lomb Periodogram

residuals do not appear random, it is not appropriate to implement a Shewhart control chart on the residuals.

Analyzing the same data using the classical periodogram with 5000 permutation resamples reveals similar results, as shown in Figure 5-3. Significant periods are indicated where the periodogram estimates, represented with '*', exceed the 99.9% upper confidence bound denoted by 'o'. The periods which were statistically significant are listed in Table 5.2. The period of 139 days, found by the periodogram, was used to account for the slow oscillation present in the data. Both the Lomb periodogram and the classical periodogram were in agreement on all of the most significant periods, with the only disagreement occurring in the last two periods detected. The agreement between the two techniques indicates that the dominant structures are consistent.

5.2.1 Additional Analysis of the Coating Process

The evolutionary spectrum can be used to analyze the behavior of the dominant frequencies found in periodogram analysis. The evolutionary spectrum of the first principal component,

Period (Approx. Days)	% Variance Explained
139	17.76
8	6.17
52	3.40
75	2.53
25	1.27
20	1.22
3	1.11
7	.92

Table 5.2 Significant Periods Found Using Classical Periodogram

shown in Figures 5-6 and 5-7, confirm that the 8 day cycle is indeed the strongest cycle with period less than 14 days. In addition, the influence of the cycle is not constantly present throughout the span of the data. Another interesting fact that is not apparent when looking at the periodogram of the entire data set is the presence of many short lived periodic cycles, especially around days 50 to 140. These cycles are active for only a relatively short length of time and therefore do not appear significant in the periodograms shown in Figures 5-2 and 5-3. For example, the 4, 5, and 11 day cycles are not determined to be significant in the periodogram, however, they appear to be signficant in the evolutionary spectrum for a brief period of time. Later, at approximately day 300, there appears to be a six day cycle. This information is definitely useful, since the manufacturing process can be affected by the presence of these cyclic behaviors.

Further analysis of this process can be conducted using the residuals that remain after extracting the periodic behavior listed in Table 5.1. As mentioned previously, the residuals do not appear to be random. Since the data is not evenly sampled, common methods for testing the randomness of the data, like the partial and autocorrelation functions, can not be used. The spectral control chart designed for unevenly sampled data and developed in

Section 4.3, can be used to analyze the residuals. The control chart confirms that periodic behavior is still present in the data. Since no prior information was provided about the process, the variance estimate of .024 was obtained using the first 76 data points from the first principal component. Monte Carlo methods using simulations of unevenly sampled, random normal data with the same variance parameter, .024, were conducted to establish a critical value of 11.2. This critical value provides the upper control limit of the control chart at a .0027 level of significance level. The spectral control chart in Figure 5-8, shows that Fisher's test statistic exceeds the critical value many times. These out of control points seem to indicate the temporary cyclic behavior illustrated by the evolutionary spectrum.

5.3 Discussion of Industrial Example

In general, irregular sampling limits our choices for statistical time series models. For example, the unevenly sampled data prevents us from using popular statistical models like the autoregressive integrated moving average (ARIMA) processes. However, we have shown that a variety of periodogram methods can be applied to unevenly sampled data to identify the true underlying periodic structure of a process.

Although periodic behavior was found for this process which seems to describe the process, we can not account for all of the determinism in the data. This is not too surprising since we have only one year's worth of data for a system that contains a 139 day cycle. However, interesting dynamical information was provided through the analysis. The information regarding the dominant periods associated with the coating process can be used to identify factors which influence the process. For instance the strong 8 day cycle could be related to the employees' schedule of working 4 days on/4 days off. The 3 day cycle could be caused by fluctuations in temperature, humidity or barometric pressure. The low frequency oscillation which dominates the process could be attributed to a factor which varies seasonally or semi-annually, like temperature or humidity or a change in suppliers. In addition, several short term periodic influences are also indicated in the analysis. Investigations into the 4, 5 and 11 day cycles which appear around day 50 and 100 could be conducted. As more data is collected, better estimates of the longer periods can be found. Consequently, a better model could be fit to the data. In particular, new information may explain for the structure remaining in the residuals.

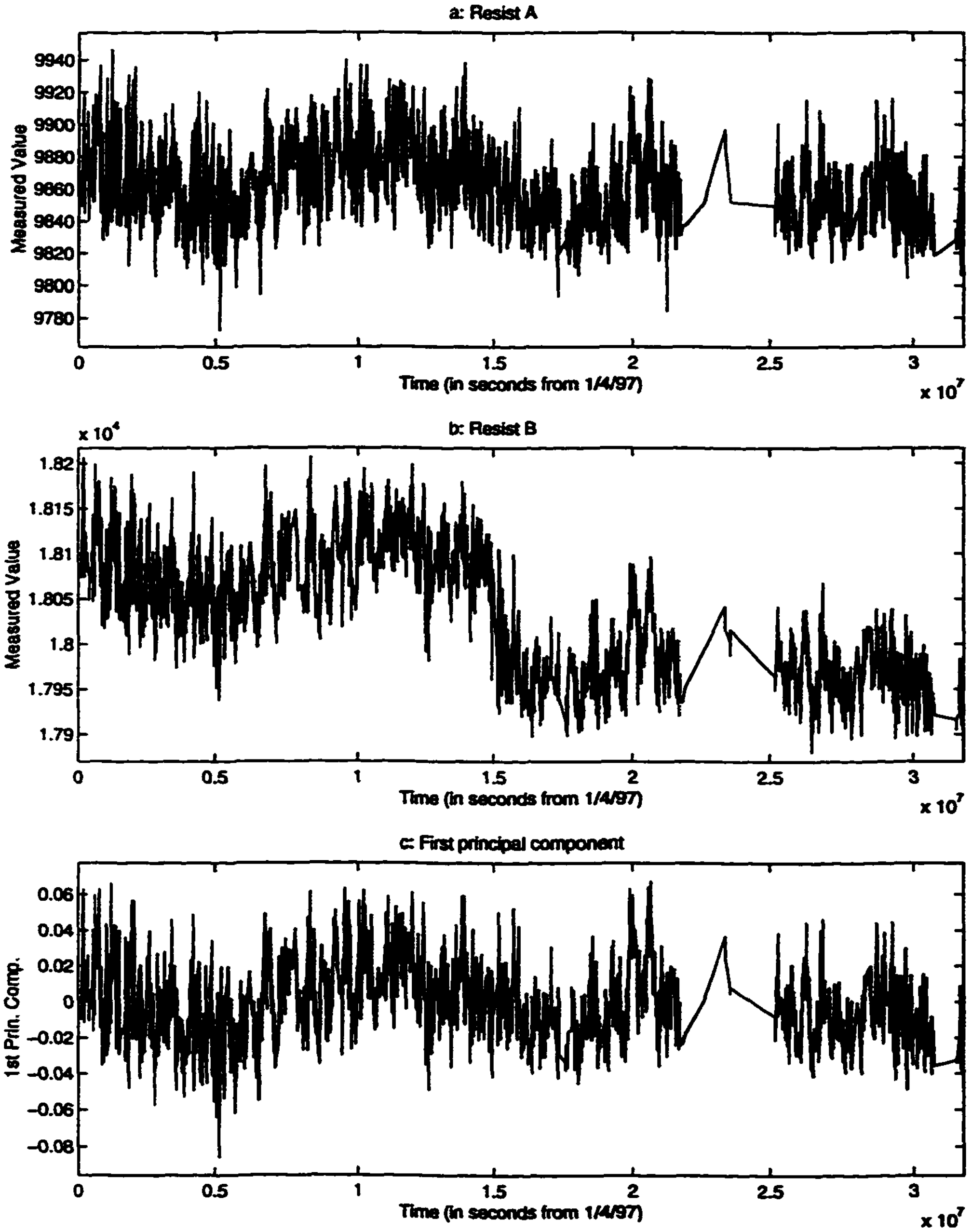


Figure 5-1 Resist data

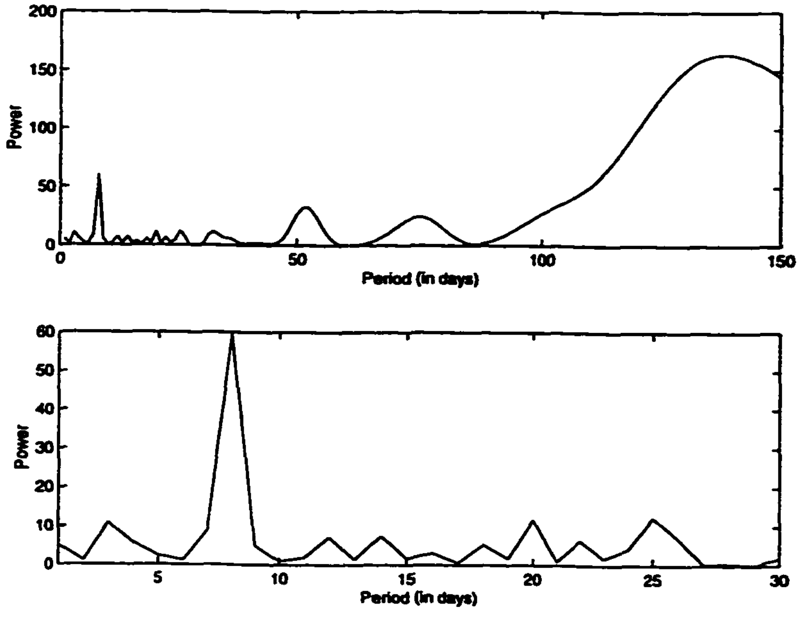


Figure 5-2 Lomb periodogram of first principal component

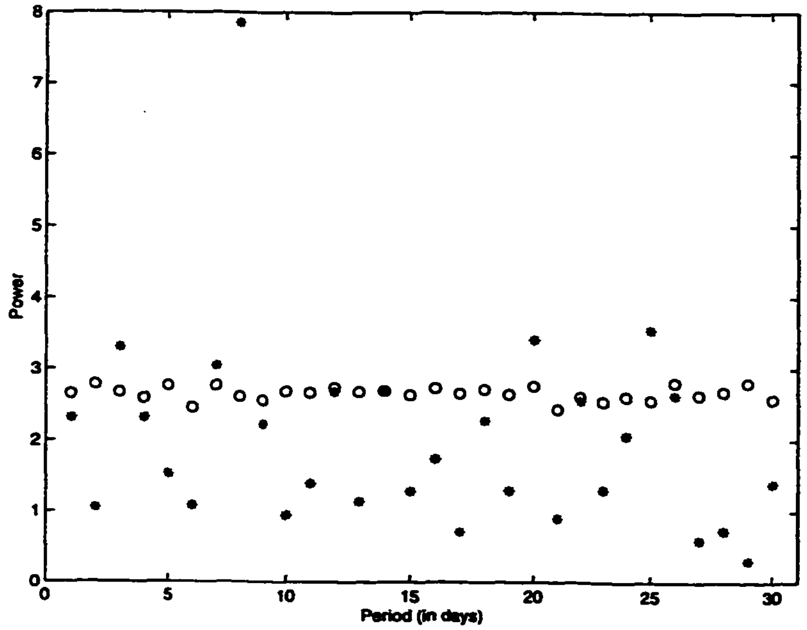


Figure 5-3 Classical periodogram with permutation resampling

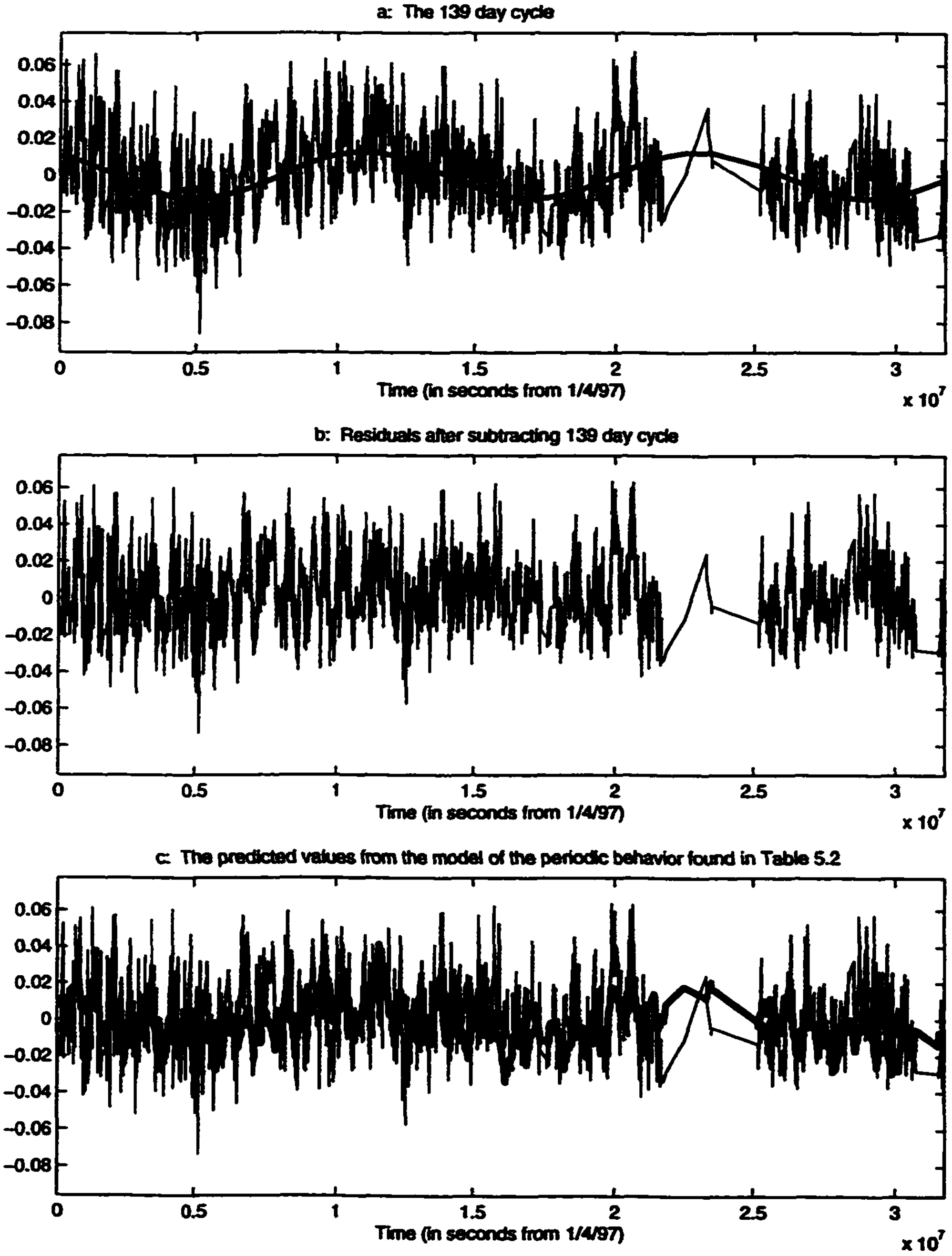


Figure 5-4 Model of periodic behavior

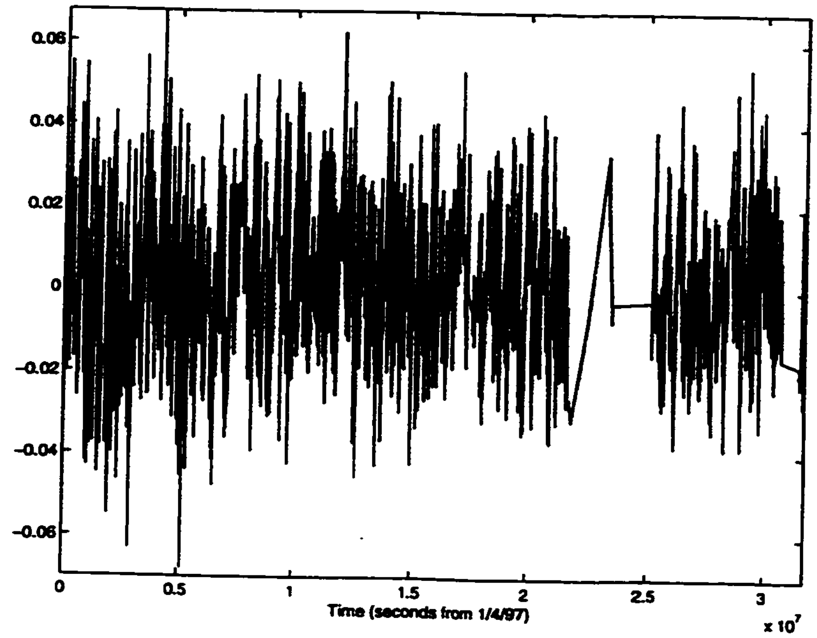


Figure 5-5 Residuals of first principal component after extracting periodic model

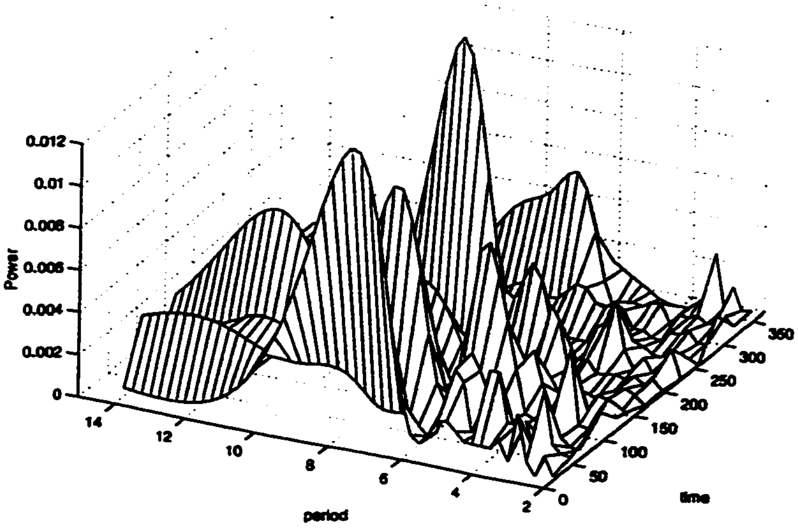


Figure 5-6 Surface plot from the evolutionary spectrum of the first principal component

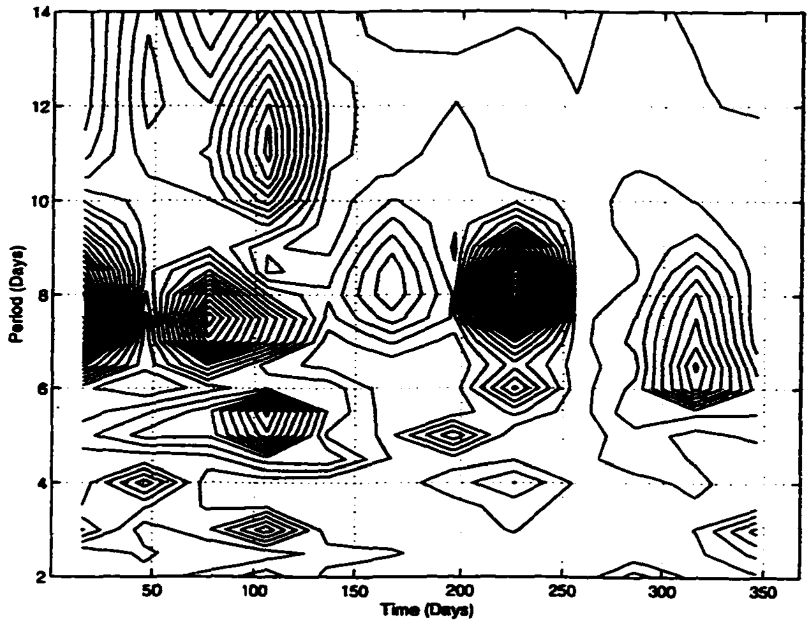


Figure 5-7 Contour plot from the evolutionary spectrum of the first principal component

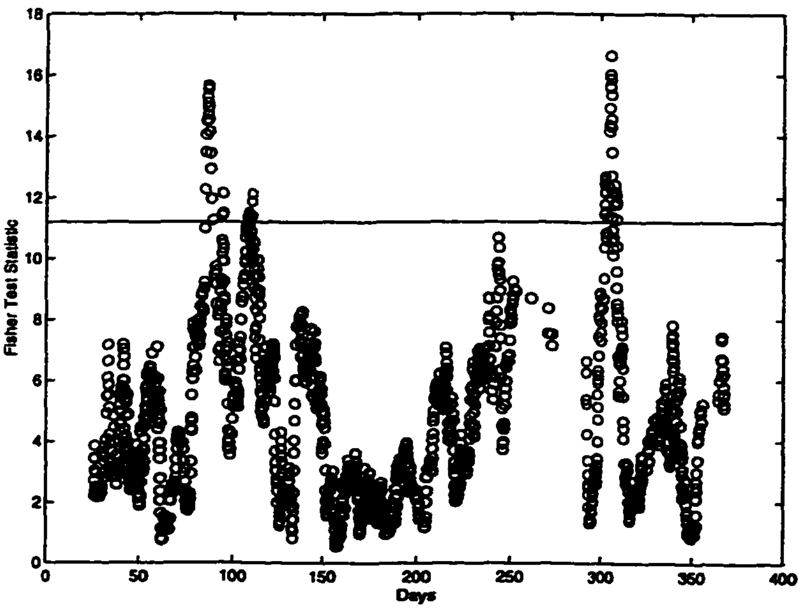


Figure 5-8 Spectral control chart of the residuals of the first principal component

Chapter 6

NONLINEAR DYNAMICS

The methods developed in the previous chapters are ideal for data that exhibit periodic or quasiperiodic behavior. The periodogram for these types of data sets have well defined spectral peaks at the dominant frequencies and their subharmonics. These frequencies can be used to model the process and can be used to obtain the residuals. On the other hand, data that is chaotic in nature tend to have spectra that have broadband behavior. The techniques developed for periodic data, therefore, will not be very helpful in predicting or modeling chaotic data since a continous interval of frequencies can not be extracted.

Stochastic models are sometimes used to model nonlinear behavior, although these techniques can have some disadvantages [35]. For example, the statistical model may not contribute any information regarding the underlying dynamics or physical properties. Also, since chaotic systems are complex, stochastic models will typically require a large number of parameters to model the behavior. Thus, the modeling procedure tends to curve fit the data rather than model the dynamical behavior.

To remedy these problems, nonlinear dynamic (NLD) analysis techniques have been developed to extract the chaotic components, and can be used for short-term forecasting. These methods have been applied to model and predict such phenomena as weather patterns, quasar emissions [9], sea clutter [34, 35], stock returns [62], sunspot activity [13, 74], measle populations [13, 71], fluid turbulence [13] and the detection of teleseismic activity [67]. In addition, there has been a surge of activity in the field of using synchronization of chaotic systems for secure communications [12, 17, 42, 52, 79].

A successful application of NLD analysis hinges upon uncovering a hidden geometry of the underlying deterministic system. The hidden structure, called an attractor, can be found with a technique that reconstructs the dynamical system that produced the observed data. Section 6.1 will discuss common methods of reconstructing the hidden attractor. In addition, a new toroidal reconstruction technique will be discussed. Predictions can be made relative to the attractor. The NLD forecasting method will be introduced in Section 6.2. Comparisons between standard reconstruction techniques and the toroidal technique will be examined using simulated data as well as data collected from a warehouse airduct.

6.1 Reconstructing Underlying Dynamics

Chaotic systems are considered deterministic because there are a set of equations, usually nonlinear differential equations, which capture the behavior of the data. For example, the Rössler attractor [57] is generated using the system of equations

$$\begin{aligned}\dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b + z(x - c)\end{aligned}$$

where a, b and c are control parameters. This simple system of equations only has one nonlinear term, but this term gives rise to various behaviors depending on the values of the control parameters. Table 6.1 lists the type of behaviors which result when parameters a and b are fixed and c is varied. Figure 6-1 shows the phase plot of the chaotic regime

$c = 2$	limit cycle
$c = 3$	period-2 cycle
$c = 4$	period-4 cycle
$c > 4.5$	chaos

Table 6.1 Behavior of Rössler System with $a = b = .2$

which was numerically generated from the differential equations. It illustrates the general shape and flow along the attractor. The stretching and folding demonstrated when the system reinjects itself keeps the attractor in a bounded region and is typical of an attractor. Clearly, predictions are possible using the geometry of the attractor.

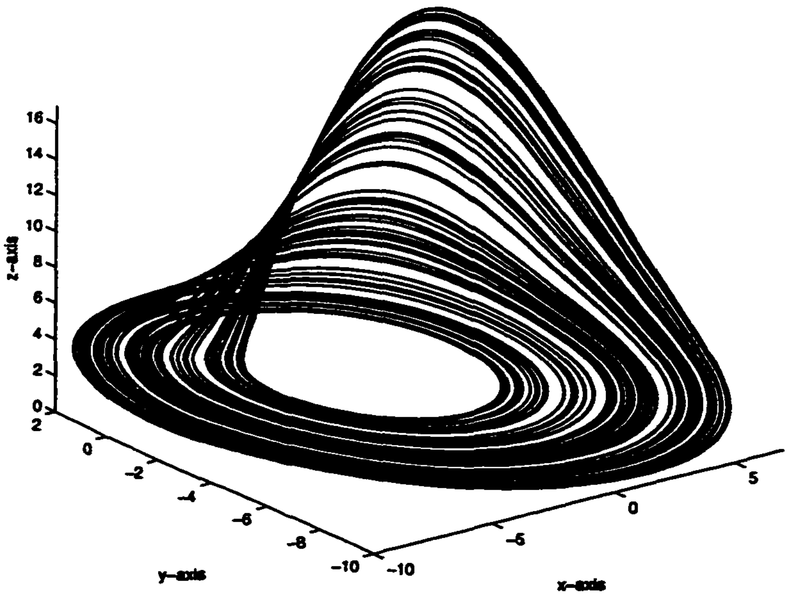


Figure 6-1 Rössler system

6.1.1 Standard Reconstruction Techniques

Unfortunately, in most situations the system of equations which governs the motion is not known. In addition, the observed data is often one-dimensional while the true attractor is multidimensional. The question of how to reconstruct the attractor from a single time series of measurements arises. For example, the trace of the x-coordinate of the Rössler system, shown in Figure 6-2, has nonlinear oscillations which look almost periodic. However, the frequencies and amplitudes are not fixed and this feature is apparent in the associated periodogram, shown in Figure 6-3. There is a strong periodic component shown by the large peak value and smaller peaks at multiples of the dominant frequency. This associated period can be considered to be related to the average period of the system. In addition to this strong cycle, there is a broadband behavior which impedes spectral prediction methods. Similar nonlinear behavior may appear in real world data of oscillations which are induced, for example, by a motor attached to an apparatus. The motor provides a driving force which would prevent the amplitudes of the oscillations from damping.

Since the geometry of the attractor provides a framework to make predictions, it would be beneficial to reconstruct the attractor from the observed data. A theorem by Takens [72] states that it is generically possible to reconstruct the attractor using the one-dimensional time trace. Other dimensions can be generated from the observed data using time delayed series or derivative estimations. Takens showed that the reconstructions based on these coordinates will be a diffeomorphism, or smooth deformation, of the underlying attractor.

Typically, the attractor is estimated using what is called a time delay reconstruction in d dimensions with a suitably chosen time delay τ. The embedding dimension can be calculated

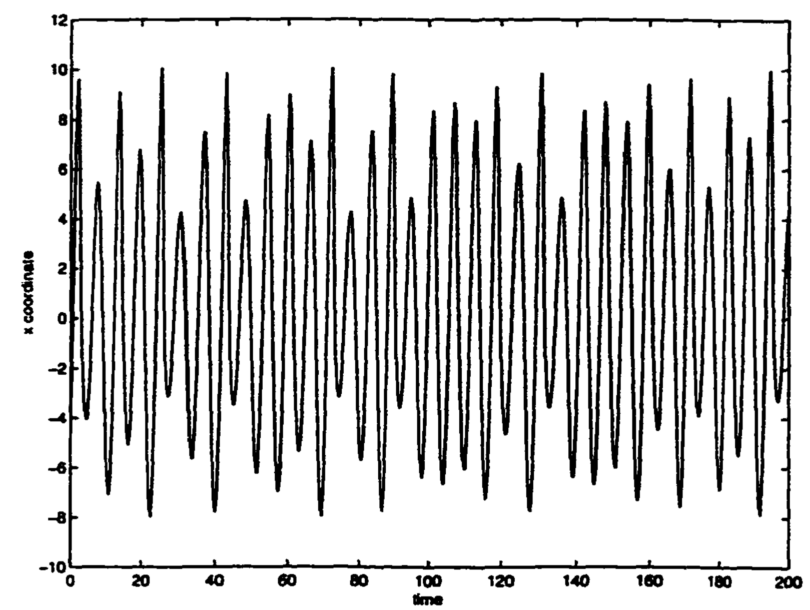


Figure 6-2 Rössler system: x-coordinate

using a variety of methods [1, 5, 9, 11, 15, 34, 39, 41, 48, 49]. Some methods include the box dimension, pointwise dimension [47], information dimension [28, 36], correlation dimension [30, 31] and the Liapunov dimension [24, 40]. Similarly, there are a variety of techniques used to calculate the time delay [1, 9, 10, 48]. The first zero crossing of the autocorrelation function [3, 45, 47] and the first minimum of the mutual information function [1, 23] are commonly used to determine an appropriate time delay. Although the values of the delay may vary slightly between different methods, there are usually several acceptable values or range of acceptable values. The uncovered attractor should have consistent behavior regardless of the methods used to calculate the time delay.

Given an embedding dimension d and time delay τ, the observed data values, $s_0, s_1, s_2, \ldots$

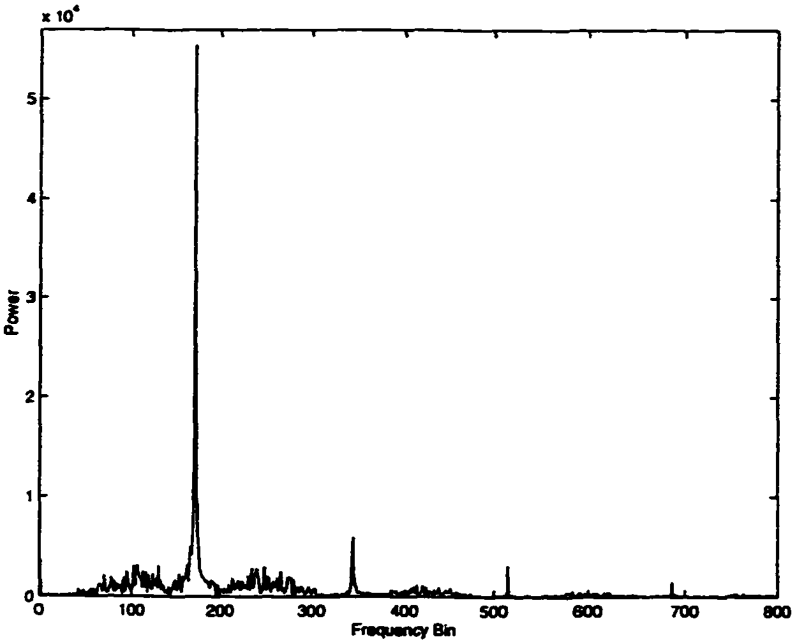


Figure 6-3 Periodogram of Rössler system

are used to reconstruct the higher dimensional vectors

$$\begin{aligned}
\vec{x}_0 &= (s_0, s_{0+\tau}, s_{0+2\tau}, \ldots, s_{0+(d-1)\tau}) \\
\vec{x}_1 &= (s_1, s_{1+\tau}, s_{1+2\tau}, \ldots, s_{1+(d-1)\tau}) \\
&\vdots \\
\vec{x}_n &= (s_n, s_{n+\tau}, s_{n+2\tau}, \ldots, s_{n+(d-1)\tau}).
\end{aligned}$$

Thus, $\vec{x}_0 \longrightarrow \vec{x}_1 \longrightarrow \ldots \longrightarrow \vec{x}_n$ form the time evolving trajectory that carves out the attractor. Suppressed in the subscript notation is the implicit time ordering of the data.

This procedure can be used to convert the one-dimensional, nonlinear oscillations of the

Rössler system into the chaotic attractor in Figure 6-4. For this example, a time delay of 80 was used so each vector has the form $(x(t), x(t+80), x(t+160))$. The reconstruction shows the same general behavior as the phase plot shown in Figure 6-1, although the reconstruction is slightly deformed; it is more peaked and the folded region is warped. However, the regular structure of the reconstructed attractor makes prediction possible.

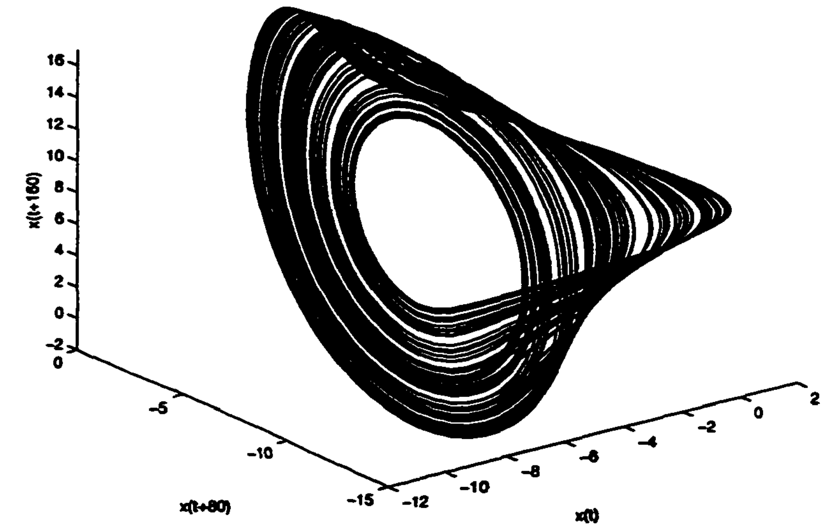


Figure 6-4 Time delay reconstruction of the Rössler system

Although the time delay reconstruction is widely used, other reconstruction techniques have been proposed [1, 27, 48]. These extensions of Takens' method involve the use of derivatives, integrals, linear filters or Fourier interpolants.

6.1.2 Toroidal Reconstruction Technique

While these reconstruction techniques are useful on simulated data, using them on real world data often yields attractors which look like steel wool pads. Although fairly accurate predictions can sometimes be made with a poor reconstruction, the reconstruction does not reveal much about the underlying dynamics. One reason for a poor reconstruction appears to be that components of the underlying dynamics occur on different scales and standard reconstruction techniques do not compensate for this. A fixed time delay value implies that the data may have a fixed scale or fixed periodic behavior. Often, though, frequencies and amplitudes are modulated. These frequency structures often appear when analyzing data from systems that are affected by vibrations or are motor driven. In these situations, the time delay reconstructions have trajectories which cross.

For example, consider the samples illustrated in Figure 6-5 that are drawn from a quasiperiodic process with the phase portrait shown in Figure 6-6. The observations follow the equation

$$X_j = [106 + 2\cos(2\pi t_j/5)]\cos(2\pi t_j/\pi).$$

The time delay reconstruction, shown in Figure 6-7, unfortunately does not reveal the underlying toroidal structure. Instead it appears that the attractor is one-dimensional. Since the process is dominated by the large oscillation, the time delay reconstruction is not capable of separating the trajectories.

To improve upon previous work, a new toroidal reconstruction technique is being developed which will be driven by the dominant frequencies in the data. The dominant frequencies, which could have amplitudes of different scales, are used to create a toroidal framework for the attractor. The actual dynamics can then be studied relative to this framework. The toroidal structure, in some cases, will separate the trajectories to prevent crossings and reveal the underlying dynamical flow. The toroidal reconstruction of the previous example shows how trajectories are separated and the underlying geometry of the attractor is uncovered, as illustrated in Figure 6-8.

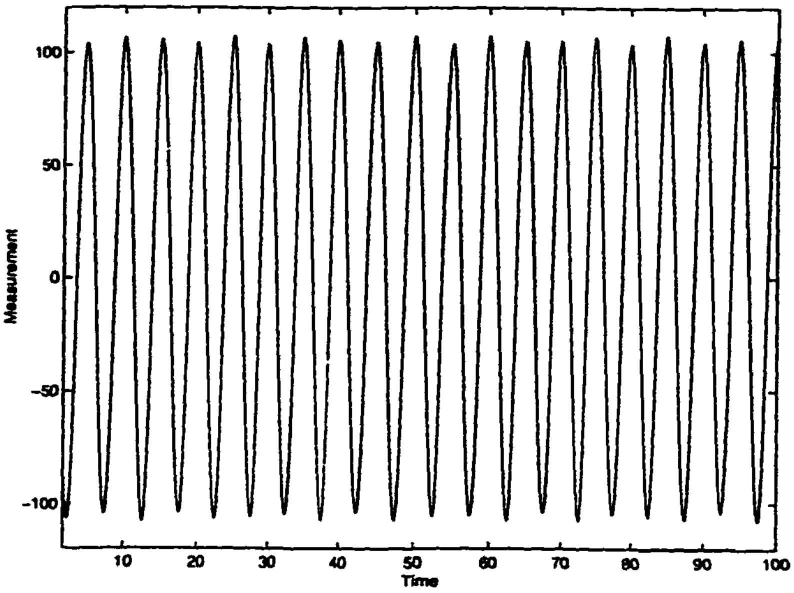


Figure 6-5 Observed Data

At any given time, the position on a torus can be defined by the coordinates $(x(t), y(t), z(t))$ with the values

$$\begin{aligned} x(t) &= [\beta + \alpha \cos(2\pi f_2 t)] \cos(2\pi f_1 t) \\ y(t) &= [\beta + \alpha \cos(2\pi f_2 t)] \sin(2\pi f_1 t) \\ z(t) &= \alpha \sin(2\pi f_2 t). \end{aligned}$$

The torus can be defined using the parameters α, β, f_1 and f_2 where β and α represent the small and large radii of the torus and f_1 and f_2 represent the wrapping frequencies associated

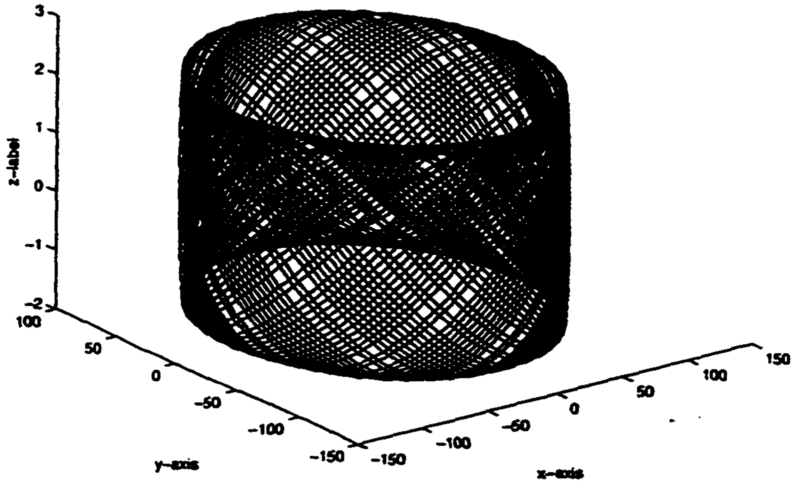


Figure 6-6 True Underlying Attractor

with those two radii. When these parameters are defined it is possible to represent any point in time in terms of these four parameters. An example of a torus with parameters $\alpha = 9, \beta = 31, f_1 = 1/(6\pi)$ and $f_2 = 1/2$ is shown in Figure 6-9.

Unfortunately, in most applications, only one-dimensional data is collected and information regarding the values of the four parameters is unknown. In order to construct a torus from the observed data, estimates of the parameters need to be found. Let us assume that the observed data corresponds to the x-projection of the torus. Let the observed data be represented by $\{(t_n, x_n)|\ n = 1, 2, \ldots, N\}$. Using the assumption and trigonometric identities the discretized expression for the x-coordinate of a torus at time t_n can be written as follows

$$x_n = [\beta + \alpha \cos(2\pi f_2 t_n)] \cos(2\pi f_1 t_n)$$

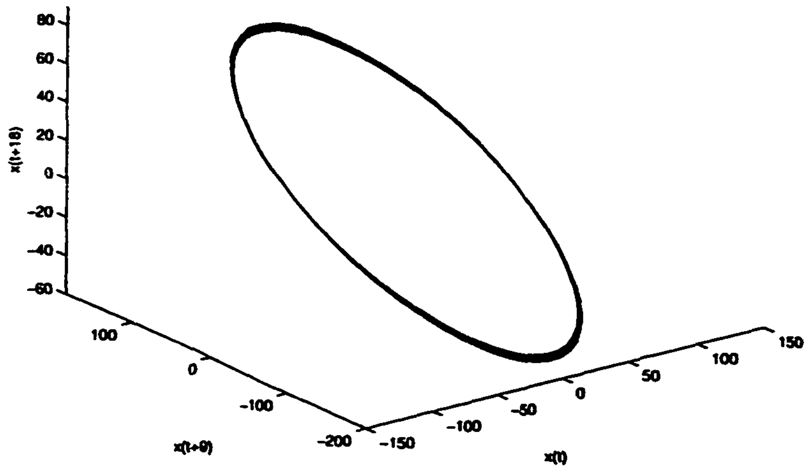


Figure 6-7 Time delay reconstruction of toroidal data

$$
\begin{aligned}
&= \beta\cos(2\pi f_1 t_n) + \alpha\cos(2\pi f_2 t_n)\cos(2\pi f_1 t_n) \\
&= \beta\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(2\pi(f_2 - f_1)t_n) + \frac{\alpha}{2}\cos(2\pi(f_2 + f_1)t_n) \\
&= \beta\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(2\pi f_L t_n) + \frac{\alpha}{2}\cos(2\pi f_U t_n)
\end{aligned}
$$

where $f_L = f_2 - f_1$ and $f_U = f_2 + f_1$. Since the cosine is an even function it is possible that $f_L = f_1 - f_2$. In all cases, $f_U \geq f_L$.

The estimate of the frequencies f_1, f_L and f_U can be obtained directly from the periodogram. Since β is assumed to be the larger radius, f_1 will correspond to the frequency with the dominant peak in the periodogram. The spectral peaks associated with f_U and f_L should have the same power, since the amplitudes of those cosine terms are equal. Depending on the relationship between f_1 and f_2, the behavior and location of f_U and f_L changes.

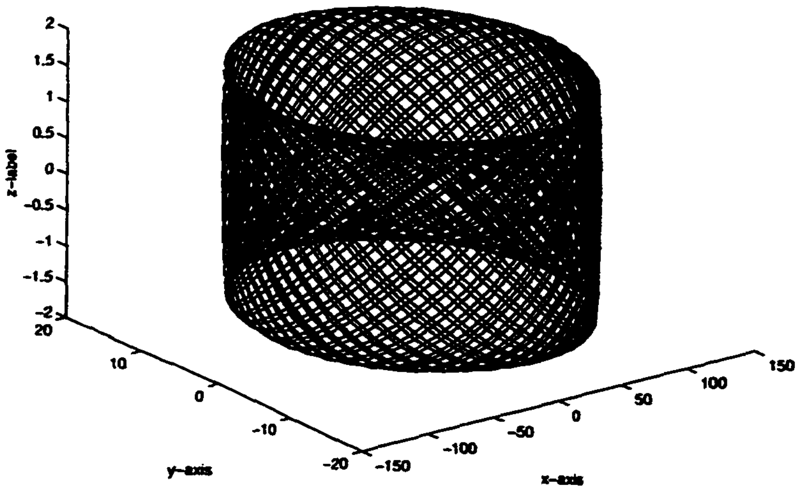


Figure 6-8 Toroidal reconstruction of toroidal data

Figure 6-10 shows the three distinct behaviors of the spectra of tori. When $f_1 > f_2$, the frequency f_1 will be located halfway between f_U and f_L, as shown in Figure 6-10a. The estimate of f_2 can be obtained using the relationship $f_2 = f_U - f_1 = f_1 - f_L$. If $f_2 > f_1$ then both f_U and f_L can be greater than f_1, as in Figure 6-10b. In this case, $f_2 = f_U + f_1 = f_1 - f_L$. Or, f_L can be less than f_1 and $f_2 = f_U - f_1 = f_L + f_1$. Notice, for all cases $f_2 = f_U - f_1$. In these situations, the periodogram has three distinct peaks.

However, whenever $f_2 = 2f_1$, a degenerate torus occurs and only two spectral peaks appear since in this case,

$$\begin{aligned} x_n &= \beta\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(\pm 2\pi(f_2 - f_1)t_n) + \frac{\alpha}{2}\cos(2\pi(f_2 + f_1)t_n) \\ &= \beta\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(\pm 2\pi(2f_1 - f_1)t_n) + \frac{\alpha}{2}\cos(2\pi(2f_1 + f_1)t_n) \\ &= \beta\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(\pm 2\pi(-f_1)t_n) + \frac{\alpha}{2}\cos(2\pi(3f_1)t_n) \end{aligned}$$

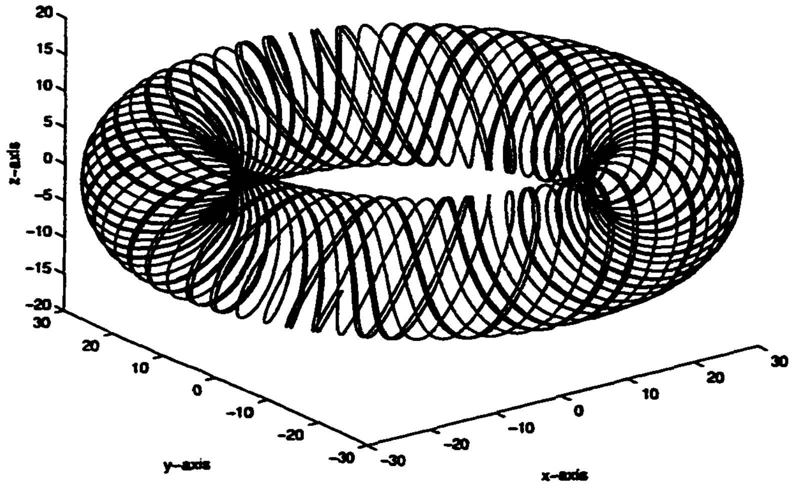


Figure 6-9 Example of a torus

$$= \left(\beta + \frac{\alpha}{2}\right)\cos(2\pi f_1 t_n) + \frac{\alpha}{2}\cos(2\pi(3f_1)t_n).$$

Therefore, the peaks of the periodogram will identify f_1 and $3f_1$, as demonstrated in Figure 6-10c.

When the frequencies have been determined, the estimates of the amplitudes β and $\alpha/2$ can be obtained using the extraction method described in section 5.1.2 with the frequencies f_1, f_U and f_L. The other dimension of the torus can then be generated using the estimated parameters and the observed data. The procedure will make use of the property that the x_n and y_n terms only differ by a trigonometric function involving f_1. The term $\cos(2\pi f_1 t_n)$

can be estimated from the observed data using the fact

$$\cos(2\pi f_1 t_n) \approx \widehat{\cos}(2\pi f_1 t_n) \equiv \frac{x_n}{\hat{\beta} + \hat{\alpha}\cos(2\pi \hat{f}_2 t_n)}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are the estimates of the amplitudes and $\hat{f}_2$ is the frequency from the periodogram. This approximation is well-defined since it is assumed that $\beta > \alpha$. Integration can be used to estimate $\widehat{\sin}(2\pi f_1 t_n)$. In particular, for any time t_n for $n = 0, 1, \ldots, N$ with $t_0 = 0$

$$\begin{aligned} \int_{t_0}^{t_n} \cos(2\pi f_1 t)dt &= \frac{1}{2\pi f_1}\sin(2\pi f_1 t)\Bigg|_{t_0}^{t_n} \\ &= \frac{1}{2\pi f_1}[\sin(2\pi f_1 t_n) - \sin(2\pi f_1 t_0)] \\ &= \frac{1}{2\pi f_1}\sin(2\pi f_1 t_n). \end{aligned}$$

So,

$$\widehat{\sin}(2\pi f_1 t_n) \approx 2\pi \hat{f}_1 \int_{t_0}^{t_n} \widehat{\cos}(2\pi f_1 t)dt.$$

where the integral of $\widehat{\cos}(2\pi f_1 t_n)$ can be calculated numerically using the trapezoidal method. Hence,

$$\hat{y}_n = \left[\hat{\beta} + \hat{\alpha}\cos(2\pi \hat{f}_2 t_n)\right] \widehat{\sin}(2\pi \hat{f}_1 t_n)$$

and

$$\hat{z}_n = \hat{\alpha}\sin(2\pi \hat{f}_2 t_n).$$

Since numerical errors are compounded during the integration estimation, a trend often appears in $\widehat{\sin}(2\pi \hat{f}_1 t_n)$. A filter using the running mean can be used to remove the trend prior to estimating $\hat{y}_n$.

Although this method supposes that the structure of the data evolves about a torus, the reconstruction technique can be applied to data for which this condition is not true. If the peaks in the periodogram reveal a pattern like the examples in Figure 6-10 and the associated frequencies can be selected to satisfy the relationships between f_1, f_2, f_U and f_L, then the toroidal method can be applied. For example, consider the one-dimensional trace from the Rössler system. The periodogram reveals a dominant frequency and subharmonics are also present. By considering the degenerate torus with $f_2 = 2f_1$ and applying the method outlined above, the underlying attractor can be reconstructed. For this particular sample, $\hat{f}_1 = 0.17099145042748$, $\hat{\alpha} = .32$ and $\hat{\beta} = 5.55$. Figure 6-11 shows the x-y projection of the toroidal reconstruction which exhibits an attractor which is just a smooth deformation of the original attractor.

6.2 Method of Prediction

Once a reconstruction from the data is obtained, predictions can be made relative to the attractor. Due to the sensitive dependence on initial conditions, long term predictions are usually not feasible. Thus, a global model is not realistic. However, the consistent flow directions in a local region can be exploited for short term predictions [1, 13, 19, 53, 60, 71].

The standard local modeling technique will be used to make a one-step prediction of a point $\vec{x}_\phi$. A set of p neighbors, $\{\vec{x}_i\}$, located around the point $\vec{x}_\phi$ are selected, usually based on euclidean distance. It is important to note that neighbors which are spatially close on the attractor may be temporally distant. It is desirable to estimate a local predictor function $\vec{F}$ such that $\vec{F}(\vec{x}_i) = \vec{x}_{i+1}$ for each of the p neighbors of $\vec{x}_\phi$. Here $\vec{x}_{i+1}$ represents the data value collected immediately after $\vec{x}_i$. A set of regressors, or basis elements $\{g_k(x)\}$, are chosen for the expansion of $\vec{F}(\vec{x}) = (f_0(\vec{x}), f_1(\vec{x}), \ldots, f_{d-1}(\vec{x}))$. Since the dynamics are not usually as complex in a localized region of the attractor, it is usually sufficient to let the expansion basis be the set of polynomials up to the second degree. For example, Figure 6-12 illustrates that the local behavior around the point $\vec{x}_\phi$ appears to be quadratic. The function space can be represented by

$$f_j(\vec{x}_i) = \sum_{k=0}^{n} a_{jk} g_k(\vec{x}_i) \quad \text{for} \quad j = 0, \ldots, d-1$$

for all the p neighbors $\{\vec{x}_i\}$. The number of summands, n, is determined by the order of the expansion basis. The coefficients a_{jk} can be estimated using least squares to satisfy the condition that $\vec{F}(\vec{x}_i) = \vec{x}_{i+1}$ for each of the neighbors. In doing this, the d components of the predictor function are completely determined and can be used to predict

$$\vec{F}(\vec{x}_\phi) = (f_0(\vec{x}_\phi), f_1(\vec{x}_\phi), \ldots, f_{d-1}(\vec{x}_\phi)).$$

Interestingly, the method of time delay reconstruction and local forecasting is related to autoregressive time series modeling. In that application, the autoregressive model of order p corresponds to a time delay reconstruction with the reconstruction vector $\vec{x}_i = (x_i, x_{i-1}, \ldots, x_{i-p+1})$. In order to find the estimates of the linear model, a neighborhood consisting of all points is used. Therefore, a global model is found by using least squares to find the coefficients of

$$\vec{x}_{i+1} = F(\vec{x}_i) = a_0 + a_1 x_i + a_2 x_{i-1} + \ldots + a_p x_{i-p+1} \quad \text{for all } i.$$

So, the NLD predictors can be considered localized autoregressive models. For obvious reasons, the standard, global autoregressive models have not been successful at modeling the complex behavior of chaotic data.

While the basic forecasting method using local properties seems powerful, the models are susceptible to errors caused by self-intersections of the attractor. Time delay reconstructions, especially on real-world, noisy data, often intersect. For instance, in the Rössler reconstruction, there is the potential to choose neighbors which have very distinct trajectory flows. This could occur when the point that is being predicted falls near the reinjection part of the attractor. Improvements to the basic local forecasting method have been proposed [66, 67]. Choosing neighbors that are close in a euclidean sense and that have similar slopes usually prevents incorrect neighbors being selected. It should be noted that this method will not work when tangential intersections occur.

Another problem with the local forecasting technique is that the results can be affected by noise. The application of noise reduction has been crucial to make the local prediction methods more robust [18, 25, 29, 32, 43, 50]. Another modification to the general noise reduction schemes, based on singular value decompostion is the use of a change of coordinates [66, 67]. After the selection of the neighbors, a new coordinate axis is created to align with the local flow directions. The set of neighbors are transformed via a rotation matrix **V** found using singular value decomposition on the local flow matrix where the ijth entry of the local flow matrix is defined to be the jth component of $(\vec{x}_{i+1} - \vec{x}_i)$. The forecasting method is carried out using the transformed coordinates of the neighbors to predict $\vec{y}_{\phi+1}$, where y_ϕ represents the transformed value of x_ϕ. This predicted value can then be transformed back into the original coordinates via the equation

$$\vec{x}_{\phi+1} = \vec{y}_{\phi+1}\mathbf{V}^T + \vec{x}_\phi.$$

The singular value decomposition can also be used to detect the principal flow direction of the attractor in the local coordinate system. Small singular values indicate a small contribution to the overall flow direction and might be attributed to extraneous noise or a hidden signal. Singular values which fall below a predetermined cutoff value are zeroed out and the the space of the embedding is collapsed to a lower dimensional subspace. The reduction of the embedding dimension helps to make the prediction algorithm more stable and forces the predictions to be more consistent with the flow of the underlying attractor.

The one-step prediction method used in the next section will include the enhancements to the local modeling algorithm. The forecasting techniques will be used to compare the reconstructions mentioned in the previous section. These methods will be applied to vibrational data collected by an accelerometer placed on large warehouse air handler unit.

6.3 Air Handler Analysis

The data to be considered in this section may represent an example of real-world chaos. The observations were taken from a vibrating air handler unit located in a large, open spaced warehouse. The unit, responsible for the temperature regulation of the warehouse room, is attached to large metal ducts. The measuring device was an accelerometer placed directly on the metal housing and it quantified the vertical movement, or vibration, of the handler. A DAT recorder collected 48000 samples per second and these measurements are shown in Figure 6-13.

The periodogram of the data reveals some interesting properties, as demonstrated in Figure 6-14. The first characteristic illustrated is that the data seems to be dominated by certain frequencies and their harmonics; however, there is also broadband structure in the spectrum. The dominant frequencies may be related to the drives of the motors or the properties of the air handler materials. The data also appears to have oscillations occuring on two scales with periods of approximately 200 and 2500 samples, but the power of the larger cycle is irregular. For example, the first plot of Figure 6-15 demonstates a region where high frequency dominates the system and the low frequency contribution has disappeared. Contrast this to the second plot where both behaviors are present. The evolutionary spectrum of the data also confirms this behavior, as shown in Figure 6-16. Notice that the 200 period cycle is consistent throughout the data, while there are many regions where the period of 2500 samples has little or no power. These aperiodic and nonstationary properties indicate that this data set may be an example of an real world chaotic process.

Another interesting property of the power spectrum is the presence of a toroidal frequency structure. This makes the data set a great candidate on which to evaluate the new toroidal reconstruction method. An initial investigation focused on a bandpassed filtered version of the data set, shown in Figure 6-17. The power spectrum illustrated in Figure 6-18 shows what frequencies remained in this analysis. Because the high frequency structures were removed, the data had little noise in it. However, there is some amplitude modulation of the central frequency. A three-dimensional time delay reconstruction of the air handler data, shown in Figure 6-19, was formed using a delay of 18. Even though the reconstruction is in three dimensions, the attractor looks two-dimensional.

The dominant frequencies shown in the periodogram correspond to the frequencies $f_1 = .005, f_U = .0054$ and $f_L = .0046$. With these estimates suggested by the data, $f_2 = .0005$. Following the methods outlined in section 6.1.2, a torus based on the 200 and 2500 sample cycle was formed. The reconstruction, in Figure 6-20, shows how the data wraps around the torus frame. Interestingly, it appears that the fit about torus is much tighter near the bottom than the top. This structure may provide some information about when amplitude and frequency modulations occur. It appears that the amplitude has a periodic behavior. This behavior is not obvious in the time delay reconstruction.

Both reconstructions were scaled down to range approximately between ±2.5. One-step predictions were made on the reconstructions using nonlinear dynamic forecasting methods. One method of evaluating and comparing the competing methods is to see how the residual variance compares to the variance of the original system. The filtered data, after scaling, had a standard deviation of approximately .6. The residuals of the time delay reconstruction are shown in Figure 6-21 along with the power spectrum of the residuals. The residuals have a standard deviation of 3.4342e-04. However, the power spectrum of the residuals still appear to have the toroidal frequency structure remaining. Contrast this to the residuals and power spectrum from the toroidal reconstruction method, shown in Figure 6-22. The standard deviation is reduced by 5 orders of magnitude to 6.1994e-06. In addition, the power spectrum seems to have noise-like behavior. These features indicate that the toroidal reconstruction method seems to model the underlying dynamics of the filtered data much better than the time delay reconstruction.

Next, the original air handler data set was considered. Again, the two reconstruction methods were used to model the data. The time delay reconstruction was based on a delay parameter of 16 while the toroidal reconstruction was based on the same frequency structure as the filtered data. The original data had a standard deviation of .656. The standard deviation of the residuals from the time delay and toroidal reconstructions were .011 and .0076, respectively. There was no overwhelming evidence to support that the toroidal method did better than the time delay reconstruction.

The new toroidal reconstruction method modeled the bandpassed air handler data much better than the standard time delay reconstruction. The toroidal method reveals some of the hidden structure, like the amplitude modulation, that may have gone unnoticed using the other method. The residuals of the time delay reconstruction still appear to have the same structure as the original data indicating that the residuals have a lot of structure remaining. Meanwhile, the toroidal reconstruction seems to have modeled that determinism.

The influence of other harmonic structures outside of the bandpassed region clearly affects the toroidal reconstruction of the original data. These other oscillations result in a reconstruction where the torus no longer has a hole. Therefore tangential intersections may be present and these types of intersections hinder the ability to make good predictions. One possible way to minimize the influences of other harmonic structures is to generalize the toroidal reconstruction method into higher dimensions. By incorporating these other dominant frequencies into the reconstruction the tangential intersections should separate.

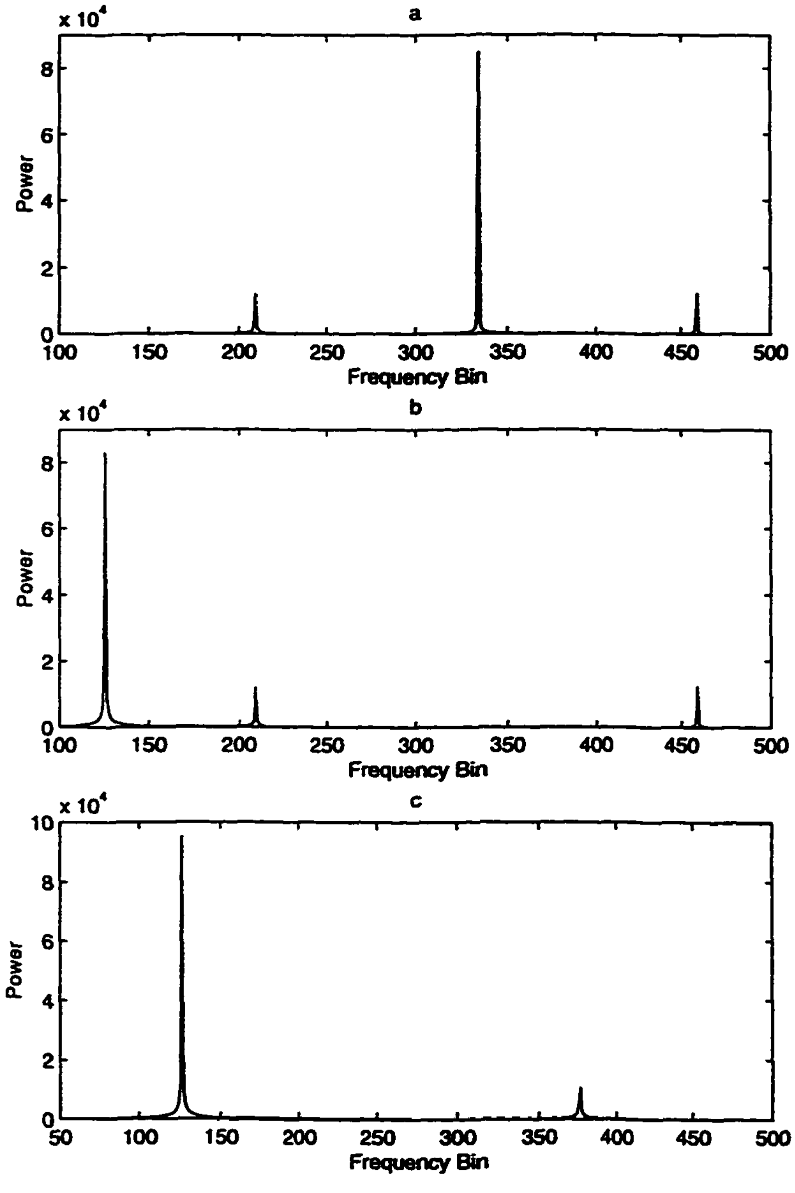


Figure 6-10 Frequency structure of a torus

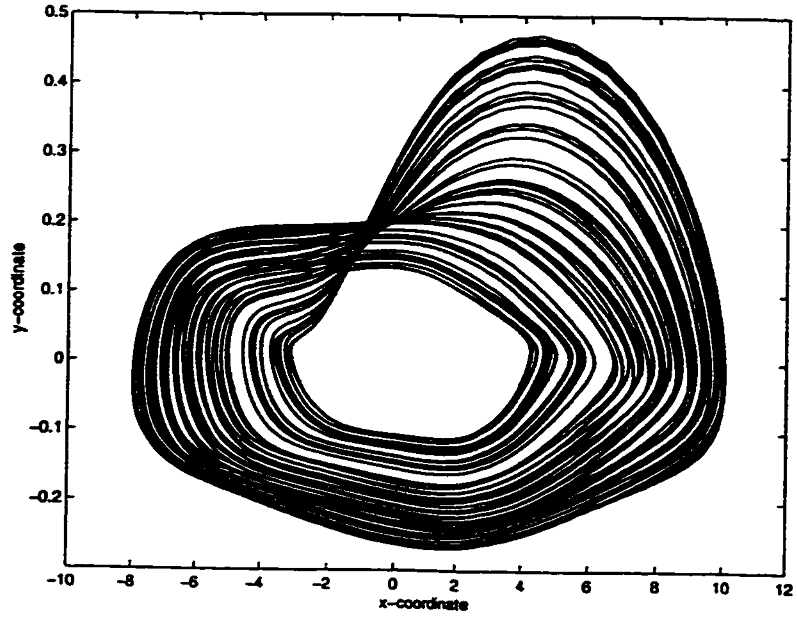


Figure 6-11 Toroidal reconstruction of the Rössler system

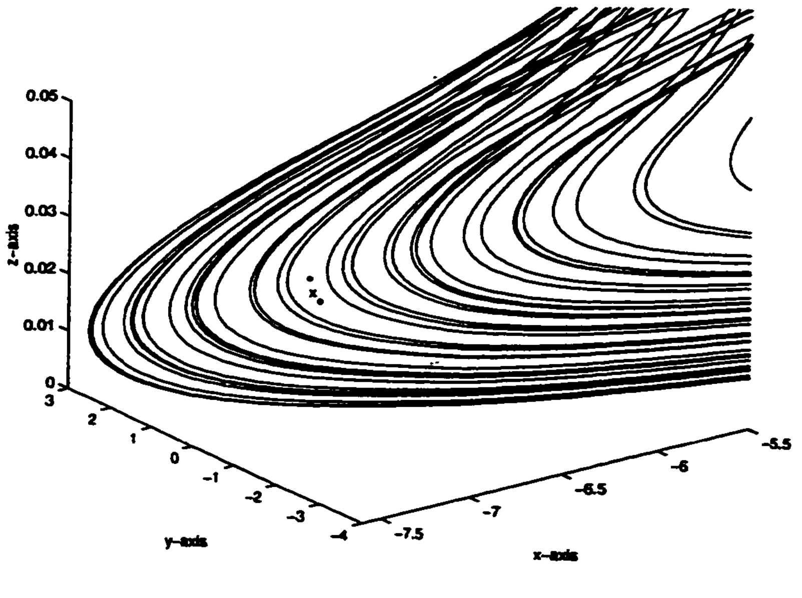


Figure 6-12 Local prediction method

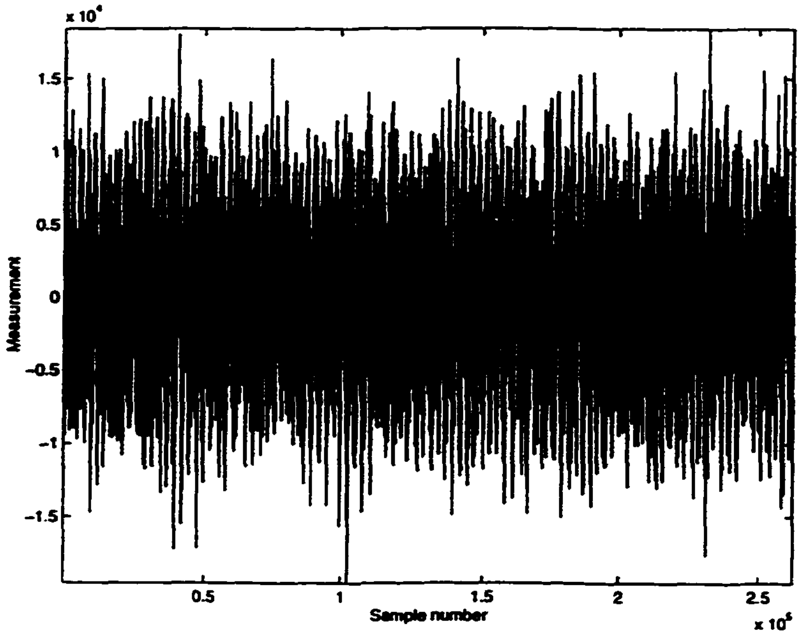


Figure 6-13 Air handler data

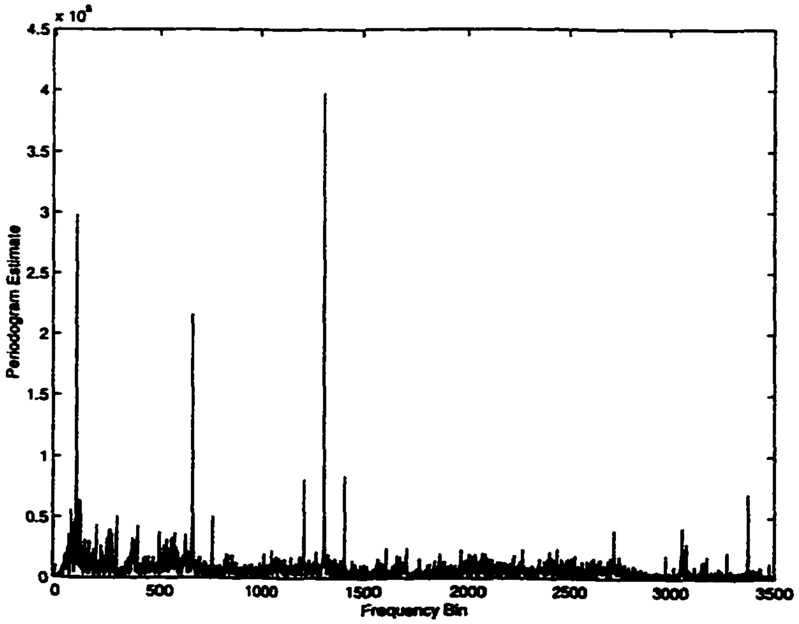


Figure 6-14 Periodogram of air handler data

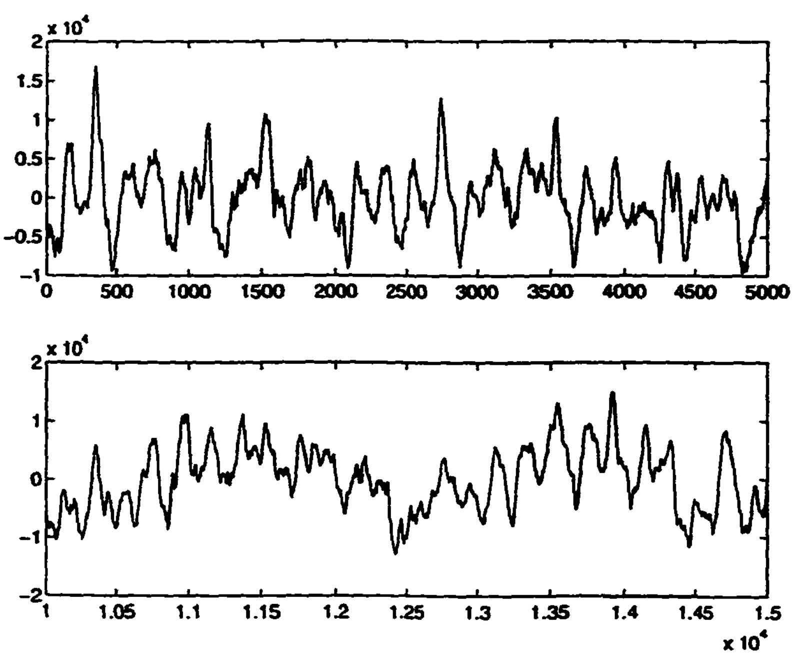


Figure 6-15 Frequency structures in the air handler data

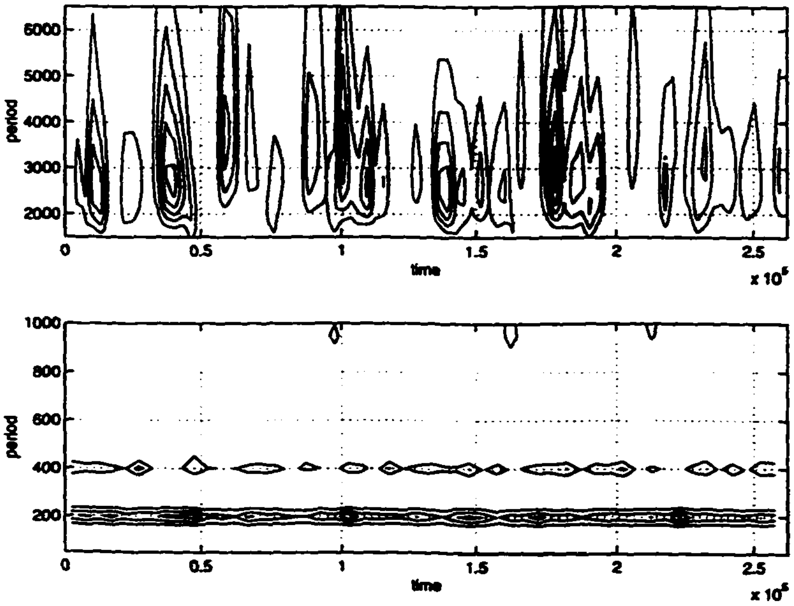


Figure 6-16 Evolutionary spectrum of air handler data

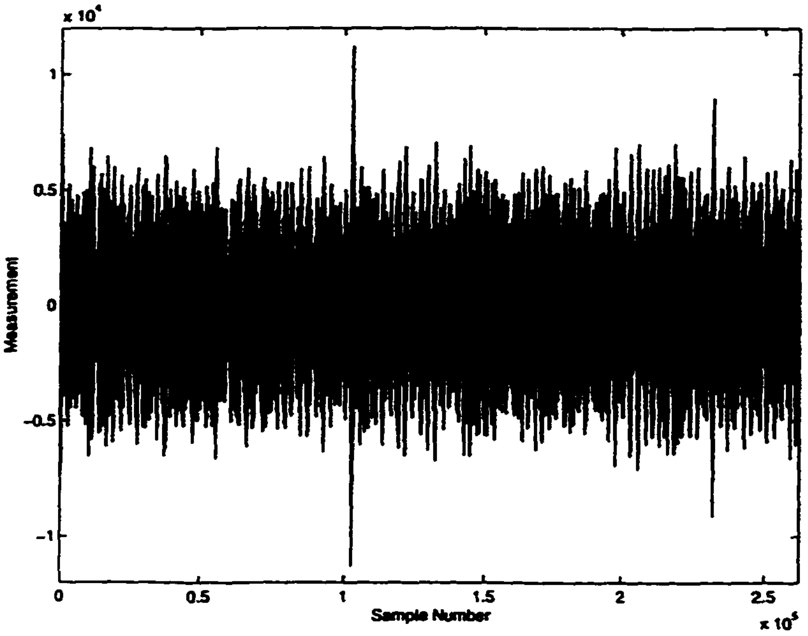


Figure 6-17 Filtered air handler data

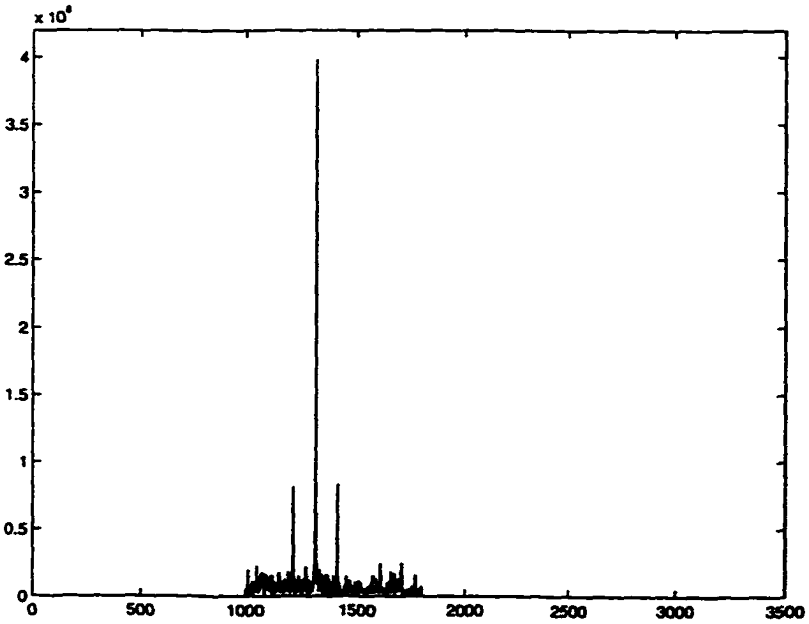


Figure 6-18 Bandpassed frequency structures

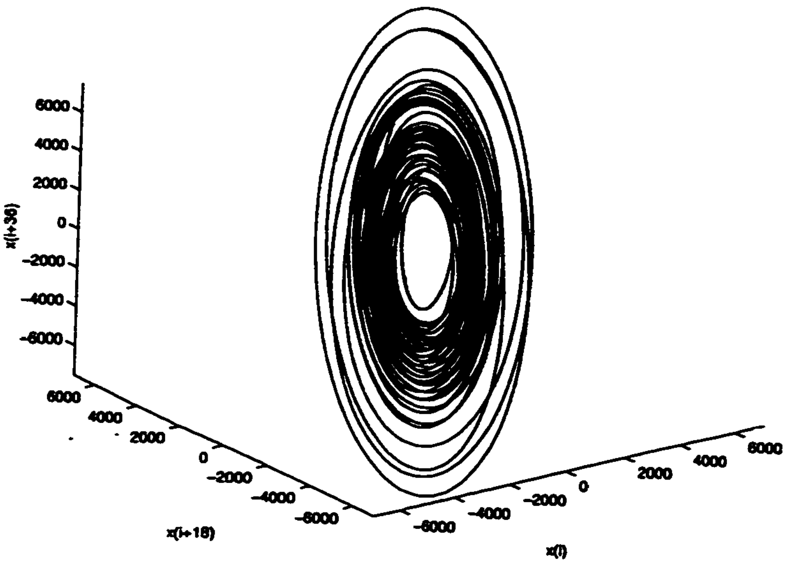


Figure 6-19 Time delay reconstruction of bandpassed data

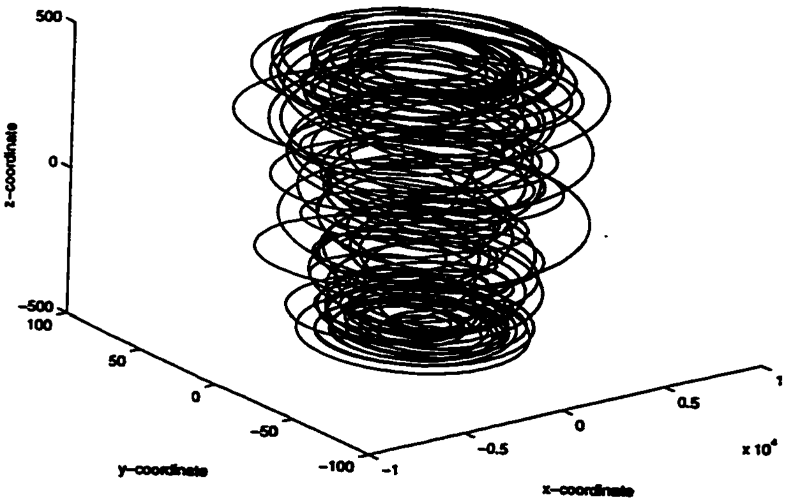


Figure 6-20 Toroidal reconstruction of bandpassed data

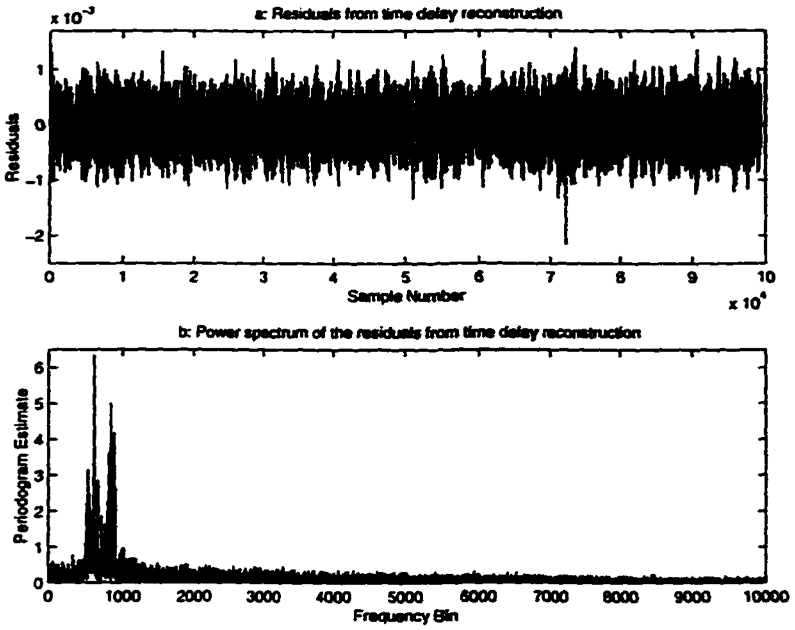


Figure 6-21 Residuals from the time delay reconstruction of the bandpassed data

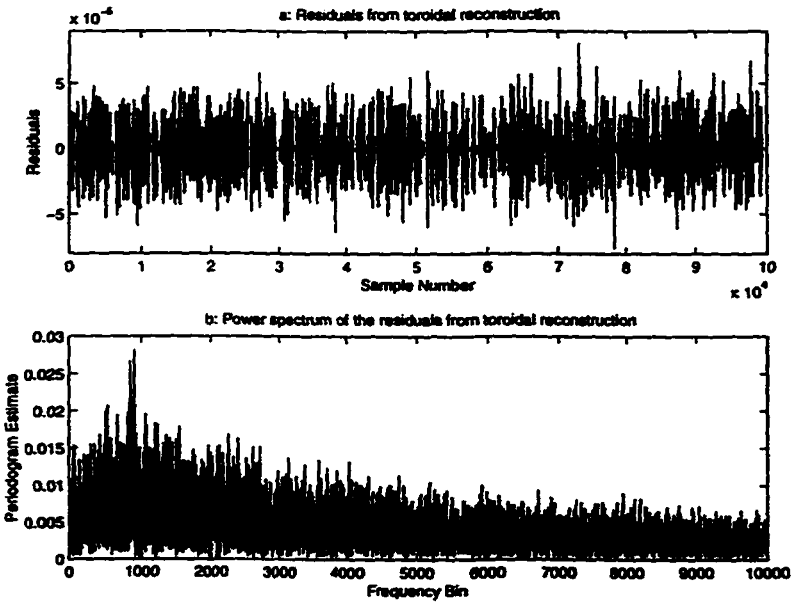


Figure 6-22 Residuals from the toroidal reconstruction of the bandpassed data

Chapter 7

FUTURE DIRECTIONS

The ultimate goal of this work is to model the underlying determinism in a process using a combination of spectral, nonlinear and statistical methods. By learning how the process changes over time, factors which influence production might be identified. New information about the process can then be incorporated to improve the process. Also, by combining dynamical modeling with statistical process control it should be possible to predict and detect process changes. Chatterjee and Yilmaz comment [15, page 50] that knowledge of chaotic systems can "... enrich the arsenal of modeling tools available to statisticians, generate new developments in their refinement and, ultimately, facilitate a better understanding of the processes being studied."

Continued work in this direction could include relaxing the assumption that the data is normally distributed. Oftentimes this assumption may not be reasonable. The development of methods to detect periodic behavior in non-normal data would be useful for these situations. There is also potential to extend the ideas of detecting periodic behavior, not only over time, but over space. Such problems are faced by researchers collecting data from satellites and the data is typically unevenly sampled. Again, the use of standard approaches may not appropriately apply to unevenly sampled temporal-spatial data.

Other extensions to statistical methods that may benefit from the exploitation of the underlying determinism require the study of the system to be carried out for a sufficient length of time so that the full range of dynamical behavior is captured. In that situation it is possible to consider the applicability of nonlinear dynamic forecasting in conjunction with Kalman filtering or state space modeling. Another possible direction which would combine the two areas is to extend transfer models to allow for nonlinear dynamic analysis to replace the predictor as the pre-whitener. Also, since various reconstruction techniques are available, there is a need to explore how to compare, in a statistical sense, competing reconstruction methods.

Lastly, improvements of the toroidal reconstruction are currently being considered. Instead of having a single toroidal model with fixed amplitudes and frequencies, methods are being developed to allow these parameters to change according to the amplitude and frequency modulation of the data. Therefore, this toroidal framework will be allowed to evolve according to the underlying dynamics of the data. It may also be possible to generalize the toroidal reconstruction technique into higher dimensions to create a framework for aperiodic data that is dominated by multiple oscillations. This may improve the results of the air handler analysis, since are there are multiple toroidal features shown in the air handler frequency spectrum.

Bibliography

[1] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring. The analysis of observed chaotic data in physical systems. *Reviews of Modern Physics*, 65(4):1331–1392, October 1993.

[2] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* National Bureau of Standards, 1965.

[3] A. M. Albano, J. Muench, C. Schwartz, A. I. Mees, and P. E. Rapp. Singular value decomposition and the Grassberger-Procaccia algorithm. *Physical Review A*, 38:3017—-3026, 1988.

[4] M. Beneke, L. M. Leemis, R. Schlegel, and B. L. Foote. Spectral analysis in quality control: A control chart based on the periodogram. *Technometrics*, 30(1):63–70, February 1988.

[5] L. M. Berliner. Statistics, probability and chaos. *Statistical Science*, 7(1):69–122, 1992.

[6] E. Bølviken. The distribution of certain rational functions of order statistics from exponential distributions. *Scandinavian Journal of Statistics*, 10(2):117–123, 1983.

[7] E. Bølviken. New tests of significance in periodogram analysis. *Scandinavian Journal of Statistics*, 10(1):1–9, 1983.

[8] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control.* Holden-Day, revised edition, 1976.

[9] J. L. Breedon and N. H. Packard. Nonlinear analysis of data sampled nonuniformly in time. *Physica D*, 58:273–283, 1992.

[10] D. S. Broomhead and G. P. King. Extracting qualitative dynamics from experimental data. *Physica D*, 20:217–236, 1986.

[11] T. Buzug, T. Reimers, and G. Pfister. Optimal reconstruction of strange attractors from purely geometrical arguments. *Europhysics Letters*, 13(7):605–610, December 1990.

[12] T. M. Carroll and L. M. Pecora. Synchronizing chaotic circuits. *IEEE Transactions: Circuits and Systems*, 38:4553–456, 191.

[13] M. Casdagli. Chaos and deterministic versus stochastic non–linear modeling. *Journal of the Royal Statistical Society, B.*, 54(2):303–328, 1991.

[14] C. Chatfield. *The Analysis of Time Series: An Introduction.* Chapman & Hall, fifth edition, 1996.

[15] S. Chatterjee and M. R. Yilmaz. Chaos, fractals and statistics. *Statistical Science*, 7(1):49–121, 1992.

[16] T. Cipra. Tests of periodicity with missing observations. *Statistics*, 22(2):233–243, 1991.

[17] K. M. Cuomo and A. V. Oppenheim. Chaotic signals and systems for communications. In *Proceedings of IEEE ICASSP*, 1993.

[18] B. De Moor. The singular value decomposition and long and short spaces of noisy matrices. *IEEE Transactions on Signal Processing*, 41(9):2826 - 2838, September 1993.

[19] J. D. Farmer and J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59:845, 1987.

[20] D. J. Finney. The joint distributions of variance ratios based on a common error mean square. *Annals of Eugenics*, 11:130, 1941.

[21] R. A. Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London: Series A*, 125:54–59, November 1929.

[22] B. Flury. *Common Principal Components and Related Multivariate Models*. John Wiley & Sons, Inc., 1988.

[23] A. M. Fraser and H. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A.*, 33:1134–1140, 1986.

[24] P. Frederickson, J. L. Kaplan, E. D. Yorke, and J. A. Yorke. The Liapunov dimension of strange attractors. *Journal of Differential Equations*, 49:185 - 207, 1983.

[25] J.-J. Fuchs. Estimating the number of sinusoids in additive white noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(12):1846–1853, December 1988.

[26] W. A. Fuller. *Introduction to Statistical Time Series*. John Wiley & Sons, Inc., 1976.

[27] R. Gilmore. Topological analysis of chaotic dynamical systems. *Reviews of Modern Physics*, 70(4):1455–1529, October 1998.

[28] P. Grassberger. Generalized dimensions of strange attractors. *Physical Letters A*, 97:227–230, 1983.

[29] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber. On noise reduction methods for chaotic data. *Chaos*, 3:127–142, 1993.

[30] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Physical Review Letters*, 50:346–349, 1983.

[31] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189–208, 1983.

[32] S. Hammel. A noise reduction method for chaotic systems. *Physics Letters A*, 148:421–428, 1990.

[33] O. H. Hartley. Tests of significance in harmonic analysis. *Biometrika*, 36:194–201, 1949.

[34] S. Haykin and Xiao Bo Li. Detection of signals in chaos. *Proceedings of the IEEE*, 83(1):94–122, January 1995.

[35] S. Haykin and S. Puthusserypady. Chaotic dynamics of sea clutter. *Chaos*, 7(4):777–802, 1997.

[36] H. G. E. Hentschel and I. Procaccia. The infinite number of generalized dimensions of fractals and strange attractors. *Physica D*, 8:435–444, 1983.

[37] J. H. Horne and S. L. Baliunas. A prescription for period analysis of unevenly sampled time series. *The Astrophysical Journal*, 302:757–763, March 15 1986.

[38] J. R. Hosking. Fractional differencing. *Biometrika*, 68(1):165–171, 1981.

[39] J. L. Jensen. Chaotic dynamical systems with a view towards statistics: a review. *Monographs on Statistics and Applied Probability*, 50:201–250, 1993.

[40] J. L. Kaplan and J. A. Yorke. Chaotic behavior of multidimensional difference equations. In H.-O. Peitgen and H.-O. Walter, editors, *Functional Differential Equations and Approximation of Fixed Points*, volume 730 of *Lecture Note in Mathematics*, pages 228 - 237. Springer-Verlag, New York, 1979.

[41] M. Kennel, R. Brown, and H. Abarbanel. Determining embedding dimension for phase-space reconstruction. *Physical Review A*, 45:3403–3411, 1992.

[42] L. Kocarev and U. Parlitz. General approach for chaotic synchronization with applications to communication. *Physical Review Letters*, 74:5028, 1995.

[43] E. J. Kostelich and J. A. Yorke. Noise reduction in dynamical systems. *Physical Review A*, 38:1649–1652, 1988.

[44] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39:447–462, 1976.

[45] A. I. Mees, P. E. Rapp, and L. S. Jennings. Singular value decomposition and embedding dimension. *Physcal Review A*, 36:340–346, 1987.

[46] D. C. Montgomery. *Introduction to Statistical Quality Control*. John Wiley & Sons, Inc., third edition, 1996.

[47] A. H. Nayfeh and B. Balachandran. *Applied Nonlinear Dynamics: Analytical, Computational, and Experimental Methods*. John Wiley & Sons, 1995.

[48] E. Ott, T. Sauer, and J. A. Yorke, editors. *Coping With Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*. John Wiley & Sons, Inc., New York, 1994.

[49] N. Packard, J. Crutchfield, D. Farmer, and R. Shaw. Geometry from a time series. *Physical Review Letters*, 45(712), 1980.

[50] M. Paluš and I Dvořák. Singular-value decomposition in attractor reconstruction: pitfalls and precautions. *Physica D*, 55:221–234, 1992.

[51] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill Book Company, second edition, 1984.

[52] L. M. Pecora and T. M. Carroll. Synchronization in chaotic systems. *Physical Review Letters*, 64:821–824, 1990.

[53] V. Petrov and K. Showalter. Nonlinear prediction, filtering and control of chemical systems from time series. *Chaos*, 7(4):614–620, 1997.

[54] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C.* Cambridge University Press, New York, 1992.

[55] A. C. Rencher. *Methods of Multivariate Analysis.* John Wiley & Sons, Inc., 1995.

[56] J. Rice. *Mathematical Statistics and Data Analysis.* Duxbury Press, second edition, 1995.

[57] O.E. Rossler. An equation for continuous chaos. *Physics Letters A*, 57:397, 1976.

[58] W. Rudin. *Principles of Mathematical Analysis.* McGraw-Hill, Inc., third edition, 1976.

[59] W. Rudin. *Real and Complex Analysis.* McGraw-Hill, Inc., third edition, 1987.

[60] T. Sauer. Time series prediction using delay coordinate embedding. In A. S. Weigend and N. A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, volume XV of *Santa Fe Institute Studies in the Science of Complexity*, page 175. Addison-Welsley, Reading, MA, 1993.

[61] J. Scargle. Studies in astronomical time series analysis. ii. statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, December 15 1982.

[62] J. Scheinkman and B. LeBaron. Nonlinear dynamics and stock returns. *Journal of Business*, 62(3):311–337, 1989.

[63] A. Schuster. On the periodicities of sunspots. *Phil. Trans. Royal Soc.*, A206, 1906.

[64] S. Selvin. *Practical Biostatistical Methods.* Duxbury Press, 1995.

[65] W. A. Shewhart. *Economic Control of Quality of Manufactured Products.* Van Nostrand Co., New York, 1931.

[66] K. M. Short. Steps toward unmasking secure communications. *International Journal of Bifurcation and Chaos*, 4:957, 1994.

[67] K. M. Short. Detection of teleseismic events in seismic sensor data using nonlinear dynamic forecasting. *International Journal of Bifurcation and Chaos*, 7(10):1833–1845, 1997.

[68] A. F. Siegel. Testing for periodicity in a time series. *Journal of the American Statistical Association*, 75(370):345–348, June 1980.

[69] J. D. Spurrier and L. A. Thombs. Control charts for detecting cyclical behavior. *Technometrics*, 32(2):163–171, May 1990.

[70] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Addison-Wesley, 1994.

[71] G. Sugihara and R. M. May. Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:734, 1990.

[72] F. Takens. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, volume 898 of *Lecture Notes in Math*, page 366. Springer, Berlin, 1981.

[73] L. G. Tatum. Control charts for the detection of a periodic component. *Technometrics*, 38(2):152 - 160, May 1996.

[74] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D*, 58, 1992.

[75] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S-Plus.* Springer–Verlag, 1997.

[76] F. Vernotte, G. Zalamansky, M. McHugh, and E. Lantz. Spectral analysis of irregularly spaced timing data: Comparison of several methods. In *Proceedings of the 1996 10th European Frequency and Time Forum*, pages 344–349, Stevenage, England, 1996. IEE.

[77] A. F. Ware. Fast approximate Fourier transforms for irregularly spaced data. *SIAM Review*, 40(4):838–856, December 1998.

[78] W. Wehlau and K.-C. Leung. The multiple periodicity of delta delphini. *Astrophysical Journal*, 139(3):843–863, 1964.

[79] C. W. Wu and L. O. Chua. A simple way to synchronize chaotic systems with applications to secure communications. *International Journal of Bifurcation and Chaos*, 3:1619–1627, 1993.

[80] Chen Zhao-Guo. An alternative consistent procedure for detecting hidden frequencies. *Journal of Time Series Analysis*, 9(3):301–317, 1988.

[81] Chen Zhao-Guo. Consistent estimates for hidden frequencies in a linear process. *Advances in Applied Probability*, 20:295–314, 1988.

Appendix

Appendix A

Polarplot.m Program

```
function r = polarplot(t, period, x)
%Written by L. McSweeney, Univ. of New Hampshire, 3/98
%This function will take time and observation vectors (t and x, respectively)
%and plots the polar plot for the specific period.
%function r = polarplot(t, period, x)


N = length(t);
x = x - mean(x);              %centers the data
theta = 2*pi*t/period;        %converts the times into an angle
polar(theta, x, 'o')          %plots the polar coordinates


%Calculating the resultant vector and its length by converting from
%polar to rectangular coordinates and finding the squared length of the
%resultants.

[x1, x2] = pol2cart(theta, x);
r = (sum(x1)^ 2 + sum(x2)^ 2)/N;
```

Appendix B

Periodogram.m Program

```
function[power] = periodogram(r, t, p)
%Written by K. Short and L. McSweeney, Univ. of New Hampshire, 3/98
%r is the measurement vector: r = (x1, x2, ..., xn) as a row vector
%t is the time vector: t = (t1, t2, ..., tn) as a row vector
%p is the period vector where you want to check for significant cycles
%p = (p1, p2, ...pm)' (a column vector)
%This program will calculate the power at each period in the p vector
%function[power] = periodogram(r, t, p);


N = length(r);
r = r - mean(r);            %centers the data to have zero mean
w = 2*pi./p;                %converts the periods to (angular) frequencies
W = w*t;                    %gets a matrix of all angular frequencies for each time
E = exp(i*W);               %matrix of complex angles
R = E*r';                   %vector of the resultant lengths
power = abs(R).^ /N;        %vector of power (for each period)
```

Appendix C

Polarmovie.m Program

```
function polarmovie = makemovie(E, resultant, r, period, axisrange)
%Written by L. McSweeney, Univ. of New Hampshire, 3/98
%E and resultant are outputs of periodogram.m
%r is the measurement vector
%period is the period vector where you want to check for significant cycles
%defined as an input to polplot.m
%axisrange defines the size of the square window for the frame
%function polarmovie = makemovie(E, resultant, r, period, axisrange)

m = mean(r);
n = length(period);
M = moviein(n);

%Creates the frames for the movie where the ith frame has the polar plot
%for the ith period

for i = 1 : n
plot((r - m).*E(i, :), 'o');
axis([ -axisrange axisrange -axisrange axisrange]);
text( -axisrange + 1, axisrange - 1, num2str(period(i)));
text( -axisrange + 1, axisrange - 2, num2str(resultant(i)));
M(: , i) = getframe;
end
polarmovie = M;
```

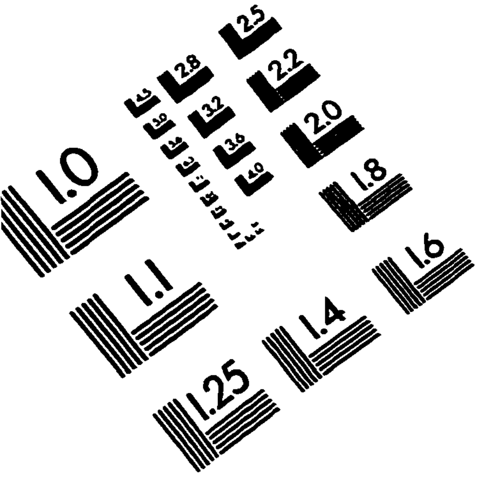
Appendix D

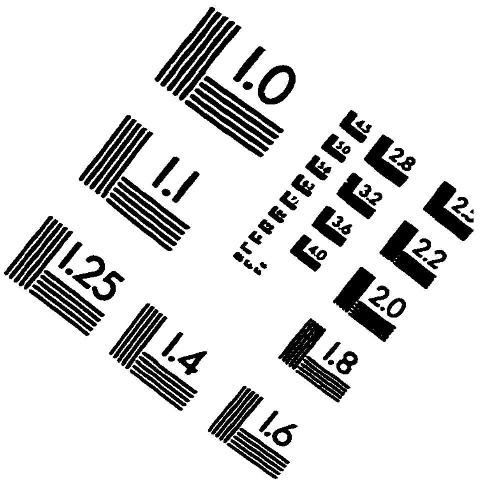
Evspecun.m program

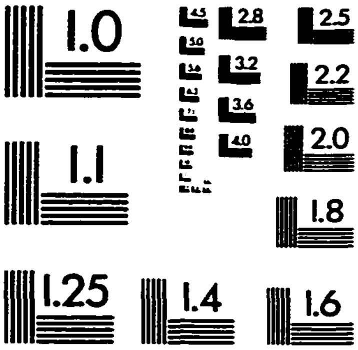
```
function[nosamp] = evspcun(x, t, prds, window, CB);
%Written by L. McSweeney, Univ. of New Hampshire, 3/98
%Adapted from a program written by L. D. Meeker, Univ. of New Hampshire
%Plots the evolutionary spectrum where
%x is the measurement vector: x = (x1, x2, ..., xn) as a row vector
%t is the time vector: t = (t1, t2, ..., tn) as a row vector
%p is the period vector where you want to check for significant cycles
%p = (p1, p2, ...pm)' (a column vector)
%window is the length (in time units) of the moving window.
%The periodogram is calculated in each one of these windows
%CB is the critical value for detecting significant periodic behavior
%Only estimates larger than CB will be plotted. If no bound is wanted, use CB = 0
%nosamp is the number of data points in each window, especially important
%for unevenly sampled data
%[ts, ES, nosamp, sig] = evspcun(x, t, prds, window, CB);

minpd = min(prds);
maxpd = max(prds);


N = length(x);
M = length(prds);
K = floor((max(t) - min(t))/window); %the number of periodograms to be calculated
ts = zeros(K, 1);                    %initializing time vector
ES = zeros(M, K);                    %initializing the periodogram matrix
nosamp = zeros(K, 1);                %initializing the number of samples vector


for i = 1 : K
        % calculating the midpoint of each window
        if i = = 1
                ts(i) = min(t) + window/2;
        else
                ts(i) = ts(i - 1) + window;
        end


        %This portion checks to see how many samples are in the ith window
        %Use for unevenly sampled data
        id = find((t <= i*window) & (t >= (i - 1)*window));
```

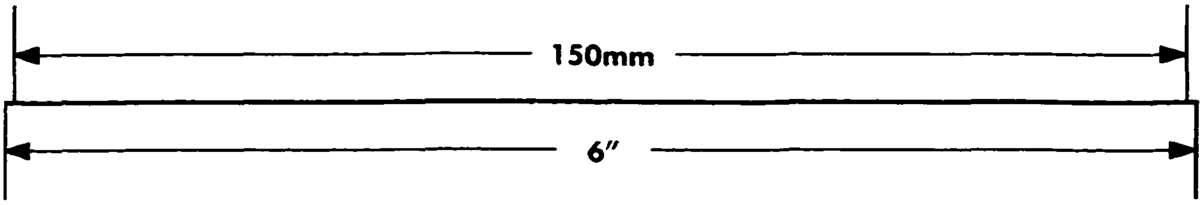
```
          nosamp(i) = length(id);


          %Calculates and records the M periodogram estimates for the ith window
          if nosamp(i) > 0
                          y = polplot(x(id)', t(id)', prds');
                          ES(1 : M, i) = y;
          else ES(1 : M, i) = = 0;
          end
end


id = find(ES < CB);
ES(id) = 0;                   %zeros all periodogram estimates which are below CB


colormap(hsv)
pcolor(ts, prds, ES)
subplot(2, 1, 1)
contour(ts, prds, ES, 20); %contour plot of the evolutionary spectrum
grid on
xlabel('time')
ylabel('period')
set(gca, 'xlim', [min(t) max(t)], 'ylim', [minpd maxpd]);
subplot(2, 1, 2)
mesh(ts, prds, ES);          %surface plot of the evolutionary spectrum
xlabel('time')
ylabel('period')
zlabel('Power')
grid on
```

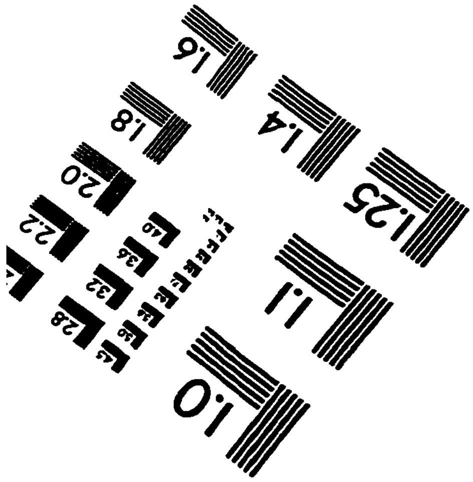
IMAGE EVALUATION
TEST TARGET (QA-3)

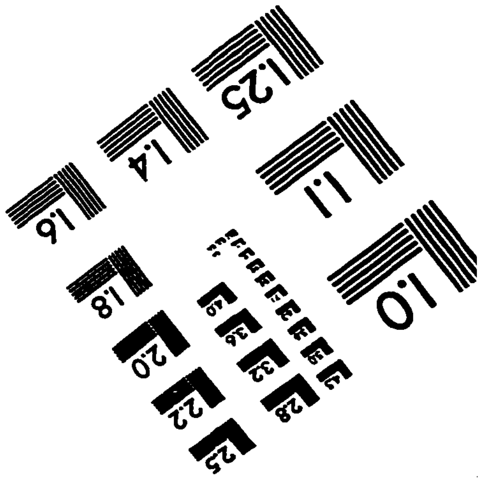












APPLIED IMAGE . Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989