1998

Time series analysis of ozone data

Delparde Raleigh Guthrey

Follow this and additional works at: http://scholarworks.lib.csusb.edu/etd-project

Part of the Computer and Systems Architecture Commons

Recommended Citation

http://scholarworks.lib.csusb.edu/etd-project/1788

This Project is brought to you for free and open access by the John M. Pfau Library at CSUSB ScholarWorks. It has been accepted for inclusion in Theses Digitization Project by an authorized administrator of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
TIME SERIES ANALYSIS OF OZONE DATA

A Project
Presented to the
Faculty of
California State University,
San Bernardino

In Partial Fulfillment
of the Requirements for the Degree
Master of Arts
in
Mathematics

by
Delparde Raleigh Guthrey
December 1998
TIME SERIES ANALYSIS OF OZONE DATA

A Project
Presented to the
Faculty of
California State University,
San Bernardino

by
Delparde Raleigh Guthrey
December 1998
Approved by:

Charles Stanton, Chair, Mathematics
J. Terry Hallett, Mathematics
Joseph Chavez, Mathematics
J. Terry Hallett, Graduate Coordinator
Peter Williams, Department Chair

12/10/98
This paper covers the basic models used in time series analysis, i.e., autoregressive, moving average, ARMA models. The autocorrelation and partial autocorrelation function are discussed, along with examples. State-space models and the associated Kalman filter are studied, as well as stationarity, invertibility, and the Wiener filter. Fourier transforms are used to convert time oriented data into models in the frequency domain by use of the periodogram. The topics and methods presented are used to analyze hourly ozone data collected at sites throughout California. The model that most accurately reflects the data is ARIMA(1,0,0)(1,0,0)_{24}.
ACKNOWLEDGEMENTS

I wish to thank Dr. Charles Stanton for all his help and support in the development and formation of this paper. I also want to thank Dr. Terry Hallett and Dr. Joseph Chavez for their comments on the paper as well. Your assistance was very much appreciated.
# TABLE OF CONTENTS

ABSTRACT ................................................................. iii

ACKNOWLEDGEMENTS ........................................ iv

CHAPTER 1. Introduction ........................................... 1
  1.1 Time Series .................................................. 1
  1.2 Types of Models ............................................ 1
  1.3 What is Covered in this Project ......................... 2
  1.4 Data Analysis .............................................. 3

CHAPTER 2. Time Series ............................................. 4
  2.1 Introduction ................................................ 4
  2.2 Stationarity ............................................... 5
  2.3 Stochastic Processes ..................................... 6
  2.4 Wold’s Decomposition .................................... 7
  2.5 Geometry of Weakly Stationary Stochastic Sequences 8

CHAPTER 3. Model Specification ................................. 13
  3.1 ARMA Models ............................................. 13
  3.2 State-Space Models ...................................... 26
  3.3 Spectral Analysis and the Periodogram ............... 39
  3.4 The Wiener Filter ....................................... 43

CHAPTER 4. Data Analysis ......................................... 48
  4.1 Choosing Between Two Models ......................... 48
  4.2 A Kalman Filter Model ................................. 49

APPENDIX A. ACF of AR(1) process ......................... 52
APPENDIX B. PACF of AR(1) process ........................................ 53
APPENDIX C. ACF of Barton Flats data, July 1994 ...................... 54
APPENDIX D. PACF of Barton Flats data, July 1994 ..................... 55
APPENDIX E. Residuals of AR(2) model ................................. 56
APPENDIX F. Residuals of seasonal model ............................... 57
APPENDIX G. Graph of AR(2) model .................................... 58
APPENDIX H. Graph of seasonal model ................................. 59
BIBLIOGRAPHY .............................................................. 60
Chapter 1. Introduction

1.1 Time Series

A time series is a collection of observations indexed by the time of each observation. A special role is played by time in the relationship between time-ordered variables. Time series analysis is the study of these relationships, and is used in many different areas to detect underlying cyclical patterns. We attempt to find these patterns, whether they are evident or more subtle. As we do so, we can attempt to achieve the goal of finding a useful way (or model) to express a time-structured relationship that gives rise to an observed series and to predict future values of a series based on the history of that series.

1.2 Types of Models

We try to use the available data to build a statistical model representing the relationship between the variable \( x \) and the time \( t \) to forecast future values, or to explain past values of the \( x \)'s. If we only applied regression techniques, we might formulate a model that does not have a time-ordered relationship between the variables, nor a measurement of correlations between the error terms. Also, the assumption of independence would not be applicable to time series. Fortunately, we have more satisfactory models at our disposal.

We will discuss three types of models: autoregressive moving average, state space, and frequency domain models. Each of these have benefits as well as obstacles involved in their usage. The first models, usually called ARMA models, have assumed great importance in modeling situations. They can be used if we
can reasonably represent the data as a weighted linear combination of present and past terms, or random shocks. Unfortunately, we never know the exact underlying process. If we are unable to observe data directly, but have to depend on a related observation that has measurement error contained in it, we may use a state space model. In some applications, it is important to have an efficient method of computing the current estimated value which requires as little storage of information as possible, such as the Kalman filter, which however is somewhat complicated. And if we wish to find hidden cycles or periodicities, we can enter the frequency domain and make use of the periodogram. However, other procedures have been designed to reduce the computational costs of computing covariances with very large numbers of observations.

1.3 What is Covered in this Project

This project will begin with an overview of time series. This section will include topics such as stationarity and invertibility. The Wold decomposition is introduced as a fundamental representation of the time series that we will be investigating. Also, we will attempt to show how estimators can be derived based on geometrical methods using Hilbert spaces.

In the next section, we discuss model specification, going into more detail concerning the three types of models. Here we will study autocorrelation functions, which describe how a current value of a series is related to its own past (or future) values by measuring the strength and direction of the relationship among the observations. The Kalman filter, which is a method to update a linear projection for the system, is mentioned as a major example of the state space representation. We will
show how Fourier transforms are used to convert time-oriented data into frequency domain models by use of the periodogram, and the Wiener filter will be discussed.

1.4 Data Analysis

To conclude the project, we will use the topics and methods presented to look for various patterns in a large database of hourly ozone measurements taken at a dozen sites throughout California over a period of four to five years during the growing season from the late spring to mid-autumn. Some patterns are clear, such as ozone level vs. time of day. Other patterns are not as obvious. The computer package SYSTAT is used to examine computational problems and analyze the data. Since there is a certain amount of missing data, the topic of how to use the remaining data is also addressed.

As we investigate the data, we choose a site (Barton Flats) as the one that had the highest amount of ozone recorded among the sites, as well as the highest variance. After comparison to several basic models, the autoregressive models appear to be best suited to fit the recorded measurements of ozone. Two models are chosen and their residuals are compared. A seasonal model: ARIMA(1,0,0),(1,0,0)_{24}, is chosen based on having a smaller residual total. The Kalman filter is also used, based on the monthly highs obtained for the same site.
Chapter 2. Time Series

2.1 Introduction

As an introduction, we first consider the probabilistic structure that underlies time series observations. We write $x_t$ for the observation made at time $t$. The units of time depend on the application; they could be years, months, weeks, days, etc. In this project, we will be primarily using hours. We assume that the observations are equally spaced in time.

In order to model the uncertainty in our observations, we assume that for each time $t$, $x_t$ is a random variable. Therefore, the behavior of $x_t$ will be determined by a probability distribution. However, a very important characteristic of time series models, as compared to models used in many other areas of statistics, is that we assume that the observations made at different points in time are statistically dependent. We seek to investigate this dependence. So, for two time points $t$ and $s$, the joint behavior of $x_t$ and $x_s$ will be determined from a joint bivariate distribution. More generally, the collection of random observations would be governed from their joint multivariate distribution. The finite set of observations $x_1, x_2, \ldots, x_n$ is considered to be a portion of a much longer sequence going indefinitely into the future, and possibly into the past.

The sequence of random variables $\{x_1, x_2, \ldots\}$ or $\{..., x_{-1}, x_0, x_1, x_2, \ldots\}$ is called a stochastic process. Although it is known that the complete probabilistic structure of such a process is determined by the set of distributions of all finite collections of $x$’s, we will not deal explicitly with these multivariate distributions. Much of the information in these joint distributions can be described in terms of means, variances
2.2 Stationarity

In order to make statistical inferences about the structure of a stochastic process on the basis of a finite observed record of the process, we need to make some simplifying assumptions. The most important such assumption is that of stationarity.

**Definition.** A stochastic process \( \{x_t\} \) is said to be strictly stationary if the joint distribution of \( x(t_1), x(t_2), \ldots, x(t_n) \) is the same as the joint distribution of \( x(t_1 - k), x(t_2 - k), \ldots, x(t_n - k) \) for all choices of time points \( t_1, t_2, \ldots, t_n \) and all choices of time lag \( k \). If a stationary process has finite variance, then the covariance must depend only on the time lag. A process \( \{x_t\} \) is said to be weakly stationary if the mean is constant over time, and \( \text{cov}(x_t, x_{t-k}) = \text{cov}(x_0, x_k) \) for all time \( t \) and lag \( k \).

The basic idea is that the probability laws governing the process do not change with time, that is, the process is in a sort of statistical equilibrium. The weak form of stationarity says that the process has a mean, variance and autocorrelation function that are constant through time. The strong form of stationarity requires that the entire probability distribution function for the process is independent of time. If the joint distributions of a process are all multivariate normal, then the concepts are identical. With stationarity, we can develop a simple theoretical framework and useful sample statistics.

If a time series is not stationary, we may be able to modify the data to get a stationary serie sand later reverse the modifications to obtain forecasts. Examples include taking a square root or using a logarithmic transformation for variance.
stabilization. Another example of a transformation to achieve stationarity is the technique of differencing. A series with a stationary mean returns fairly quickly to a constant mean. If a series does not return quickly to a constant mean, we may be able to create a new series with a constant mean by differencing the data, that is, finding successive changes in the series. This technique will be explained more fully in section 3.1.2.

2.3 Stochastic Processes

Although it is possible to study time series analysis without using Hilbert space terminology and techniques, it does have advantages, since results obtained from Euclidean geometry can often be used to make complicated algebraic results geometrically easier to understand.

If we have a set of random vectors of dimension $p$ with mean zero and finite variances, we may define an inner product of two members $X$ and $Y$ by $<X,Y> = E(XY')$, a matrix value. This forms a Hilbert space, since the inner product space is complete, as will be the case here. If two random vectors have an inner product of zero, then they are orthogonal (and uncorrelated):

**Theorem 1.** [5] Let $(\Omega, P)$ be a probability space. Then the set $H = \{X \mid E \|XX'\| < \infty\}$ is a Hilbert space under the inner product $<X,Y>$.

The inner product is the covariance, so it is reasonable to use projections. A $p$-component stochastic process is a parameterized family $\{X_t\}, t \in T$, where the parameter set $T$ represents time, of random variables defined on some probability space $(\Omega, P)$. A random vector is a column vector, $X = (X_1, \ldots, X_p)'$, each of
whose components is a random variable. Since a stochastic process is a function \( X : \Omega \times \mathbb{R} \rightarrow \mathbb{R} \), we may use \( X_t \) instead of \( X(\omega, t) \).

A process that is weakly stationary is also called stationary in the wide sense. If \( X \) is stationary, \( E(X_t) = E(X_0) \), and \( E(X_{t+k}, X_t) = E(X_k, X_0) \) for all \( t \).

2.4 Wold’s Decomposition

In this section we discuss Wold’s decomposition, which provides a basis for decomposing a process into a series derived from a “white noise” process, so that any covariance-stationary series can be written in this manner.

**Definition.** The Gaussian white noise process with parameter \( \sigma^2 \) is a stationary process defined as a sequence of independent, identically distributed random variables \( \{\varepsilon_t\}_{t=-\infty}^{\infty} \), where each \( \{\varepsilon_t\} \) has a \( N(0, \sigma^2) \) distribution.

The term white noise comes from the fact that a frequency analysis of the model shows that, in analogy with white light, all frequencies enter equally.

**Theorem 2.** [9] (Wold’s decomposition): Let \( X_t \) be a covariance-stationary stochastic process. Then there exists a Gaussian white noise process \( \varepsilon_t \), and a deterministic process \( \kappa_t \), such that:

\[
X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t, \tag{1}
\]

where \( \psi_0 = 1 \) and \( \sum_{j=0}^{\infty} \psi_j^2 < \infty \).

One might think that we were able to write this type of process in the form of (1) just because we are restricting our focus to a convenient class of models. In fact, this representation is fundamental for any covariance-stationary time series.
The term $\varepsilon_t$ represents the error made in forecasting $X_t$ on the basis of a linear function of lagged $X$:

$$\varepsilon_t = X_t - \hat{E}(X_t|X_{t-1}, X_{t-2}, \ldots).$$

This equation uses conditional notation, with the predicted value of $X_t$, given $X_{t-1}, X_{t-2}, \ldots$ that have come beforehand. The value of $\kappa_t$ is uncorrelated with $\varepsilon_{t-j}$ for any $j$, though $\kappa_t$ can be predicted arbitrarily well from a linear function of past values of $X$:

$$\kappa_t = \hat{E}(\kappa_t|X_{t-1}, X_{t-2}, \ldots)$$

The term $\kappa_t$ is called the linearly deterministic component of $X_t$, while $\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-1}$ is called the linearly indeterministic component. If $\kappa_t = 0$, then the process is called purely linearly indeterministic.

This proposition was first proved by Wold in 1938. It relies on stable second moments of $X$ but makes no use of higher moments. Therefore it describes optimal linear forecasts of $X$, as opposed to forecasts obtained by fitting a quadratic model to the data.

2.5 Geometry of Weakly Stationary Stochastic Sequences

**Definition.** The backshift operator $B$ is defined as follows: when $B$ operates on any time-subscripted variable, the time subscript is shifted by $i$ time units. So, $B^i x_t = x_{t-i}$, but $B^i C = C$. This notation can also be used to write differenced series. Such as series may be written as: $w_t = x_t - x_{t-1} = x_t - B x_t = (1 - B) x_t$. 

8
Suppose that \( \{\varepsilon_t\} \) is the innovation process for \( x_t \) and the \( \kappa_t \) in the Wold decomposition is equal to zero. Then we may write the Wold decomposition as:

\[
x_t = \psi(B)\varepsilon_t
\]

and

\[
\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j, \quad \psi_0 = 1.
\]

We may predict \( x_{t+m}, m > 0 \), by using a linear combination of current and past \( x \)'s (called an information set): \( I_t = \{x_t, x_{t-1}, x_{t-2}, \ldots\} \).

**Definition.** \( \psi(B) \) is invertible if \( \psi(B)^{-1} = \phi(B) \) with \( \phi(B) = \sum_{j=0}^{\infty} \phi_j B^j \).

If \( \psi(B) \) is invertible, the \( \{x_t\} \) process can also be expressed in this way:

\[
\phi(B)x_t = \varepsilon_t. \quad (2)
\]

For example, a process defined by \( x_t = \varepsilon_t - \psi_1 \varepsilon_{t-1} \) is invertible (this will be defined later in section 3.1.1 as an MA(1) process). This process is \( \varepsilon_t - \psi_1 B \varepsilon_t = (1 - \psi_1 B) \varepsilon_t \).

So we see that \( \varepsilon_t = (1 - \psi_1 B)^{-1} x_t \). This shows that \( \varepsilon_t = \phi(B)^{-1} x_t \), so that (2) is satisfied. Since \( |\psi_1| < 1 \), to find \( (1 - \psi_1 B)^{-1} \) we use a geometric series, and obtain

\[
1 + \psi_1 B + \psi_1^2 B^2 + \ldots. \quad \text{Therefore,} \quad \varepsilon_t = (1 + \psi_1 B + \psi_1^2 B^2 + \ldots) x_t = \sum_{j=0}^{\infty} \psi_1^j B^j x_t = \sum_{j=0}^{\infty} \psi_j B^j x_t = \psi(B) x_t, \quad \text{if we let } \psi_1^j = \psi_j. \quad \text{However, not all processes are invertible: for example, the differenced series } x_t = \varepsilon_t - \varepsilon_{t-1}. \]

Notice that the set \( I_t \) is equivalent to the set \( \tilde{I}_t = \{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\} \) in the sense that the spans are the same on Hilbert space, that is, invertibility implies that span \( (x_t, x_{t-1}, \ldots) = \text{span} (\varepsilon_t, \varepsilon_{t-1}, \ldots) \).
Definition. A linear predictor $x_{t+m|t}$ of $x_{t+m}$ based on $I_t$ is a random variable of the form

$$x_{t+m|t} = A(B)\varepsilon_t,$$

where

$$A(B) = a_0 + a_1 B + a_2 B^2 + \ldots.$$

By definition, the best linear prediction minimizes the prediction error variance, which is given by:

$$E[(x_{t+m} - x_{t+m|t})^2|I_t] = E[(\psi(B)\varepsilon_{t+m} - A(B)\varepsilon_t)^2|I_t].$$

We separate out the future $\varepsilon$'s from those in $I_t$ by writing:

$$\psi(B)\varepsilon_{t+m} = \sum_{j=0}^{m-1} \psi_j \varepsilon_{t+m-j} + \sum_{j=m}^{\infty} \psi_j \varepsilon_{t+m-j},$$

where the first term on the right side represents the future $\varepsilon$'s, and the second term represents those in $I_t$. Notice that for $j = m$, this particular term is $\psi_m \varepsilon_t$.

We may also write the equation as: $\sum_{j=0}^{m-1} \psi_j \varepsilon_{t+m-j} + \psi_m(B)\varepsilon_t$, where $\psi_m(B) = \sum_{i=0}^{\infty} \psi_{m+i} B^i$. Also notice that for $i = 0$, the right side of the equation becomes $\psi_m B^0 \varepsilon_t = \psi_m \varepsilon_t$, as before. Then the variance in equation (3) is equal to:

$$\sum_{j=0}^{m-1} \psi_j^2 \sigma^2 + \{[\psi_m(B) - A(B)]\varepsilon_t\}^2.$$

If $A(B) = \psi_m(B)$, then the variance is minimized, and the best least-squares predictor of $x_{t+m}$ based on $I_t$ is

$$x_{t+m|t} = \psi_m(B)\varepsilon_t.$$
Let $F(B)$ be a Laurent series in $B$. We introduce the notation $[F(B)]_+$, to denote the series where only positive powers of $B$ are kept. Thus, we express $\psi_m(B)$ as $[\psi(B)/B^m]_+$, where only non-negative powers of $B$ are kept, that is, the current and past values. For example, if $m = 2$, then $\psi_2(B) = [\theta(B)/B^2]_+$. All expressions with negative powers (the future values) are dropped. Then the best least squares predictor of $x_{t+m}$ is: $\psi(B)/B^m \varepsilon_t$. The operation $[F(B)]_+$ chooses the realizable portion of the function $[F(B)]$, called the transfer function. When basic or fundamental noise sequences are involved, $[F(B)]_+$ realizes the orthogonal projection onto the subspace spanned by the noise sequences.

If we look further at (4), we see that the variable to be predicted, $x_{t+m}$, can be written as $\psi(B)B^{-m}\varepsilon_t$ since $B^{-m}\varepsilon_t = \varepsilon_{t+m}$. The best predictor, $[\psi(B)/B^m]_+ \varepsilon_t$, eliminates the random variables $\varepsilon_{t+m}, \varepsilon_{t+m-1}, \ldots, \varepsilon_{t+1}$ from $\psi(B)B^{-m}\varepsilon_t$. Since these $\varepsilon$'s are uncorrelated with the $\varepsilon$'s that are in $I_t$, dropping these uncorrelated random variables is equivalent to taking the orthogonal projection of them onto the subspace spanned by the $\varepsilon$'s in $I_t$. So the key point here is that the operator $[\psi(B)/B^m]_+$ is, therefore, the Hilbert space projection operator.

Another way to say this is that the best predictor $x_{t+m|t}$ is such that the prediction error $x_{t+m} - x_{t+m|t}$ is orthogonal to, i.e., uncorrelated with, all the $\varepsilon$'s in $I_t$, or to $x_t, x_{t-1}, \ldots$, by the equivalence of the subspace spanned by $\varepsilon_t, \varepsilon_{t-1}, \ldots$ with that spanned by $x_t, x_{t-1}, t = 0, 1, \ldots$.

This fact is known as the orthogonality principle. As an illustration, we look at the least squares approach from classical regression. We consider random variables $x_1, \ldots, x_n$, which are used to estimate another random variable $y$ in the minimum
mean square sense, that is:

\[ \min_{\{a_i\}} E(y - \sum_{i=1}^{n} a_i x_i)^2. \]

If we look at a slight change in \(a_j\) from its optimal value \(a_j^0\), we see that the coefficients are optimal if and only if:

\[ 0 = (y - \sum_{i=1}^{n} a_i^0 x_i, x_j), \quad j = 1, \ldots, n. \]

Therefore, this means that the error of the optimal estimator, \(y - \sum_{i=1}^{n} a_i^0 x_i\), is orthogonal to every vector, \(x_j, \ j = 1, \ldots, n\).

In principle, finding a Wold representation requires fitting an infinite number of parameters \((\psi_1, \psi_2, \ldots)\) to the data. With a finite number of observations on \((x_1, x_2, \ldots, x_T)\), of course this will never be possible. To be practical, we therefore need to make some additional assumptions about the nature of \((\psi_1, \psi_2, \ldots)\). A typical assumption is that \(\psi(B)\) can be expressed as the ratio of two finite-order polynomials:

\[ \sum_{j=0}^{\infty} \psi_j B^j = \frac{\theta(B)}{\phi(B)} = \frac{1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q}{1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p}. \]

These give what are known as ARMA models.
Chapter 3. Model Specification

This section introduces univariate autoregressive moving average (ARMA) processes, which provide a very useful class of models for describing the dynamics of an individual time series. We discuss the models first, and then proceed to an overview of estimating and formulation of the models, defining the autocorrelation and partial autocorrelation functions along with graphical examples. Next, we describe state-space models and the associated Kalman filter. Finally, we switch from the time domain to the frequency domain using spectral analysis, the periodogram, and the Wiener filter.

3.1 ARMA Models

We now look at the autoregressive moving average models, and also we provide an overview of the Box and Jenkins methodology for estimating parameters of the models.

3.1.1 Stationary ARMA Processes

There are three major types of ARMA processes: autoregressive, moving average, and mixed processes, which are discussed in this section[9].

Autoregressive processes are as their name implies: regressions on themselves. The autoregressive (AR) model is given by equations such as:

\[ x_t = \phi_1 x_{t-1} + a_t, \]

or:

\[ x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + a_t. \]
The autoregressive terms are the past values of $x$ with their associated coefficients. The order ($p$) of an AR model or process is the highest lag length of the AR terms. The first equation has $p = 1$ and is denoted AR(1). The second equation has as its highest lag length $p = 2$, and is AR(2). We could intuitively think of this model representing the height of ocean waves above the bottom of the ocean as a random variable $x_t$, measured every half second. Then our predicted value, $\hat{x}_t$, would be based on a measured value from .5 seconds before multiplied by an appropriate constant $\phi_1$, in addition to a measured value from 1 second earlier multiplied by a constant $\phi_2$. The random error (or shock) $a_t$ would be assumed to be zero for the prediction. Values obtained more than 1 second beforehand would not be used. This is a second order difference equation, whose behavior can be compared with a second order differential equation.

The general AR($p$) model is given by:

$$x_t = \phi(B)x_t + a_t = \sum_{j=1}^{p} \phi_j x_{t-j} + a_t$$

The $\phi$ coefficients satisfy certain conditions if the mean of the process is stationary. For example, in an AR(1) process, stationarity requires that $|\phi_1| < 1$. For an AR(2) process, these conditions must all be true: $|\phi_2| < 1, \phi_2 + \phi_1 < 1$, and $\phi_2 - \phi_1 < 1$. In practice, we don't observe the AR process coefficients, instead we see if the estimated model coefficients satisfy the stationarity conditions.

We next investigate the moving average process. The moving average (MA) model is given by equations such as:

$$x_t = a_t - \theta_1 a_{t-1},$$
or:

\[ x_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}. \]

The terminology moving average comes from the fact that \( x_t \) is obtained by applying the weights \( 1, -\theta_1, -\theta_2 \) to the variables \( a_t, a_{t-1}, a_{t-2} \) and then moving the same weights one unit of time forward and applying them to \( a_{t+1}, a_t, a_{t-1} \) to obtain \( x_{t+1} \).

The past “random shocks” \( (a_{t-k}) \) with their associated coefficients are called moving average terms. A past random shock \( a_{t-k} \) is not a past value of \( x_t \), but is the random shock component of \( x_{t-k} \), so an MA term represents part of a past value of \( x_t \). As an example, the MA term in the first equation is \( -\theta_1 a_{t-1} \). Here the \( a_{t-1} \) term is not the past value \( x_{t-1} \), but is a part of \( x_{t-1} \). To see this, we can subtract 1 from each time subscript to get \( x_{t-1} = -\theta_1 a_{t-2} + a_{t-1} \). So, \( a_{t-1} \) is a component of \( x_{t-1} \). MA terms are conventionally written with negative signs.

The order \( (q) \) of an MA process is the highest lag length of the MA terms, so the first equation is a first-order \( (q = 1) \) MA process, known as MA(1). The second equation has a lag length of 2, and is called MA(2). A possible use of this model might be to let the random variable \( x_t \) represent the height of a river in a tropical area that receives rain nearly every day. \( x_t \) would be influenced by the prior day’s height plus the day’s rainfall, minus water that flows away from the area measured. We could predict the height based on the current random shock, minus the random shock from the previous day multiplied by a constant \( \theta_1 \) (the first MA term). We would also remove the random shock from two days earlier multiplied by a constant \( \theta_2 \), to obtain \( \hat{x}_t \).
In general, MA($q$) processes can be written as:

$$x_t = a_t - \sum_{j=1}^{q} \theta_j a_{t-j}.$$ 

A model that combines both AR and MA processes is given by:

$$x_t = \phi_1 x_{t-1} - \theta_1 a_{t-1} + a_t.$$ 

This is called a mixed process with $p = 1$ and $q = 1$, referred to as an ARMA(1,1) process. Generally, this type of ARMA($p,q$) process is represented as:

$$x_t = \sum_{j=1}^{p} \phi_j x_{t-j} - \sum_{j=1}^{q} \theta_j a_{t-j} + a_t.$$ 

Earlier, we discussed stationarity conditions for AR processes. MA processes have similar conditions that are required for invertibility. Unless we are careful, ambiguous results may be obtained, such as $\theta$ and $1/\theta$ yielding the same correlation (for an MA(1) process). The reason for this is because an MA process has an equivalent AR form. To see this, consider an MA(1) model, $x_t = a_t - \theta a_{t-1}$. First we rewrite this as $a_t = x_t + \theta a_{t-1}$, and then replace $t$ by $t-1$ to get $a_{t-1} = x_{t-1} + \theta a_{t-2}$. We substitute for $a_{t-1}$ to get:

$$a_t = x_t + \theta (x_{t-1} + \theta a_{t-2}) = x_t + \theta x_{t-1} + \theta^2 a_{t-2}.$$ 

If $|\theta| < 1$, we may continue this substitution into the past and obtain:

$$a_t = x_t + \theta x_{t-1} + \theta^2 x_{t-2} + \ldots$$

or

$$x_t = (-\theta x_{t-1} - \theta^2 x_{t-2} - \ldots) + a_t.$$ 

16
If $|\theta| < 1$, we see that the MA(1) can be “inverted” into an infinite-order autoregressive process, AR(∞), and so the MA(1) process is invertible. For an MA(2) process, invertibility requires that $|\theta_2| < 1$, $\theta_2 + \theta_1 < 1$, and $\theta_2 - \theta_1 < 1$. Invertibility ensures that the absolute values of the implied weights on the x’s in this equivalent AR form become smaller as the lag length on the past x’s increases. The more recent data is then given more importance for understanding the present than data from the distant past.

For a general MA($q$) model, we define the MA characteristic polynomial as:

$$\theta(y) = 1 - \theta_1 y - \theta_2 y^2 - \ldots - \theta_q y^q,$$

and the corresponding MA characteristic equation:

$$1 - \theta_1 y - \theta_2 y^2 - \ldots - \theta_q y^q = 0. \quad (5)$$

It can then be shown that the MA($q$) model is invertible, that is, there are constants $\pi_j$ such that

$$x_t = \sum_{j=1}^{\infty} \pi_j x_{t-j} + a_t$$

if and only if the roots of the MA characteristic equation (5) exceed 1 in absolute value. The uniqueness problem is solved if we restrict attention to the physically realizable class of invertible models.

We may look at this in terms of zeros and poles of polynomials. Recall that if $f(x)$ has a pole at $x_0$, then as $x \to x_0$, $f(x) \to \pm \infty$. For example, if $f(x) = 1/(1-x)$, then there is a pole at $x = 1$.

For $|\theta| < 1$, we may use a geometric series to express $(1 - \theta B)^{-1} = 1 + \theta B + \theta^2 B^2 + \ldots$ where $B^i a_t = a_{t-i}$. There are poles where $1/(1-\theta x) < 0$. When $1-\theta x =$
0, \( x = 1/\theta \), so there are poles when \( |1/\theta| > 1 \). We see that with the MA(1) process:
\[
x_t = a_t - \theta a_{t-1} = (1 - \theta B)a_t.
\]
Then, \((1 - \theta B)^{-1}x_t = a_t\). Substituting, we find:
\[
(1 + \theta B + \theta^2 B^2 + \ldots)x_t = a_t
\]
and
\[
x_t = -\theta x_{t-1} - \theta^2 x_{t-2} - \ldots + a_t,
\]
an AR(\(\infty\)) process as before.

3.1.2 Estimating Model Parameters

As we formulate the models to be used, and estimate the parameters involved in them, we will follow the strategy given by Box and Jenkins (see[2]). In model specification, the classes of time series models are selected that may be appropriate for a given observed series. We look at the time plot of the series, compute various statistics, and apply knowledge from the subject area involved. The model chosen is tentative at this point and subject to revision later. We shall try to use the principle of parsimony; that is, the model should require the smallest possible number of parameters that will adequately represent the data. The more parameters we estimate, the more room there is to go wrong. Although complicated models can track the data very well over the period for which the parameters are estimated, they often perform poorly when used for forecasting. For example, the 1960s saw the development of a number of large macroeconomic models supposedly describing the economy using hundreds of macroeconomic variables and equations. It was subsequently discovered that univariate ARMA models with small values of \( p \) or \( q \) often produced better forecasts than the big models.[9]
The next step, model fitting, consists of finding the best possible estimates of the unknown parameters that are used in the model. Next, model diagnostics is concerned with analyzing the quality of the model specified. We see if the model fits the data and check if the model’s assumptions are reasonably satisfied. If no inadequacies are found, then the modeling may be assumed to be complete, and the model can be used to forecast future values. Otherwise, we choose another model by returning to model specification.

In this section we cover several topics: autocorrelation, partial autocorrelation coefficients, integrated processes, more on backshift notation, the general ARIMA\((p, d, q)\) process, and seasonal models. Since we are investigating the dependence between observations made at different points in time, we now look at the concept of autocorrelation. We are able to measure how any current value of a series \(x_t\) is related to its own future values \((x_{t+1}, x_{t+2}, \ldots)\), or its own past values \((x_{t-1}, x_{t-2}, \ldots)\). This measurement is called autocorrelation, and it measures the strength and the direction of the relationship among observations within a single time series \(x_t\) when the observations are separated by \(k\) time periods \((k = 1, 2, \ldots, n)\), where \(n\) is the number of total time periods observed. We treat the terms of a series \(x_t\) as one random variable, the terms \(x_{t+k}\) as another, and consider the correlation coefficient between them. The study of autocorrelation patterns thereby obtained can lead to identifying the appropriate ARMA model for the series.

Some definitions follow. By the assumption of stationarity, the mean is independent of \(t\). We denote this as \(\mu_x = E(x_t)\). The population variance is \(\sigma_x^2 = E(x_t - \mu_x)^2\).
which is also independent of $t$ by the hypothesis of stationarity. Also,

$$\text{cov}(x_t, x_{t+k}) = E[(x_t - \mu_x)(x_{t+k} - \mu_x)].$$

We define the population autocorrelation coefficient at lags $k = 1, 2, \ldots, K$ to be:

$$\rho_k = \frac{\text{cov}(x_t, x_{t+k})}{\sigma_x^2}$$

Suppose we observe $x_1, \ldots, x_n$. Then the sample autocorrelation coefficient is defined as:

$$r_k = \frac{\sum_{i=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{i=1}^{n} (x_t - \bar{x})^2},$$

where $n$ is the sample size. Values of $r_k$ for $k = 1, 2, \ldots, K$ are computed and the resulting set of values is the sample autocorrelation function (SACF).

Any $r_k$ is only a sample value that could differ from zero only because of sampling variation. To determine the importance of this sample statistic, we compare it with its standard error. An approximate standard error for $r_k$ is:

$$s(r_k) = \sqrt{\frac{1 + 2 \sum_{j=1}^{k-1} r_j^2}{n}}$$

To test for a linear association in the population between $x_t$ and $x_{t+k}$, we test $H_0 : \rho_k = 0$, versus the alternate hypothesis $H_a : \rho_k \neq 0$, with the t-statistic:

$$t = \frac{(r_k - \rho_k)}{s(r_k)}.$$  

Notice that $t$ is the ratio of the statistic $r_k$ to its standard error $s(r_k)$, since $\rho_k$ is hypothesized to be zero. If we have a stationary mean, the SACF decays quickly toward zero. This means that the autocorrelation coefficients should be within two standard error limits from zero by about lag 5 or 6.
We now introduce the concept of partial autocorrelation coefficients. To do so, we consider the set of $n$ regression equations on time $t$:

\[ x_t = C_1 + \phi_{11}x_{t-1} + a_{1,t} \]
\[ x_t = C_2 + \phi_{21}x_{t-1} + \phi_{22}x_{t-2} + a_{2,t} \]
\[ \vdots \]
\[ x_t = C_n + \phi_{n1}x_{t-1} + \phi_{n2}x_{t-2} + \cdots + \phi_{nn}x_{t-n} + a_{n,t} \]

Note that if $x_t$ is positively correlated with $x_{t-1}$, then by the assumption of stationarity, $x_{t-1}$ is positively correlated with $x_t$, and then $x_t$ is likely to be positively correlated with $x_{t-2}$. The population partial autocorrelation coefficient at lag $k = 1, 2, \ldots, n$ is defined to be the coefficient ($\phi_{kk}$) in each equation. Each population coefficient is estimated for a given data set by its sample $\hat{\phi}_{kk}$, resulting in a set of estimates called a sample partial autocorrelation function (SPACF). The partial autocorrelation function (PACF) at lag $k$, denoted by $\phi_{kk}$, is the correlation between $x_t$ and $x_{t-n}$ after removing the effect of the intervening variables $x_{t-1}, x_{t-2}, \ldots, x_{t-n+1}$. Computationally efficient formulas for computing $\hat{\phi}_{kk}$ are available.

Each ARMA process has a theoretical autocorrelation function (ACF) and a partial autocorrelation function associated with it. In order to identify a particular ARMA model, we construct the SACF and SPACF for a given series of data. We then compare them with common ACF's and PACF's. If there is a reasonable match, then we have a tentative ARMA model.

Characteristics of stationary AR processes include the fact that the theoretical ACF decays exponentially, or with a damped sine wave pattern, or both. The PACF has spikes through lag $p$, and is zero afterwards. In Appendices A and B, an AR(1)
model with $\phi_1 = .8$ is shown. Appendix A demonstrates the decaying pattern of the ACF, and Appendix B shows the PACF with a spike at lag 1, and virtually zero after that.

A stationary MA process has an ACF that has spikes through lag $q$, and is zero after lag $q$. Here, the PACF decays. In a mixed process, both the ACF and PACF decay.

Since for MA$(q)$ series the autocorrelation function is zero for lags beyond $q$, the ACF is a good indicator of the order of the process. However, the autocorrelations of AR$(p)$ series do not remain zero after a certain number of lags, and the PACF is useful for determining the order $p$.

We now turn to nonstationary series, where we begin by writing the model for the differenced series $w_t$. We say that $d = 1$, where $d$ is the number of differences taken. This means that we compute the successive changes in the series for all $t$: $w_t = x_t - x_{t-1}$. This is first differencing. If the mean is not yet constant, we can take the first differences of the first differences (known as the second differences): $w_t^* = w_t - w_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2})$ where $w_t^*$ are the second differences of $x_t$. In this case $d = 2$, but first differences are usually adequate.

If $d = 1$ and a differenced series $w_t$ is ARMA $(0,1)$, we may then define a model for this nonstationary series as an ARIMA($p,d,q$) = ARIMA $(0,1,1)$ model for the original series $x_t$. Here ARIMA is an abbreviation for Autoregressive Integrated Moving Average. This integrated process occurs very often in practice, and each forecast from this process is an exponentially weighted moving average.

To observe the ARIMA$(0,1,1)$ we notice that with $d = 1$, we have $w_t = x_t - x_{t-1}$. With $p = 0$ and $q = 1$, $w_t = -\theta_1 a_{t-1} + a_t$. This implies a model for the original
Here we are integrating, which is the opposite of differencing. So, a process that includes differencing is called an integrated process, since we must integrate to get back to the process for the original series.

We next consider the ARIMA(1,1,1) process given by: \( w_t = \phi_1 w_{t-1} - \theta_1 a_{t-1} + a_t \).

If we substitute \( x_t - x_{t-1} \) for \( w_t \), and add \( x_{t-1} \) to both sides of the equation, we obtain: \( x_t = x_{t-1} + \phi(x_{t-1} - x_{t-2}) - \theta_1 a_{t-1} + a_t \), which is an awkward way to express a fairly simple process. To simply the writing of ARIMA processes, we often use backshift notation, as in section 2.5.

To rewrite the AR(1) process using the backshift notation defined by \( B^t x_t = x_{t-i} \), we obtain:

\[
x_t = \phi_1 x_{t-1} + a_t,
\]

\[
x_t - \phi_1 x_{t-1} = a_t = x_t - \phi_1 B x_t = (1 - \phi_1 B) x_t.
\]

Other common results using backshift notation are:

AR(2): \( (1 - \phi_1 B - \phi_2 B^2) x_t = a_t \)

MA(1): \( x_t = (1 - \theta_1 B) a_t \)

MA(2): \( x_t = (1 + \theta_1 B - \theta_2 B^2) a_t \)

ARMA(1,1): \( (1 - \phi_1 B) x_t = (1 - \theta_1 B) a_t \)

ARIMA(0,1,1): \( (1 - B) x_t = (1 - \theta_1 B) a_t \)
ARIMA(1,1,1): \((1 - \phi_1 B)(1 - B)x_t = (1 + \theta_1 B)a_t\).

Although most ARIMA processes are of low order, the general process can be given using these definitions:

\[\nabla^d = (1 - B)^d\] (the d-order differencing operator)
\[
\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)\] (the p-order AR operator)
\[
\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)\] (the q-order MA operator).

If we generalize the examples given in the previous section, we see that the general ARIMA\((p, d, q)\) process is:

\[(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)(1 - B)^d x_t = (1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)a_t.\]

This now becomes: \(\phi(B)\nabla^d x_t = \theta(B)a_t.\)

In the data analysis chapter of this project, we will discuss various models that exhibit cyclical patterns due to a seasonal effect. In general, for a seasonal pattern of length \(s\), there should be a relationship between \(x_t, x_{t-s}, x_{t-2s}, \ldots\). For example, monthly data (such as sales, rainfall, etc.) has length \(s = 12\). Seasonal differencing may be used, perhaps in addition to nonseasonal differencing. This idea is also useful for daily, weekly, yearly or other time periods that may be appropriate. A purely seasonal process is expressed as ARIMA\((P, D, Q)_s\), where \(P_s\) is the maximal lag length on seasonal AR terms, and \(Q_s\) is the same on seasonal MA terms. For example, an ARIMA\((1, 0, 2)_{12}\) has an AR term at lag \(s = 12\), and MA terms at lags \(s = 12\) and \(s = 24\). The most commonly used seasonal model is ARIMA\((0, 1, 1)_s\). This model is written as: \(z_t = z_{t-s} - \theta_s a_{t-s} + a_t\). If we let \(\nabla_s = 1 - B^s\), then an equivalent form is: \(\nabla_s z_t = (1 - \theta_s B^s)a_t\).

A combined nonseasonal and seasonal process is given by
Suppose we have the nonseasonal part of an ARIMA process given by

\[ \phi(B) \nabla^d x_t = \theta(B) b_t, \]

but the seasonal part is not represented yet. The random shock series is a series \( b_t \) with a seasonal pattern. If this pattern can be represented by AR and MA terms we define:

\[ \nabla^D_s = (1 - B^s)^D \] (the D-order seasonal differencing operator)

\[ \phi(B^s) = (1 - \phi_1 B^s - \phi_2 B^{2s} - \cdots - \phi_p B^{ps}) \] (the P-order seasonal AR operator)

\[ \theta(B^s) = (1 - \theta_1 B^s - \theta_2 B^{2s} - \cdots - \theta_q B^{qs}) \] (the Q-order seasonal MA operator).

Suppose the seasonal behavior of \( b_t \) is described as:

\[ \phi(B^s) \nabla^D_s b_t = \theta(B^s) a_t. \]

If we solve for \( b_t \), we get

\[ b_t = \frac{\theta(B^s)}{\phi(B^s) \nabla^D_s} \cdot a_t. \]

If we substitute for \( b_t \) into (6) we have:

\[ \phi(B^s) \nabla^d x_t = \theta(B) \frac{\theta(B^s)}{\phi(B^s) \nabla^D_s} a_t. \]

Rearranging the terms, we obtain the combined multiplicative seasonal and nonseasonal

ARIMA\((p, d, q)(P, D, Q)_s\) process:

\[ \phi(B^s) \phi(B) \nabla^D_s \nabla^d x_t = \theta(B^s) \theta(B) a_t. \]

For example, ARIMA (2, 1, 0)(0, 1, 1)\(_s\) has a nonseasonal portion \((p, d, q) = (2, 1, 0)\) and a seasonal portion \((P, D, Q)_s = (0, 1, 1)_s\). The model says that after
both nonseasonal and seasonal differencing of degree one \((d = D = 1)\), the data has a nonseasonal AR(2) pattern \((p = 2)\) and a seasonal MA(1)\(_{12}\) pattern \((Q = 1)\). Other cases will be explored later as we investigate the ozone data.

3.2 State-Space Models

We now introduce some useful tools named for the contributions of R. E. Kalman. We will express a dynamic system in a particular form called the state-space representation. Data processing algorithms which separate desired signals from unwanted disturbances or noise are referred to as filters. We will look at the Kalman filter, which is an algorithm for sequentially updating a linear projection for the system (see[3]). The Kalman filter can be applied to different types of problems, such as filtering, interpolation, smoothing and extrapolation of time series. It is used to combine measurement data provided by different measuring devices, each of which has its own type of errors (e.g. in navigation problems: gyroscopes, accelerometers, doppler radar, etc.). It is used in parameter estimation by restating a particular problem with parameters instead of a state vector. Also, this algorithm provides a way to calculate exact finite-sample forecasts, to factor matrix autocovariance-generating functions or spectral densities, and to estimate vector autoregressions with coefficients that change over time. An example of the usage of the Kalman filter is to describe a missile’s location (with an associated random vector) along with an indirect satellite measurement of the location that is affected by a random noise vector.
3.2.1 Definitions

In contrast with our earlier scalar processes, we turn now to a multivariate state-space with a vector-valued model that gives enough information about the state of the system.[13] One difference between state-space and conventional linear model representations is that the state of nature is not assumed to be a constant, but may change with time. This comparison is addressed further in the next section. The main point of state space representation is that knowing $X_t$, known as the state of nature, is enough to predict $X_{t+1}$. This is known as the Markovian property. This is a property possessed by many physical systems. This dynamic feature is incorporated by way of the dynamic (or system) equation:

$$X_{t+1} = F_t X_t + V_t.$$  \hspace{1cm} (7)

Here, $F_t$ is a known quantity, and the dynamic equation error, $V_t$, has zero mean and known variance. This is one of the fundamental state equations, which may or may not be stationary. This equation describes the evolution of a state $X_t$ (an unobservable quantity represented as a $v \times 1$ vector) of a system at time $t$ in terms of a known sequence of $v \times v$ matrices $F_1, F_2, \ldots$ and the sequence of random vectors $X_1, V_1, V_2, \ldots$.

We proceed from the state variables to the observation variables, since in order for a vector-valued time series model to be represented in linear state-space form, we let the data $Y_t, Y_{t-1}, \ldots, Y_1$ denote the observed values of a variable of interest at times $t, t-1, \ldots, 1$. We assume that $Y_t$ depends on $X_t$, and we would like to make inferences about $X_t$, whose dimension may be different from the dimension of $Y_t$. We assume the relationship between $Y_t$ and $X_t$ is linear and satisfies a vector
equation of the form:

\[ Y_t = G_t X_t + W_t, \]  

where \( G_t \) is a known quantity. This is the observation equation, one of the state equations. The observation error, \( W_t \), is assumed to be normally distributed with mean zero and a known variance. This equation then defines a sequence of observations \( Y_t \) obtained by applying a linear transformation to \( X_t \) and adding a random noise vector \( W_t \).

As mentioned previously, we may think of the location of a missile at time \( t \) as \( X_t \). The observation equation gives the actual observation \( Y_t \) at time \( t \). We look at \( Y_t \), which may represent a radar track which does not fully represent \( X_t \), to find \( X_t \) indirectly, since it is not directly observable.

Another example is that of tracking a satellite's orbit around the earth. The unknown state of nature, \( X_t \), could be the position and speed of the satellite at time \( t \), with respect to a spherical coordinate system with the origin at the center of the earth. Since we cannot measure these quantities directly, we may obtain measurements of distance to the satellite and the accompanying angles of measurement from tracking stations around the earth; these are the \( Y_t \)'s. The principles of geometry, mapping \( Y_t \) into \( X_t \), would be incorporated in \( G_t \), while \( W_t \) would reflect the measurement error. \( F_t \) would indicate how the position and speed change in time according to physical laws governing orbiting bodies, while \( V_t \) would allow for deviations from these laws due to such factors as nonuniformity of the earth's gravitational field, etc.

It is possible to formulate a great variety of time series models in state-space
form. When a state-space representation can be found, the simple structure of the
system equation permits relatively simple analysis of the process \( \{X_t\} \). The behavior
of \( \{Y_t\} \) is then easy to determine from the observation equation.

We make the following assumptions:

1. \( F_1, F_2, \ldots \) is a sequence of known \( v \times v \) matrices: \( v = \text{dim} X \).
2. \( G_1, G_2, \ldots \) is a sequence of known \( w \times v \) matrices: \( w = \text{dim} Y \).
3. \( \{X_1, (V'_t, W'_t)\} \) is an orthogonal sequence of random vectors with finite second
   moments. The random vectors \( X \) and \( Y \) are orthogonal, written \( X \perp Y \), if the
   matrix \( E(XY_t) = 0 \). Note that the vectors need not be the same length to be
   orthogonal.
4. \( EW_t = 0 \) and \( EW_t = 0 \) for all \( t \).
5. \( E(V_tV'_t) = Q_t, E(W_tW'_t) = R_t, E(V_tW'_t) = S_t \), where \( \{Q_t\}, \{R_t\}, \text{and} \{S_t\} \)
   are specified sequences of \( v \times v \), \( w \times w \), and \( v \times w \) matrices, respectively. \( Q_t, R_t, \text{and} \)
   \( S_t \) are covariances of the noise process.

Since \( X_t \) and \( Y_t \) have functional forms, it follows from the third assumption
that we have the orthogonality relations, \( V_t \perp X_s, V_t \perp Y_s, W_t \perp X_s, \) and \( W_t \perp Y_s, 1 \leq s < t \). Also, when the matrices are independent of \( t \), the subscripts will be
suppressed. The processes are related to linear time-invariant filters.

**Definition.** The process \( \{Y_t, t = 0, \pm 1, \ldots \} \) is said to be obtained from \( \{X_t, t = 0, \pm 1, \ldots \} \) by application of the linear filter \( \{c_{t,k}, t, k = 0, \pm 1, \ldots \} \) if

\[
Y_t = \sum_{k=-\infty}^{\infty} c_{t,k} X_k, t = 0, \pm 1, \ldots
\]

The coefficients \( c_{t,k} \) are called the weights of the filter. [3]
The filter is time-invariant if \( c_{t,k} \) depends only on \( t - k \). We have the following (see [5]):

**Theorem 3.** A linear time-invariant stochastic state-space system has the description given by state equations (7) and (8) and hence:

\[
X_{t'} = F^{t'-t}X_t + \sum_{i=t}^{t'-1} F^{t'-i-1}V_i = F^{t'-t}X_t + \sum_{j=1}^{t'-t} F^{t'-1}V_{t'-j},
\]

where \( t > t' \), and \( \sum_{j=1}^{0} \) is interpreted as an empty sum.

### 3.2.2 Comparison to ARMA Processes

There is an intimate connection between ARMA and state-space models; in fact, they can be shown to be equivalent. The following theorem shows that for each ARMA system, there exists a state-space system that can reproduce its behavior. In other words, each system can dynamically realize a system of the other type (see [5]).

**Theorem 4.** Given any ARMA system on \( \mathcal{T} = \{k_0, \infty\} \),

\[
y_k + a_{k,1}y_{k-1} + \ldots + a_{k,n_1}y_{k-n_1} = c_{k,0}v_k + \ldots + c_{k,n_2}v_{k-n_2}
\]

which relates the sequences \( y = \{y_k \in \mathbb{R}^p; k \in \mathcal{T}\} \) and \( \{v_k \in \mathbb{R}^m; k \in \mathcal{T}\} \), with initial conditions at \( k_0 \), there exists a linear state-space system

\[
X_{k+1} = F_kX_k + G_kV_k, \quad (9)
\]

\[
\bar{v}_k = H_kX_k + D_kV_k, k \geq k_0 \quad (10)
\]

and an initial state \( x_{k_0} \in X = \mathbb{R}^n \) for some \( n \in Z_1 \), where \( Z_1 \) represents integers \( \geq 1 \), such that \( \bar{y}_k = y_k \) for all \( k \geq k_0 \). This means that there exists a
state-space system dynamically realizing the ARMA system on \( [k_0, \infty) \). Conversely, whenever \( \{H_{k+1}, F_{k+1}; k \geq k_0\} \) are uniformly observable with observability index \( \delta \), i.e., whenever \( [H_{k+1}^\top, (H_{k+2}F_{k+1})^\top, \ldots, (H_{k+\delta}[F]_{k+1})^\top]^\top \) has full rank for all \( k \geq k_0 \), any system of the form (9) can be dynamically realized by an ARMA system with \( \max(n_1, n_2) \leq \delta + 1 \).

These statements are true for stochastic systems when \((y, v)\), or \((y, x, v)\), respectively, are defined on some probability space \((\Omega, P)\). If the matrices appearing in either case are time invariant, that is, if either the ARMA system or the state-space system is time invariant, then there exists a corresponding time invariant dynamical realization of the other system.

### 3.2.3 The Kalman Recursions

We next consider some problems associated with the state-space model. Here we would like the best (minimum mean-square error) estimates of the state vector \( X_t \) in terms of the observations \( Y_1, Y_2, \ldots \) and a random vector \( Y_0 \), under the conditions \( Y_0 \perp V_t \) and \( Y_0 \perp W_t \) for all \( t \) (that is, the matrices \( E(\{Y_0V'_t\}) \) and \( E(\{Y_0W'_t\}) \) are zero). The vector \( Y_0 \) depends on the type of estimates required, and in many applications, it is convenient to take \( Y_0 = (1, 1, \ldots, 1)' \). If we try to estimate \( X_t \) in terms of the data \( Y_0, \ldots, Y_{t-1} \) known before time \( t \), this is called a prediction problem. If we estimate \( X_t \) in terms of the data known as of time \( t: Y_0, \ldots, Y_t \), we have a filtering problem, and if we estimate \( X_t \) in terms of \( Y_0, \ldots, Y_n \), where \( n > t \), we have a smoothing problem. Each of these problems can be solved using Kalman recursions. We will look at the prediction problem.
**Definition.** The best one-step linear predictor $\hat{X}_t$, of $X_t = (X_{t1}, \ldots, X_{tv})'$ is the random vector whose $i$th component, $i = 1, \ldots, v$, is the best linear predictor (minimizing prediction error variance, as discussed earlier in the discussion of Hilbert spaces) of $X_{ti}$ in terms of all the components of the $t$ vectors, $Y_0, Y_1, \ldots, Y_{t-1}$.

**Definition.** Generally, the best estimator $X_{t|i}^k$ of $X_t$ is defined to be the random vector whose $i$th component is the best linear estimator of $X_{ti}$ in terms of all the components of $Y_0, Y_1, \ldots, Y_k$. This notation covers the problems of prediction, filtering and smoothing with $k = t - 1$, $t$, and $n$ respectively. In particular, we denote $\hat{X}_t = X_{t|t-1}$. The corresponding error covariance matrices are defined to be $\Omega_{t|i} = E[(X_t - X_{t|i})(X_t - X_{t|i})']$.

Using Hilbert space techniques, we may also define $P(X|Y_0, \ldots, Y_t)$ to be the random $v$-vector whose $i$th component is the orthogonal projection $P(X_i|S)$ of $X$ onto the span, $S$, of all the components $Y_0, \ldots, Y_t$. We shall abbreviate $P(X|Y_0, \ldots, Y_t)$ by writing it as just $P_t(X)$. By the definition of $P(X_i|S)$, $P_t(X)$ is the unique random vector with components in $S$ such that $[X - P_t(X)] \perp Y_s$, $s = 0, \ldots, t$.

We define a generalized inverse of $A$ as a matrix $A^{-1}$ such that $AA^{-1}A = A$. Since $P$ is linear,

$$P(X|Y) = E(XY') [E(YY')]^{-1},$$

where $[E(YY')]^{-1}$ is any generalized inverse of $E(YY')$ (see [3]).

The best estimator $X_{t|i}s = P_s(X_t)$ and in particular, the best one-step linear predictor $\hat{X}_t = P_{t-1}(X_t)$. Now, suppose that our state equations (7) and (8) hold,

$$EU_t = E \begin{bmatrix} V_t \\ W_t \end{bmatrix} = 0, \quad E(U_tU'_t) = \begin{bmatrix} Q_t & S_t \\ S'_t & R_t \end{bmatrix},$$

from the assumptions in the prior
section, where \( U_t \) is defined as the combined noise vector for the two processes, \( X_1, U_1, U_2, \ldots \) are uncorrelated, and \( Y_0 \perp V_t \) and \( Y_0 \perp W_t \) for all \( t \).

We will now use orthogonality concerning a sequence \( \{I_t\} \). We define the innovations \( I_t \) as:

\[
I_0 = Y_0 \text{ and } I_t = Y_t - P_{t-1}Y_t.
\]

(Note: Innovations are not the information set \( I_t \) of section 2.5). By the definition of the one-step predictor, \( \hat{Y}_t = P_{t-1}Y_t \), so \( I_t = Y_t - \hat{Y}_t \), which equals \( Y_t - G_t\hat{X}_t \) by the observation equation (8), with \( \hat{W}_t = 0 \). Using the same equation, \( I_t \) is equal to \( G_tX_t + W_t - G_t\hat{X}_t = G_t(X_t - \hat{X}_t) + W_t \).

By the definition of the projection \( P(X_t|S) \), the sequence \( \{I_t\} \) is orthogonal. Now, if in general, \( M_1 \) and \( M_2 \) are two closed subspaces of a Hilbert space \( H \), and if \( M_1 \perp M_2 \), (that is, \( x \perp y \) for all \( x \in M_1 \) and \( y \in M_2 \)), then it follows that \( P_{M_1 \oplus M_2} = P_{M_1} + P_{M_2} \), where \( M_1 \oplus M_2 \) is the closed subspace \( \{x + y : x \in M_1 \text{ and } y \in M_2\} \). Thus, the following relation holds:

\[
P_t(\cdot) = P_{t-1}(\cdot) + P(\cdot|I_t).
\]

Now, from the projection definition, we find that

\[
\hat{X}_{t+1} = P_tX_{t+1} = P_{t-1}X_{t+1} + P(X_{t+1}|I_t).
\]

Since

\[
P_{t-1}X_{t+1} = P_{t-1}(F_tX_t + V_t) = F\hat{X}_t,
\]

and thus

\[
\hat{X}_{t+1} = F_t\hat{X}_t + K_t(Y_t - G_t\hat{X}_t),
\]

33
where we need to find an “optimal gain matrix” $K$. To do so, we define

$$\Delta_t = E(I_t I_t^t) = E\{[G_t(X_t - \hat{X}_t) + W_t][(X_t - \hat{X}_t)'G_t' + W_t']\}$$

which becomes $G_t \Omega_t G_t' + R_t$, where $\Omega_t$ is the error covariance matrix. Also, we define

$$\Theta_t = E(X_{t+1} I_t^t) = E\{(F_t X_t + V_t)[(X_t - \hat{X}_t)'G_t' + W_t']\}$$

which becomes $F_t \Omega_t G_t' + S_t$. We can now see that due to (7)

$$\hat{X}_{t+1} = F_t \hat{X}_t + \Theta_t \Delta_t^{-1} I_t,$$ ...

where $\Delta_t^{-1}$ is any generalized inverse of $\Delta_t$, and now we have found $K = \Theta_t \Delta_t^{-1}$.

So, to update for $\hat{X}_{t+1}$, we get $F_t \hat{X}_t$, and then look for the best predictor based on the difference between $Y_t$ and its prediction, $G_t \hat{X}_t$.

The one-step predictors, $\hat{X}_t = P_{t-1}X_t$, and matrices $\Omega_t$ are uniquely determined by the initial conditions:

$$\hat{X}_1 = P(X_1|Y_0),$$ ...

Here, $\Pi_{t+1}$ is the uncorrelated total error covariance, and is equal to $E(\bar{X}_t \bar{X}_t)$.

The covariance of the optimal predictor is given by $\Psi_t$.

To evaluate $\Delta_t, \Theta_t$ and $\Omega_t$ recursively, we notice that

$$\Omega_{t+1} = E(X_{t+1}X_{t+1}^t) - E(\bar{X}_{t+1} \bar{X}_{t+1}^t) = \Pi_{t+1} - \Psi_{t+1}$$
from the initial conditions (14-16). From the dynamic equation (7), and (11), \( \Pi_{t+1} = F_t \Pi_t F'_t + Q_t \) and \( \Psi_{t+1} = F_t \Psi_t F'_t + \Theta_t \Delta_t^{-1} \Theta'_t \).

The recursions are given by:

\[
\begin{align*}
\Delta_t &= G_t \Pi_t G'_t + R_t \\
\Theta_t &= F_t \Pi_t G'_t + S_t \\
\Pi_{t+1} &= F_t \Pi_t F'_t + Q_t \\
\Psi_{t+1} &= F_t \Psi_t F'_t + \Theta_t \Delta_t^{-1} \Theta'_t \\
\Omega_{t+1} &= \Pi_{t+1} - \Psi_{t+1}.
\end{align*}
\]

So, given \( \Omega_t \), we can calculate \( \Delta_t \) and \( \Theta_t \) by (17) and (18). With \( \Pi_t \) from the state equations, we can easily obtain \( \Pi_{t+1} \). And with \( \Theta_t \), we can obtain \( \Psi_t \).

Example:

We now present an example[3] of a non-stationary state-space model defined by

\[ X_{t+1} = 2X_t + V_t \quad \text{and} \quad Y_t = X_t + W_t, \quad t = 1, 2, \ldots \],

where \( X_1 = 1, Y_0 = 1 \), and

\[
U_t = \begin{bmatrix} V_t \\ W_t \end{bmatrix} \sim WN \left( 0, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).
\]

\( U_t \) is the error matrix containing the dynamic equation error and the observation error. \( V_t \) and \( W_t \) are independent white noise processes.

This model essentially doubles each value of \( X_{t+1} \) in the system equation relative to its prior value. The observation equation is basically unchanged, so that it appears like an exponential function. So, \( F_t = 2, G_t = 1, Q_t = 1, R_t = 1 \), and \( S_t = 0 \). We would like state estimates in terms of the \( y \)'s, and therefore choose \( y_0 = 1 \). We also let \( \Pi_1 = \Psi_1 = 1, \Omega_1 = 0 \).
In this case, we have:

\[ \Delta_t = \Omega_t + 1 \]  
\[ \Theta_t = 2\Omega_t \]  
\[ \Pi_{t+1} = 4\Pi_t + 1 \]

(This is a geometric series with the first term, \( a_1 = \Pi_1 = 1 \), and common ratio, \( r = 4 \). The sum of the first \( n \) terms is, then: \( \frac{a_1 r^n - a_1}{r - 1} \). Therefore, \( \Pi_{t+1} = \frac{1}{3}(4^{t+1} - 1) \).)

Also:

\[ \Psi_{t+1} = 4\Psi_t + \Theta_t \Delta_t^{-1} \Theta_t' = 4\Psi_t + 4\Theta_t^2 \Delta_t^{-1} = 4\Psi_t + \frac{4\Omega_t^2}{1 + \Omega_t} \]  
\[ \Omega_{t+1} = \Pi_{t+1} - \Psi_{t+1} = \frac{1}{3}(4^{t+1} - 1) - \Psi_{t+1}. \]

If we set \( \Psi_{t+1} = -\Omega_{t+1} + \frac{1}{3}(4^{t+1} - 1) \), then from (25):

\[-\Omega_{t+1} + \frac{1}{3}(4^{t+1} - 1) = 4[-\Omega_t + \frac{1}{3}(4^t - 1)] + \frac{4\Omega_t^2}{1 + \Omega_t} \]

\[-\Omega_{t+1} + \frac{4^{t+1} - 1}{3} = -4\Omega_t + \frac{4(4^{t+1}) - 4}{3} + \frac{4\Omega_t^2}{1 + \Omega_t} \]

\[-\Omega_{t+1} = -4\Omega_t - 1 + \frac{4\Omega_t^2}{1 + \Omega_t} \]

\[ \Omega_{t+1} = \frac{(1 + \Omega_t)(1 + 4\Omega_t) - 4\Omega_t^2}{1 + \Omega_t}. \]

This yields the recursion:

\[ \Omega_{t+1} = \frac{1 + 5\Omega_t}{1 + \Omega_t}. \]
From this it can be shown that:

\[
\Omega_t = \frac{4 + 2\sqrt{5} - (\sqrt{5} - 1)c^{2-t}}{2 + (\sqrt{5} + 3)c^{2-t}}, \quad c = \frac{1}{2}(7 + 3\sqrt{5}) \approx 6.854.
\]

The solution of the recursion equations is, then:

\[
\Delta_t = \Omega_t + 1,
\]

\[
\Theta_t = 2\Omega_t,
\]

\[
\Pi_t = \frac{1}{3}(4^t - 1),
\]

\[
\Psi_t = \frac{1}{3}(4^t - 1) - \Omega_t.
\]

The equations for the estimators and mean square errors can now be found. We see that the one-step predictor of \(X_{t+1}\) satisfies the recursions:

\[
\hat{X}_{t+1} = 2\hat{X}_t + \frac{2\Omega_t}{1 + \Omega_t} (Y_t - \hat{X}_t), \quad \hat{X}_1 = 1,
\]

with mean squared error, \(\Omega_{t+1}\). This compares favorably with the intuitive solution offered earlier, with each new value of \(\hat{X}_{t+1}\) essentially twice that of \(\hat{X}_t\), along with an appropriate adjustment.

The one-step predictor of \(Y_{t+1}\) is \(P_t Y_{t+1} = \hat{X}_{t+1}\). If we define the mean squared error of \(Y_{t+1}\) to be \(\Sigma_t^{(i)} = G_{t+1}\Omega_t^{(i)}G_{t+1}^t + R_{t+1}\), and \(\Omega_t^{(i)} = \Omega_{t+1} + 1\), then \(\Sigma_t^{(i)} = \Omega_{t+1} + 1\). The mean squared error, \(\Omega_t\), of the one-step predictor of the state \(X_t\) converges as \(t \to \infty\). So, then we have:

\[
\Omega_{t|t-1} \to \lim_{t \to \infty} \Omega_t = 2 + \sqrt{5} \approx 4.236.
\]

Kalman recursions are ideally suited to the precise analysis of data with missing values (see[3]). We now look at \(\{Y_{i_1}, \ldots, Y_{i_r}\}\) where \(i_1, i_2, \ldots, i_n\) are positive integers.
such that \( 1 \leq i_1 \leq i_2 \leq \ldots \leq i_r \leq n \). The process \( \{Y_t\} \) can be observed at irregular intervals, or equivalently, we may have data such that we have only \( r \) observations out of the \( n \) total possible observations, so that \( n - r \) observations are missing from the sequence \( \{Y_1, \ldots, Y_n\} \)

Example: [3]

We look at an AR(1) series with one missing observation, namely, we have observations \( y_0, y_1, y_3, y_4, y_5 \). Let \( \{Y_t\} \) be the process defined by: \( Y_t = X_t, X_{t+1} = \phi X_t + Z_{t+1} \). A corresponding model for a series with missing observations may be called \( \{Y^*_t\} \) (to distinguish it from one without missing values). This is given by:

\[
X_{t+1} = F_t X_t + V_t, Y^*_t = G^*_t X_t + W^*_t,
\]

where \( F_t = \phi; G^*_t = 1 \) if \( t \neq 2, 0 \) if \( t = 2 \); \( V_t = Z_{t+1}; W^*_t = 0 \) if \( t \neq 2, N_t \) if \( t = 2 \) (where \( \{N_t\} \) is a \( N(0, 1) \) white noise sequence); \( Q_t = \sigma^2; R^*_t = 0 \) if \( t \neq 2, 1 \) if \( t = 2; S^*_t = 0 \); and we assume that \( \{Y^*_t\} \) is stationary. Starting from the initial conditions,

\[
x_1 = 0, \Pi_1 = \Omega_1 = \sigma^2/(1 - \phi^2),
\]

and applying the recursions, we find:

\[
\Theta_t \Delta_t^{-1} = \begin{cases} 
\phi & \text{if } t = 1, 3, 4, 5 \\
0 & \text{if } t = 2,
\end{cases}
\]

\[
\Omega_t = \begin{cases} 
\sigma^2/(1 - \phi^2) & \text{if } t = 1 \\
\sigma^2 & \text{if } t = 3 \\
\sigma^2 & \text{if } t = 2, 4, 5.
\end{cases}
\]
And,

\[ \hat{x}_1 = 0, \hat{x}_2 = \phi y_1, \hat{x}_3 = \phi^2 y_1, \hat{x}_4 = \phi y_3, \hat{x}_5 = \phi y_4. \]

It can be shown that:

\[ \hat{y}^*_1 = y^*_2 = 0, \hat{y}^*_3 = \phi^2 y_1, \hat{y}^*_4 = \phi y_3, \hat{y}^*_5 = \phi y_4, \]

with corresponding mean squared errors:

\[ \Sigma^*_1 = \sigma^2/(1 - \phi^2), \Sigma^*_2 = 1, \Sigma^*_3 = \sigma^2(1 + \phi^2), \Sigma^*_4 = \sigma^2, \Sigma^*_5 = \sigma^2. \]

Now, applying \( X_{t+1} = F_t X_t + V_t \), we find that:

\[
P_1 x_2 = P_2 x_2 = \phi y_1, P_3 x_2 = P_4 x_2 = P_5 x_2 = \frac{\phi(y_1 + y_3)}{1 + \phi^2};
\]

\[
\Omega_{2,2} = \sigma^2; \Omega_{2,3} = \phi \sigma^2; \Omega_{2,t} = 0, t \geq 4.
\]

Also, \( \Omega_{2|1} = \Omega_{2|2} = \sigma^2; \Omega_{2|t} = \frac{\sigma^2}{1 + \sigma^2}, t \geq 3. \)

Since \( E(V_t W_t^*) = S_t = 0, t = 1, \ldots, n \), we find from the fact that

\[ P(Y_t|Y_0, Y_1, \ldots, Y_n) = G_t P(X_t|Y_0^*, Y_1^*, \ldots, Y_n^*) \]

that the minimum mean squared error estimator of the missing value \( y_2 \) is:

\[ P_5 y_2 = P_5 x_2 = \frac{\phi(y_1 + y_3)}{1 + \phi^2}, \]

with mean squared error: \( \Omega_{2|5} = \frac{\sigma^2}{1 + \phi^2}. \)

3.3 Spectral Analysis and the Periodogram

Up until now, we have been concerned with the time domain. We can also analyze the value of a scalar process \( x_t \) as a weighted sum of periodic functions of
the form $\cos(\omega t)$ and $\sin(\omega t)$, where $\omega$ denotes a particular frequency:

$$x_t = \mu + \int_0^\infty \alpha(\omega) \cos(\omega t) d\omega + \int_0^\infty \beta(\omega) \sin(\omega t) d\omega + \varepsilon.$$  

We express $t$ as continuous time in this context, since the theory is then easier to formulate than with discrete time.

We will attempt to determine how important cycles of different frequencies are in accounting for the behavior of $x$. This is known as frequency domain or spectral analysis. The two kinds of analysis are not mutually exclusive. Any covariance-stationary process can be expressed by either the time domain or frequency domain representations. [9]

**Definition.** Let $\{x_t\}_{t=-\infty}^{\infty}$ be a covariance-stationary process with mean $E(x_t) = \mu$ and $j$th autocovariance $E(x_t - \mu)(x_{t-j} - \mu) = \gamma_j$. The time $t$ is discrete time. If the autocovariances are absolutely summable, the autocovariance-generating function is given by:

$$g_x(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j,$$

where $z$ denotes a complex scalar. If the prior equation is divided by $2\pi$ and evaluated at some $z$ represented by $z = e^{-i\omega}$ for a real scalar $\omega$, the result is called the population spectrum of $x$:

$$s_x(\omega) = \frac{1}{2\pi} g_x(e^{-i\omega}) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}.$$  

This spectrum is a function of $\omega$, since given any value of $\omega$ and a sequence of autocovariances: $\{\gamma_j\}_{j=-\infty}^{\infty}$, we can calculate the value of $s_x(\omega)$. According to De Moivre’s Theorem, $e^{-i\omega j} = \cos(\omega j) - i \sin(\omega j)$, meaning that the spectrum can be
\[ s_x(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j \left[ \cos(\omega j) - i \sin(\omega j) \right]. \]

Now, given an observed sample of \( n \) observations denoted \( x_1, x_2, \ldots, x_n \), we can calculate up to \( n - 1 \) sample autocovariances from the formulas:

\[ \hat{\gamma}_j = n^{-1} \sum_{n=n+1}^{n} (x_t - \bar{x})(x_{t-j} - \bar{x}), \quad \text{for } j = 0, 1, 2, \ldots, n - 1, \]

\[ = \hat{\gamma}_{-j}, \quad \text{for } j = -1, -2, \ldots, -n + 1. \]

where \( \bar{x} \) is the sample mean, equal to \( n^{-1} \sum_{t=1}^{n} x_t \).

**Definition.** For any given \( \omega \) we can construct the sample analog of \( s_x(\omega) \), which is known as the sample periodogram:

\[ \hat{s}_x(\omega) = \frac{1}{2\pi} \sum_{j=-n+1}^{n-1} \hat{\gamma}_j e^{-i\omega j}. \]

This can be expressed as:

\[ \hat{s}_x(\omega) = \frac{1}{2\pi} \left[ \hat{\gamma}_0 + 2 \sum_{j=1}^{n-1} \hat{\gamma}_j \cos(\omega j) \right] \]

It can be shown that the area under the periodogram is the sample variance of \( y \):

\[ \int_{-\pi}^{\pi} \hat{s}_x(\omega) d\omega = \hat{\gamma}_0. \]

Like the population spectrum, the sample periodogram is symmetric around \( \omega = 0 \), so that we equivalently write:
\[ \hat{\gamma}_0 = 2 \int_0^\pi \hat{S}_x(\omega) \, d\omega. \]

Now, given any \( n \) observations on a process \( (x_1, x_2, \ldots, x_n) \), there exist frequencies \( \omega_1, \omega_2, \ldots, \omega_m \) and coefficients \( \hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_m, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m \) such that the value for \( x \) at date \( t \) can be expressed as:

\[ x_t = \hat{\mu} + \sum_{j=1}^{m} \{ \hat{\alpha}_j \cos[\omega_j(t-1)] + \hat{\beta}_j \sin[\omega_j(t-1)] \}, \quad (27) \]

where the variable \( \hat{\alpha}_j \cos[\omega_j(t-1)] \) is orthogonal in the sample to \( \hat{\alpha}_k \cos[\omega_k(t-1)] \), for \( j \neq k \), the variable \( \hat{\beta}_j \sin[\omega_j(t-1)] \) is orthogonal to \( \hat{\beta}_k \sin[\omega_k(t-1)] \) for \( j \neq k \), and the variable \( \hat{\alpha}_j \cos[\omega_j(t-1)] \) is orthogonal to \( \hat{\beta}_k \sin[\omega_k(t-1)] \) for all \( j \) and \( k \). Orthogonality is taken in the sense that if \( f \) and \( g \) are orthogonal, then \( \int_0^{2\pi} fg = 0 \).

The sample variance of \( x \) is \( n^{-1} = \sum_{t=1}^{n} (x_t - \bar{x})^2 \), and the portion of this variance that can be attributed to cycles with frequency \( \omega_j \) can be inferred from the sample periodogram \( \hat{S}_x(\omega_j) \).

Now, let us assume that sample size \( n \) is an odd number. Then, \( x_t \) will be expressed in terms of periodic functions with \( m = (n - 1)/2 \) different frequencies in (27). The frequencies \( \omega_1, \omega_2, \ldots, \omega_m \) are specified as follows:

\[ \omega_1 = 2\pi/n, \omega_2 = 4\pi/n, \ldots, \omega_m = 2m\pi/n. \]

Therefore, the highest frequency considered is:

\[ \omega_m = \frac{2(n - 1)\pi}{2n} = \left(1 - \frac{1}{n}\right)\pi < \pi. \]

If we perform an ordinary least squares regression of the value of \( x_t \) on a constant
and on the various cosine and sine terms, we obtain:

\[ x_t = \mu + \sum_{j=1}^{m} \{ \alpha_j \cos[\omega_j(t-1)] + \beta_j \sin[\omega_j(t-1)] \} + u_t. \]

We may view this as a standard regression model of the form:

\[ x_t = \Lambda'Z_t + u_t, \]  

(28)

where

\[ Z_t = [1, \cos(\omega_1(t-1)), \sin(\omega_1(t-1)), \cos(\omega_2(t-1)), \sin(\omega_2(t-1)), \ldots, \cos(\omega_m(t-1)), \sin(\omega_m(t-1))]' \]

\[ \Lambda' = [\mu, \alpha_1, \beta_1, \alpha_2, \beta_2, \ldots, \alpha_m, \beta_m]. \]

Note that \( Z_t \) has \( 2m+1 = n \) elements, so there are as many explanatory variables as observations. It can be shown that the elements of \( Z_t \) are linearly independent, meaning that an ordinary least squares regression of \( x_t \) on \( Z_t \) yields a perfect fit. So, the fitted values for this regression are of the form of (28) with no error term \( u_t \). Also, the coefficients of this regression have the property that \( \frac{1}{2}(\hat{\alpha}_j^2 + \hat{\beta}_j^2) \) represents the portion of the sample variance of \( x \) that can be attributed to cycles with frequency \( \omega_j \). This magnitude \( \frac{1}{2}(\hat{\alpha}_j^2 + \hat{\beta}_j^2) \) further turns out to be proportional to the sample periodogram evaluated at \( \omega_j \). In other words, any observed series \( x_1, x_2, \ldots, x_n \) can be expressed in terms of periodic functions as in (28), and the portion of the sample variance that is due to cycles with \( \omega_j \) can be found from the sample periodogram.

3.4 The Wiener Filter

The problem of parameter estimation is closely related with that of optimal filtering and state estimation. We turn now to the contributions of N. Wiener,
who, during the second world war did his fundamental studies on interpolation, extrapolation and smoothing of time series. To introduce the Wiener filter, we first make the following definitions. Let $x(t)$ be the information-carrying terms (where $t$ represents continuous time), and $n(t)$ be the observation noise. There are three situations: let $x_d$ stand for one particular case (the desired case), where $x(t - \tau)$ is interpolated, $x(t)$ is filtered, and $x(t + \tau)$ is extrapolated, with $\tau > 0$. We will consider the situation where $x(t)$ is filtered. The others can also be approached by these methods. We let $\hat{x}_d(t)$ be the estimation of $x_d(t)$, obtained by means of the filter from $x + n$. Lastly, $e(t) = \hat{x}_d(t) - x_d(t)$ is the error that has to be minimized by choice of filters in the least squares sense. The filter has to be chosen based on a priori knowledge about $x$ and $n$.[6]

Wiener assumed that the signals $x(t)$ and $n(t)$ are stationary, that is, their statistical properties don’t change with time, and that they have known power spectra. The filter is restricted to be linear, time invariant and physically realizable, characterized by a process $h(t)$. Both the Weiner filter and the Kalman filter are considered optimal, and should lead to the same result under appropriate conditions. Assuming that $x(t)$ and $n(t)$ are sample functions of an ergodic (i.e., $\bar{x} \rightarrow E(x_t)$, time average) stochastic process, the criterion of optimality is to minimize the expected squared error, i.e., $\min E[e^2(t)]$. For $x_d(t) = x(t)$ and $\hat{x}_d(t) = \hat{x}(t)$ this results in a minimization of

$$E[e^2(t)] = E[(\hat{x}(t) - x(t))^2]$$

with

$$\hat{x}(t) = \int_{-\infty}^{\infty} h(\theta)y(t - \theta)d\theta.$$
The desired terms \( x(t) \) and the input terms \( u(t) \) make up the output, \( y(t) \). Due to linearity, we find the convolution integral holds:

\[
y(t) = \int_0^\infty h(\theta) u(t - \theta) d\theta + n(t) = h(t) * u(t) + n(t) \quad (29)
\]

An input \( u(t) \) and a disturbance \( n(t) \) are both stochastic, so the output \( y(t) \) is as well. Since \( u(t) \) is stationary, the implication is that for processes with a memory \( h(t) \) of finite duration, \( y(t) \) will be stationary also.

We use a correlation technique, multiplying both sides of the prior equation by \( u(t - \tau) \), and taking the mathematical expectation:

\[
E[u(t - \tau)y(t)] = \int_0^\infty h(\theta)E[u(t - \tau)u(t - \theta)]d\theta + E[u(t - \tau)n(t)].
\]

Defining \( \psi_{uy}(t_1, t_2) = E[U(t_1)Y(t_2)] \) as the cross-correlation function between two processes \( u \) and \( y \), this becomes:

\[
\psi_{uy}(\tau) = \int_0^\infty h(\theta)\psi_{uu}(\tau - \theta)d\theta + 0 = h(\tau) \ast \psi_{uu}(\tau) + 0, \quad (30)
\]

if \( n(t) \) and \( u(t) \) are independent. Equation (29) relates the stochastic functions \( u(t) \) and \( y(t) \); equation (30) is the relation between deterministic functions, describing important characteristics of those stochastic functions.

To determine a function \( h(t) \), we consider the average-squared error:

\[
E = \frac{1}{T} \int_0^T e^2 dt
\]

with

\[
e(t) = y(t) - \int_0^\infty h(\theta)u(t - \theta)d\theta.
\]
Let us assume that \( \hat{h}(\theta) \) is the function that minimizes \( E \). Then any \( h(t) \) can be written as

\[
h(t) = \hat{h}(\theta) + \alpha h_a(\theta)
\]

where \( h_a \) is an arbitrary function with \( h_a(\theta) \equiv 0 \) for \( \theta < 0 \). By definition, \( E \) will be larger for \( \alpha \neq 0 \) than for \( \alpha = 0 \). Therefore,

\[
\frac{\partial E}{\partial \alpha} \bigg|_{\alpha=0} = 0,
\]

or:

\[
\frac{2}{T} \int_0^T \left[ y(t) - \int_0^\infty \hat{h}(\theta)u(t - \theta)d\theta \right] \left[ \int_0^\infty h_a(\tau)u(t - \tau)d\tau \right] dt = 0.
\]

This can be written as:

\[
\int_0^\infty h_a(\tau) \left[ \frac{1}{T} \int_0^T y(t)u(t - \tau)dt - \int_0^\infty \hat{h}(\theta)d\theta \frac{1}{T} \int_0^T u(t - \theta)u(t - \tau)dt \right] d\tau = 0.
\]

Since \( h_a(\tau) \) is an arbitrary function, we may infer that:

\[
\frac{1}{T} \int_0^T y(t)u(t - \tau)dt = \int_0^\infty \hat{h}(\theta)d\theta \frac{1}{T} \int_0^T u(t - \theta)u(t - \tau)dt,
\]

or:

\[
\tilde{\psi}_{y\psi}(\tau, T) = \int_0^\infty \hat{h}(\theta)\tilde{\psi}_{uu}(\tau - \theta, T)d\theta.
\]

The correlation measurement has to be over a finite time interval, resulting in an approximation (indicated by \( \tilde{\cdot} \)) of the true correlation function. As \( T \to \infty \):

\[
\psi_{y\psi}(\tau) = \int_0^\infty \hat{h}(\theta)\psi_{uu}(\tau - \theta)d\theta.
\]

This is known as the \textit{Wiener-Hopf equation}, which is essentially the same as equation (30). For physical realizability of the filter, \( h(t) = 0 \) for \( t < 0 \). The
solution of this equation for $h(\theta)$ is in general not simple. Usually it is done by transformation to the frequency domain; here the condition of physical realizability adds a serious complication.
Chapter 4. Data Analysis

In this section we summarize an analysis of a set of data using the topics discussed. In a cooperative venture between the U. S. Forest Service and the University of California at Davis, ozone measurements were taken hourly from June through October, for a period of four to five years concluding in 1994, at a dozen sites throughout California. The sites ranged from Mt. Lassen in the north to Barton Flats in the south. These two sites also had the lowest and highest overall levels of ozone, respectively. Cyclical patterns were evident, with levels invariably peaking in the afternoon, and with lowest levels occurring just before sunrise.

4.1 Choosing Between Two Models

The ozone measurements for all twelve sites were investigated, and it was decided to use the site at Barton Flats in the San Bernardino National Forest. This was picked since it is the closest to the CSUSB campus, and since it usually has the highest recorded amounts of ozone. At first, ozone level was plotted vs. days (expressed in Julian dates). Due to the large number of hours recorded, it was more useful to focus on smaller windows of time, such as a month. We focused on July of 1994, which had the highest mean amount of ozone for the months recorded there, although there were a few isolated higher months found at other sites.

We next obtained the ACF and PACF plots (Appendices C and D) and found that, along with nearly all of the sites, it seemed to fit the patterns of an AR(2) model established earlier. That is, the ACF decayed exponentially and the PACF had spikes through the second lag, and was close to zero afterwards. The parameters obtained from SYSTAT were $\phi_1 = 1.467$ and $\phi_2 = -0.484$. 
However, since levels of ozone exhibit a daily pattern, we looked at a seasonal model for a 24 hour “season”, ARIMA(1,0,0)(1,0,0)_{24}. The $\phi_1$ parameter had a value of 0.976, and the seasonal parameter was 0.515. We used the model: 

$$(1 - \phi_1)(1 - \phi_{24})x_t = a_t,$$

and it also seemed to fit the data fairly well. Then we looked at the residuals obtained by subtracting predicted values from observed values. In both cases, the residuals had a zero mean, as expected. Appendix E shows the residuals of the AR(2) model that had only two spikes near 50, where the predicted value was incorrect by that amount. Appendix F has the residuals of the seasonal model. One spike on this graph exceeded 50, however, it still wasn’t clear which model had the best fit. We then decided to compare mean squared residuals, to take into account the different number of observations. The seasonal model had an average of 91.969 squared residuals, compared to 104.657 with the AR(2) model. Therefore, the seasonal model appears to be the best model to represent this particular case.

This seems to be substantiated by graphs of the measured ozone as compared to the predicted values for a four day period from July 15 to 18, 1994. This time period was chosen because the high ozone values and strong fluctuations might have tended to discourage a good fit by any model. Also, narrowing the period gives us a better view of the graphs. Whereas the AR(2) model has trouble forecasting the level at the beginning of day 198 (July 17) in Appendix G, it seems as if the seasonal model in Appendix H has few misses in this period of time.

### 4.2 A Kalman Filter Model

Finally, we generated a list of the monthly highs for each of the sites and chose to investigate the Barton Flats site again by using the Kalman Filter. This data set
showed characteristics of a stationary AR(2) model, given by $x_{t+1} = \phi_1 x_t + \phi_2 x_{t-1} + \varepsilon_t$. This model can be written in state-space form:

$$
\begin{pmatrix}
  x_{t+1} \\
  x_t
\end{pmatrix} =
\begin{pmatrix}
  \phi_1 & \phi_2 \\
  1 & 0
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  x_{t-1}
\end{pmatrix} +
\begin{pmatrix}
  \varepsilon_t \\
  0
\end{pmatrix}
$$

with observation equations:

$$
\begin{pmatrix}
  y_t \\
  y_{t-1}
\end{pmatrix} =
\begin{pmatrix}
  G
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  x_{t-1}
\end{pmatrix}
$$

where $G$ is used to model missing data, and is equal to either

$$
P1 = \begin{pmatrix}
1 & 0 \\
0 & 0
\end{pmatrix}, \quad P2 = \begin{pmatrix}
0 & 0 \\
0 & 1
\end{pmatrix},
$$

or the identity or zero matrices. Since there is no randomness in the model, $W_t = 0$.

The computer package MINITAB was used, and during the months that ozone levels were not available, a background value of 50 parts per billion was substituted. For the observation equation (8), the matrix for $G$ was first inputted, consisting of 2×2 identity matrices used during the main growing season, zero matrices during the non-growing season when we didn’t get anything observable, as well as $P1$ and $P2$, used to transition from one to another in June and October. Values of $\phi_1$ and $\phi_2$ were found to be 0.8354 and -0.1781, respectively. These values were included in the $F$ matrix used in the dynamic equation (7). The $R$ and $S$ matrices are zero, since they are covariances of the noise process: $E(W_t W_t')$ and $E(V_t W_t')$ respectively, where $V_t$ is the dynamic equation error, and $W_t$ is the observation error, both with zero mean.

By use of what are known as Yule-Walker equations for stationarity, a value of $\sigma^2_e$ was found to be 2069.73, and $\sigma^2_x = 4299.42$ was obtained from the data. These values
make up the II matrix (19), which remains constant and equals $\sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, where $\rho = \frac{\phi_1 \sigma^2}{1-\phi_2}$. $\sigma^2$ is the only non-zero component of the $Q$ matrix. As with equation (17), the $\Delta$ matrix is found to be equal to

$$Q = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 0 \end{pmatrix}$$

during the growing season, and then,

$$\Delta^{-1} = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 0 \end{pmatrix}.$$ 

During the non-growing season, when $\Delta$ is the zero matrix, we define $\Delta^{-1}$ to be the same. By way of (18), $\Theta$ is also found. The optimal gain matrix $K$ is obtained by $\Theta \Delta^{-1}$.

After the recursions went through 27 cycles, a results page was printed which showed that the predicted values improved towards the end of the growing or non-growing cycles, but had a tough time explaining the transition between them. If 12-month data had been obtained, it appears likely that the recursions would have done a better job of prediction.
APPENDIX A

Autocorrelation Plot
AR(1) Process

Correlation

Lag
APPENDIX B
Partial Autocorrelation Plot
AR(1) Process
APPENDIX C
Autocorrelation Plot
Barton Flats, July 1994

Correlation

Lag

0 10 20 30
APPENDIX D
Partial Autocorrelation Plot
Barton Flats, July 1994
APPENDIX E
RESIDUALS OF AR(2) MODEL

![Graph of residuals over time](image)
APPENDIX F
RESIDUALS OF SEASONAL MODEL

![Graph showing residuals of a seasonal model over time. The x-axis represents time from 180 to 205, and the y-axis represents residuals from -75 to 50. The graph displays a series of fluctuating residuals with some distinct peaks and troughs.]
APPENDIX G

GRAPH OF OBSERVED VS. PREDICTED OZONE
AR(2) MODEL

Value

TIME
APPENDIX H
GRAPH OF OBSERVED VS. PREDICTED OZONE
SEASONAL MODEL

![Graph of Observed vs. Predicted Ozone Seasonal Model](image)

- Value
- Time

**Legend:**
- Solid line: Observed Ozone Levels
- Dashed line: Predicted Ozone Levels
BIBLIOGRAPHY


