

MULTIVARIATE TIME SERIES MODELING  
AND FORECASTING OF WINNIPEG'S  
ELECTRICAL LOAD - TEMPERATURE  
RELATIONSHIP

BY

SCOTT KLIPPENSTEIN

A Practicum Submitted to the Faculty of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Statistics

University of Manitoba

Winnipeg, Manitoba

©Scott Klippenstein, June 2005

**THE UNIVERSITY OF MANITOBA**  
**FACULTY OF GRADUATE STUDIES**  
\*\*\*\*\*  
**COPYRIGHT PERMISSION**

**Multivariate Time Series Modeling and Forecasting of Winnipeg's Electrical Load-Temperature Relationship**

**BY**

**Scott Klippenstein**

**A Thesis/Practicum submitted to the Faculty of Graduate Studies of The University of  
Manitoba in partial fulfillment of the requirement of the degree**

**Of**

**M. Sc.**

**Scott Klippenstein © 2005**

**Permission has been granted to the Library of the University of Manitoba to lend or sell copies of this thesis/practicum, to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film, and to University Microfilms Inc. to publish an abstract of this thesis/practicum.**

**This reproduction or copy of this thesis has been made available by authority of the copyright owner solely for the purpose of private study and research, and may only be reproduced and copied as permitted by copyright laws or with express written authorization from the copyright owner.**

## Abstract

Manitoba Hydro has approximately 5000 megawatts of hydroelectric generating capacity. It generates, transmits and distributes almost all the electricity consumed in Manitoba, and also sells and purchases electricity under agreements with neighboring systems in Canada and the United States. Power demand, especially in space heating and cooling, is linked to several weather variables, mainly the air temperature. This process involves fitting univariate time series ARIMA models to the temperature data and then fitting multivariate models to load and temperature. In this practicum, we look at the average temperature models and its extremes at monthly and daily time intervals. Next, we model hourly temperature and examine its behaviour throughout the day with comparison to average load's behaviour. Later, we extend this into the multivariate case for load and temperature.

In the problems that were investigated, it is shown that the temperature and load relationship are fairly consistent throughout the day on average for each month. It was shown for a given week, there was no relationship that existed between the two series. On the monthly time intervals, load was strongly dependent on temperature, but decreased as we moved to daily time intervals. This was due to a drop in demand for electrical load on the weekends than during the week. Because of this, partitioning of the daily model was required. Further investigation on other weather variables may provide a better understanding on the noise factors. It was also shown that ARIMA models gave fairly accurate results on all different time intervals. For smaller time units, only a small fraction of temperature data was needed to achieve optimal results. Future research on this area using state space models may provide a more up-to-date forecasting on load and weather or Bayesian VAR models to reduce overparameterization.

## Acknowledgements

I would like to thank my initial supervisor Dr. John F. Brewster for introducing me to this project and my current supervisor, Dr. Liqun Wang for all their help and encouragements. I would like to thank my committee, Alex LeBlanc and Gady Jacoby for their helpful and crucial comments and suggestions throughout the preparation of this practicum. I wish to thank all the Professors at the Department of Statistics who have enriched my knowledge in Statistics through my undergraduate and graduate courses. I'm very grateful for my parents and friends who gave me the needed support that aided me during my studies. Finally, I would like to thank Manitoba Hydro with awarding a research grant to the Institute of Industrial Mathematical Sciences at the University of Manitoba. The financial support was greatly appreciated.

# Contents

<b>1</b>	<b>Introduction and Overview</b>	<b>1</b>
1.1	Preamble . . . . .	1
1.2	Scope of the Practicum . . . . .	2
1.3	Summary Statistics for Temperature Data . . . . .	3
<b>2</b>	<b>Description of the Univariate Box-Jenkins ARIMA Method</b>	<b>10</b>
2.1	Introduction . . . . .	10
2.2	Methods for a Non-Seasonal ARIMA Model . . . . .	11
2.2.1	Model Identification . . . . .	12
2.2.2	Estimation . . . . .	18
2.2.3	Diagnostic Checking . . . . .	20
2.2.4	Forecasting . . . . .	22
2.3	Methods for a Seasonal ARIMA Model . . . . .	23
2.3.1	Identification . . . . .	24
2.3.2	Estimation . . . . .	26
2.3.3	Diagnostic stage . . . . .	27
2.4	Conclusion . . . . .	27
<b>3</b>	<b>Results and Interpretations</b>	<b>28</b>
3.1	Introduction . . . . .	28
3.2	Fitting an ARIMA Model to Monthly Temperature . . . . .	29

3.2.1	An ARIMA Model for Mean Temperature . . . . .	29
3.2.2	An ARIMA Model for Maximum Temperature . . . . .	39
3.2.3	An ARIMA Model for Minimum Temperature . . . . .	50
3.3	ARIMA models for Daily Temperature . . . . .	58
3.3.1	Model for Mean Temperature . . . . .	59
3.3.2	Model for Minimum Temperature . . . . .	67
3.3.3	Model for Maximum Temperature . . . . .	74
3.3.4	Using ARIMA Modeling for a Subset of Daily Temperature . . . . .	82
3.4	Fitting an ARIMA Model for Hourly Data . . . . .	87
3.5	Conclusion . . . . .	94
<b>4</b>	<b>Multivariate Time Series Analysis</b>	<b>95</b>
4.1	Introduction . . . . .	95
4.2	Vector Time Series Models . . . . .	96
4.2.1	Vector Moving Average Model . . . . .	98
4.2.2	Vector Autoregressive AR(p) Model . . . . .	99
4.2.3	Vector Autoregressive Moving Average Process (VARMA(p,q)) . . . . .	100
4.2.4	Nonstationary Vector ARMA Models . . . . .	100
4.2.5	Granger-Causality . . . . .	101
4.2.6	Defining the Order of a VAR Model . . . . .	102
4.3	Fitting Multivariate Models to Monthly Data . . . . .	103
4.4	Fitting Multivariate Models to Daily data . . . . .	120
4.4.1	Fitting Multivariate Models to Weekday Daily Data . . . . .	121
4.4.2	Fitting Multivariate Models to Weekend Daily Data . . . . .	138
4.5	Fitting Multivariate Models to Hourly Data . . . . .	155
4.6	Conclusion . . . . .	156

<b>5 Conclusion</b>	<b>157</b>
<b>A The Data</b>	<b>160</b>
A.1 Monthly and Daily Data . . . . .	161
<b>B SAS and R code</b>	<b>163</b>
B.1 PROC ARIMA code for ARIMA models in SAS . . . . .	163
B.2 Using the ESACF method in SAS . . . . .	165
B.3 Summary Statistics in SAS . . . . .	165
B.4 PROC VARMAX Code for Multivariate Time Series . . . . .	166
B.5 The ACF, PACF and CCF plots in R . . . . .	167
B.6 Spline Smoothing in R . . . . .	168
<b>Bibliography</b>	<b>170</b>

# List of Figures

1.1	Boxplots of mean temperature for each month cumulated over 50 years	3
1.2	Boxplots of extreme temperature for each month cumulated for 50 years	4
1.3	Temperature varying throughout the day for each month . . . . .	6
1.4	Temperature varying throughout the day for each month . . . . .	7
1.5	Load behaviour throughout the day . . . . .	8
3.1	Time series plot of the monthly mean temperature . . . . .	30
3.2	Sample ACF and PACF plots of the mean temp . . . . .	31
3.3	Time series plot of the monthly mean temperature after seasonally diff	31
3.4	Sample ACF and PACF plots of the mean temp seasonal diff of 12 . .	32
3.5	Table of parameter estimates . . . . .	33
3.6	Residual sample ACF and PACF plots of the mean temp . . . . .	33
3.7	Table of parameter estimates . . . . .	34
3.8	Check for white noise for an ARIMA(2,0,0)(0,1,1) . . . . .	35
3.9	Residual sample ACF and PACF plots of the ARIMA(2,0,0)(0,1,1) .	35
3.10	Table of parameter estimates . . . . .	36
3.11	Check for white noise for Nonseasonal MA(1,2,23) and seasonal MA(12)	37
3.12	Forecasted values from January, 2002 to June 2003 . . . . .	38
3.13	Plot of Forecasted (stars) and actual (dots) values from Jan, 2000 to Dec, 2003 . . . . .	39
3.14	Time series plot of the monthly maximum temperature . . . . .	40
3.15	Sample ACF and PACF plots of the max temp . . . . .	41



3.16	Time series plot of the seasonally differenced max temp . . . . .	42
3.17	Sample ACF and PACF plots of the max temp seasonal diff of 12 . .	42
3.18	Table of parameter estimates for seasonal MA(1) . . . . .	43
3.19	Residual sample ACF and PACF plots of the mean temp . . . . .	44
3.20	Table of parameter estimates for a nonseasonal AR(1) . . . . .	44
3.21	Check for white noise for an ARIMA(1,0,0)(0,1,1)S=12 . . . . .	45
3.22	Data set with forecasted values of the adequate model . . . . .	46
3.23	Plot of monthly forecasted (stars) and actual (dots) max temp values from Jan, 2000 to Dec, 2003 . . . . .	50
3.24	Time series plot of the monthly minimum temperature . . . . .	51
3.25	Time series plot of the seasonally differenced min temperature . . . .	52
3.26	Sample ACF and PACF plots of the min temp seasonal diff of 12 . .	52
3.27	Table of parameter estimates for a ARIMA(1,0,0)(0,1,1) . . . . .	53
3.28	Check for white noise for a ARIMA(1,0,0)(0,1,1) . . . . .	54
3.29	Table of parameter estimates for a AR(1,28) and seasonal MA(1) with diff=12 . . . . .	54
3.30	Check for white noise for a nonseasonal AR(1,28) and seasonal MA(1) with diff=12 . . . . .	55
3.31	Forecasting future min temperature from Jan, 2002 to Jun, 2003 . . .	57
3.32	Plot of Forecasted (stars) and actual (dots) values from Jan, 2000 to Dec, 2003 . . . . .	58
3.33	Time series plot of the daily mean temperature . . . . .	59
3.34	Sample ACF and PACF plots of the mean temp of the original series	60
3.35	Time series plot of the daily differenced mean temperature . . . . .	61
3.36	Sample ACF and PACF plots of the mean temp of the seasonal difference	62
3.37	Table of parameter estimates for daily mean temp . . . . .	63
3.38	Check for white noise from the given model . . . . .	64
3.39	Forecast values of mean temp from 20Dec02 to 06Jan03 . . . . .	66

3.40	Plot of daily Forecasted (stars) and actual (dots) mean temp values .	67
3.41	Time series plot of the daily minimum temperature . . . . .	68
3.42	Time series plot of the differenced minimum temperature . . . . .	69
3.43	Sample ACF and PACF plots of the min temp of the seasonal difference	69
3.44	ESACF table of order selection . . . . .	70
3.45	Table of parameter estimates for daily minimum temperature . . . . .	71
3.46	Check for white noise for the above model . . . . .	72
3.47	Plot of daily Forecasted (stars) and actual (dots) min temp values . .	73
3.48	Forecast values of minimum temp from Dec 20/02 to Jan 06/03 . . .	74
3.49	Time series plot of the daily minimum temperature . . . . .	75
3.50	Time series plot of the differenced maximum temperature . . . . .	76
3.51	Sample ACF and PACF plots of the max temp of the seasonal difference	76
3.52	ESACF table of order selection . . . . .	77
3.53	Table of parameter estimates for daily max temp . . . . .	78
3.54	Check for white noise for the above model . . . . .	79
3.55	Forecast values of max temperature from Dec 20/02 to Jan 06/03 . .	81
3.56	Plot of daily Forecasted (stars) and actual (dots) max temp values . .	82
3.57	Table of parameter estimates for daily mean temp of reduced data set	83
3.58	Check for white noise of the reduced data set . . . . .	84
3.59	Forecasted values for the reduced data from Dec 20/02 to Jan 06/03 .	85
3.60	Plot of daily Forecasted (stars) and actual (dots) min temp values . .	86
3.61	Time series plot of the hourly temperature . . . . .	87
3.62	Sample ACF and PACF plots of the hourly temperature for Dec, 2002	88
3.63	Sample ACF and PACF plots of the nonseasonal first diff . . . . .	89
3.64	Sample ACF and PACF plots of the nonseasonal and seasonal diff . .	89
3.65	Table of parameter estimates for hourly temp . . . . .	90
3.66	Check for white noise for hourly temperature . . . . .	91
3.67	Forecasted values for hourly temp . . . . .	92

3.68	Plot of hourly Forecasted (stars) and actual (dots) temp values . . . . .	93
4.1	Time series plot of the monthly mean load and temperature . . . . .	104
4.2	Time series plot of the differenced monthly mean load and temperature	105
4.3	Sample auto and cross correlation of differenced data $Z_t$ . . . . .	107
4.4	Prediction error plot of an ARX(4,12) model . . . . .	115
4.5	Forecast plot with forecasted monthly values for 2003 . . . . .	116
4.6	Prediction error plot of an ARX(3,12) model . . . . .	118
4.7	Daily time plot of the original series . . . . .	120
4.8	Behaviour of load throughout the week . . . . .	121
4.9	Time plots of load and temp without weekends . . . . .	122
4.10	ACF plots of load and temp without weekends . . . . .	123
4.11	Time plots of seasonally differenced load and temp without weekends	124
4.12	Sample ACF and PACF of differenced data $Z_t$ without weekends . . .	125
4.13	Prediction error plot of a re-estimated VAR(6) model . . . . .	134
4.14	Forecast plot of a simplified VAR(6) model . . . . .	136
4.15	Prediction error plot of a unrestricted VAR(6) model . . . . .	137
4.16	Prediction error plot for the unrestricted VAR(6) model . . . . .	138
4.17	Time plots of the daily weekend data for load and temp . . . . .	139
4.18	ACF plots of the daily weekend data for load and temp . . . . .	140
4.19	Time plots of the first difference daily weekend data . . . . .	141
4.20	ACF plots of the daily weekend data $Z_t$ . . . . .	142
4.21	Prediction error plot of a VAR(8) model . . . . .	148
4.22	Forecasts plot of a VAR(8) model . . . . .	149
4.23	Forecasts plot of a VAR(8) model . . . . .	151
4.24	Prediction error plot of a VAR(8) model . . . . .	153
4.25	Time plot of load and temp across one week by the hour . . . . .	155

# List of Tables

2.1	Primary distinguishing characteristics of theoretical acf's and pacf's for stationary process . . . . .	15
2.2	ESACF Table . . . . .	17
2.3	Theoretical ESACF Table for an ARMA(1,2) . . . . .	17
2.4	Summary of Stationarity Conditions . . . . .	19
2.5	Summary of Invertibility Conditions . . . . .	20
2.6	Primary distinguishing characteristics of theoretical acf's and pacf's for stationary process for the seasonal model . . . . .	26
4.1	Sample correlation matrices for the differenced series $Z_t$ . . . . .	106
4.2	Minic Information criterion using AIC method to identify possible VARMA model . . . . .	108
4.3	Granger-Causality test for exogeneous variables . . . . .	109
4.4	Parameter Estimates for ARX(3,3) . . . . .	110
4.5	Parameter Estimates for ARX(3,3) with XL2 dropped . . . . .	111
4.6	Residuals diagnostics for ARX(3,3) . . . . .	112
4.7	Check for white noise in the residuals . . . . .	112
4.8	Parameter Estimates for ARX(4,12) model . . . . .	113
4.9	Residuals diagnostics for ARX(4,12) . . . . .	114
4.10	Forecasted values from March 2002 to December 2003 . . . . .	117
4.11	Forecasted values from March 2002 to December 2003 for the alterna- tive model . . . . .	119

4.12	Minic Information criterion using AIC method to identify possible VARMA model . . . . .	126
4.13	Granger-Causality test for exogeneous variables . . . . .	127
4.14	Parameter Estimates for VAR(6) . . . . .	128
4.15	Residuals diagnostics for VAR(6) . . . . .	129
4.16	Parameter Estimates for a simplified VAR(6) model . . . . .	131
4.17	Residuals diagnostics for a simplified VAR(6) . . . . .	132
4.18	Forecasted values for daily weekday values of load and temp . . . . .	135
4.19	Minic Information criterion using AIC method to identify possible VARMA model . . . . .	143
4.20	Granger-Causality test for exogeneous variables . . . . .	144
4.21	Parameter Estimates for a VAR(8) model . . . . .	145
4.22	Residuals diagnostics for VAR(6) . . . . .	146
4.23	Forecasted values for daily weekend values of load and temp . . . . .	150
4.24	Temperature forecasts from multivariate vs. univariate . . . . .	154
A.1	Partial data set collected by Manitoba Hydro . . . . .	160
A.2	Partial data set of monthly temp . . . . .	161
A.3	Partial data set of daily temperature . . . . .	162

# Chapter 1

## Introduction and Overview

### 1.1 Preamble

Over the last 30 years, Manitoba Hydro has been exporting power to the United States. While Manitoba Hydro currently has an abundance of electricity for export, increasing domestic demand is slowly eating away at that excess capacity. The electric load forecasts is one of the key drivers of Manitoba Hydro's planning activities. The primary purpose of the electric load forecast is to address the key questions of "when, where, why, and how much" electricity will be required on the Manitoba Hydro system to allow Manitoba Hydro to evaluate planning alternatives, as well as to provide a basis for forecasting domestic revenues.

It is very important to be able to accurately forecast the amount of power that Manitoba Hydro requires to produce at a particular point of time. To determine this, we will need to accurately forecast what the future temperatures will be like at a particular time in order to determine the anticipated demand of electricity and maximizing the value of any excess capacity for exports. Underestimating daily temperature forecasts result in costly spot market payments for electricity, or the risks of unscheduled brown-outs or rolling blackouts in markets that are unable to meet the additional, unplanned demand. If longer lead time forecasts are available

for temperature, energy needs and supplies can be estimated more accurately.

The Research Program of Manitoba Hydro awarded a research grant to the Institute of Mathematical Sciences at the University of Manitoba for a project titled "Robust Modeling of the Demand for Electrical Power". The focus is understanding temperature data and its influence on load. This practicum is an added component to Mark Silva's Practicum on this research [10].

## 1.2 Scope of the Practicum

Electrical load forecasting plays a central role in the operation and planning of electric power. The provincial energy estimation, the planning of a new plant, the routine of maintaining and scheduling of daily electrical generation are all dependent on accurate load forecasting in the future. Many factors are influential to the electrical power generation and consumption and one of these factors is weather.

In this practicum we will not be examining the load in detail for the univariate case. For further discussions on load, see Mark Silva's Practicum [10]. Rather, we will look at the most important factor which is related to weather. Among those factors is temperature. Temperature is the most important because it has direct influence on many kinds of electrical consumption, such as air conditioners, heating and refrigeration. Peak demands in both summer and winter typically occur during the periods of extreme weather. Unfortunately, the occurrence and timing of extreme weather is impossible to accurately forecast far in advance.

In this practicum, our primary focus will be examining multivariate time series modeling for load and temperature. We will also examine the relationship between the two series for different time intervals and the univariate ARIMA models for temperature data and its associated probability distribution.

As we mentioned before, this practicum is only a piece of the larger project mentioned in the previous section. The intention is to forecast future temperature values and the corresponding probability distribution for the extreme temperatures, incorporating the required forecasted load needed for Manitoba, and then selling the excess to the neighboring Provinces and States.

### 1.3 Summary Statistics for Temperature Data

In order to understand the probability distribution for the temperature data, we first look at the summary statistics for the average monthly data across a span of 50 years. Figure 1.1 shows the boxplots of each month cumulated from the 50 years, starting from the beginning of 1953 to the end of 2002. A portion of the data set is shown in Appendix A, along with the variables used.

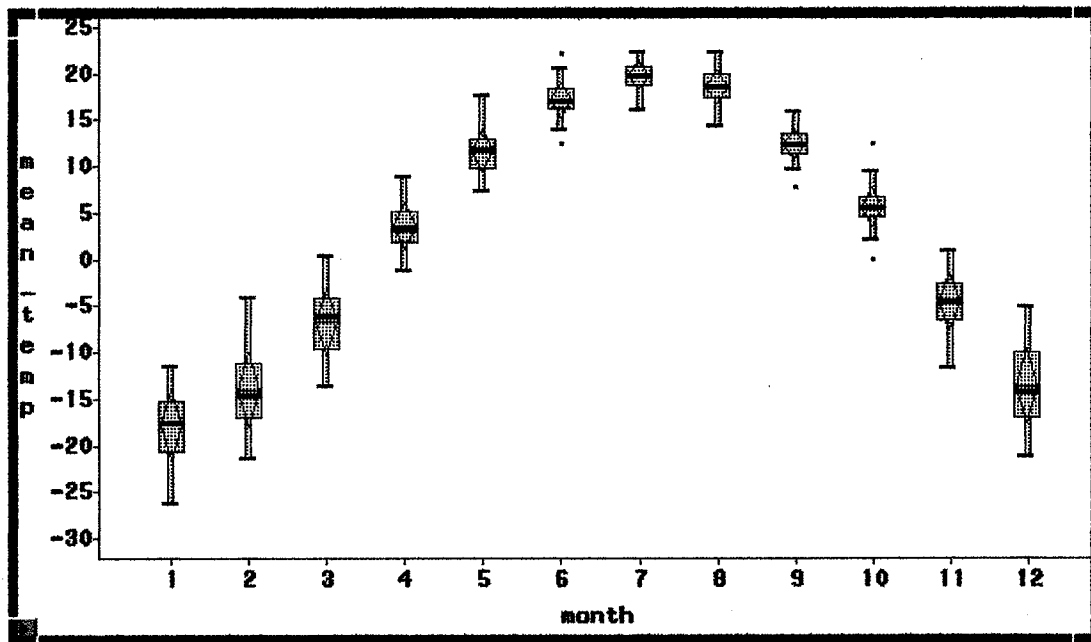


Figure 1.1: Boxplots of mean temperature for each month cumulated over 50 years



By looking at the boxplots, we can get a quick summary of both the center and spread at the middle half of the data with the median in the center. We see that the summer months are the most consistent with the least variability in temperature followed by the most variability given in the winter months.

The Boxplots based on the 5 number summary are useful for comparing several distributions. The box spans the quartiles and shows the spread of the central half of the data. The median is marked within the box, lines extend from the box to the extremes and show the full spread of the data. Since the IQR (interquartile range) of the boxplots are useful for assessing symmetry, its values are not affected by a few extreme outliers at either end of the distribution. It shows that most of the boxplots appear to be fairly symmetric with some light skewness even with the outliers present. This implies that the means are fairly close to the medians which indicates that the distribution is fairly symmetric. Using the mean would be more appropriate because the mean are computed from actual values of the observations and contains more information than the median. Figure 1.2 shows the boxplots of the extreme temperature for each month, cumulated for the 50 years.

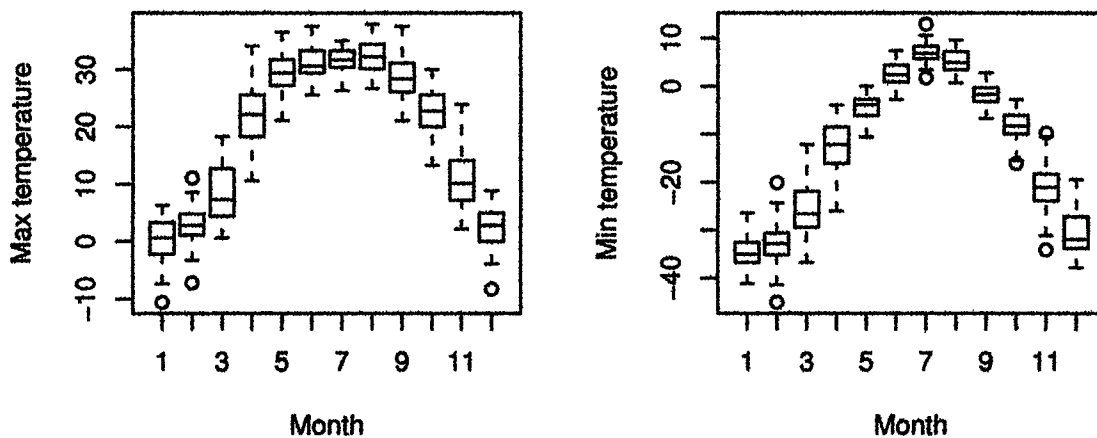


Figure 1.2: Boxplots of extreme temperature for each month cumulated for 50 years

The extreme temperatures, maximum and minimum, appear to follow the same pattern as the average temperature boxplots shown in the previous figure. The box-

plots of max and min temperature appear to have the same pattern as the mean temp, while the spring and fall seasons have the most variability present. While we have a few outliers present for both the extremes, they both appear to be symmetric in nature. What does this mean? Well, we will probably expect to see the model fit the data reasonably well with the exception of more variation in the change of seasons. We see that the monthly temperature data seems to be following a normal distribution, indicating that we can find the weather forecast probability distribution with the given model for the mean temperature and its extremes.

If we now look at the behaviour of temperature throughout the day for each month, we see in Figure 1.3 and 1.4 that temperature has a cyclical pattern for the 24 hours. Generally, the temperature is the coldest around 7 and 8am and the warmest around 3pm. This seems to shift only a little as we progress into the summer months where the coldest temperature is at about 5 or 6am, possibly from the time change. The summer months now has a more broader peak on the curve, meaning it stays warmer for a longer period of time.

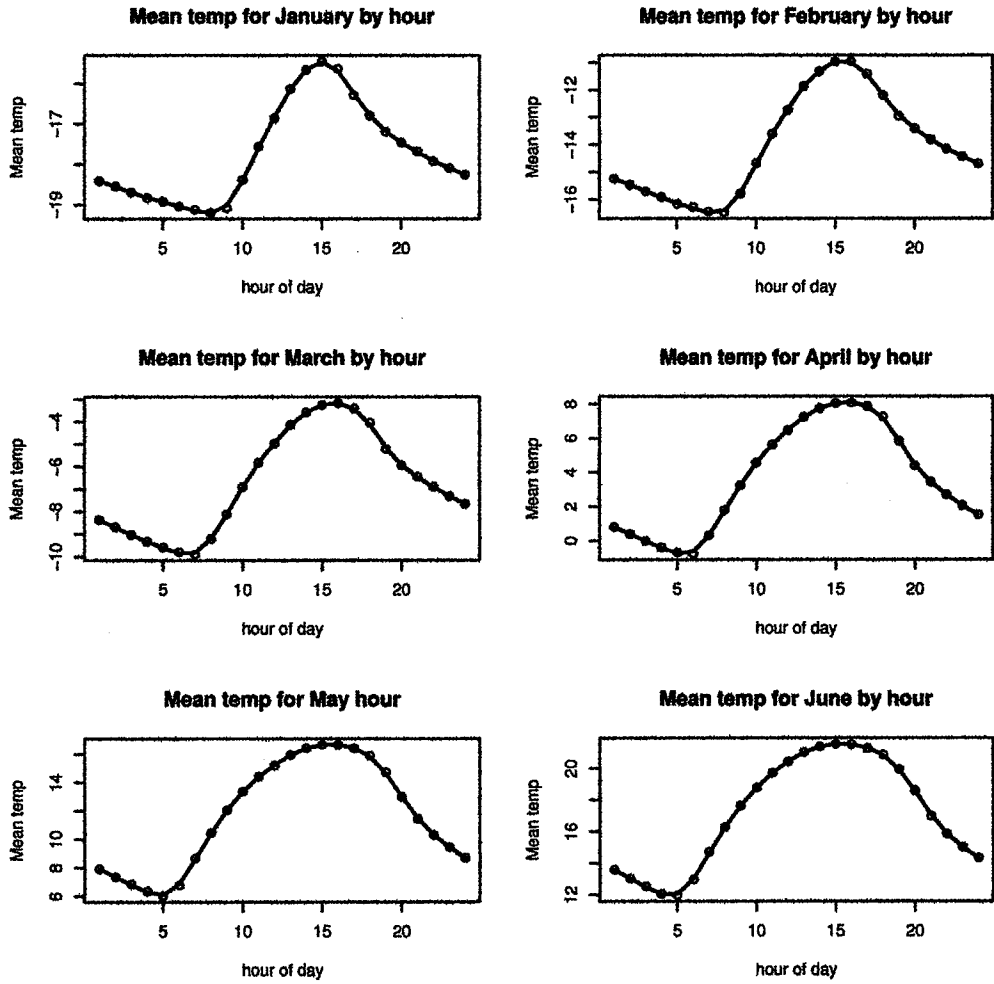


Figure 1.3: Temperature varying throughout the day for each month

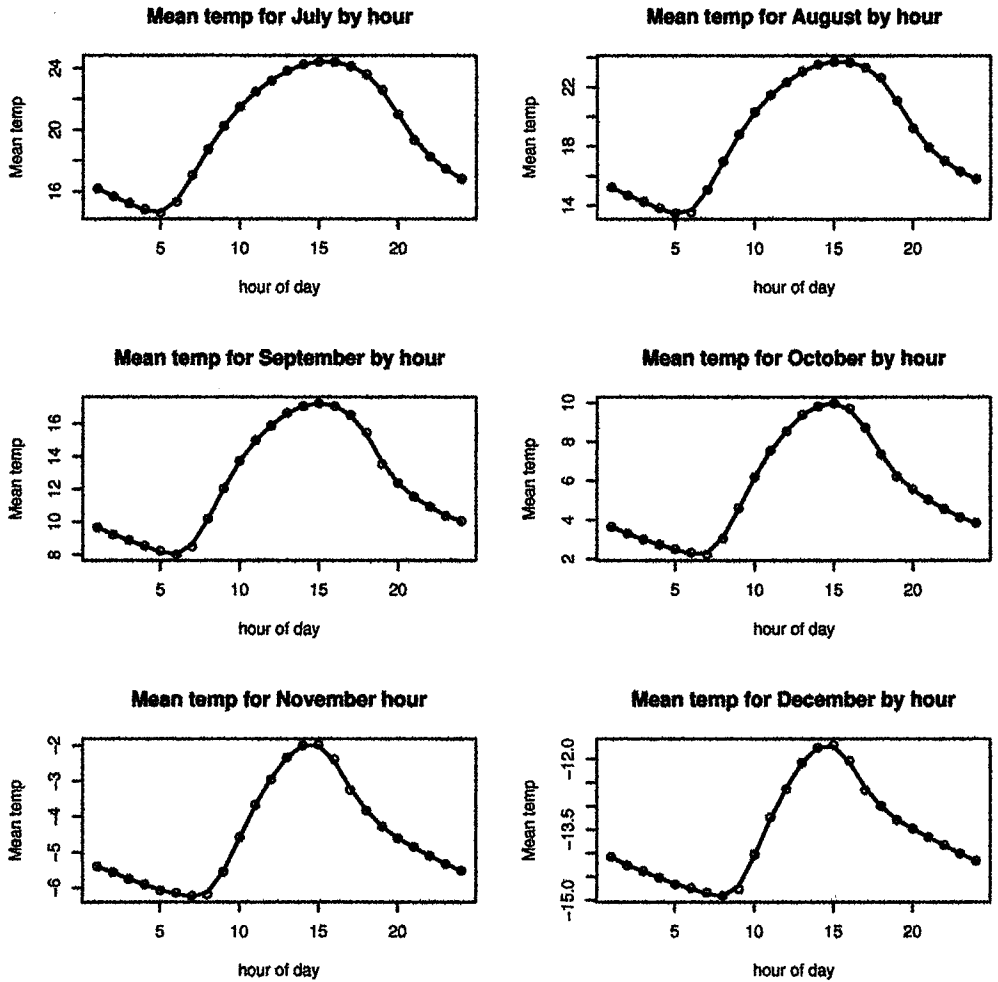


Figure 1.4: Temperature varying throughout the day for each month

How does this relate to the load throughout the day? To get a better idea on how the load behaves throughout the day, Figure 1.5 shows the mean and max plots of load taken from April 01, 1992 to December 31, 2002.

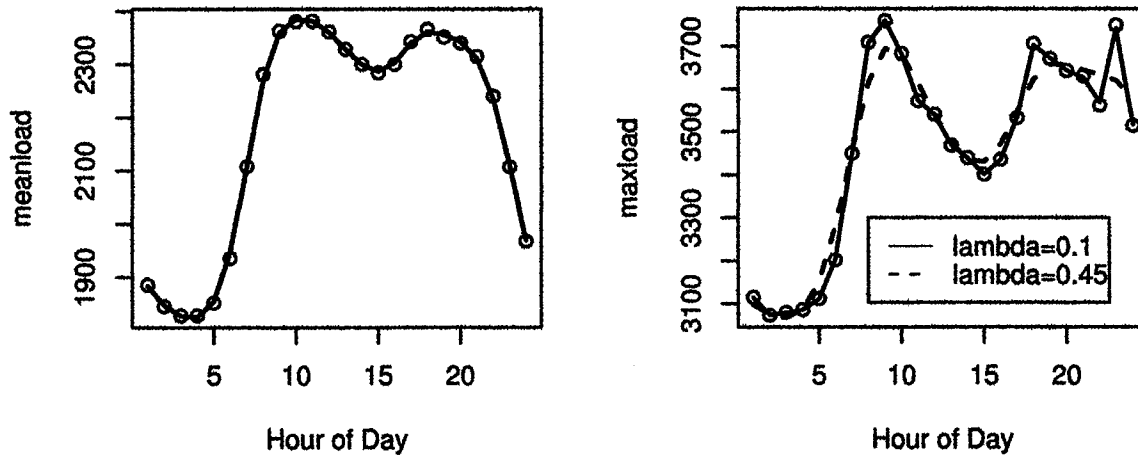


Figure 1.5: Load behaviour throughout the day

The mean load roughly follows a similar pattern/relationship to the mean temperature except the mean load has two peaks, once at about 9am and the other at 6pm. This means that when the temperature is at its lowest around 8am, the load has reached its peak. The max load has a similar effect as the mean load, except for the more significant drop at 3pm and the possible outlier at the 23rd hour. Using a smoothing spline, we can get an overview of the pattern for max load. The R code is given in Appendix B. If we put more emphasis on the goodness of fit (the time plot with the straight lines), we have very rigid interpolated lines for the max load with  $\lambda = 0.1$  and it is difficult to see the pattern. Taking a larger value for lambda gives us a more smoother pattern of the data ( $\lambda = 0.45$  with the dashed lines on the time plot). This shows the same pattern as described before with the two peaks at 9am and 6pm, indicating that most of the load is being influenced during work hours. The load-temperature relationship appears to be consistent throughout the hour of the day. We see that the mean temperature has a seasonal pattern that occurs throughout the day. Do we need to incorporate this seasonal pattern into the model? We will see later how our temperature model performs for hourly data.

In Chapter 2 we introduce the univariate ARIMA model and how it's used. We will look at both nonseasonal and seasonal cases with the three stages of ARIMA modeling. In Chapter 3 we will fit these ARIMA models to the temperature data. We will find a model for monthly, daily and hourly temperature data set. In chapter 3 we also discuss taking a smaller sample of past temperature data which produces equal or better results compared to using the large past temperature data. In Chapter 4 we will expand to a multivariate time series model for load and temperature. We will examine and identify the model between the two time series in order to perform adequate forecasts for each series. We will also try to understand the relationship that exists between them at the monthly, daily and hourly time intervals or frequencies. Finally in Chapter 5 we give some concluding remarks and discuss some of the further work that can be done.

# Chapter 2

## Description of the Univariate Box-Jenkins ARIMA Method

### 2.1 Introduction

One of the goals in this study is to find a model for temperature in the Manitoba Hydro data. In this chapter we will discuss the univariate ARIMA model and how it is used. ARIMA models (Autoregressive Integrated Moving Average) are an algebraic statement describing how temperature observations on a time series are statistically related to past temperature observations and past residual terms from the same time series. In the next two sections we introduce the ARIMA procedure, which is useful for representing data with/without trends in time series data. In Section 2.2 and 2.3 we introduce the nonseasonal and seasonal ARIMA procedures. The focus will be to show you how to identify an ARIMA model followed by estimations and checking the model adequacy in order to forecast future values.

## 2.2 Methods for a Non-Seasonal ARIMA Model

In theory, ARIMA (AutoRegressive Integrated Moving Average) is a general class of models for describing and forecasting a time series. Lags of the differenced series appearing in the forecasting equation are called *Autoregressive*, lags of the forecasts errors are called *Moving Average*, and the time series which needs to be differenced to be made stationary is said to be an *Integrated* version of a stationary series [7].

A non-seasonal ARIMA Model is classified as an "ARIMA( $p, d, q$ )" model [1]:

$$(1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p)(1 - B)^d(Z_t - \mu) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)a_t$$

or

$$\phi(B)(1 - B)^d(Z_t - \mu) = \theta(B)a_t \quad (2.1)$$

where

$Z_t$  is the observed series

$\mu$  is the mean of the series

$\phi(B)$  is the autoregressive polynomial of order  $p$

$\theta(B)$  is the moving average polynomial of order  $q$

$d$  is the number of nonseasonal differences

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process  $WN(0, \sigma_a^2)$

The autoregressive part of the model  $\phi(B)$  is simply a linear regression of the current value of the series against one or more prior values of the series. The value  $p$  is called the order of the AR model. The moving average part of the model  $\theta(B)$  is conceptually a linear regression of the current value of the series against the white noise or random shocks associated with one or more prior values of the series [2]. The random shocks are assumed to come from the same distribution, typically a normal distribution with mean at zero and constant variance [1, 2]. The distinction



in the MA model is that these random shocks are propagated to future values of the time series. Fitting the MA estimates is more complicated than with AR models because the error terms are not observable. This means that iterative non-linear fitting procedures need to be used in place of ordinary least squares.

ARIMA modeling involves three stages [14]:

1. Identification of the initial  $p$ ,  $d$ , and  $q$  parameters, using autocorrelation and partial autocorrelation methods.
2. Estimation of the  $p$  (autoregressive) and  $q$  (moving average) components to see if they contribute significantly to the model or if one or the other should be dropped.
3. Diagnosis of the residuals to see if they are random and normally distributed, indicating a good model.

### 2.2.1 Model Identification

In order to identify the appropriate ARIMA model for a time series the first step is to check if the time series is **stationary** [7]. A stationary process has the property that the mean, variance and autocorrelation structure does not change over time. Stationarity can be detected from an *autocorrelation plot*. Specifically, non-stationarity is often indicated by an autocorrelation plot with very slow decay [4, 7]. If the time series data is not stationary, usually the series first needs to be *differenced* until it is stationary. This also sometimes require log transformations to the data in order to stabilize the variance [14]. The number of times the series needs to be differenced to achieve stationarity is reflected in the  $d$  parameter [1]. In order to determine the necessary level of differencing, one should examine the plot of the data and the autocorrelation plot. Significant changes in level (strong upward or downward changes) usually require first order non-seasonal (lag=1) differencing; strong changes of slope usually require second order non-seasonal differencing [7]. If  $Z_t$  denotes the values

of the time series  $Z$  at period  $t$ , then the first difference is equal to  $Z_t - Z_{t-1}$  [7]. If the lag 1 autocorrelation is zero or even negative, then the series doesn't require further differencing. However, one should keep in mind that some time series may require little or no differencing, and that *over differenced* series produces less stable coefficient estimates [1].

After a time series has been stationarized by differencing, the next step in fitting an ARIMA model is to determine whether AR (Auto-Regressive of order  $p$ ) or MA (Moving Average of order  $q$ ) terms are needed to correct any autocorrelation that exists in the differenced series. By looking at the autocorrelation function (ACF) and partial autocorrelation (PACF) plots of the differenced series, you can get an idea of the number of AR and/or MA terms needed [5].

The **ACF plot** is a bar chart of the coefficients of correlation between a time series and the lags of itself [1]. The autocorrelation,  $\rho_k$  is defined as:

$$\rho_k = \frac{\text{cov}(Z_t, Z_{t-k})}{[v(Z_t)v(Z_{t-k})]^{1/2}} = \frac{\gamma_k}{\gamma_0}, \quad k = 0, \pm 1, \pm 2, \dots \quad (2.2)$$

$\rho$  is considered a function of the lag  $k$ . Since  $\rho_{(-k)} = \rho_k$ , only a nonnegative  $k$  has to be considered.

The **PACF plot** is a plot of the partial correlation coefficients between the series and lags of itself [1].

"It measures the additional correlation between  $Z_t$  and  $Z_{t-k}$  after adjustments have been made for the intermediate variables  $Z_{t-1}, \dots, Z_{t-k-1}$ ." [1]

The partial autocorrelation of lag  $k$  can be thought of as the partial regression coefficient  $\phi_{kk}$  in the following representation:

$$Z_t = \phi_{k1}Z_{t-1} + \cdots + \phi_{kk}Z_{t-k} + a_t$$

### Order of the AR Process ( $p$ )

Specifically, for an AR(1), the sample autocorrelation function should either have an exponentially decreasing or damped sinusoidal components [7]. For higher-order autoregressive processes, the sample autocorrelation needs to be supplemented with a partial autocorrelation plot. The partial autocorrelation of an AR( $p$ ) process becomes zero at lag  $p+1$  and greater, so we examine the sample partial autocorrelation function to see if there is evidence of a departure from zero [5]. This is usually determined by placing a 95% confidence interval (CI) on the sample partial autocorrelation plot. Most software programs will plot this CI when generating a sample autocorrelation plot. The CI is approximately  $\pm \frac{2}{\sqrt{n}}$ , with  $n$  denoting the sample size [1, 5].

### Order of the MA Process ( $q$ )

The autocorrelation function of a MA( $q$ ) process becomes zero at lag  $q+1$  or greater, so we examine the sample autocorrelation function to see where it essentially becomes zero after a certain lag [7]. We also do this to the partial autocorrelation plot by placing the 95% confidence interval for the sample autocorrelation function on the sample autocorrelation plot.

The following table 2.1 summarizes how we use the sample autocorrelation function and partial autocorrelation for model identification of nonseasonal models [7].

Model	ACF	PACF
AR(1)	decays exponentially (alternating signs if $\phi < 0$ )	single spike at lag 1
MA(1)	single spike at lag 1	decays exponentially (alternating signs if $\theta < 0$ )
AR(p)	decays exponentially, may contain damped oscillations	$p$ spikes (cuts off to zero after lag $p$ )
MA(q)	$q$ spikes (cuts off to zero after lag $q$ )	decays exponentially, may contain damped oscillations
ARMA(p,q)	tails off after $(q - p)$	tails off $(p - q)$

Table 2.1: Primary distinguishing characteristics of theoretical acf's and pacf's for stationary process

Although experience is helpful, developing good models using these sample plots can involve much trial and error. For this reason, in recent years information based criteria such as AIC (Akaike Information Criterion) and BIC have been preferred and used [12]. These techniques can help automate the model identification process [12]. These techniques require computer software to use such as SAS, that provide ARIMA seasonal capabilities [9]. One of these identification procedures is the Extended Sample Autocorrelation Function (ESACF).

### Extended Sample Autocorrelation Function (ESACF)

The ESACF method can tentatively identify the orders of a stationary or nonstationary ARMA process based on iterated least squares estimates of the autoregressive parameters [11].

Given a stationary or nonstationary time series  $\{Z_t : 1 \leq t \leq n\}$  with mean corrected form  $\tilde{Z}_t = Z_t - \mu_z$ , with a true autoregressive order of  $p + d$  and with a true

moving average order of  $q$ , you can use the ESCAF method to estimate the unknown orders  $p + d$  and  $q$  by analyzing the autocorrelation functions associated with filtered series of the form

$$W_t^{(m,j)} = \hat{\phi}_{(m,j)}(B)\tilde{Z}_t = \tilde{Z}_t - \sum_{i=1}^m \hat{\phi}_i^{(m,j)}\tilde{Z}_{t-i} \quad (2.3)$$

Where

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$m = p_{min}, \dots, p_{max}$  are the autoregressive test orders

$j = q_{min} + 1, \dots, q_{max} + 1$  are the moving average test orders

$\hat{\phi}_i^{(m,j)}$  are the autoregressive parameter estimates under the assumption that the series an ARMA( $m,j$ ) process.

The  $j$ th lag of the sample autocorrelation function of the filtered series,  $W_t^{(m,j)}$ , is the extended sample autocorrelation function, denoted as  $r_{j(m)} = r_j(W_t^{(m,j)})$  [11]. The standard errors of  $r_{j(m)}$  are computed using Bartlett's approximation of the variance of the sample autocorrelation function [11],

$$var(r_{j(m)}) \approx 1 + \sum_{t=1}^{j-1} r_j^2(W^{(m,j)})$$

An ESACF table is then constructed using the  $r_{j(m)}$  for  $m = p_{min}, \dots, p_{max}$  and  $j = q_{min} + 1, \dots, q_{max} + 1$  to identify the ARMA orders given in table 2.2 [12]. It is useful to arrange  $r_{j(m)}$  in a two way table where the first row corresponding to  $r_{j(0)}$  gives the standard sample ACF of  $\tilde{Z}_t$ , the second row gives the first ESACF  $r_{j(1)}$ , and so on [12]. The rows are numbered 0, 1, ... to specify the AR order, and the columns are numbered in a similar way for the MA order.

		MA				
AR	0	1	2	3	...	
0	$r_{1(0)}$	$r_{2(0)}$	$r_{3(0)}$	$r_{4(0)}$	...	
1	$r_{1(1)}$	$r_{2(1)}$	$r_{3(1)}$	$r_{4(1)}$	...	
2	$r_{1(2)}$	$r_{2(2)}$	$r_{3(2)}$	$r_{4(2)}$	...	
.	.	.	.	.	...	
.	.	.	.	.	...	

Table 2.2: ESACF Table

The orders are tentatively identified by finding a right triangular pattern with vertices located at  $(p + d, q)$  and  $(p + d, q_{max})$  and in which all elements are insignificant (based on asymptotic normality of the autocorrelation function) [11]. The vertex  $(p + d, q)$  identifies the order as shown in table 2.3, which depicts the pattern associated with an ARMA(1,2) series [12].

		MA							
AR	0	1	2	3	4	5	6	7	
0	*	X	X	X	X	X	X	X	
1	*	X	0	0	0	0	0	0	
2	*	X	X	0	0	0	0	0	
3	*	X	X	X	0	0	0	0	
4	*	X	X	X	X	0	0	0	
		X = significant terms							
		0 = insignificant terms							
		* = no pattern							

Table 2.3: Theoretical ESACF Table for an ARMA(1,2)

The ESACF table can then be constructed using indicator symbols with X referring to values beyond the range of 2 standard deviations of the mean and 0 for values within  $\pm 2$  standard deviations [12]. Due to sample correlations among the ACF, the pattern in the ESACF table from most time series is usually not that simple. Models can usually be identified without that much difficulty through using a combination of the sample ACF, PACF, and ESACF.

## 2.2.2 Estimation

At the identification stage we tentatively select one or more models that seem likely to provide statistically adequate representations of the available data according to some criterion like the AIC and BIC [12]. A general approach to estimation is the *Maximum Likelihood* (ML) approach. But, finding exact ML estimates of ARIMA models can be long and cumbersome. The alternative is using the *Least Squares* (LS) approach for AR models [2, 7]. If the random shocks are normally distributed (as we assume they are), then the least squares point estimates are either exactly or very near the ML estimates. For the rest of the practicum, we will be using least squares for calculating point estimates in SAS.

Associated with the point estimate of each parameter in a ARMA model is its *standard error* and *t-value* [2]. Let  $\theta$  denote any particular parameter in a ARMA model, let  $\hat{\theta}$  denote the point estimate of  $\theta$  and  $s_{\hat{\theta}}$  denotes the standard error of the point estimate of  $\theta$ . The t-value is then given as:

$$t_{\hat{\theta}} = \frac{\hat{\theta}}{s_{\hat{\theta}}} \quad (2.4)$$

If the absolute value of  $t_{\hat{\theta}}$  is large then we tend to reject  $H_0 : \theta = 0$  and conclude that parameter  $\theta$  is significant and should include it into the model [2]. If the  $t_{\hat{\theta}} > 2$  or the corresponding P-value is less than  $\alpha = 0.05$ , we will reject  $H_0 : \theta = 0$  and include the parameter into the model [2].

## Checking Coefficients for Stationarity and Invertibility

The Box-Jenkins methodology requires that the model to be used in describing and forecasting a time series be both *stationary* and *invertible* [6]. We already have discussed the meaning of stationarity but not formally have discussed the meaning of invertibility. A non-invertible ARIMA model implies that the weights placed on past  $Z$  observations when expressing  $Z_t$  as a function of these observations do not decline as we move further into the past [12]. Common sense tells us that an invertible model should have larger weights attached to more recent observations than to more distant observations. Each of the stationarity and invertibility conditions implies that the parameters of the operators  $\phi(B)$  and  $\theta(B)$  used in the model should satisfy certain conditions [6]. Table 2.4 and 2.5 summarize the stationarity and invertibility conditions on the parameters of nonseasonal models.

Model Type	Stationarity Conditions
ARMA(0, q)	Always stationary
AR(1) or ARMA(1, q)	$ \phi_1  < 1$
AR(2) or ARMA(2, q)	$ \phi_2  < 1$ $\phi_2 + \phi_1 < 1$ $\phi_2 - \phi_1 < 1$
When $p > 2$ for ARMA(p,q)	$\phi(z) \neq 0$ , for all $ z  \leq 1$ where $z$ is a complex variable

Table 2.4: Summary of Stationarity Conditions



Model Type	Invertibility Conditions
ARMA(p, 0)	Always invertible
MA(1) or ARMA(p, 1)	$ \theta_1  < 1$
MA(2) or ARMA(q, 2)	$ \theta_2  < 1$ $\theta_2 + \theta_1 < 1$ $\theta_2 - \theta_1 < 1$
When $q > 2$ for ARMA(p,q)	$\theta(z) \neq 0$ , for all $ z  \leq 1$ where $z$ is a complex variable

Table 2.5: Summary of Invertibility Conditions

### 2.2.3 Diagnostic Checking

After the parameters in the model have been estimated, next we should check to see whether the model assumptions are satisfied [1]. The assumptions for an ARIMA model is that the  $a_t$ 's are white noise [7]. That is, the  $a_t$ 's are uncorrelated random variables with mean zero and constant variance. For any estimated model, the residuals  $\hat{a}_t$ 's are estimates of these unobserved white noise  $a_t$ 's. To check whether the mean of the residuals is zero and the variance is constant, we can check the plot of residuals. This plot can also be used to detect possible outliers and any other systematic patterns. To check if the residuals are white noise, we can compute the sample ACF and PACF of the residuals to see whether they do not form any pattern and are all statistically insignificant (within two standard errors for  $\alpha = 0.05$ ) [12]. The sample autocorrelation of the residuals is given as

$$r_k = \frac{\sum_{t=1}^{n-k} (\hat{a}_t - \bar{a})(\hat{a}_{t+k} - \bar{a})}{\sum_{t=1}^n (\hat{a}_t - \bar{a})^2} \quad (2.5)$$

and compare them with their standard errors using Bartlett's approximate formula [7], we get

$$s[r_k] = \left(1 + 2 \sum_{j=1}^{k-1} r_j^2\right) n^{-\frac{1}{2}} \quad (2.6)$$

If any residual ACF value is beyond  $\pm 2$  standard errors, we conclude that the residuals are correlated and that the estimated model may be inadequate [12]. We then identify an new model and estimate it once again until our model is adequate.

One way to use the residuals to check the adequacy of the overall models is to examine a statistic that determines whether the first  $K$  sample autocorrelations of the residuals are adequate. This useful test is called the *Chi-squared test* or sometimes referred to as the *Ljung-Box Q-Test* [1, 7]. This test uses all the  $K$  residual autocorrelations as a set to check the joint null hypothesis about the correlations among the random shocks [7], where the choice of  $K$  is somewhat arbitrary.

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_K = 0 \quad (2.7)$$

with the test statistic

$$Q^* = n(n+2) \sum_{k=1}^K \frac{r_k^2}{(n-k)} \sim \chi^2(K-m) \quad (2.8)$$

where  $n$  is the number of observations to estimate the model. The test statistic  $Q^*$  follows a chi-squared distribution with  $(K - m)$  degrees of freedom, where  $m$  is the number of parameters in the ARIMA model [7]. The modeling process is suppose to account for the relationship between the time series observations. If it does account for these relationships, the residuals should be unrelated and therefore the autocorrelations of the residuals should be small. Thus, the larger that  $Q^*$  is, the larger autocorrelations of the residuals and more related are the residuals. Therefore, a large value of  $Q^*$  indicates that the model is inadequate [2].

We can reject the adequacy of the model under consideration if either of the following equivalent conditions hold [2]:

1.  $Q^* > \chi_\alpha^2(K - m)$  at  $\alpha = 0.05$
2. P-value is less than  $\alpha = 0.05$

If the p-value is larger than 0.05 or if  $Q^*$  is smaller than the critical value  $\chi_\alpha^2(K - m)$ , then it is reasonable to conclude that the model is adequate and is ready to forecast [12].

## 2.2.4 Forecasting

Without going into depth, forecasting is one of the most important objectives in time series. The most convenient way to produce point forecasts from a ARIMA model is to write the model in a differenced equation form [1].

Let  $t$  be the current time period. When forecasting, we're interested in future values of a time series variable, denoted  $Z_{t+\ell}$ , where  $\ell > 0$ . Period  $t$  is called the forecast origin and  $\ell$  is called the forecast lead time.

In ARIMA analysis, forecasts depend on the available observations on variable  $Z$  up through period  $t$ . Let the information contained in the set of available observations  $(Z_t, Z_{t-1}, \dots)$  be designated  $I_t$ . Then the forecast of  $Z_{t+\ell}$  is designated as  $\hat{Z}_t(\ell)$ , which is the conditional expectation of  $Z_{t+\ell}$  [1]

$$\hat{Z}_t(\ell) = E[Z_{t+\ell} | Z_t, Z_{t-1}, \dots] = E[Z_{t+\ell} | I_t]$$

since  $a_{t+\ell}$  is unknown at time  $t$ , we assign its expected value of zero [7].

$$E[a_{t+\ell} | I_t] = \begin{cases} a_{t+\ell}, & \ell \leq 0 \\ 0, & \ell > 0 \end{cases}$$

A forecast error for predicting  $Z_t$  with lead time  $\ell$  (i.e., forecasting ahead  $\ell$  time periods) is denoted by  $e_t(\ell)$  and defined as the difference between an observed  $Z_t$

and its forecasts counterpart  $\hat{Z}_t(\ell)$  [1]:

$$e_t(\ell) = Z_t - \hat{Z}_t(\ell) \quad (2.9)$$

This forecast error has variance

$$Var[e_t(\ell)] = \sigma_a^2(1 + \psi_1^2 + \dots + \psi_{\ell-1}^2) \quad (2.10)$$

where the  $\psi_i$  coefficients are the coefficients in the random shock (infinite MA) form of the ARIMA model [1].

If the random shocks ( $a_t$  values) are normally distributed and if we have an appropriate ARIMA model, then our forecasts and associated forecast errors are approximately normally distributed [7]. Then an approximate 95% prediction interval around any forecasts will be

$$\hat{Z}_t(\ell) \pm 1.96\sqrt{Var[e_t(\ell)]} \quad (2.11)$$

The forecasted values, standard errors and 95% confidence intervals can be easily calculated through the use of SAS or other statistical software [9].

## 2.3 Methods for a Seasonal ARIMA Model

Sometimes there is a cyclical or seasonal component in a time series. By this we mean the recurrence of some recognizable pattern after some regular interval that we can call the *seasonal period* denoted by  $S$  [5]. As an example, for monthly temperature of the Hydro data, there is clearly a recurring pattern with seasonal pattern of 12. A pure seasonal model is characterized by nonzero correlations only at lags that are multiples of the seasonal period  $S$  [7]. This means that the time series at time  $t$ ,  $Z_t$ , depends on  $Z_{t-S}, Z_{t-2S}, Z_{t-3S}, \dots$  only. In reality, few time series are just purely seasonal. We need to take into account the correlations between the time series values within each period. This can be done by combining the nonseasonal and seasonal

effects into a single model. A *Multiplicative Seasonal ARIMA Model* is classified as an "ARIMA( $p, d, q$ )( $P, D, Q$ ) $_S$ " model [1]:

$$\phi(B)\Phi(B^S)(1 - B)^d(1 - B^S)^D(Z_t - \mu) = \theta(B)\Theta(B^S)a_t \quad (2.12)$$

where

$Z_t$  is the observed series

$\phi(B) = \phi_1 B + \dots + \phi_p B^p$  is the nonseasonal autoregressive polynomial of order  $p$

$\theta(B) = \theta_1 B + \dots + \theta_q B^q$  is the nonseasonal moving average polynomial of order  $q$

$\Phi(B) = \Phi_1 B + \dots + \Phi_P B^{PS}$  is the seasonal autoregressive polynomial of order  $P$

$\Theta(B) = \Theta_1 B^S + \dots + \Theta_Q B^{QS}$  is the seasonal moving average polynomial of order  $Q$

$d$  is the number of nonseasonal differences

$D$  is the number of seasonal differences

$S$  is the period of seasonality

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process  $WN(0, \sigma^2)$

Our previous discussion focused on conceptual construction of nonseasonal ARIMA models. The same procedures also apply to the seasonal models, except the main difference is that in seasonal series the seasonal ARIMA models are only an extension of the nonseasonal models [1]. Multiplicative Seasonal modelling involves three stages: *Identification, Estimation, and Diagnostic Checking*.

### 2.3.1 Identification

A seasonal time series is by nature non-stationary. Therefore the first step is to eliminate the seasonality. This is usually done through seasonal differencing. Though theoretically the ACF and PACF are not formally defined, in practice, the sample ACF and PACF can always be computed from a given sample. Often, the sample ACF and PACF shows certain seasonal patterns and therefore are useful tools. If the time series data at the seasonal lags (12, 24, 36,..) fail to die out quickly, this confirms

the nonstationary pattern and calls for Seasonal differencing [7]. For example, if we have a seasonal period of  $S = 12$  then we will need a seasonal differencing of order  $D = 1$ , denoted as  $Z_t - Z_{t-12}$  or  $(1 - B^{12})^1 Z_t$  using the backshift operator. It may be useful to apply a seasonal difference to the data and regenerate the sample ACF and PACF plots. This may help in the identification of the nonseasonal component of the model. In some cases, the seasonal differencing may remove most or all of the seasonality effect [7].

After a time series has been deseasonalized by differencing, the next step in fitting an ARIMA model is to determine whether the seasonal AR (Auto-Regressive of order  $P$ ) or seasonal MA (Moving Average of order  $Q$ ) terms are needed to correct any autocorrelation that exists in the seasonally differenced series. This also applies to the nonseasonal AR and MA terms of order  $p$  and  $q$  discussed in the previous section. By looking at the ACF and PACF plots of the seasonally differenced series, you can get an idea of the number of seasonal AR and/or MA terms needed at every seasonal lag  $S, 2S, 3S, \dots$  as well as the nonseasonal AR and MA terms [1, 12].

The following table 2.6 summarizes how we use the sample autocorrelation function and partial autocorrelation for model identification of seasonal terms [7].

Model	ACF	PACF
Seasonal AR(1)	decays exponentially (alternating signs if $\Phi < 0$ )	single spike at lag $S$
Seasonal MA(1)	single spike at lag $S$	decays exponentially (alternating signs if $\Theta < 0$ )
Seasonal AR( $P$ )	decays exponentially, may contain damped oscillations	$P$ spikes (cuts off to zero after lag $PS$ )
Seasonal MA( $Q$ )	$Q$ spikes (cuts off to zero after lag $Q$ )	decays exponentially, may contain damped oscillations

Table 2.6: Primary distinguishing characteristics of theoretical acf's and pacf's for stationary process for the seasonal model

The computation of the ESACF for seasonal models are time consuming and the patterns are very complicated [11]. Since the ESACF provides only information about the maximum orders of  $p$  and  $q$ , its use in modeling seasonal time series is very limited [12]. We have found that the ACF is the most useful method.

### 2.3.2 Estimation

In the estimation stage, the Multiplicative Seasonal ARIMA models are not fundamentally different than the estimation of nonseasonal ARIMA models discussed in the Section 2.2.2. The parameter of the seasonal models are usually estimated by a least squares or Maximum likelihood method and will not be further discussed here.

#### Stationarity and Invertibility Conditions

With a multiplicative model, the stationary and invertibility conditions apply separately to the seasonal and nonseasonal coefficients [1, 7]. The stationarity conditions apply only to the AR coefficients, treating the seasonal and nonseasonal separately

because they are multiplied. Likewise, the invertibility conditions apply only to the MA coefficients, treating the seasonal and nonseasonal separately [7]. The conditions for the seasonal part are the same conditions that apply to the nonseasonal models are similar to the tables discussed in table 2.4 and 2.5 except that it's multiplied by the seasonal term  $S$ .

### **2.3.3 Diagnostic stage**

We adopt the same diagnostic procedure as in the nonseasonal case and examine whether the residuals are uncorrelated. We compare the residual autocorrelations with their corresponding standard errors given in the nonseasonal section. Once the model is adequate we can use the model to forecast future observations. The forecast procedure is the same as the nonseasonal component and will not be discussed here.

## **2.4 Conclusion**

In this chapter we introduced ARIMA models. We first introduced the nonseasonal ARIMA model, what it is, and the three stages in order to forecast future values. We also introduced the seasonal ARIMA model for the time series data . This is an extension of the nonseasonal model except we look at the seasonal component for the three stages of seasonal. In Chapter 3 we will look at this topic in more detail using the temperature data given by Manitoba Hydro to see if we can build a model for monthly, daily and hourly temperature.



# Chapter 3

## Results and Interpretations

### 3.1 Introduction

In chapter 2 we studied the nonseasonal and seasonal ARIMA procedures for the time series data. In this chapter, we will present the results of various ARIMA models for the mean, maximum and minimum temperature for monthly, daily, and finally hourly data. All of the statistical models and tests were performed using the SAS package (Version 8.2). See Appendix B for SAS codes. In section 3.2 we introduce the ARIMA models for monthly temperature data. We will look at the model for mean temperature and the two extremes, maximum and minimum. In section 3.3 we will use the ARIMA model for daily temperature. Again, we will look at the model for mean temperature and the two extremes. We will also look at seasonal daily temperature using only 6 years of past data, as opposed to the 50 years and compare the two models. Finally, in section 3.4 we examine an ARIMA model for hourly temperature and see if our model is valid.

## **3.2 Fitting an ARIMA Model to Monthly Temperature**

In this section, we are examining the monthly temperature taken from the last 50 years of hourly data. This amounts to 600 observations from the Hydro data. We will first look at the mean temperature models, followed by maximum and minimum temperature models.

### **3.2.1 An ARIMA Model for Mean Temperature**

Let us start with the time series plot of monthly mean temperatures shown in Figure 3.1. We see that the winter months mean temperature are regularly lower than those in other months within the same year, while the summer months mean temperature values are regularly higher. This suggest that mean temperature values in any given month are similar to the mean temperature values in the corresponding month of other years, indicating that we have a seasonal pattern occurring every 12 months. We will see this more clearly by looking at the sample ACF and PACF plots.

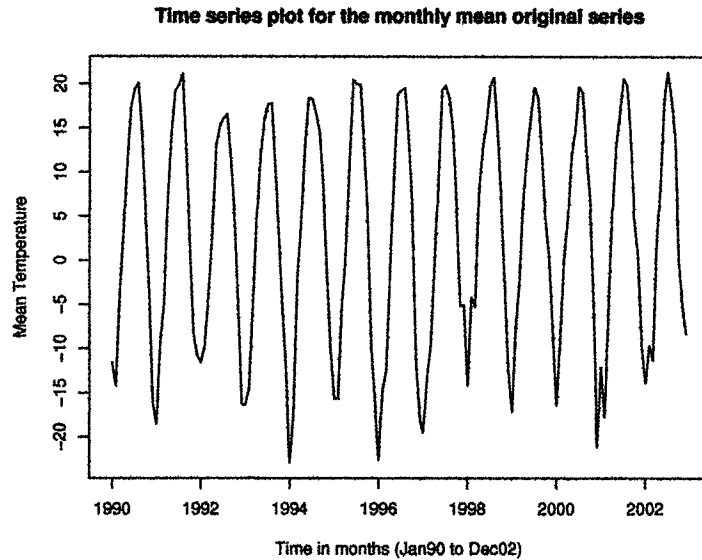


Figure 3.1: Time series plot of the monthly mean temperature

If we take a look at the sample ACF and PACF of the original series given in Figure 3.2, we see at the seasonal lags that it has a slow decay every 12 months indicating that we need a span of 12 difference in order to transform the nonstationary series into a stationary one. Usually first calculating the nonseasonal first difference is appropriate, but sometimes strong seasonality can make the nonseasonal pattern in the sample ACF appear nonstationary when, infact, the nonseasonal element is stationary. In this case, taking the nonseasonal first difference only, so differencing once ( $D=1$ ) by the seasonal length of  $S = 12$  is needed.

The dashed lines show the range in which the sample autocorrelation is not significantly different from 0.

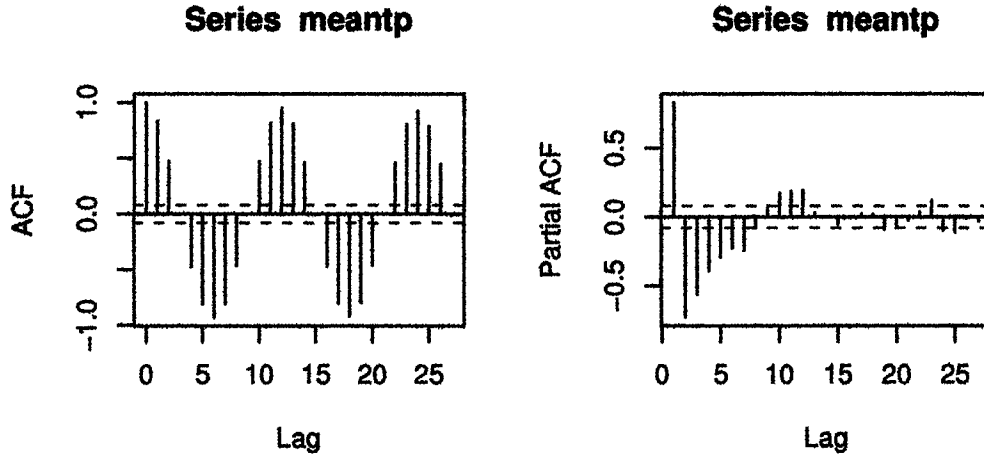


Figure 3.2: Sample ACF and PACF plots of the mean temp

The seasonally differenced data are plotted in Figure 3.3. The plot suggests that the seasonal differencing has removed any obvious peak seasonal variation appearing in the original data.

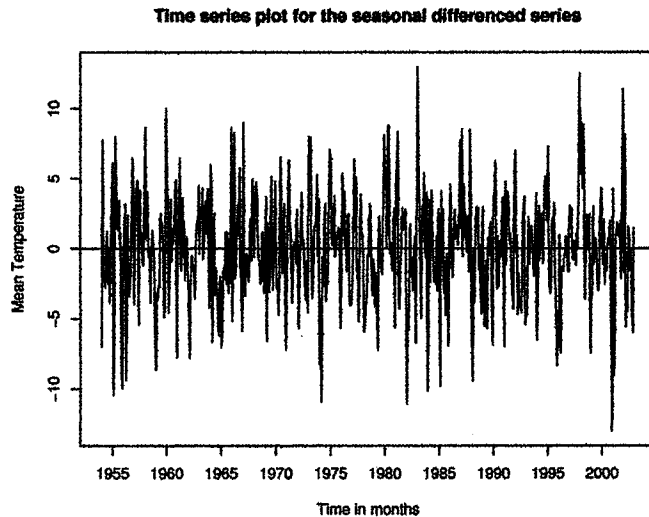


Figure 3.3: Time series plot of the monthly mean temperature after seasonally diff

Now if we take a sample ACF and PACF of the seasonally differenced data we obtain the results given in Figure 3.4:

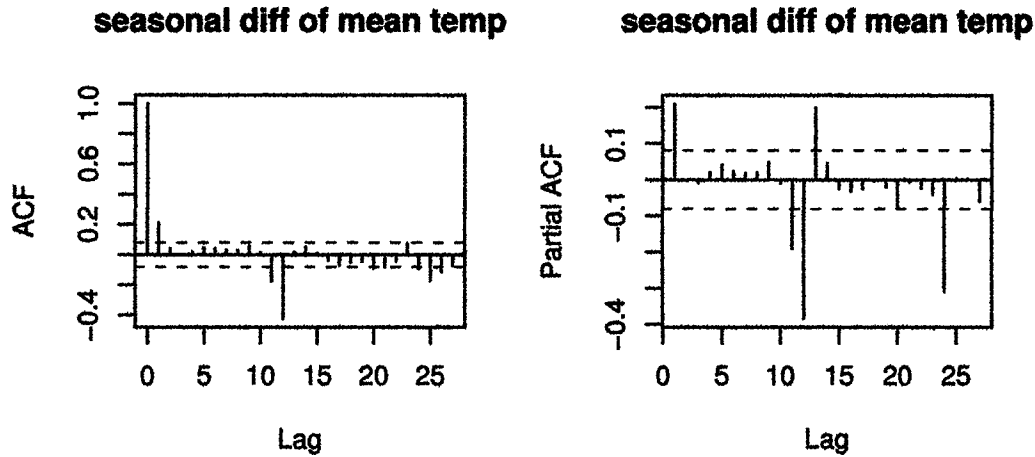


Figure 3.4: Sample ACF and PACF plots of the mean temp seasonal diff of 12

Taking the seasonal first difference now induces a stationary mean. In Figure 3.4, the estimated sample ACF confirms that the seasonally differenced series is stationary since the sample ACF moves rapidly to zero at the short lags. The large spike at lag 12 is followed by an insignificant autocorrelation at lag 24 indicates that there is still a seasonal pattern in the data. The pattern seems to be of the MA variety since the sample ACF cuts off to zero at lags 24, 36, . . . etc. The decay (rather than the cutoff) to zero on the negative side at lags 12, 24, and 36 in the sample PACF is consistent with an  $MA(1)_{12}$  model for the seasonal part of the data.

The strong spike at lag 12 in the sample ACF could influence the values of the adjacent autocorrelations at lag 11. Recall that the estimated autocorrelation coefficients can be correlated (positive or negative) with each other, thus it is difficult to identify the nonseasonal pattern. It is wise to estimate a purely seasonal model first, letting the residual sample ACF and PACF guide us in identifying the nonseasonal pattern. We tentatively entertain an  $ARIMA(0, 0, 0)(0, 1, 1)_{12}$ , realizing it could be incomplete since it has no seasonal element:

$$(1 - B^{12})(Z_T - \mu) = (1 - \Theta_1 B^{12})a_T \quad (3.1)$$

Using PROC ARIMA [9] to estimate the seasonal MA(1) term we obtain the results in Figure 3.5.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MU	0.0016816	0.01634	0.10	0.9181	0
MA1,1	0.88308	0.02026	43.59	<.0001	12

Figure 3.5: Table of parameter estimates

The t-ratios provide significance tests for the parameter estimates and indicate whether some terms in the model may be necessary. In this case from Figure 3.5, the t-ratio for the seasonal moving average parameter is 43.59, meaning that  $\hat{\Theta}_1$  is highly significant and we can leave it in our model. Our constant term  $\mu$  has a large P-value, suggesting it could be omitted from the model. Now looking at the residual sample ACF and PACF for the model in Figure 3.6, the seasonal decay at lag 12 and 24 has disappeared.

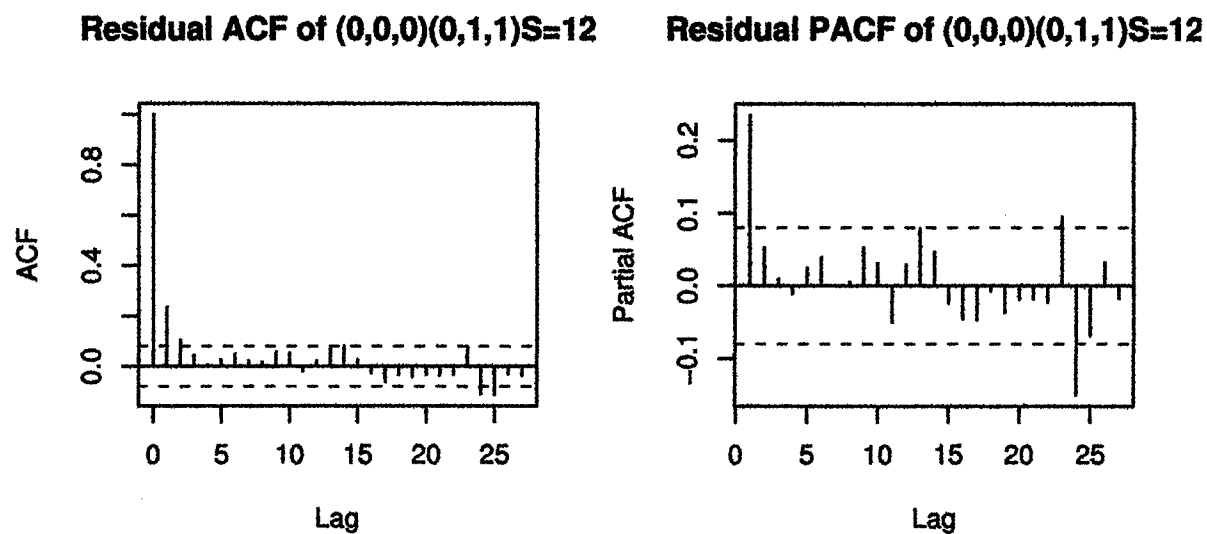


Figure 3.6: Residual sample ACF and PACF plots of the mean temp

The most striking feature of the residual sample ACF is the significant spike at lag 1 and possibly lag 2, at this point, it cuts off to zero at lag 3. The sample ACF spike at lag 23 is greater than the 95% standard error bands, but the principle of parsimony [7] says adding one coefficient at a time.

By adding MA(1) and MA(2) into the model in Figure 3.7, we see that the parameter estimates look good.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MA1,1	-0.22377	0.04148	-5.39	<.0001	1
MA1,2	-0.10648	0.04133	-2.58	0.0102	2
MA2,1	0.90014	0.01876	47.97	<.0001	12

Figure 3.7: Table of parameter estimates

Each of the estimated coefficients are significant with each satisfying its respective invertibility condition [6]. The former spike at lag 1 and 2 has disappeared and the variance estimates are smaller. After the parameters in a model have been estimated it is necessary to check whether the model assumptions are satisfied. If the assumptions are not met, the model must be respecified. This phase in the model building is usually referred to as diagnostic checking which relies heavily on the analysis of the residuals. In Figure 3.8 our  $\chi^2$  test statistics shows that our model is inadequate because our  $\chi^2$  values are still too large and P-values are small indicating that the residuals are correlated. The SAS code used is given in Appendix B.

### Autocorrelation Check of Residuals

To	Chi-	Pr >		-----Autocorrelations-----					
Lag	Square	DF	ChiSq						
6	2.47	3	0.4799	0.006	0.010	0.052	-0.012	0.017	0.030
12	7.98	9	0.5361	0.014	-0.008	0.058	0.063	-0.035	0.022
18	17.43	15	0.2939	0.075	0.079	0.022	-0.026	-0.047	-0.020
24	34.67	21	0.0306	-0.019	-0.014	-0.024	-0.029	0.133	-0.092
30	40.08	27	0.0503	-0.087	0.018	-0.017	0.000	0.004	-0.023
36	49.90	33	0.0298	-0.018	0.088	0.022	-0.049	0.051	-0.046
42	57.33	39	0.0293	0.045	0.061	0.054	-0.014	0.051	0.013
48	61.55	45	0.0509	-0.007	-0.002	0.060	0.024	0.009	-0.047

Figure 3.8: Check for white noise for an ARIMA(2,0,0)(0,1,1)

Figure 3.9 shows us there still remains a residual sample ACF spike at lag 23, with its spike outside the error bands indicating that a MA term at lag 23 is appropriate.

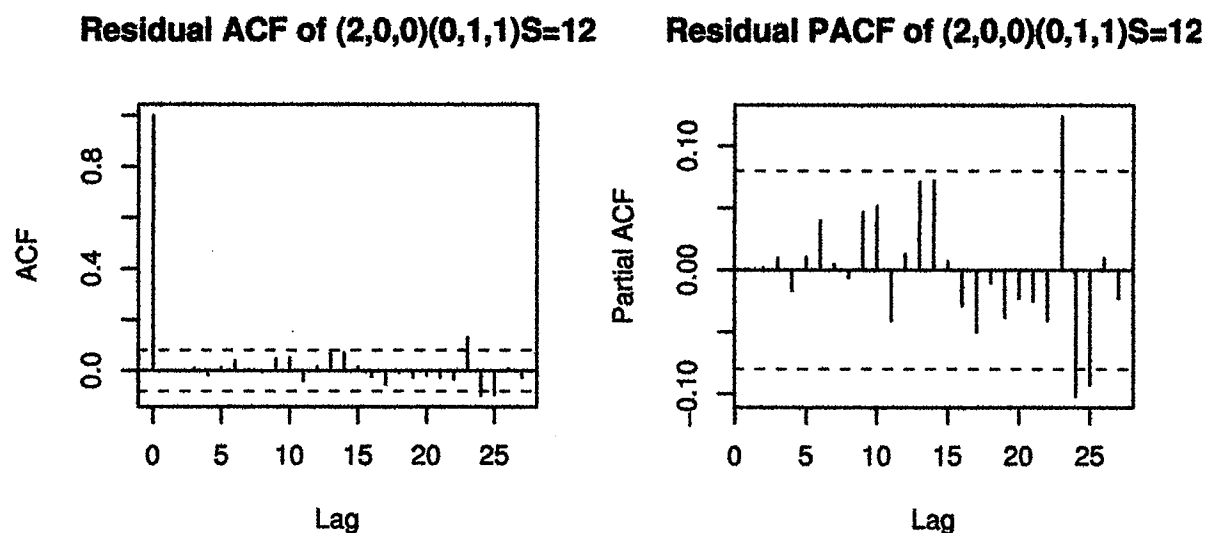


Figure 3.9: Residual sample ACF and PACF plots of the ARIMA(2,0,0)(0,1,1)



In Figure 3.10, all the estimated coefficients are significant after adding the MA term at lag 23, even at the 1% level of significance.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MA1,1	-0.24569	0.04057	-6.06	<.0001	1
MA1,2	-0.13109	0.04065	-3.22	0.0013	2
MA1,3	-0.15394	0.04057	-3.79	0.0002	23
MA2,1	0.90238	0.01840	49.04	<.0001	12

Figure 3.10: Table of parameter estimates

The variance estimates is smaller than its previous model ( $=7.84$ ) and it appears none of the coefficients in Figure 3.10 needs to be removed. Now that the model is significant, we need to perform an autocorrelation check of the residuals. All the  $\chi^2$  values given in Figure 3.11 appears to be small with large P-values ( $> 0.2$ ), which indicates that our model is adequate and is ready to forecast.

Autocorrelation Check of Residuals

To	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
Lag 6	3.05	2	0.2177	-0.000	0.005	0.060	-0.007	0.016	0.035
12	6.36	8	0.6070	0.021	-0.010	0.033	0.054	-0.028	0.014
18	15.71	14	0.3311	0.075	0.071	0.016	-0.037	-0.048	-0.027
24	19.78	20	0.4717	-0.019	-0.021	-0.022	-0.039	-0.012	-0.061
30	24.00	26	0.5759	-0.068	0.007	-0.032	0.009	-0.005	-0.031
36	34.80	32	0.3360	-0.010	0.090	0.017	-0.046	0.056	-0.060
42	42.33	38	0.2894	0.023	0.066	0.061	-0.011	0.056	0.004
48	48.38	44	0.3004	-0.001	-0.000	0.081	0.034	0.002	-0.041

Figure 3.11: Check for white noise for Nonseasonal MA(1,2,23) and seasonal MA(12)

The estimated ARIMA model for mean temp is

$$(1 - B^{12})Z_T = (1 - \Theta_1 B^{12})(1 - \theta_1 B - \theta_2 B^2 - \theta_{23} B^{23})a_T \quad (3.2)$$

where

$Z_t$  is the observed series

$\theta_1$ ,  $\theta_2$  and  $\theta_{23}$  are nonseasonal Moving Average parameters

$\Theta_1$  is the seasonal movingaverage coefficients

$(1 - B^{12})$  is the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

Given the estimated coefficients using PROC ARIMA from the SAS output in Figure 3.10 the above equation becomes:

$$(1 - B^{12})Z_T = (1 - 0.9023B^{12})(1 + 0.2456B + 0.131B^2 + 0.1539B^{23})a_t \quad (3.3)$$

Now that we have an adequate model, we can now turn to forecasting. The forecasted values from the model given in Figure 3.12 are obtained from January, 2002 to June, 2003.

	mean_					
Obs	temp	FORECAST	STD	L95	U95	RESIDUAL
589	-13.8540	-14.5815	2.80080	-20.0710	-9.0921	0.72751
590	-9.6363	-11.3721	2.80080	-16.8616	-5.8827	1.73584
591	-11.2934	-4.9966	2.80080	-10.4861	0.4929	-6.29681
592	1.9562	2.7285	2.80080	-2.7610	8.2180	-0.77230
593	8.5075	10.6887	2.80080	5.1992	16.1781	-2.18114
594	18.2143	16.6261	2.80080	11.1366	22.1156	1.58822
595	21.3070	19.7725	2.80080	14.2831	25.2620	1.53445
596	18.2610	19.2881	2.80080	13.7986	24.7775	-1.02704
597	13.6212	12.8260	2.80080	7.3366	18.3155	0.79522
598	0.2243	5.3941	2.80080	-0.0954	10.8836	-5.16982
599	-5.2710	-6.8012	2.80080	-12.2907	-1.3117	1.53025
600	-8.3499	-12.5839	2.80080	-18.0733	-7.0944	4.23394
601	.	-16.1349	2.80080	-21.6244	-10.6455	.
602	.	-11.7737	2.88410	-17.4264	-6.1210	.
603	.	-5.5976	2.90737	-11.2960	0.1007	.
604	.	3.9517	2.90737	-1.7466	9.6501	.
605	.	11.5895	2.90737	5.8911	17.2878	.
606	.	17.4977	2.90737	11.7993	23.1960	.

Figure 3.12: Forecasted values from January, 2002 to June 2003

To get a better idea of how well the forecast values perform for mean temperature, Figure 3.13 plots the forecasted and actual mean temperature values from January, 2000 to December, 2003.

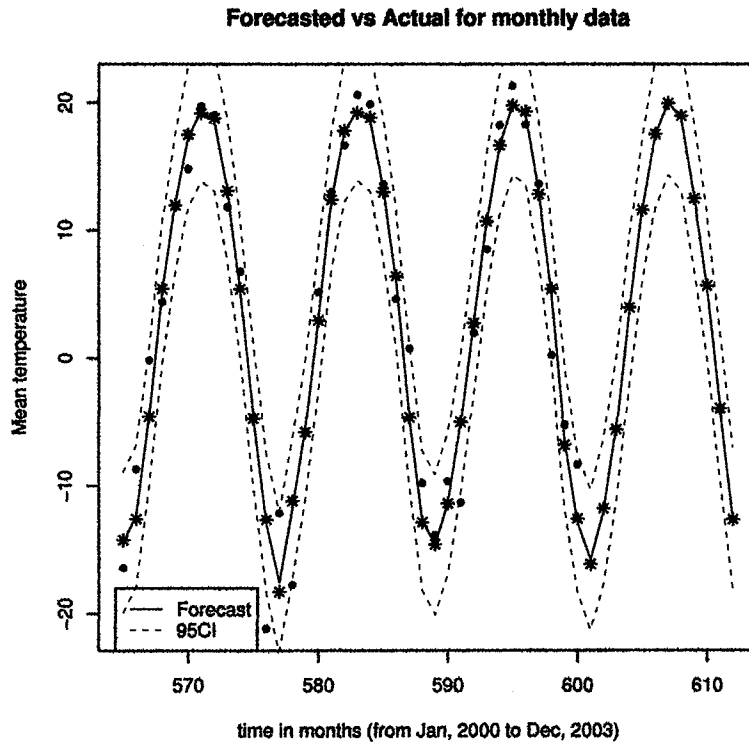


Figure 3.13: Plot of Forecasted (stars) and actual (dots) values from Jan, 2000 to Dec, 2003

The forecasted values in Figure 3.13 are fairly close to the actual values in the given plot. We see that all the values are inside the 95% confidence interval implying that we have a good model. The calculations for the forecasted values will have a more in depth explanation given in the next section for the Maximum temperature.

### 3.2.2 An ARIMA Model for Maximum Temperature

Lets consider the time series plot of monthly max temperatures shown in Figure 3.14. Again, we see that the winter months max temperature are regularly lower

than those in other months within the same year, while the summer months max temperature values are regularly higher. This suggests that max temperature values in any given month are similar to the max temperature values in the corresponding month in other years, indicating that we have a seasonal pattern occurring every 12 months. We will see this more clearly by looking at the sample ACF and PACF plots.

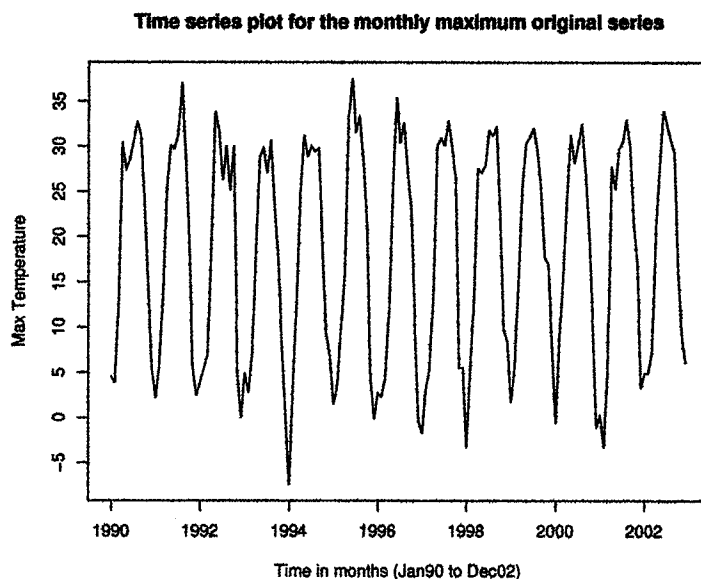


Figure 3.14: Time series plot of the monthly maximum temperature

If we observe the sample ACF and PACF of the original series for maximum temperature in Figure 3.15, we see the estimated sample ACF drops toward zero quite slowly. This implies that the series has a nonstationary mean and needs to be differenced. The series also shows a decay at the seasonal lags indicating a span of 12 difference is needed to transform the data into a stationary mean.

As we have done in the subsection for mean temperature, we will first use the first seasonal difference by the seasonal length of  $S=12$  in hopes to induce a stationary mean.

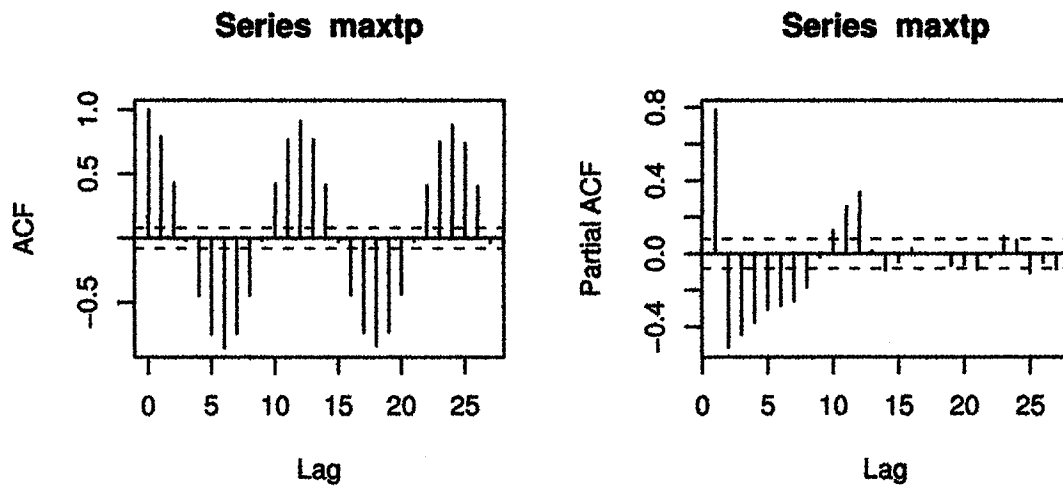


Figure 3.15: Sample ACF and PACF plots of the max temp

In Figure 3.16, the plot suggests that the seasonal differencing has removed any obvious peak seasonal variation appearing in the original data.

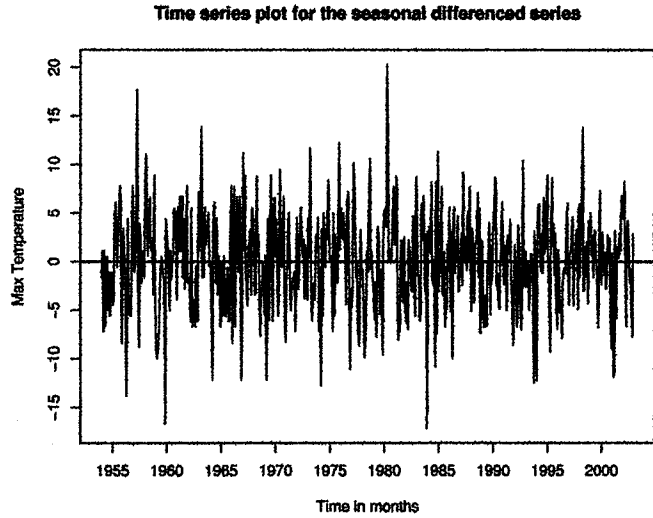


Figure 3.16: Time series plot of the seasonally differenced max temp

Now if we take the sample ACF and PACF of the seasonal differencing of 12 from the original data we obtain the output in Figure 3.17.

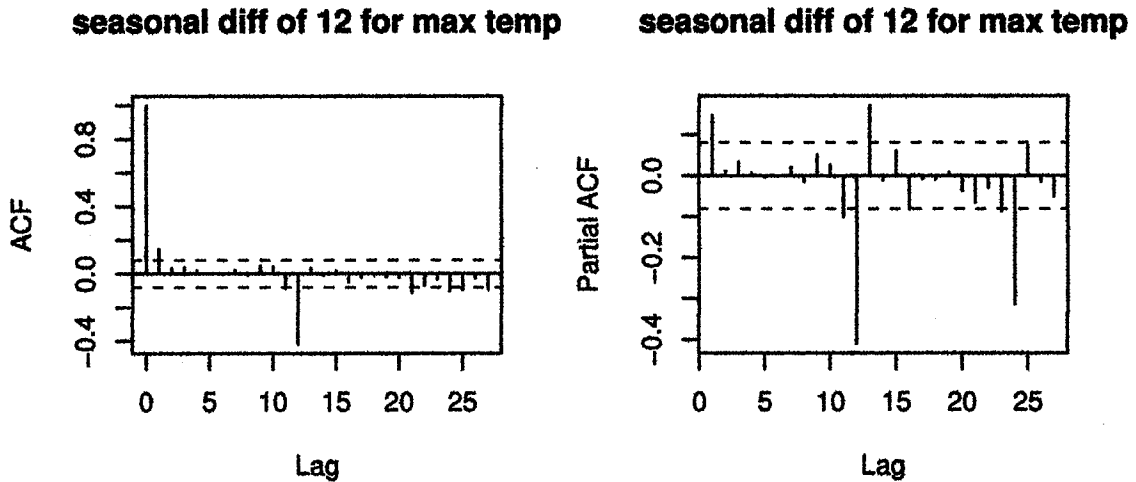


Figure 3.17: Sample ACF and PACF plots of the max temp seasonal diff of 12

The seasonal first differencing induces a stationary mean. The estimated sample ACF in Figure 3.17 moves rapidly to zero at the short lags cutting off after the first lag. On the seasonal component of the sample ACF, there is a spike at lag 12 followed by an insignificant autocorrelation at lag 24 while the seasonal spikes on the sample PACF has decayed slowly. This calls for a seasonal MA(1) coefficient.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MU	-0.0041959	0.01937	-0.22	0.8286	0
MA1,1	0.89620	0.01865	48.05	<.0001	12

Figure 3.18: Table of parameter estimates for seasonal MA(1)

Looking at the parameter estimates from Figure 3.18, the t-ratio for the seasonal moving average parameter is 48.05. This implies that this term is highly significant, but the constant term  $\mu$  is not and therefore will be removed. Looking at the residual sample ACF and PACF for the model in Figure 3.19, the seasonal decay at lag 12 and 24 has now disappeared, but a large spike is still visible at lag 1 on the sample ACF.



residual acf for ARIMA(0,0,0)(0,1,1)

residual pacf for ARIMA(0,0,0)(0,1,1)

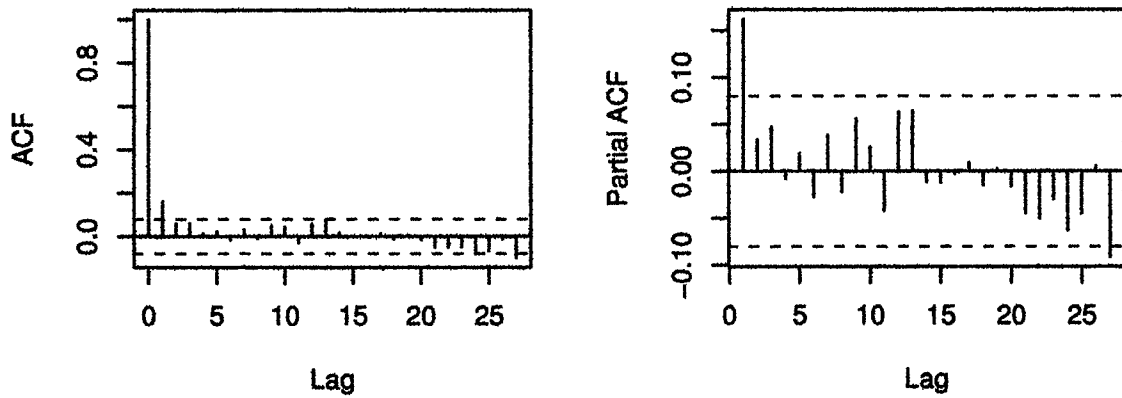


Figure 3.19: Residual sample ACF and PACF plots of the mean temp

If we try an AR(1) into the model, we see in Figure 3.20 that the model is significant with the new added coefficient as well as the seasonal MA term at lag 12. The model is invertible since  $\Theta_1$  satisfies  $|\hat{\Theta}_1| < 1$ . It is also stationary because  $\hat{\phi}_1 < 1$ .

Conditional Least Squares Estimation

Parameter	Estimate	Standard		Approx	
		Error	t Value	Pr >  t	Lag
MA1,1	0.88954	0.01913	46.49	<.0001	12
AR1,1	0.14924	0.04094	3.65	0.0003	1

Figure 3.20: Table of parameter estimates for a nonseasonal AR(1)

Our next step after the parameters in a model have been estimated is to check whether the model assumptions are satisfied. This is performed in our  $\chi^2$  test statistics displayed in Figure 3.21. It shows that our model is adequate because all of the  $\chi^2$  values are small with P-values very large, indicating that the residuals could be considered as white noise (uncorrelated).

Autocorrelation Check of Residuals									
To	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	3.91	4	0.4177	-0.007	0.038	0.059	-0.012	0.034	-0.017
12	13.25	10	0.2098	0.049	-0.033	0.054	0.051	-0.060	0.055
18	15.65	16	0.4779	0.056	0.014	0.001	-0.010	0.022	-0.007
24	21.41	22	0.4957	0.020	-0.020	-0.034	-0.032	-0.044	-0.067
30	37.48	28	0.1086	-0.071	0.020	-0.109	0.084	0.038	0.007
36	42.22	34	0.1573	0.026	-0.010	0.074	-0.020	-0.023	-0.018
42	46.85	40	0.2120	-0.046	0.027	-0.015	-0.005	0.064	0.015
48	50.37	46	0.3045	0.002	0.010	0.025	0.008	-0.057	0.038

Figure 3.21: Check for white noise for an ARIMA(1,0,0)(0,1,1)S=12

The estimated ARIMA model for maximum temperature becomes

$$(1 - \phi_1 B)(1 - B^{12})Z_t = (1 - \Theta_1 B^{12})a_t \quad (3.4)$$

or with the estimated coefficients substituted into the model, we now get:

$$(1 - 0.14924B)(1 - B^{12})Z_t = (1 - 0.88954B^{12})a_t \quad (3.5)$$

where

$Z_t$  is the observed series

$\Theta_1$  is the seasonal moving average parameter of order 1

$\phi_1$  is the nonseasonal autoregressive parameter of order 1

$(1 - B^{12})$  is the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

Since the model seems adequate, we can now forecast maximum temperature future values from January, 2002 to June, 2003 displayed in Figure 3.22.

Obs	max_temp	FORECAST	STD	L95	U95	RESIDUAL
589	4.9	0.4541	3.81344	-7.0201	7.9284	4.44585
590	4.9	3.9789	3.81344	-3.4953	11.4531	0.92111
591	7.3	9.8802	3.81344	2.4060	17.3544	-2.58017
592	21.0	22.5769	3.81344	15.1027	30.0511	-1.57690
593	27.8	29.2540	3.81344	21.7798	36.7282	-1.45397
594	33.9	30.5135	3.81344	23.0393	37.9877	3.38652
595	32.4	31.3318	3.81344	23.8576	38.8060	1.06823
596	30.7	32.4990	3.81344	25.0248	39.9732	-1.79898
597	29.3	28.2924	3.81344	20.8182	35.7666	1.00760
598	16.0	21.7782	3.81344	14.3040	29.2524	-5.77819
599	9.3	9.7762	3.81344	2.3020	17.2504	-0.47621
600	6.1	2.9385	3.81344	-4.5357	10.4127	3.16152
601	.	1.3631	3.81344	-6.1111	8.8373	.
602	.	3.5528	3.85568	-4.0042	11.1098	.
603	.	9.3941	3.85661	1.8353	16.9529	.
604	.	22.7152	3.85663	15.1564	30.2741	.
605	.	29.3493	3.85663	21.7905	36.9082	.
606	.	31.1188	3.85663	23.5599	38.6776	.

Figure 3.22: Data set with forecasted values of the adequate model

The first thing one may ask is how are the forecasted values generated? To answer this, lets use the  $ARIMA(1, 0, 0)(0, 1, 1)_{S=12}$  model for max temp, Equation (3.4) can be rewritten as

$$(1 - \phi_1 B)(Z_t - Z_{t-12}) = a_t - \Theta_1 a_{t-12} \quad ,$$

which is further reduced to

$$Z_t - \phi_1 Z_{t-1} - Z_{t-12} + \phi_1 Z_{t-13} = a_t - \Theta_1 a_{t-12} \quad ,$$

or

$$Z_t = \phi_1 Z_{t-1} + Z_{t-12} - \phi_1 Z_{t-13} - \Theta_1 a_{t-12} + a_t \quad .$$

Given the estimated coefficients from Figure 3.20 where  $\hat{\phi}_1 = 0.14924$  and  $\hat{\Theta}_1 = 0.88954$ , as well as the actual data and residuals from Figure 3.22, we can now forecast for January, 2003:

$$\begin{aligned} \hat{Z}_t(1) &= z_{601} \\ &= \hat{\phi}_1 Z_{600} + Z_{589} - \hat{\phi}_1 Z_{588} - \hat{\Theta}_1 \hat{a}_{589} + 0 \\ &= 0.14924(6.1) + 4.9 - 0.14924(3.3) - 0.88954(4.44585) \\ &= 1.3631 \end{aligned}$$

for February, 2003:

$$\begin{aligned} \hat{Z}_t(2) &= z_{602} \\ &= \hat{\phi}_1 Z_{601} + Z_{590} - \hat{\phi}_1 Z_{589} - \hat{\Theta}_1 \hat{a}_{590} \\ &= 0.14924(1.3631) + 4.9 - 0.14924(4.9) - 0.88954(0.92111) \\ &= 3.5528 \end{aligned}$$

... and so on

If we keep repeating this, then the forecasted max temperature value for June, 2003 is:

$$\begin{aligned}
 \hat{Z}_t(6) &= z_{606} \\
 &= \hat{\phi}_1 Z_{605} + Z_{594} - \hat{\phi}_1 Z_{593} - \hat{\Theta}_1 \hat{a}_{594} \\
 &= 0.14924(29.3493) + 33.9 - 0.14924(27.8) - 0.88954(3.38652) \\
 &= 31.1188
 \end{aligned}$$

Note that when the residuals reach  $\hat{a}_{t+l}$  where  $l \geq 0$ , we assign its expected value of zero.

The next step is to find the variance of the forecasted values. Before we start, we need to first write the model in the *moving average* form of the white noise because they are especially useful for estimating the variance of the forecasts [1, 7]. The coefficients of the random shock form are denoted by  $\psi_i$ .

$$\begin{aligned}
 Z_t &= \psi(B)a_t = (1 + \psi_1 B + \psi_2 B^2 + \dots)a_t \\
 &\Rightarrow (1 - \phi_1 B - B^{12} + \phi_1 B^{13})(1 + \psi_1 B + \psi_2 B^2 + \dots + \psi_k B^k) = 1 - \Theta B^{12} \\
 &\Rightarrow \psi_0 + (\psi_1 - \phi_1 \psi_0)B + (\psi_2 - \phi_1 \psi_1)B^2 + (\psi_3 - \phi_1 \psi_2)B^3 + \dots
 \end{aligned}$$

Now for  $k = 1, 2, \dots, 11$  we set the coefficients of  $B^k$  on the LHS equal to the coefficients of the same power of  $B^k$  on the RHS.

$B^k$	$\psi_k$
$B^0$	$\psi_0 = 1$
$B^1$	$\psi_1 - \phi_1 \psi_0 = 0 \Rightarrow \psi_1 = \phi_1$
$B^2$	$\psi_2 - \phi_1 \psi_1 = 0 \Rightarrow \psi_2 = \phi_1^2$
$B^3$	$\psi_3 - \phi_1 \psi_2 = 0 \Rightarrow \psi_3 = \phi_1^3$
$\vdots$	$\vdots$
$B^k$	$\psi_k - \phi_1 \psi_{k-1} = 0 \Rightarrow \psi_k = \phi_1^k$

The variance of the forecast error is then given by [1]

$$V[e_t(\ell)] = \sigma_a^2(1 + \psi_1^2 + \psi_2^2 + \dots + \psi_{\ell-1}^2) \quad (3.6)$$

Now given the estimated coefficients  $\phi_1$  and  $\Theta_1$  from Figure 3.20, as well as the forecast standard error  $\hat{\sigma}_a = 3.81344$  from Figure 3.22, we can find the variance for January, February and March:

$$\begin{aligned}\psi_0 &= 1 \\ \psi_1 &= \phi_1 = 0.14924 \\ \psi_2 &= \phi_1^2 = 0.0222726 \\ \psi_3 &= \phi_1^3 = 0.00332396 \\ &\vdots\end{aligned}$$

Using the variance in (3.6), we obtain the standard errors of each forecast error

$$SE[e_t(\ell)] = \sqrt{V[e_t(\ell)]} = \hat{\sigma}_a(1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{\ell-1}^2)^{\frac{1}{2}},$$

So that

$$SE[e_t(1)] = \hat{\sigma}_a\sqrt{1} = 3.81344\sqrt{1} = 3.81344$$

$$SE[e_t(2)] = \hat{\sigma}_a\sqrt{1 + \psi_1^2} = 3.81344\sqrt{1 + 0.14924^2} = 3.85567$$

$$SE[e_t(3)] = \hat{\sigma}_a\sqrt{1 + \psi_1^2 + \psi_2^2} = 3.81344\sqrt{1 + 0.14924^2 + 0.0222726^2} = 3.85661$$

Thus the 95% confidence interval for the forecast error becomes  $\hat{Z}_t(\ell) \pm 1.96SE[e_t(\ell)]$

$$\text{for } \ell = 1: \quad \hat{Z}_t(1) \pm 1.96(3.81344) = (-6.11124, 8.8374)$$

$$\text{for } \ell = 2: \quad \hat{Z}_t(2) \pm 1.96(3.85567) = (-4.0043, 11.1099)$$

$$\text{for } \ell = 3: \quad \hat{Z}_t(3) \pm 1.96(3.85661) = (1.8351, 16.9530)$$

The plot in Figure 3.23 displays the max temperature forecasted values from Jan, 2000 to Dec, 2003. The forecasted values are fitted with a smoothing spline with  $\lambda = 0.2$  to get a general pattern of the data.

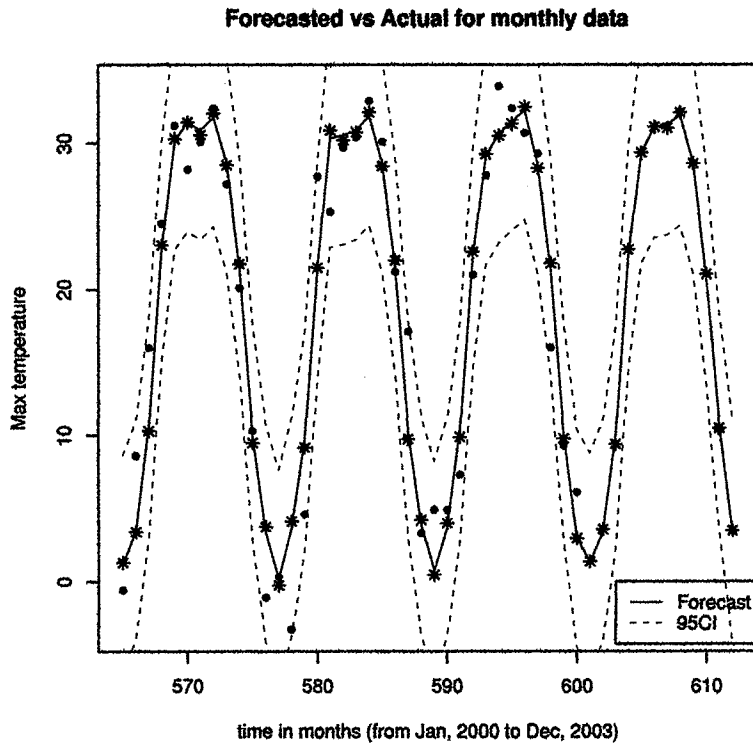


Figure 3.23: Plot of monthly forecasted (stars) and actual (dots) max temp values from Jan, 2000 to Dec, 2003

The model captures the seasonal pattern quite well with no values outside the 95% confidence limits.

### 3.2.3 An ARIMA Model for Minimum Temperature

Lets consider the time series plot of monthly minimum temperatures shown in Figure 3.24. Again, we see that the winter months minimum temperature are regularly lower than those in other months within the same year, while the summer months min temperature values are regularly higher. This suggest that min temperature values

in any given month are similar to the min temperature values in the corresponding month in other years, indicating that we have a seasonal pattern modeling every 12 months.

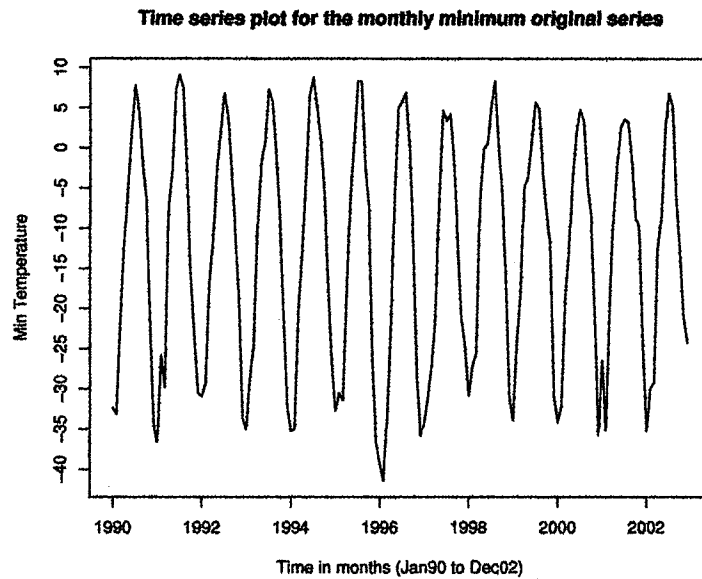


Figure 3.24: Time series plot of the monthly minimum temperature

The original series for minimum temperature has a nonstationary mean since the estimated sample ACF drops toward zero quite slowly. Just like the sample ACF for mean and maximum temperature, it shows a decay at the seasonal lags indicating a span of 12 difference is needed. Now if we take a sample ACF and PACF of the seasonal differencing  $D = 1$  from the original data, we get the following output shown in Figure 3.25 and Figure 3.26.



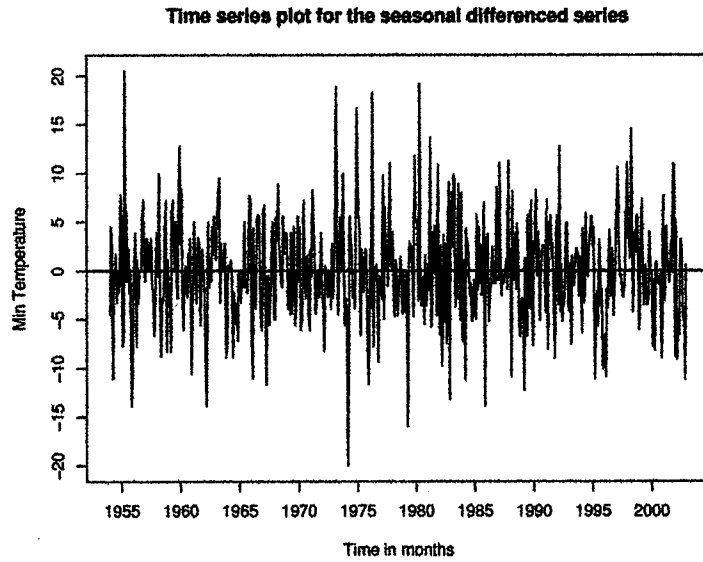


Figure 3.25: Time series plot of the seasonally differenced min temperature

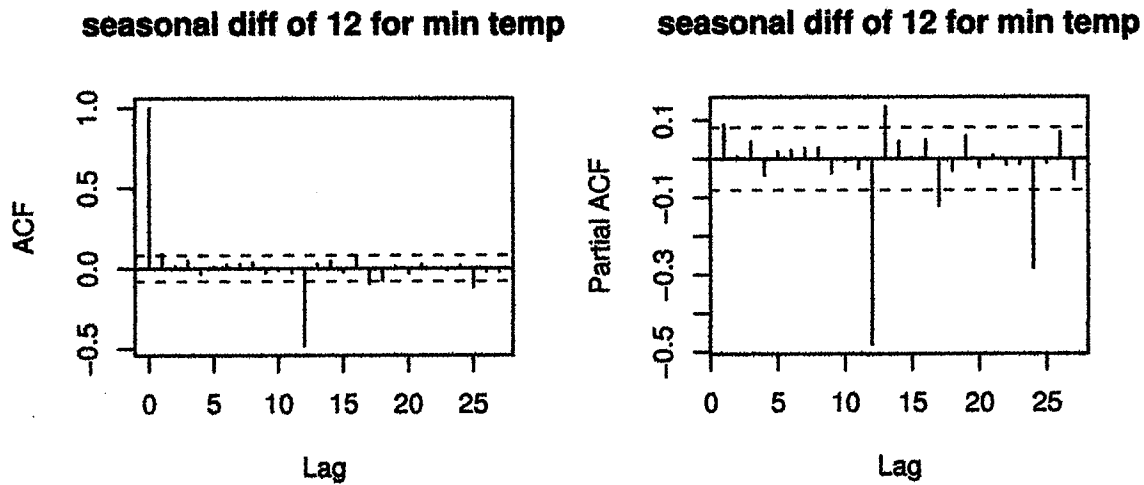


Figure 3.26: Sample ACF and PACF plots of the min temp seasonal diff of 12

After taking the seasonal first differencing, the plot in Figure 3.25 suggest that the seasonal differencing has removed any obvious peak seasonal variation appearing in the original data, inducing a stationary mean.

The estimated sample ACF in Figure 3.26 moves rapidly to zero at the short lags, and the spike at lag 12 is followed by an insignificant autocorrelation at lag 24 which calls for a seasonal MA(1) coefficient since there is a slowly seasonal decay on the sample PACF.

Now if we skip through a few procedures, we get almost the same results as we did for the maximum temperature model with an  $ARIMA(1, 0, 0)(0, 1, 1)_{s=12}$ . The estimates using PROC ARIMA in SAS are shown in Figure 3.27.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Pr >  t	Lag
MA1,1	0.86829	0.02090	41.54	<.0001	12
AR1,1	0.14203	0.04101	3.46	0.0006	1

Figure 3.27: Table of parameter estimates for a  $ARIMA(1,0,0)(0,1,1)$

Both the nonseasonal AR(1) and seasonal MA(1) coefficients are significant in the model with the P-values of less than 1%, but we're still missing a final component into the model. In our autocorrelation check of residuals displayed in Figure 3.28, not all the  $\chi^2$  values are small and not all of the P-values are large. We have lags 6 through 24 that are significant, but the rest of the lags are still not with pvalues < 0.05.

Autocorrelation Check of Residuals

To	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	5.22	4	0.2658	-0.006	0.035	0.072	-0.037	-0.026	0.017
12	7.83	10	0.6450	0.002	0.016	0.031	0.006	0.054	-0.012
18	17.29	16	0.3673	0.042	0.077	0.018	0.002	-0.084	-0.021
24	19.94	22	0.5869	-0.007	-0.005	0.056	-0.009	0.032	0.001
30	44.64	28	0.0240	-0.106	0.033	0.032	-0.137	0.057	0.069
36	52.26	34	0.0235	-0.070	0.056	0.012	-0.023	0.021	-0.055
42	58.18	40	0.0315	-0.003	-0.010	0.044	0.057	0.050	-0.040
48	63.37	46	0.0454	0.019	-0.008	0.001	0.032	-0.080	0.015

Figure 3.28: Check for white noise for a ARIMA(1,0,0)(0,1,1)

Looking under the *autocorrelation column* shown in the previous Figure, we can see that the 28th lag (-0.137) is larger than the other lags of the residuals indicating a possible AR(28) term might need to be added into the model. By adding the AR(28) term into the model, we should hopefully eliminate the large autocorrelation value given at lag 28.

Conditional Least Squares Estimation

Parameter	Estimate	Standard		Approx		Lag
		Error	t Value	Pr >  t		
MA1,1	0.86789	0.02094	41.45	<.0001		12
AR1,1	0.14440	0.04072	3.55	0.0004		1
AR1,2	-0.13131	0.04166	-3.15	0.0017		28

Figure 3.29: Table of parameter estimates for a AR(1,28) and seasonal MA(1) with diff=12

Figure 3.29 shows us that all of the terms with AR(28) added into the model appear to be significant with small p-values. The model is invertible since  $\hat{\Theta}_1$  satisfies  $|\hat{\Theta}_1| < 1$ . The model is also stationary because  $\hat{\phi}_1$  and  $\hat{\phi}_{28}$  meet the necessary conditions [7].

Now we can check whether the model assumptions are satisfied. Looking at the  $\chi^2$  test statistics shown in Figure 3.30, it shows that our model is now adequate because all P-values are larger than 0.05, indicating that the residuals could be uncorrelated.

Autocorrelation Check of Residuals									
To	Chi-		Pr >	-----Autocorrelations-----					
Lag	Square	DF	ChiSq						
6	4.09	3	0.2515	0.004	0.049	0.053	-0.032	-0.022	0.015
12	6.44	9	0.6954	0.011	0.009	0.030	0.003	0.052	-0.006
18	16.86	15	0.3273	0.051	0.086	0.026	0.001	-0.079	-0.017
24	19.13	21	0.5765	-0.012	-0.003	0.051	0.004	0.030	0.004
30	33.95	27	0.1674	-0.101	0.039	0.035	-0.017	0.070	0.077
36	40.50	33	0.1731	-0.056	0.057	0.016	-0.020	0.023	-0.054
42	45.62	39	0.2160	0.005	-0.003	0.048	0.047	0.054	-0.026
48	50.47	45	0.2663	0.024	-0.010	-0.005	0.030	-0.075	0.018

Figure 3.30: Check for white noise for a nonseasonal AR(1,28) and seasonal MA(1) with diff=12

And the estimated ARIMA model for minimum temperature is,

$$(1 - \phi_1 B - \phi_{28} B^{28})(1 - B^{12})Z_t = (1 - \Theta_1 B^{12})a_t \quad (3.7)$$

where

$Z_t$  is the observed series

$\Theta_1$  is the seasonal moving average coefficients of order 1

$\phi_1$  and  $\phi_{28}$  are the nonseasonal autoregressive parameters

$(1 - B^{12})$  are the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

or with the estimated coefficients substituted into the model, we now get the model,

$$(1 - 0.1444B + 0.13131B^{28})(1 - B^{12})Z_t = (1 - 0.86789B^{12})a_t \quad (3.8)$$

Now that the model is adequate, we can forecast the minimum temperature future values from January, 2002 to June, 2003 displayed in Figure 3.31. Each of the observations represent one month. We forecasted for the year 2002 to see how our forecasts compare to the actual data.

Obs	min_temp	FORECAST	STD	L95	U95	RESIDUAL
589	-35.2	-31.8720	3.89695	-39.5099	-24.2342	-3.3280
590	-29.9	-31.8355	3.89695	-39.4734	-24.1976	1.9355
591	-29.2	-24.5618	3.89695	-32.1997	-16.9240	-4.6382
592	-12.6	-12.7972	3.89695	-20.4351	-5.1593	0.1972
593	-8.6	-3.7453	3.89695	-11.3831	3.8926	-4.8547
594	2.4	2.0879	3.89695	-5.5499	9.7258	0.3121
595	6.8	5.3020	3.89695	-2.3359	12.9398	1.4980
596	5.2	5.3768	3.89695	-2.2611	13.0147	-0.1768
597	-6.8	-2.0042	3.89695	-9.6420	5.6337	-4.7958
598	-12.5	-9.5487	3.89695	-17.1866	-1.9108	-2.9513
599	-21.0	-19.1370	3.89695	-26.7748	-11.4991	-1.8630
600	-24.3	-30.7991	3.89695	-38.4370	-23.1612	6.4991
601	.	-32.0412	3.89695	-39.6791	-24.4033	.
602	.	-31.0843	3.93736	-38.8014	-23.3672	.
603	.	-24.1638	3.93820	-31.8825	-16.4450	.
604	.	-11.4267	3.93822	-19.1455	-3.7080	.
605	.	-5.2283	3.93822	-12.9471	2.4905	.
606	.	3.0100	3.93822	-4.7088	10.7287	.

Figure 3.31: Forecasting future min temperature from Jan, 2002 to Jun, 2003

Figure 3.32 gives a better visual representation of the forecasted values from January, 2000 to December, 2003 with the smoothing spline  $\lambda = 0.2$ .

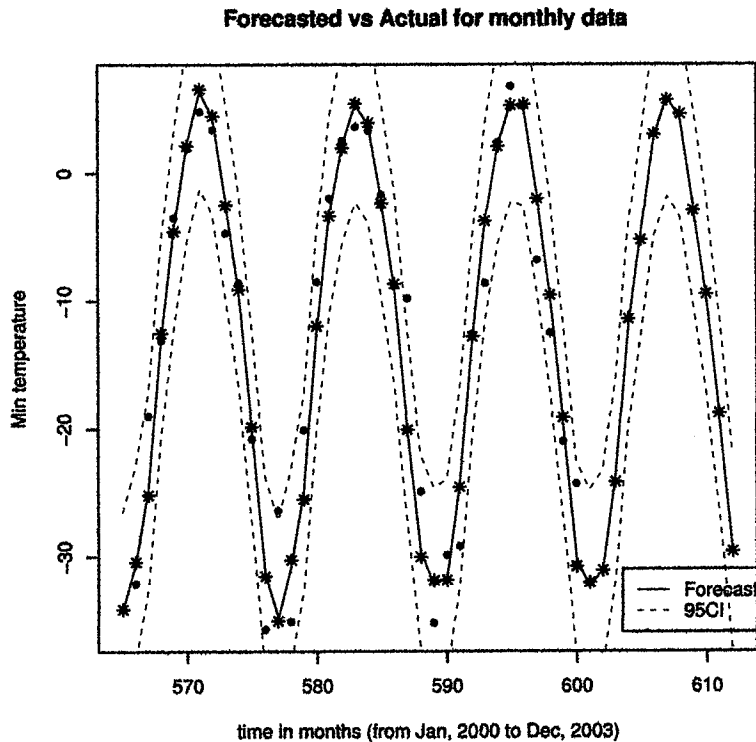


Figure 3.32: Plot of Forecasted (stars) and actual (dots) values from Jan, 2000 to Dec, 2003

We see that the forecasted values are fairly close to the data, indicating that this is a pretty good model to use. In the next section, we will look at the temperature in more detail by trying to find a model for daily temperature instead of monthly.

### 3.3 ARIMA models for Daily Temperature

In this section, we extend our discussion of ARIMA models for daily temperature. Specifically, instead of 600 observations used for monthly data, we will use 18262 observations of daily temperature taken from 01Jan53 to 31Dec02. Each of these observations consists of the mean, maximum and minimum taken for each day, month and year. This means that we're taking an average temperature of the 24 hours for

each day, as well as taking the extremes, maximum and minimum for each day. We will first discuss the mean temperature, followed by the extremes minimum and maximum.

### 3.3.1 Model for Mean Temperature

Lets consider the time series plot of daily mean temperatures shown in Figure 3.33, but we will only plot the last 6 years from 2003 to get an idea of the temperature pattern. We see that the mean temperature on winter days are regularly lower than those in other days within the same year, while the summer days mean temperatures are regularly higher. This suggests that mean temperature values in any given day are similar to the mean temperature values in the corresponding day in other years, indicating that we have a seasonal pattern modeling every 365 days.

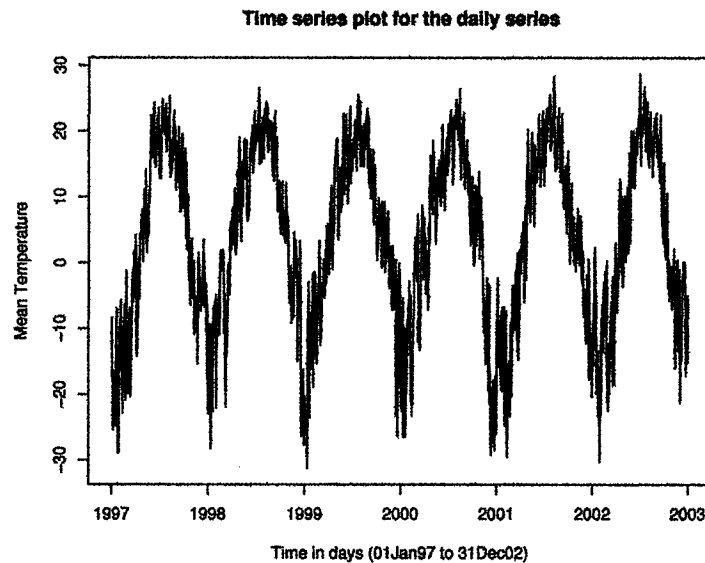


Figure 3.33: Time series plot of the daily mean temperature

The original series for the mean temperature displayed in Figure 3.33 has a non-stationary mean because the estimated sample ACF drops toward zero quite slowly (see Figure 3.34).



Since the temperature data oscillates on a yearly fashion, taking a seasonal difference of 365 will eliminate any of the seasonal pattern and transform the data into a stationary series.

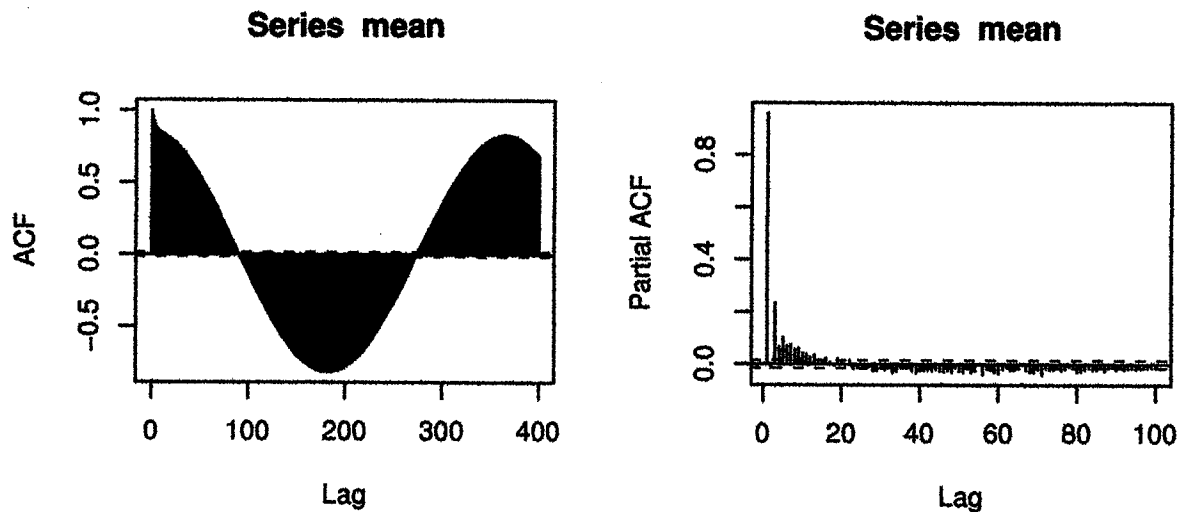


Figure 3.34: Sample ACF and PACF plots of the mean temp of the original series

Once we take the seasonal difference of 365 days, the seasonal pattern in Figure 3.35 has been removed and the daily mean temperature looks stationary with constant mean and variance.

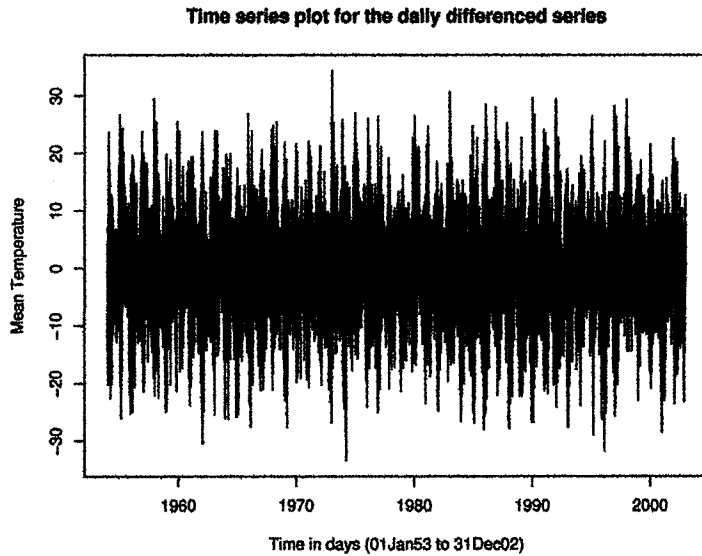


Figure 3.35: Time series plot of the daily differenced mean temperature

Now looking at sample ACF and PACF plots for the seasonally differenced data in Figure 3.36, we see there is a slow exponential decay for the ACF series and the pattern for the PACF dies off quickly after lag 2. Even though the sample ACF looks nonstationary, the mean of the differenced series is zero indicating that the series has a stationary mean. There appears to be no apparent seasonal pattern, therefore no seasonal AR or MA parameters will be needed for this model, only nonseasonal ones.

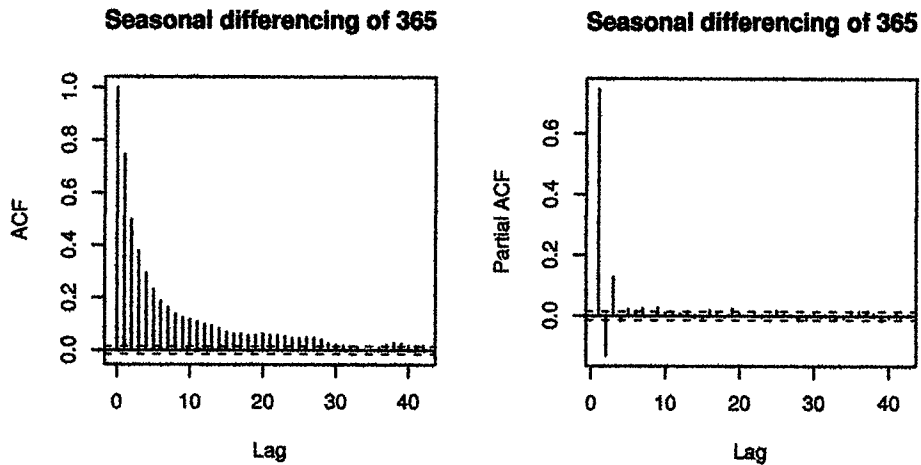


Figure 3.36: Sample ACF and PACF plots of the mean temp of the seasonal difference

Using the same method described in the monthly data, we get precise estimates of the coefficients of the model chosen at the identification stage, and use some diagnostic checks to help determine if the model is statistically adequate. This will help to find a reasonable model for the daily data and how it could be improved.

After many trial and errors to find a reasonable model, the estimated model in the Table of Estimates is shown in Figure 3.37. All of the coefficients appear to have large t-ratios and P-values  $< 0.01$ , indicating that the model parameters look good. It is now necessary to check whether the model assumptions are satisfied from the given parameters in the model.

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Pr >  t	Lag
MA1,1	-0.28780	0.0098600	-29.19	<.0001	1
MA1,2	0.59586	0.04103	14.52	<.0001	3
MA1,3	0.37409	0.03267	11.45	<.0001	4
AR1,1	0.57509	0.0088043	65.32	<.0001	1
AR1,2	0.67314	0.03912	17.21	<.0001	3
AR1,3	-0.13774	0.03733	-3.69	0.0002	4
AR1,4	-0.17475	0.02436	-7.17	<.0001	5

Figure 3.37: Table of parameter estimates for daily mean temp

Our  $\chi^2$  test statistics in Figure 3.38 shows that our model is adequate because our  $\chi^2$  values are small and the P-values are large for all the lags except for the first row which has a small  $\chi^2$  value and a small P-value. This occurred because of the fact that its a test statistic based on all the residual autocorrelations as a set [11]. We're given  $K = 6$  residual autocorrelations. The  $\chi^2$  test  $\sim \chi^2_{(K-m)}df$ , where  $m$  is the number of parameters estimated in the ARIMA model. Since the number of parameters in the model is  $m = 7$ , we obtain  $\chi^2 = 0$  and the P-value becomes small as a default. If we try reducing the model to 5 parameters or less, it doesn't produce an adequate model. Therefore, the model estimated in Figure 3.37 is the most reasonable model indicating that the residuals are white noise (uncorrelated) and are ready to forecast.

Autocorrelation Check of Residuals

To	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	0.00	0	<.0001	-0.001	-0.000	-0.002	0.001	0.006	0.002
12	4.03	5	0.5453	-0.001	-0.007	0.009	-0.006	0.004	0.003
18	15.28	11	0.1700	0.004	0.007	-0.013	-0.010	0.009	-0.014
24	22.35	17	0.1718	-0.003	0.010	-0.002	0.006	0.007	-0.014
30	30.57	23	0.1337	0.001	0.007	0.009	0.010	-0.015	-0.002
36	40.44	29	0.0770	-0.005	0.001	-0.001	-0.015	-0.016	-0.006
42	49.05	35	0.0578	0.005	0.013	0.011	-0.013	0.001	0.002
48	49.55	41	0.1691	0.002	-0.003	-0.000	-0.002	0.003	0.001

Figure 3.38: Check for white noise from the given model

The estimated ARIMA model for daily mean temp becomes,

$$(1 - \phi_1 B - \phi_3 B^3 - \phi_4 B^4 - \phi_5 B^5)(1 - B^{365})Z_t = (1 - \theta_1 B - \theta_3 B^3 - \theta_4 B^4)a_t \quad (3.9)$$

where

$Z_t$  is the observed series

$\theta_1, \theta_3$  and  $\theta_4$  are the nonseasonal moving average parameters

$\phi_1, \phi_3, \phi_4$  and  $\phi_5$  are the nonseasonal autoregressive parameters

$(1 - B^{365})$  is the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

Given estimated coefficients from the SAS output given from Figure 3.37, the above equation now becomes:

$$\begin{aligned} &(1 - 0.5750B - 0.6731B^3 + 0.1377B^4 + 0.1747B^5)(1 - B^{365})Z_t \\ &= (1 + 0.2878B - 0.5958B^3 - 0.3740B^4)a_t \end{aligned}$$

Since the model is now adequate, we can now forecast future values for the mean temperature from Dec 20, 2002 to Jan 06, 2003. The forecasted values are given in Figure 3.39.

date	mean_temp	FORECAST	STD	L95	U95	RESIDUAL
20DEC02	-5.38	-7.6183	5.28287	-17.9725	2.7359	2.2383
21DEC02	-6.86	-0.5834	5.28287	-10.9376	9.7708	-6.2766
22DEC02	-9.83	-9.5579	5.28287	-19.9121	0.7963	-0.2721
23DEC02	-11.75	-15.9224	5.28287	-26.2766	-5.5682	4.1724
24DEC02	-17.16	-12.2541	5.28287	-22.6083	-1.8999	-4.9059
25DEC02	-13.22	-19.2673	5.28287	-29.6216	-8.9131	6.0473
26DEC02	-11.21	-7.1334	5.28287	-17.4876	3.2208	-4.0766
27DEC02	-5.86	-11.2447	5.28287	-21.5989	-0.8904	5.3847
28DEC02	-10.68	-8.8136	5.28287	-19.1678	1.5406	-1.8664
29DEC02	-5.01	-10.5390	5.28287	-20.8932	-0.1847	5.5290
30DEC02	-8.59	-14.1572	5.28287	-24.5114	-3.8030	5.5672
31DEC02	-15.13	-8.7159	5.28287	-19.0701	1.6384	-6.4141
01JAN03	.	-13.0852	5.28287	-23.4394	-2.7309	.
02JAN03	.	-14.0812	6.97775	-27.7573	-0.4050	.
03JAN03	.	-11.1658	7.45397	-25.7753	3.4438	.
04JAN03	.	-8.4884	7.69624	-23.5728	6.5960	.
05JAN03	.	-5.2569	7.83470	-20.6126	10.0989	.
06JAN03	.	-14.6609	7.90566	-30.1558	0.8339	.

Figure 3.39: Forecast values of mean temp from 20Dec02 to 06Jan03

Using a smoothing spline of  $\lambda = 0.3$  to get a overview pattern of the data, the forecasted values in Figure 3.40 fit the data fairly well with all of the mean temperature values lying inside the 95% confidence interval. Thus we have found a reasonable model to forecast daily mean temperature.

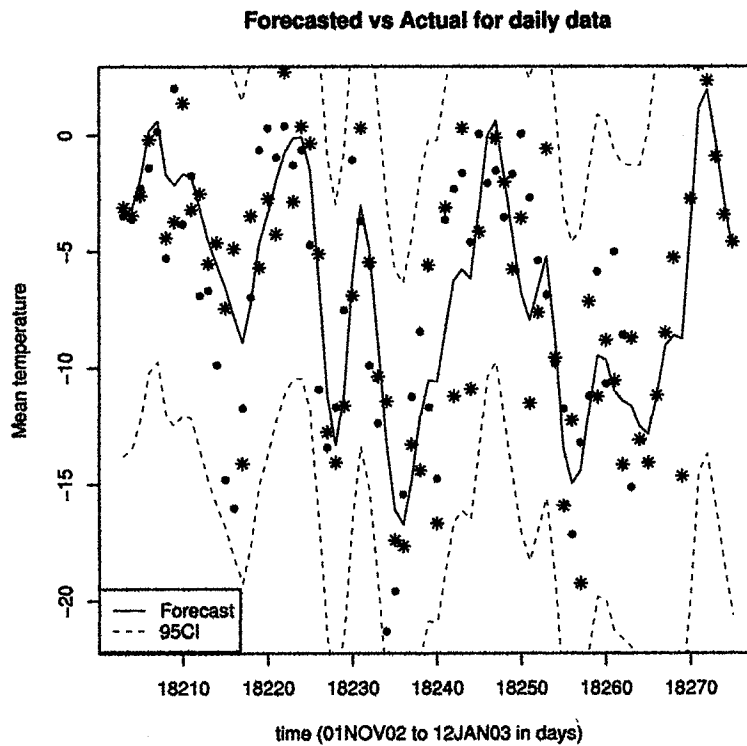


Figure 3.40: Plot of daily Forecasted (stars) and actual (dots) mean temp values

### 3.3.2 Model for Minimum Temperature

Lets consider the time series plot of daily minimum temperatures shown in Figure 3.41, but we will only plot the last 6 years from 2003 to get an idea of the temperature pattern. We see that the winter days for minimum daily temperature are regularly lower than those in other days within the same year.



This suggest that min temperature values in any given day are similar to the min temperature values in the corresponding day in other years, indicating that we have a seasonal pattern modeling every 365 days.

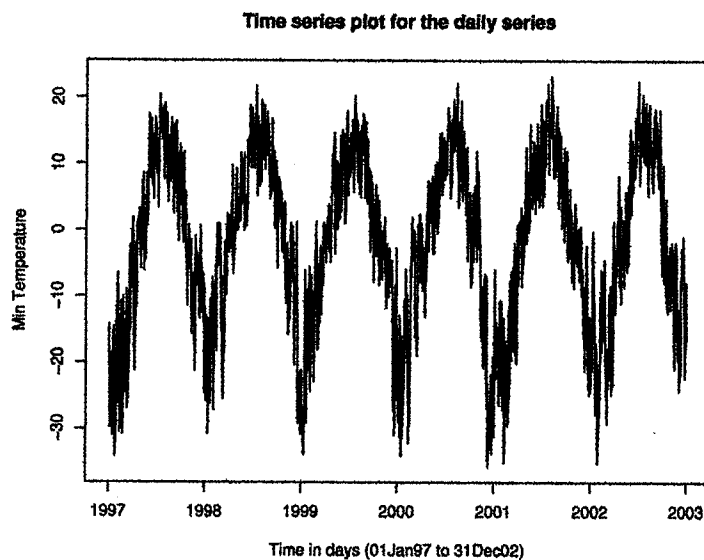


Figure 3.41: Time series plot of the daily minimum temperature

If we observe the original minimum temperature and proceed like the above procedure done for mean temperature in Figure 3.34, we will need to use seasonal difference of 365 to eliminate the nonstationarity from the original series. The differenced data displayed in Figure 3.42 and 3.43 has now been transformed into a stationary one due to the fact that the mean is now virtually zero.

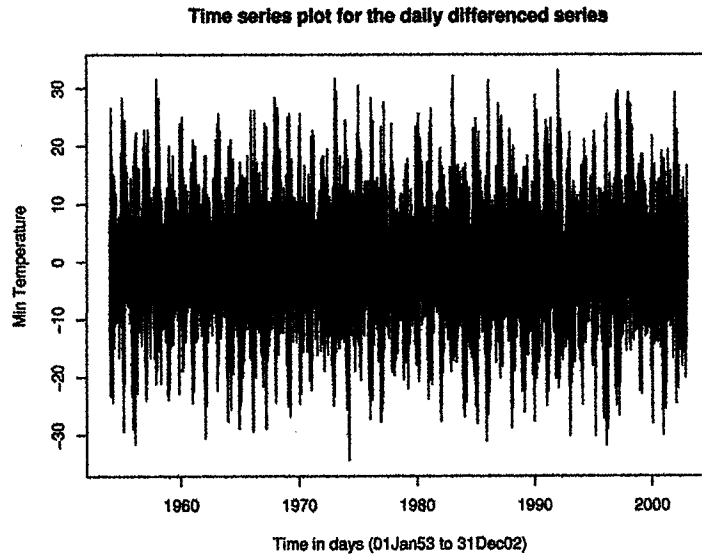


Figure 3.42: Time series plot of the differenced minimum temperature

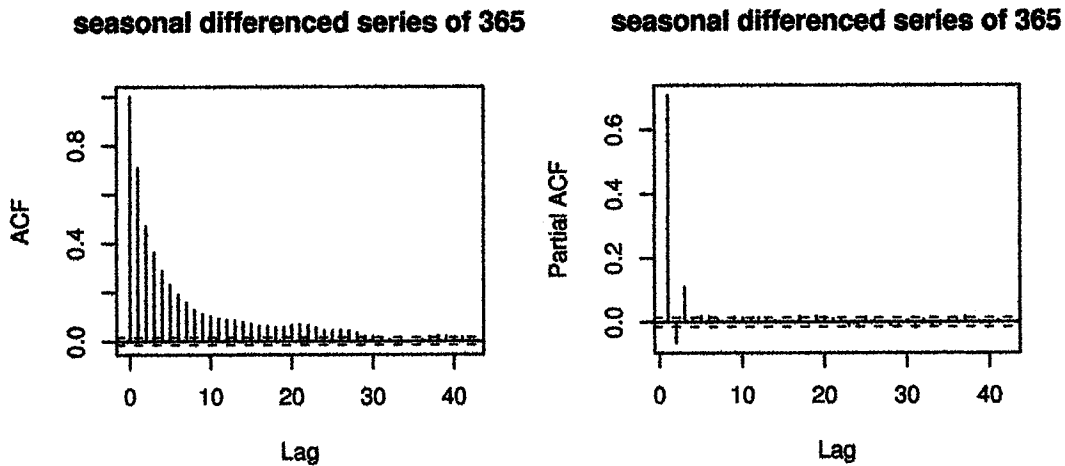


Figure 3.43: Sample ACF and PACF plots of the min temp of the seasonal difference

The sample ACF in Figure 3.43 has an exponential decay while the sample PACF cuts off quickly indicating that a possible AR(3) might be needed. Unfortunately if we take the AR(3) model into account, the diagnostic checking of the residuals leads to the model being inadequate. Further addition of estimates to the model only causes more problems in the model in which case they become more insignificant and need

to be reformulated.

There are some more complex functions, the *Extended Sample Autocorrelation Function*(ESACF) that can give some preliminary ideas about what  $p$  and  $q$  might be [12]. Each of these consists of a table with  $q$  listed across the top and  $p$  down the side, where the practitioner looks for a pattern of insignificant ESACF values on the table. Different results can be obtained depending on the number of user specified rows and columns in the table being searched. In addition, a method called *MINIC* is available in which every possible series in the aforementioned table is fit to the data and an information criterion computed. The fitting is based on an initial autoregressive approximation and thus avoids some of the non-identifying problems normally associated with fitting large numbers of autoregressive and moving average parameters. One of the non-identifying problems is there are still some  $(p, q)$  combinations that can often show failure to converge. The  $BIC(p, q)$  uses a Bayesian information criterion to select the number of  $p$  and  $q$  lags appropriate for the data, based on an initial long autoregressive approximation. The ESACF table's complex diagnostics are summarized in Figure 3.44.

ARMA(p+d, q) Tentative		
Order Selection Tests		
-----ESACF-----		
p+d	q	BIC
1	3	3.549616
4	5	3.551698
5	5	3.55212
(5% Significance Level)		

Figure 3.44: ESACF table of order selection

Notice that the BIC values are included and that  $p + d$  rather than  $p$  is indicated in the autoregressive part of the table. The  $d$  refers to differencing. Reading the table above, one of the optimal BIC model would be with  $p + d = 4$  and  $q = 5$ . This suggests that the first difference (perhaps seasonal) satisfy an ARIMA(3,5) model. Taking that model and eliminating any insignificant estimates or possibly adding new estimates if required, we get the following table given in Figure 3.45. Again, a portion of the SAS code used is given in Appendix B.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Pr >  t	Lag
MA1,1	1.22999	0.04270	28.81	<.0001	1
MA1,2	-0.28516	0.02071	-13.77	<.0001	3
MA1,3	0.03395	0.01123	3.02	0.0025	5
MA1,4	0.0042791	0.0021394	2.00	0.0455	29
AR1,1	1.98940	0.04477	44.43	<.0001	1
AR1,2	-1.08277	0.04077	-26.56	<.0001	2
AR1,3	0.08957	0.0077269	11.59	<.0001	4

Figure 3.45: Table of parameter estimates for daily minimum temperature

The estimates in the model appear to be significant. The model is both invertible and stationary since the sum of the each MA and AR coefficients are less than one. Now that we have obtained precise estimates of the coefficients in an ARIMA model, we need to perform diagnostic checking to check if the model is statistically adequate.

Autocorrelation Check of Residuals									
To	Chi-	Pr >		-----Autocorrelations-----					
Lag	Square	DF	ChiSq						
6	0.00	0	<.0001	0.000	-0.000	0.001	-0.002	-0.002	0.003
12	3.92	5	0.5603	0.010	-0.009	-0.004	0.002	-0.003	0.002
18	13.11	11	0.2865	0.007	0.002	0.003	-0.017	0.007	-0.011
24	25.40	17	0.0862	-0.006	0.007	0.007	0.013	0.002	-0.019
30	32.87	23	0.0834	-0.002	0.004	0.005	0.004	-0.016	0.011
36	40.41	29	0.0775	-0.013	0.000	0.001	-0.012	-0.008	-0.007
42	45.91	35	0.1027	0.012	0.009	0.007	-0.003	0.000	0.004
48	49.18	41	0.1784	-0.002	0.007	-0.001	-0.002	0.011	-0.001

Figure 3.46: Check for white noise for the above model

Figure 3.46 shows the overall model is adequate since all the P-values are large, except for the first row of lags. This was explained earlier with mean temp so our model for minimum temperature becomes:

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_4 B^4)(1 - B^{365})Z_t = (1 - \theta_1 B - \theta_3 B^3 - \theta_5 B^5 - \theta_{29} B^{29})a_t$$

where

$Z_t$  is the observed series for minimum temperature

$\theta_1, \theta_3, \theta_5$  and  $\theta_{29}$  are the nonseasonal moving average parameters

$\phi_1, \phi_2$  and  $\phi_4$  are the nonseasonal autoregressive parameters

$(1 - B^{365})$  is the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

Given estimated coefficients from the SAS output in Figure 3.45, the above equation now becomes:

$$(1 - 1.9894B + 1.0827B^2 - 0.0895^4)(1 - B^{365})Z_t$$

$$= (1 - 1.2299B + 0.2851B^3 - 0.0339B^5 - 0.0043B^{29})a_t$$

We are now ready to forecast future values for minimum temperature. Forecasted values of min temperature taken from *Dec20/02* to *Jan06/03* are shown in Figure 3.48. The forecast plot for the daily minimum temperature values are shown in Figure 3.47, with a smoothing spline of  $\lambda = 0.3$  for general behaviour of the data.

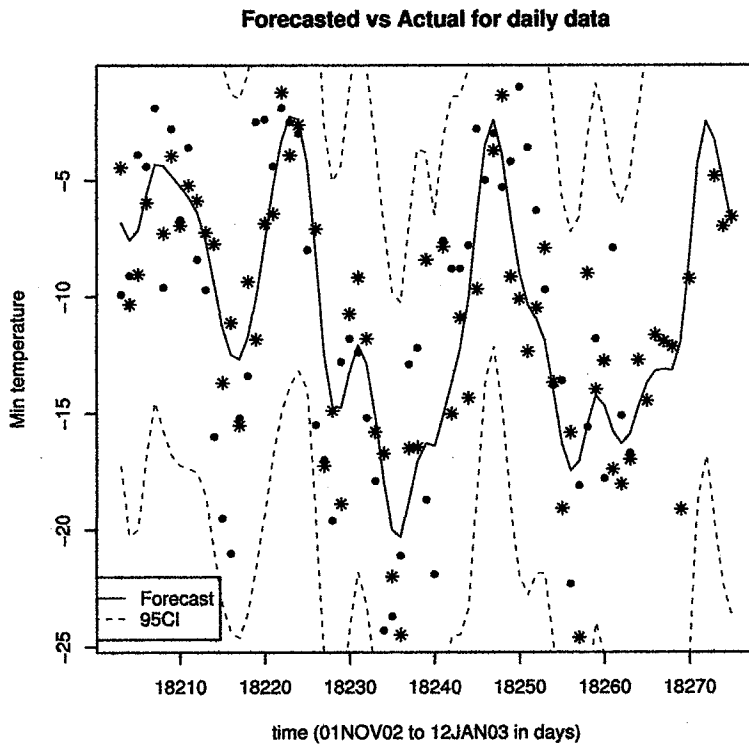


Figure 3.47: Plot of daily Forecasted (stars) and actual (dots) min temp values

date	min_temp	FORECAST	STD	L95	U95	RESIDUAL
20DEC02	-6.3	-10.4762	5.89426	-22.0287	1.0763	4.1762
21DEC02	-9.7	-7.9122	5.89426	-19.4647	3.6404	-1.7878
22DEC02	-13.8	-13.6612	5.89426	-25.2138	-2.1087	-0.1388
23DEC02	-13.6	-19.0695	5.89426	-30.6220	-7.5169	5.4695
24DEC02	-22.3	-15.8353	5.89426	-27.3879	-4.2828	-6.4647
25DEC02	-18.1	-24.6078	5.89426	-36.1604	-13.0553	6.5078
26DEC02	-15.6	-8.9972	5.89426	-20.5498	2.5553	-6.6028
27DEC02	-11.8	-13.9758	5.89426	-25.5284	-2.4233	2.1758
28DEC02	-17.8	-12.7556	5.89426	-24.3081	-1.2030	-5.0444
29DEC02	-7.9	-17.4016	5.89426	-28.9542	-5.8491	9.5016
30DEC02	-15.1	-18.0340	5.89426	-29.5865	-6.4814	2.9340
31DEC02	-16.7	-16.9777	5.89426	-28.5302	-5.4252	0.2777
01JAN03	.	-12.7128	5.89426	-24.2654	-1.1603	.
02JAN03	.	-14.4564	7.40125	-28.9625	0.0498	.
03JAN03	.	-11.6369	7.81940	-26.9626	3.6889	.
04JAN03	.	-11.9143	8.03596	-27.6645	3.8358	.
05JAN03	.	-12.1457	8.17160	-28.1617	3.8704	.
06JAN03	.	-19.1185	8.25130	-35.2907	-2.9462	.

Figure 3.48: Forecast values of minimum temp from Dec 20/02 to Jan 06/03

The forecasts from the model shows that it mimics the data quite well with fairly close temperature values.

### 3.3.3 Model for Maximum Temperature

Lets consider the time series plot of daily maximum temperatures shown in Figure 3.49, but we will only plot the last 6 years from 2003 to get an idea of the temperature

pattern. We see that the winter days for daily max temperature are regularly lower than those in other days within the same year. This suggests that max temperature values in any given day are similar to the max temperature values in the corresponding day in other years, indicating that we have a seasonal pattern modeling every 365 days.

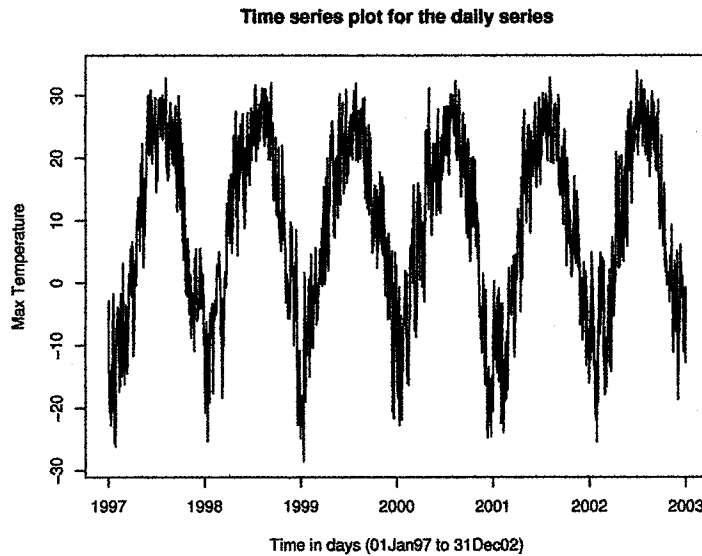


Figure 3.49: Time series plot of the daily minimum temperature

Taking the seasonal difference of 365 given in Figures 3.50 and 3.51, we transformed our nonstationary series into a stationary one with the mean now virtually zero.



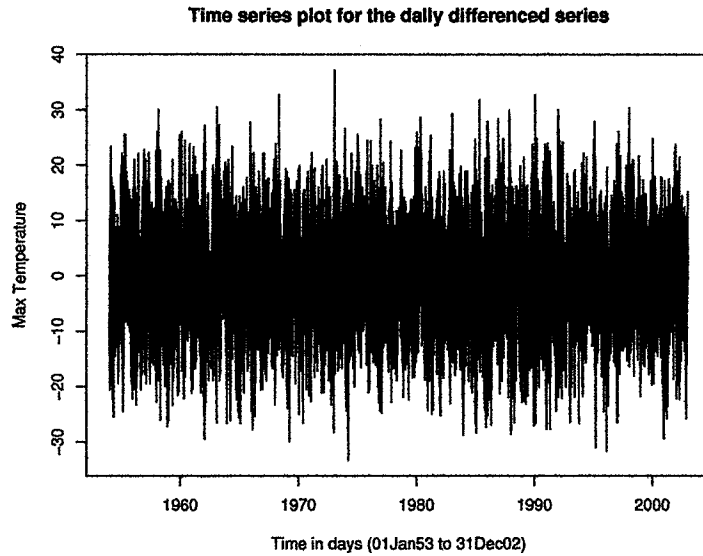


Figure 3.50: Time series plot of the differenced maximum temperature

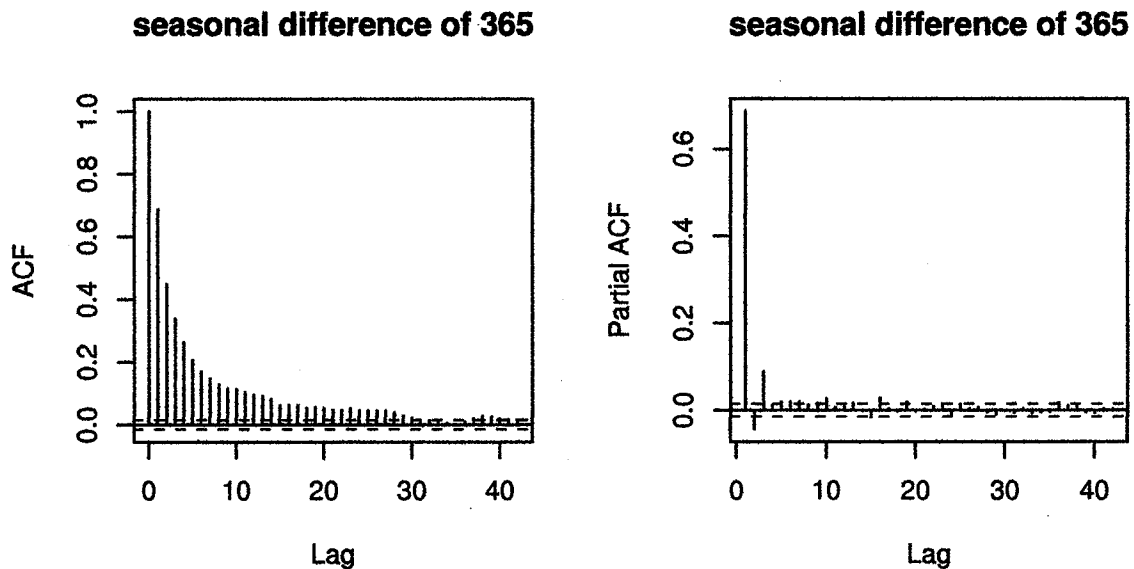


Figure 3.51: Sample ACF and PACF plots of the max temp of the seasonal difference

The differenced series has an exponential decay on the sample ACF plot and it cuts off to zero after two lags on the sample PACF plot indicating a possible AR(2) model might be needed.

Using PROC ARIMA in SAS, Figure 3.52 shows the ESACF table of complex diagnostics for possible order selection.

```
ARMA(p+d,q) Tentative
Order Selection Tests
-----ESACF-----
p+d      q      BIC
  2      3      3.712802
  1      5      3.713292
(5% Significance Level)
```

Figure 3.52: ESACF table of order selection

Reading the ESACF table, one of the optimal BIC model would be with  $p+d = 2$  and  $q = 3$ . This means that the first difference (perhaps seasonal) satisfy an ARIMA(1,3) model. There is no guarantee that this table will successfully capture the structure of the data, but it will give us a good place to start. Taking that model into account, we don't require one of the lower MA term and we needed more AR terms added into the model. In doing so, our estimates are shown in Figure 3.53.

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MA1,1	0.39948	0.02956	13.52	<.0001	3
MA1,2	0.29185	0.02266	12.88	<.0001	4
AR1,1	0.71693	0.0073439	97.62	<.0001	1
AR1,2	-0.10671	0.0091114	-11.71	<.0001	2
AR1,3	0.48016	0.02871	16.72	<.0001	3
AR1,4	-0.16076	0.01617	-9.94	<.0001	5
AR2,1	-0.02927	0.0076445	-3.83	0.0001	15

Figure 3.53: Table of parameter estimates for daily max temp

The estimates in the model all appear to be significant and lead to an, invertible and stationary model, indicating the parameters are adequate. Now that we have obtained precise estimates of the coefficients in an ARIMA model, we need to perform diagnostic checking to check if the model is statistically adequate.

Autocorrelation Check of Residuals

To	Chi-	Pr >		-----Autocorrelations-----						
Lag	Square	DF	ChiSq							
6	0.00	0	<.0001	0.001	-0.001	-0.001	0.002	0.002	0.004	
12	3.99	5	0.5512	-0.008	-0.004	-0.008	0.007	0.001	-0.003	
18	13.34	11	0.2715	0.007	0.013	-0.000	-0.002	0.008	-0.016	
24	22.02	17	0.1839	0.006	0.000	-0.005	-0.009	0.017	-0.008	
30	25.60	23	0.3202	0.002	0.001	0.009	0.008	-0.007	-0.001	
36	35.22	29	0.1973	-0.010	-0.001	-0.003	-0.005	-0.020	-0.003	
42	40.04	35	0.2564	0.002	0.012	0.009	-0.004	-0.002	-0.003	
48	42.98	41	0.3865	0.010	-0.004	-0.001	-0.000	-0.006	0.002	

Figure 3.54: Check for white noise for the above model

Figure 3.54 shows that all our  $\chi^2$  values small and all the P-values are large, so our model is adequate overall. Our ARIMA model for maximum temperature becomes,

$$\begin{aligned}
 (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_5 B^5)(1 - \Phi_1 B^{15})(1 - B^{365})Z_t \\
 = (1 - \theta_3 B^3 - \theta_4 B^4)a_t
 \end{aligned}$$

where

$Z_t$  is the observed series for maximum temperature

$\theta_3$  and  $\theta_4$  are the nonseasonal moving average parameters

$\phi_1, \phi_2, \phi_3$  and  $\phi_5$  are the nonseasonal autoregressive parameters

$\Phi_{15}$  is the seasonal autoregressive parameter

$(1 - B^{365})$  is the seasonal difference operator

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

Given estimated coefficients from the SAS output displayed in Figure 3.53, the above equation becomes:

$$\begin{aligned} (1 - 0.7169B + 0.1067B^2 - 0.4801B^3 + 0.1607B^5)(1 + 0.0292B^{15})(1 - B^{365})Z_t \\ = (1 - 0.3994B^3 - 0.2918B^4)a_t \end{aligned}$$

We are now ready to forecast future values for maximum temperature. Figure 3.55 shows the forecasted values of max temperature from *Dec20/02* to *Jan06/03*.

date	max_temp	FORECAST	STD	L95	U95	RESIDUAL
20DEC02	-3.7	-6.9053	6.39192	-19.4333	5.6226	3.2053
21DEC02	-5.1	2.6329	6.39192	-9.8951	15.1608	-7.7329
22DEC02	-7.1	-5.6439	6.39192	-18.1718	6.8841	-1.4561
23DEC02	-10.8	-12.3732	6.39192	-24.9011	0.1547	1.5732
24DEC02	-11.0	-11.2503	6.39192	-23.7782	1.2777	0.2503
25DEC02	-7.6	-11.4008	6.39192	-23.9288	1.1271	3.8008
26DEC02	-6.4	-5.4678	6.39192	-17.9958	7.0601	-0.9322
27DEC02	-0.9	-6.1724	6.39192	-18.7003	6.3555	5.2724
28DEC02	-6.1	-6.4936	6.39192	-19.0215	6.0343	0.3936
29DEC02	-1.2	-7.3229	6.39192	-19.8508	5.2050	6.1229
30DEC02	-0.7	-9.1913	6.39192	-21.7192	3.3366	8.4913
31DEC02	-12.8	-3.4164	6.39192	-15.9443	9.1115	-9.3836
01JAN03	.	-11.6332	6.39192	-24.1611	0.8948	.
02JAN03	.	-11.7829	7.86489	-27.1978	3.6320	.
03JAN03	.	-8.2493	8.28454	-24.4867	7.9881	.
04JAN03	.	-2.8524	8.49809	-19.5083	13.8036	.
05JAN03	.	-1.7035	8.61498	-18.5886	15.1815	.
06JAN03	.	-11.0932	8.67686	-28.0996	5.9131	.

Figure 3.55: Forecast values of max temperature from Dec 20/02 to Jan 06/03

With  $\lambda = 0.3$ , all the forecasted values displayed in Figure 3.56 are fairly close to the actual max temperatures and all the max temperatures lies inside the 95% confidence intervals indicating that the model is adequate for forecasting.

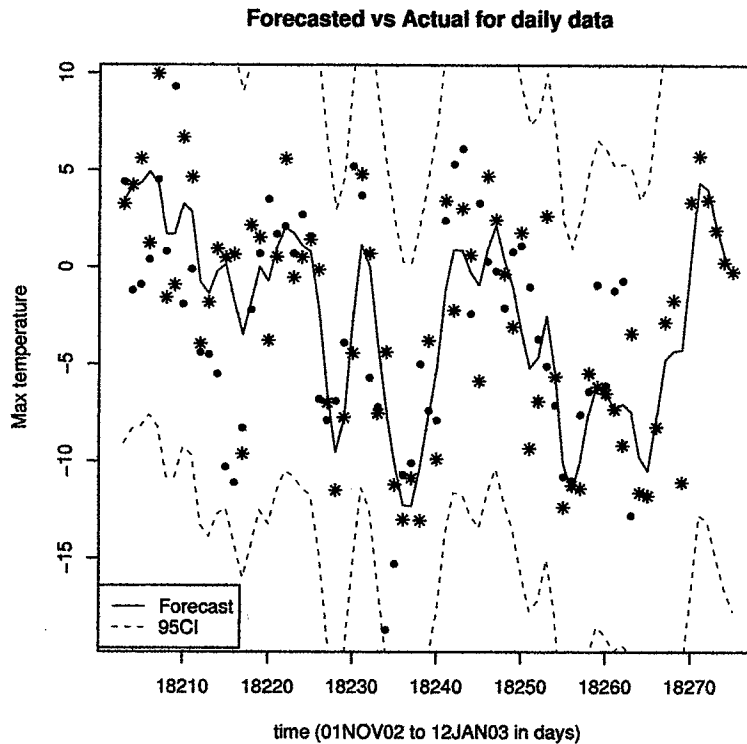


Figure 3.56: Plot of daily Forecasted (stars) and actual (dots) max temp values

### 3.3.4 Using ARIMA Modeling for a Subset of Daily Temperature

We introduced the ARIMA model for daily temperature taken from the beginning of 1953 until the end of 2002 which resulted in 18262 observations. The first thing one may ask is why are we using the whole data of 50 years to find an ARIMA model for mean, max and min temperature. Can we use less amount of time in order to forecast daily temp, say 5 or 10 years instead of the 50 years? How does it compare to the larger data set? To answer this, lets us consider the last 6 years ranging from the beginning of 1997 to the end of 2002 with the time series plot being the same as displayed in Figure 3.33.

The procedure will be the same as before, except the finer details will not be given. Instead of using the seasonal difference of 365, we will use the nonseasonal first difference for mean, max and min temperature.

If we take a look at the ARIMA model for mean temperature after it has been differenced by lag=1, we get the following results in Figure 3.57.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard		Approx	
		Error	t Value	Pr >  t	Lag
MA1,1	0.16037	0.02164	7.41	<.0001	3
MA1,2	0.15682	0.02158	7.27	<.0001	4
MA1,3	0.08506	0.02090	4.07	<.0001	5
AR1,1	-0.08692	0.02139	-4.06	<.0001	1
AR1,2	-0.28707	0.02137	-13.43	<.0001	2

Figure 3.57: Table of parameter estimates for daily mean temp of reduced data set

We see from the table that all the parameters are significant which leads to a, stationary and invertible model. After the parameters have been estimated, it is necessary to check the adequacy of the model. Figure 3.58 shows all of the P-values are larger than 0.05, indicating that we have an adequate model.



Autocorrelation Check of Residuals

To	Chi-	Pr >		-----Autocorrelations-----						
Lag	Square	DF	ChiSq							
6	1.32	1	0.2499	0.000	0.001	0.004	0.011	0.014	0.017	
12	11.60	7	0.1143	-0.020	-0.028	-0.046	-0.027	0.021	0.015	
18	19.24	13	0.1159	-0.020	0.001	-0.021	0.002	0.049	-0.016	
24	26.47	19	0.1176	-0.030	-0.031	0.002	0.027	0.023	0.011	
30	35.95	25	0.0723	0.017	-0.009	-0.007	0.029	0.008	0.054	
36	42.62	31	0.0799	0.014	0.034	0.010	0.023	0.009	0.031	
42	45.66	37	0.1554	0.026	0.013	-0.016	0.001	-0.013	0.010	
48	50.05	43	0.2138	-0.014	-0.001	0.024	0.014	-0.013	0.029	

Figure 3.58: Check for white noise of the reduced data set

Our model for the reduced data set for mean temperature becomes,

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)Z_t = (1 - \theta_3 B^3 - \theta_4 B^4 - \theta_5 B^5)a_t \quad (3.10)$$

or

$$(1 + 0.0869B + 0.287B^2)(1 - B)Z_t = (1 - 0.1603B^3 - 0.1568B^4 - 0.085B^5)a_t$$

Notice that using the reduced data for mean temperature has only 5 parameters instead of 7 that was used for the larger data set. Let us see how well it forecasts compare to the model obtained with the full data set. Displayed in Figure 3.59 are the forecasted values of mean temperature from Dec 20, 2002 to Jan 06, 2003.

	mean_					
date	temp	FORECAST	STD	L95	U95	RESIDUAL
20DEC02	-5.38	-3.3046	3.60133	-10.3631	3.7539	-2.0754
21DEC02	-6.86	-4.9725	3.60133	-12.0310	2.0859	-1.8875
22DEC02	-9.83	-6.1760	3.60133	-13.2344	0.8825	-3.6540
23DEC02	-11.75	-8.6731	3.60133	-15.7316	-1.6146	-3.0769
24DEC02	-17.16	-9.9309	3.60133	-16.9893	-2.8724	-7.2291
25DEC02	-13.22	-15.0800	3.60133	-22.1385	-8.0216	1.8600
26DEC02	-11.21	-10.7824	3.60133	-17.8409	-3.7239	-0.4276
27DEC02	-5.86	-10.5631	3.60133	-17.6216	-3.5046	4.7031
28DEC02	-10.68	-5.8049	3.60133	-12.8634	1.2535	-4.8751
29DEC02	-5.01	-11.4051	3.60133	-18.4636	-4.3466	6.3951
30DEC02	-8.59	-4.9646	3.60133	-12.0230	2.0939	-3.6254
31DEC02	-15.13	-9.8259	3.60133	-16.8844	-2.7674	-5.3041
01JAN03	.	-14.1949	3.60133	-21.2534	-7.1364	.
02JAN03	.	-12.4055	4.87672	-21.9637	-2.8473	.
03JAN03	.	-11.9543	5.38408	-22.5069	-1.4017	.
04JAN03	.	-11.3670	5.70335	-22.5454	-0.1887	.
05JAN03	.	-11.0964	5.93459	-22.7280	0.5351	.
06JAN03	.	-11.2886	6.11393	-23.2716	0.6945	.

Figure 3.59: Forecasted values for the reduced data from Dec 20/02 to Jan 06/03

Figure 3.60 displays the plots of forecasted values for the reduced data with a smoothing spline of  $\lambda = 0.3$  for better representation of the forecasted model.

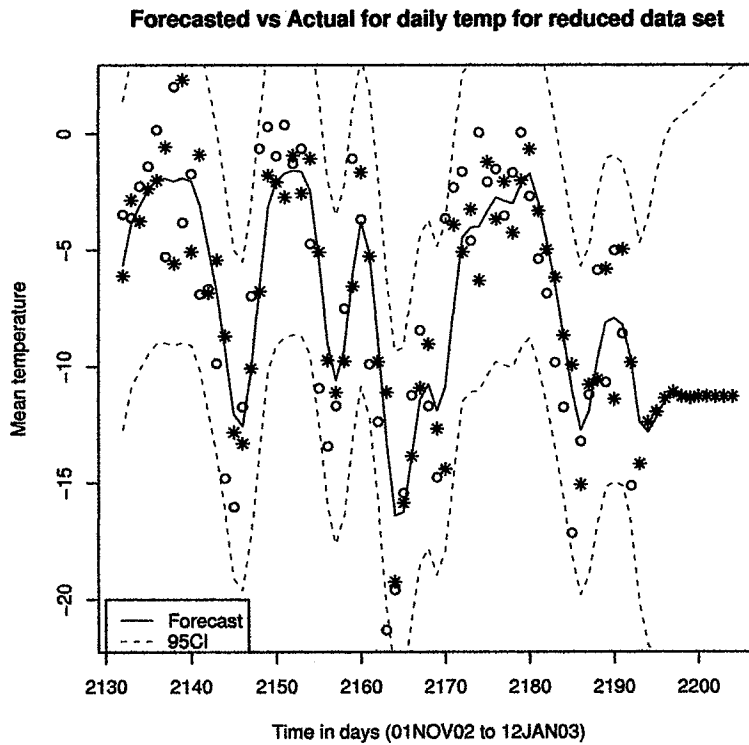


Figure 3.60: Plot of daily Forecasted (stars) and actual (dots) min temp values

If we compare the forecasts to one another, we see that the model given in Figure 3.60 is closer to the actual values compared to the ARIMA model given in Figure 3.40. We also see that the standard error estimates are smaller for the reduced data shown in Figure 3.59, with  $\text{Std}=3.601$ . The same approach was also applied to the maximum and minimum temperatures as well, but will not be shown here since we obtained the same result. Therefore, we can use smaller amount of past data such as 6 years rather than 50 years in order to forecast daily temperature with better results. Too much data will only cause the model to become more complex. Taking a smaller amount of data still conveys the same information as the full 50 years except the model will become much more simpler.

### 3.4 Fitting an ARIMA Model for Hourly Data

In subsection 3.3.4, we showed how we don't need to use such a large data set in order to achieve a good model. We will repeat the method here and use the month of December, 2002 to find an ARIMA model for hourly data. This contains all of the hourly observations for the month of December, 2002 except the temperature value for the 24th hour of Dec 31. This amounts to 743 temperature observations for the hourly data.

Let us consider the time series plot of the temperature by the hour shown in Figure 3.61. The plot for the hourly data do not have any obvious pattern present. There appears to be a roughly constant variance. However, the series appears to change level, suggesting that a nonseasonal differencing may be needed to achieve a stationary mean. From the graphs displayed in Figure 1.3 and Figure 1.4, we see there is some seasonal pattern of 24 hours apparent, which suggest that seasonality may exist. This will be shown in the sample ACF and PACF plots.

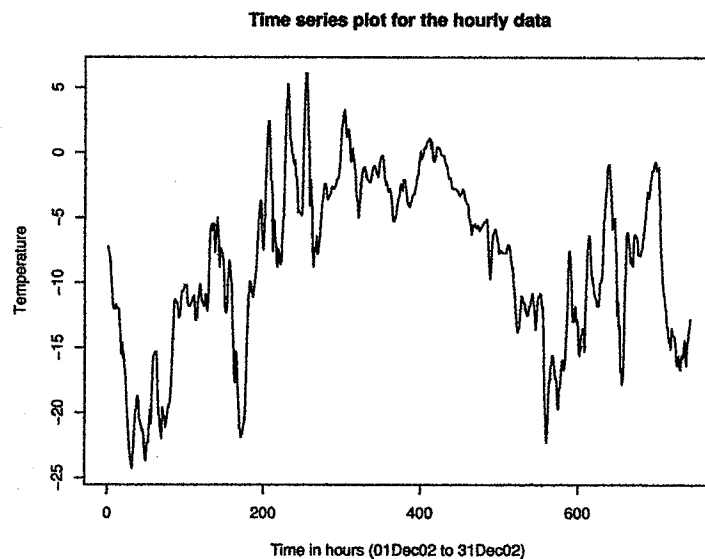


Figure 3.61: Time series plot of the hourly temperature

Taking a look at Figure 3.62 we see that the sample ACF has a slow exponential decay and suggests nonstationarity. There appears to be a slight rise at every 24th lag, indicating that a possible seasonal pattern may exist. We will first try a nonseasonal difference of one and check if a seasonal differencing is needed.

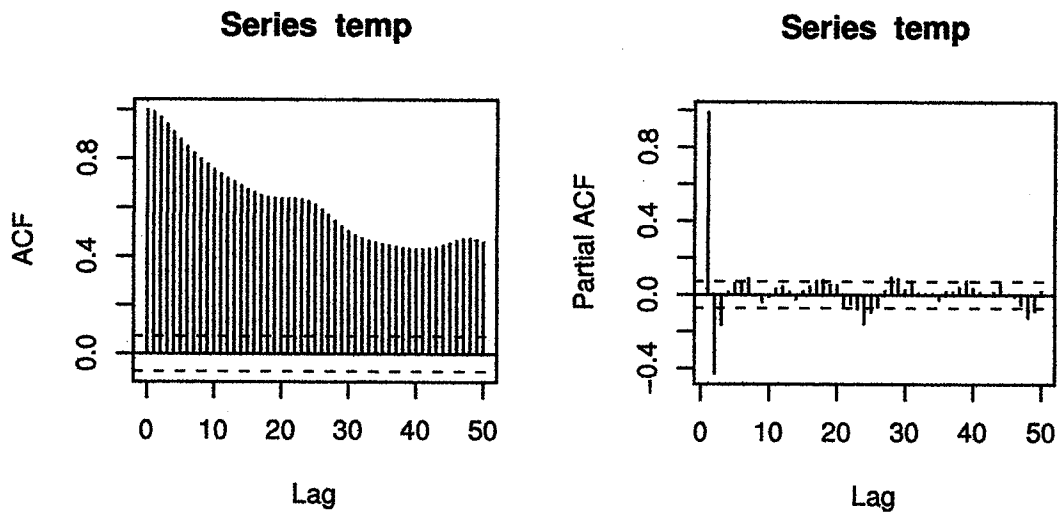


Figure 3.62: Sample ACF and PACF plots of the hourly temperature for Dec, 2002

Once the nonseasonal first differencing is used, the sample ACF in Figure 3.63 has revealed the seasonal element more clearly. The slow decay of the autocorrelations at lags 24 and 48 suggests that seasonal differencing seems appropriate.

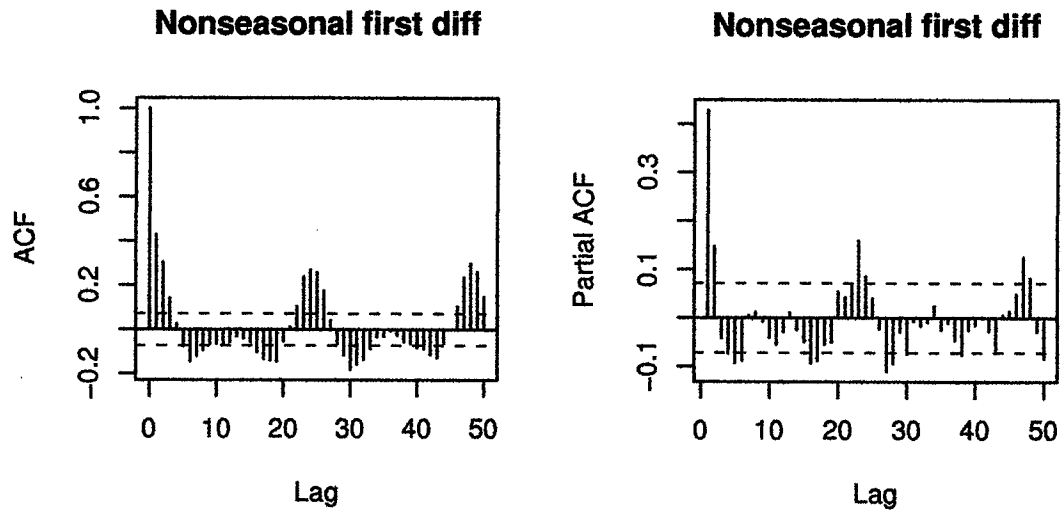


Figure 3.63: Sample ACF and PACF plots of the nonseasonal first diff

Figure 3.64 shows us that taking the seasonal differencing now induces a stationary mean. The large sample ACF spike at the 24th lag represents a possible seasonal MA(1) time structure. The noticeable spikes around lag 24 are best ignored for now, since the strong spike could influence adjacent autocorrelations at lag 25. There is slow decay for the nonseasonal spikes on the sample ACF and a cutoff to zero after lag 2 on the sample PACF indicates that a nonseasonal AR(2) is needed.

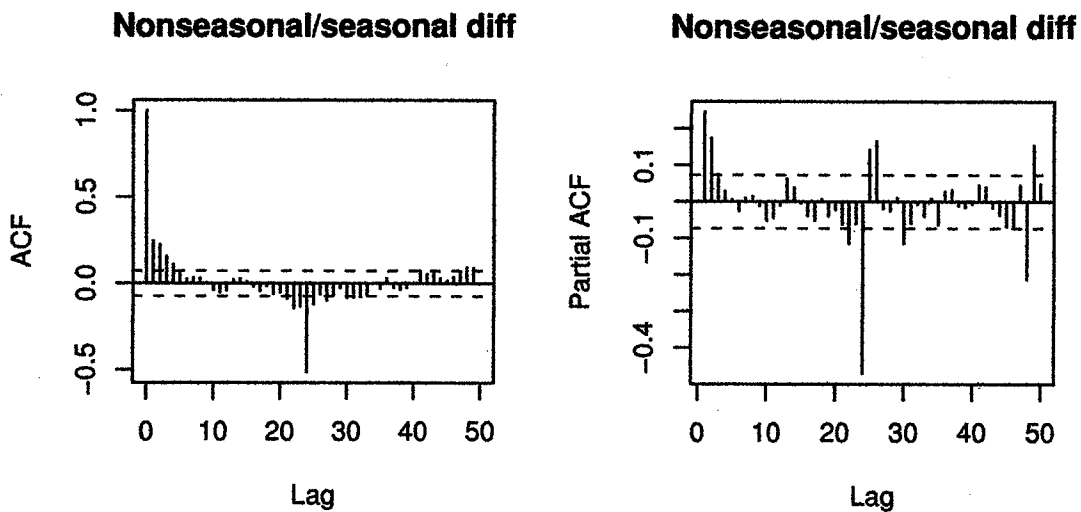


Figure 3.64: Sample ACF and PACF plots of the nonseasonal and seasonal diff

The parameters shown in Figure 3.65 all appear to be significant. The model is invertible since  $\hat{\Theta}_1$  satisfies  $|\hat{\Theta}_1| = 0.8556 < 1$ . The model is also stationary because  $\hat{\phi}_1$  and  $\hat{\phi}_2$  meet the necessary conditions [7].

$$|\hat{\phi}_2| = 0.21248 < 1$$

$$\hat{\phi}_2 + \hat{\phi}_1 = 0.21248 + 0.24236 = 0.4548 < 1$$

$$\hat{\phi}_2 - \hat{\phi}_1 = 0.21248 - 0.24236 = -0.0298 < 1$$

#### Conditional Least Squares Estimation

Parameter	Estimate	Standard		Approx	
		Error	t Value	Pr >  t	Lag
MA1,1	0.85569	0.01985	43.10	<.0001	24
AR1,1	0.24236	0.03655	6.63	<.0001	1
AR1,2	0.21248	0.03656	5.81	<.0001	2

Figure 3.65: Table of parameter estimates for hourly temp

Now we can check whether the model assumptions are satisfied. Looking at the  $\chi^2$  test statistics shown in Figure 3.66, it shows that our model is adequate because all our P-values are greater than 5%. This means that the residuals are uncorrelated.

Autocorrelation Check of Residuals

To	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	4.58	3	0.2050	-0.010	-0.019	0.035	0.046	0.023	-0.045
12	7.80	9	0.5541	-0.008	0.006	0.003	-0.018	-0.043	-0.046
18	8.79	15	0.8882	0.024	0.016	0.006	-0.019	-0.005	0.009
24	11.54	21	0.9512	-0.025	0.036	-0.019	0.004	0.038	0.001
30	25.23	27	0.5617	0.050	0.110	-0.005	-0.002	0.038	-0.047
36	28.88	33	0.6724	-0.028	-0.041	-0.028	0.025	-0.002	0.031
42	33.10	39	0.7350	0.011	-0.035	-0.023	-0.004	0.057	0.021
48	38.66	45	0.7359	-0.010	0.014	-0.033	0.007	0.037	0.066

Figure 3.66: Check for white noise for hourly temperature

Our model for the hourly temperature becomes,

$$(1 - \phi_1 B - \phi_2 B^2)(1 - B)(1 - B^{24})Z_t = (1 - \Theta_1 B^{24})a_t \quad (3.11)$$

where

$Z_t$  is the observed series for hourly temperature

$\Theta_1$  is the seasonal moving average parameter of order 1

$\phi_1$  and  $\phi_2$  are the nonseasonal autoregressive parameters

$(1 - B)$  is the nonseasonal difference operator ( $d=1$ )

$(1 - B^{24})$  is the seasonal difference operator ( $D=1$ )

$B$  is the backshift operator ( $B^k Z_t = Z_{t-k}$ )

$a_t$  is a white noise process ( $WN(0, \sigma^2)$ )

or with the estimated coefficients, we get,

$$(1 - 0.24236B - 0.21248B^2)(1 - B)(1 - B^{24})Z_t = (1 - 0.85569B^{24})a_t$$



Now that the model for hourly temperature is adequate, we can forecast the future values from the 16th hour, Dec31, 2002 until the 6th hour, Jan01, 2003 displayed in Figure 3.67. The plot of the forecasted values for the model are given in Figure 3.68.

Obs	TEMP	FORECAST	STD	L95	U95	RESIDUAL
736	-14.4	-15.3279	0.85549	-17.0047	-13.6512	0.92793
737	-15.9	-14.8817	0.85549	-16.5584	-13.2049	-1.01834
738	-16.4	-16.2464	0.85549	-17.9231	-14.5696	-0.15363
739	-14.6	-16.9183	0.85549	-18.5951	-15.2416	2.31833
740	-14.6	-14.0813	0.85549	-15.7580	-12.4045	-0.51875
741	-13.9	-14.1571	0.85549	-15.8338	-12.4803	0.25707
742	-13.5	-13.8614	0.85549	-15.5381	-12.1847	0.36139
743	-12.8	-12.9188	0.85549	-14.5955	-11.2420	0.11876
744	.	-12.6379	0.85549	-14.3146	-10.9612	.
745	.	-12.7361	1.36435	-15.4102	-10.0620	.
746	.	-13.1474	1.88098	-16.8341	-9.4608	.
747	.	-12.8846	2.34191	-17.4746	-8.2945	.
748	.	-13.1671	2.76437	-18.5852	-7.7490	.
749	.	-13.6753	3.14889	-19.8471	-7.5036	.

Figure 3.67: Forecasted values for hourly temp

Forecasted vs Actual for hourly temp for reduced data set

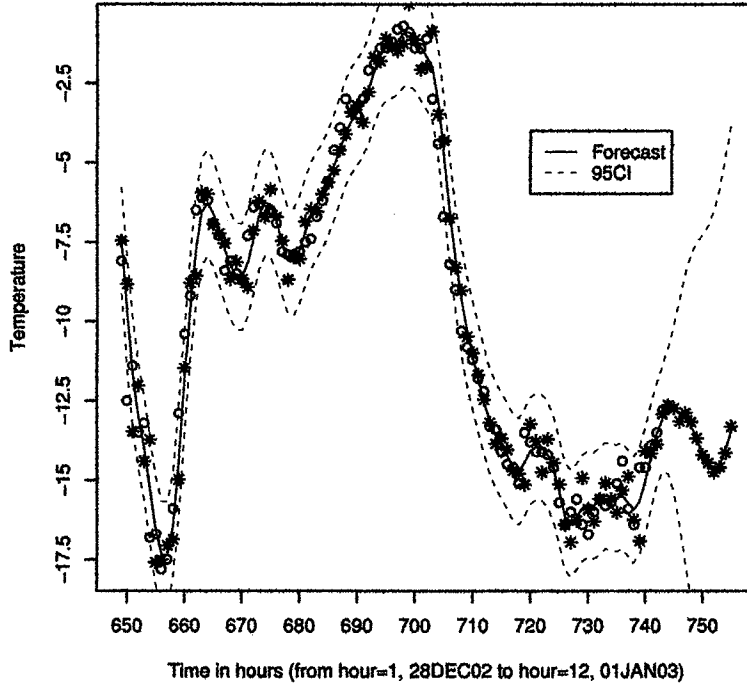


Figure 3.68: Plot of hourly Forecasted (stars) and actual (dots) temp values

The forecasted values for the model fits the hourly data quite well with a low standard error and all the values inside the 95% confidence interval. We see that no more than a months amount of past data is needed in order to forecast the hourly data accurately.

## 3.5 Conclusion

In this chapter we have examined our ARIMA models a bit further. This time we looked at the ARIMA models for temperature at different time intervals. We looked at the monthly temperature, as well as the daily and hourly temperature to find a model for forecasting future values. We also showed that taking a large amount of past data for smaller time frequencies such as 50 years are usually not required. We can take a smaller sample of past temperature to produce equal or better results compared to using the large past temperature data.

# Chapter 4

## Multivariate Time Series Analysis

### 4.1 Introduction

In Chapter 1 we discussed temperature and load behaviors throughout the day and saw how they are fairly similar in pattern. In Chapters 2 and 3 we examined the univariate cases by modeling and forecasting temperature at different time intervals. Observations are often taken simultaneously on two or more time series. Given multivariate data like temperature and load, it may be helpful to develop a multivariate model to describe the interrelationships between the series. In this chapter we will look at examining and forecasting a bivariate or multivariate time series at three different time intervals: monthly, daily and hourly. In section 4.2 we will discuss the models and procedures for multivariate time series that will aid in determining possible models. In sections 4.3, 4.4 and 4.5, we will apply the methods of multivariate models to the data at monthly, daily and hourly time intervals. The focus is to identify the model and the relationship between the two series in order to attain better understanding and optimal forecasts.

## 4.2 Vector Time Series Models

The idea behind the vector time series model procedure is similar to that used in the univariate case. Suppose we have a two dimensional process  $Z_t$  with

$$Z_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}, \quad t = 1, \dots, n$$

that is a stationary bivariate time series if the mean  $E(Z_t) = \mu$  is constant for each  $X_t$  and  $Y_t$  and the covariance between  $X_t$  and  $Y_t$  are functions of the time difference lag ( $k$ ) [12]. Therefore, we have the mean vector

$$E(Z_t) = \mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$$

and the covariance matrix is given as

$$\begin{aligned} \Gamma(k) &= \text{Cov}\{Z_t, Z_{t-k}\} = E[(Z_t - \mu)(Z_{t-k} - \mu)'] \\ &= E \begin{bmatrix} X_t - \mu_x \\ Y_t - \mu_y \end{bmatrix} [X_{t-k} - \mu_x, Y_{t-k} - \mu_y] \\ &= E \begin{bmatrix} (X_t - \mu_x)(X_{t-k} - \mu_x) & (X_t - \mu_x)(Y_{t-k} - \mu_y) \\ (Y_t - \mu_y)(X_{t-k} - \mu_x) & (Y_t - \mu_y)(Y_{t-k} - \mu_y) \end{bmatrix} \\ &= \begin{bmatrix} \gamma_{xx}(k) & \gamma_{xy}(k) \\ \gamma_{yx}(k) & \gamma_{yy}(k) \end{bmatrix} \end{aligned}$$

$\Gamma(k)$  is called the covariance matrix function for the vector process  $Z_t$  [8, 12]. Although for the autocovariance function of a univariate stationary time series is symmetric about zero, the same is not true for the cross autocovariance for a multivariate stationary time series. For the multivariate time series the covariance matrix is  $\Gamma(k) \neq \Gamma(-k)$ , but  $\Gamma(-k) = \Gamma'(k)$  and  $\gamma_{xy}(-k) = \gamma_{yx}(k)$  because it is *not* an even function. The matrix  $\Gamma(0)$  can be easily seen as the variance-covariance matrix of the process.

The diagonal matrix, denoted  $D$ , which is defined by the variances of  $x$  and  $y$  respectively, is

$$D = \text{diag}[\gamma_{xx}(0), \gamma_{yy}(0)]$$

and the off-diagonal element of  $\Gamma_k$ , ie.  $\gamma_{xy}(k)$  is the cross-covariance function between  $X_t$  and  $Y_t$ .

The size of the cross-covariance coefficients depends on the units in which  $X_t$  and  $Y_t$  are measured. Thus for interpretative purposes, it is useful to standardize the cross-covariance function to produce a function called **cross – correlation function**,  $\rho(k)$  [3, 12], which is defined by

$$\begin{aligned} \rho(k) &= D^{-1/2}\Gamma(k)D^{-1/2} \\ &= \begin{bmatrix} \frac{\gamma_{xx}(k)}{\sqrt{\gamma_{xx}(0)\gamma_{xx}(0)}} & \frac{\gamma_{xy}(k)}{\sqrt{\gamma_{xx}(0)\gamma_{yy}(0)}} \\ \frac{\gamma_{yx}(k)}{\sqrt{\gamma_{xx}(0)\gamma_{yy}(0)}} & \frac{\gamma_{yy}(k)}{\sqrt{\gamma_{yy}(0)\gamma_{yy}(0)}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\gamma_{xx}(k)}{\sigma_x^2} & \frac{\gamma_{xy}(k)}{\sigma_x\sigma_y} \\ \frac{\gamma_{yx}(k)}{\sigma_x\sigma_y} & \frac{\gamma_{yy}(k)}{\sigma_y^2} \end{bmatrix} \end{aligned}$$

where  $\sigma_x = \sqrt{\gamma_{xx}(0)}$  denotes the standard deviation of the X-process, and similarly for  $\sigma_y$ . This function measures the correlation between  $X_t$  and  $Y_{t-k}$  and has these two properties [12]:

1.  $\rho_{xy}(k) = \rho_{yx}(-k)$
2.  $|\rho_{xy}(k)| \leq 1$

Whereas  $\rho_x(0)$ ,  $\rho_y(0)$  are both equal to one, the value of  $\rho_{xy}(0)$  is usually not equal to one. Assuming that the same number of observations have been collected on the two variables over the same time period, the sample cross-correlation of  $X_t$  and  $Y_t$  at lag  $k$  can be calculated by

$$\hat{\rho}(k) = \begin{bmatrix} \frac{\hat{\gamma}_{xx}(k)}{s_x^2} & \frac{\hat{\gamma}_{xy}(k)}{s_x s_y} \\ \frac{\hat{\gamma}_{yx}(k)}{s_x s_y} & \frac{\hat{\gamma}_{yy}(k)}{s_y^2} \end{bmatrix} \quad (4.1)$$

where  $s_x = \sqrt{\hat{\gamma}_{xx}(0)}$  denotes the sample standard deviation of the X-process and the sample cross-covariance are given as

$$\begin{aligned}\hat{\Gamma}(k) &= \begin{bmatrix} \hat{\gamma}_{xx}(k) & \hat{\gamma}_{xy}(k) \\ \hat{\gamma}_{yx}(k) & \hat{\gamma}_{yy}(k) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t-k} - \bar{x}) & \frac{1}{n} \sum_{t=1}^{n-k} (x_t - \bar{x})(y_{t-k} - \bar{y}) \\ \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(x_{t-k} - \bar{x}) & \frac{1}{n} \sum_{t=1}^{n-k} (y_t - \bar{y})(y_{t-k} - \bar{y}) \end{bmatrix}\end{aligned}$$

Where

$$\epsilon_t = \begin{bmatrix} \epsilon_{xt} \\ \epsilon_{yt} \end{bmatrix} \sim WN(0, \Sigma)$$

#### 4.2.1 Vector Moving Average Model

A pure moving average MA(q) model is given by

$$\begin{aligned}Z_t &= \mu + (I - \Theta_1 B - \dots - \Theta_q B^q) \epsilon_t \\ &= \mu + \Theta(B) \epsilon_t \\ &= \mu + \epsilon_t - \sum_{j=1}^q \Theta_j \epsilon_{t-j}\end{aligned}$$

Where

$$\begin{aligned}\epsilon_t &\sim WN(0, \Sigma) \\ \Theta_j &= \begin{pmatrix} \Theta_{j,11} & \Theta_{j,12} \\ \Theta_{j,21} & \Theta_{j,22} \end{pmatrix} \quad \text{is a } 2 \times 2 \text{ matrix of real numbers for } j = 1, \dots, q\end{aligned}$$

The finite order moving average MA(q) process is always stationary and causal because its representation

$$Z_t = \mu + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} = \mu + \psi(B) \epsilon_t \quad \text{where } \psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$$

has  $\psi(B) = \Theta(B) = I - \Theta_1 B - \dots - \Theta_q B^q$  which is automatically convergent.

The vector MA(q) process is invertible [8] if it can be represented in the form

$$Z_t = \mu + \sum_{j=1}^{\infty} \Pi_j (Z_{t-j} - \mu) + \epsilon_t \quad \text{with } \sum_{j=1}^{\infty} \|\Pi_j\| < \infty$$

Let  $d(z) = \det\{\Theta(z)\} = \det\{I - \Theta_1 z - \dots - \Theta_q z^q\}$ , where  $z$  is a complex variable. For the MA(q) to be invertible and be expressible as a convergent infinite form, it is required that  $d(z)^{-1}$  form a convergent series for  $|z| \leq 1$ , and hence that all roots of  $d(z) = \det\{\Theta(z)\} = 0$  be greater than one in absolute value, referred as the invertibility condition [8, 12].

#### 4.2.2 Vector Autoregressive AR(p) Model

The vector autoregressive AR(p) model is given by [8]

$$(Z_t - \mu) - \sum_{j=1}^p \Phi_j (Z_{t-j} - \mu) = \epsilon_t \quad , \quad \text{where } \Phi_j = \begin{pmatrix} \Phi_{j,11} & \Phi_{j,12} \\ \Phi_{j,21} & \Phi_{j,22} \end{pmatrix}$$

or  $\Phi(B)(Z_t - \mu) = \epsilon_t \quad , \quad \text{where } \Phi(B) = I - \Phi_1 B - \dots - \Phi_p B^p$

For the process to be stationary and causal, we require it to be expressible in the causal infinite MA form [8] as

$$Z_t - \mu = \sum_{j=0}^{\infty} \Psi_j \epsilon_{t-j} = \Psi(B) \epsilon_t \quad ,$$

where  $\Psi(B) = \sum_{j=0}^{\infty} \Psi_j B^j$ ,  $\sum_{j=0}^{\infty} \|\Psi_j\| < \infty$

This is similar to the discussion of invertibility for the MA process, from the expression

$$\Phi(z)^{-1} = \left[ \frac{1}{\det\{\Phi(z)\}} \right] \text{Adj}\{\Phi(z)\}$$

we find that  $\Psi(z) = \Phi(z)^{-1}$  will be a convergent series for  $|z| \leq 1$  if all the roots of  $\det\{\Phi(z)\} = 0$  are greater than one in absolute value [3].

Under the stationarity condition, the matrix weights  $\Psi_j$  can be obtained from the relation  $\Phi(B)\Psi(B) = I$  [8] since

$$\begin{aligned} \Phi(B)\Psi(B) &= (I - \Phi_1 B - \dots - \Phi_p B^p)(I + \Psi_1 B + \Psi_2 B^2 + \dots) \\ &= I + (\Psi_1 - \Phi_1)B + (\Psi_2 - \Phi_1 \Psi_1 - \Phi_2)B^2 + \dots \\ &\quad + (\Psi_j - \Phi_1 \Psi_{j-1} - \Phi_p \Psi_{j-p})B^j + \dots \end{aligned}$$



by equating the coefficient matrices of various powers  $B^j$  in the equation  $\Phi(B)\Psi(B) = I$  we have

$$\begin{aligned}\Psi_1 &= \Phi_1 \\ &\vdots \\ \Psi_j &= \Phi_1\Psi_{j-1} - \Phi_p\Psi_{j-p}, \quad j \geq p\end{aligned}$$

### 4.2.3 Vector Autoregressive Moving Average Process (VARMA(p,q))

We now briefly consider general properties of the mixed vector autoregressive moving average VARMA(p,q) model

$$\begin{aligned}\Phi_p(B)(Z_t - \mu) &= \Theta_q(B)\epsilon_t \\ \Rightarrow (Z_t - \mu) - \sum_{j=1}^p \Phi_j(Z_{t-j} - \mu) &= \epsilon_t - \sum_{j=1}^q \Theta_j\epsilon_{t-j}\end{aligned}$$

where

$$\text{cov}(\epsilon_t, \epsilon_{t-k}) = \begin{cases} \Sigma & \text{if } k = 0 \\ 0 & \text{if } k \neq 0 \end{cases}$$

The conditions for stationarity and invertibility of the vector ARMA process are clearly the same as in the pure MA and AR cases [12]. A vector ARMA model is stationary if the determinant of its AR matrix polynomial is constant or has zeros lying outside the unit circle [12]. It is invertible if the MA matrix polynomial is either constant or zeros lying outside the unit circle.

### 4.2.4 Nonstationary Vector ARMA Models

In time series analysis, it is common to observe series that exhibit nonstationary behaviour. One useful way to eliminate nonstationarity is by differencing. Often, in univariate ARIMA time series models, a nonstationary time series can be reduced to

stationary through differencing the series by  $(1 - B)^d Z_t$  [8][12]. In the multivariate case, a natural extension of the ARIMA is

$$\Phi_p(B)(1 - B)^d Z_t = \Theta_q(B)\epsilon_t$$

The problem with the extension is it implies that the component series  $X_t$  and  $Y_t$  are differenced the same number of times. To alleviate this problem, we assume that even though  $Z_t$  may be nonstationary, it can be transformed to stationarity by applying the differencing operator  $D(B)$  such that

$$D(B) = \begin{bmatrix} (1 - B)^{d1} & \\ & (1 - B)^{d2} \end{bmatrix} \quad (4.2)$$

Therefore we have the vector ARMA model

$$\Phi_p(B)D(B)Z_t = \Theta_q(B)\epsilon_t \quad (4.3)$$

where the zeros of  $\det\{\Phi_p(z)\}$  and  $\det\{\Theta_q(z)\}$  are assumed to be outside the unit circle [8].

#### 4.2.5 Granger-Causality

One of the key questions that can be addressed to VAR models is how useful some variables are for forecasting others. If the history of  $X$  does not help to predict the future values of  $Y$ , we say that  $X$  does not Granger-cause  $Y$ . Usually the prediction ability is measured in terms of the MSE (Mean Squared Error). Hence,  $X$  fails to Granger-cause  $Y$ , if for all  $k > 0$  [12]

$$MSE(\hat{Y}_{t+k}|Y_t, Y_{t-1}, \dots) = MSE(\hat{Y}_{t+k}|Y_t, Y_{t-1}, \dots, X_t, X_{t-1}, \dots)$$

In a bivariate context, If  $Y_t$  fails to Granger-cause  $X_t$  then all the  $\Phi_i$  matrices are lower triangular [8],

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} \Phi_{11,1} & 0 \\ \Phi_{21,1} & \Phi_{22,1} \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \dots + \begin{bmatrix} \Phi_{11,p} & 0 \\ \Phi_{21,p} & \Phi_{22,p} \end{bmatrix} \begin{bmatrix} X_{t-p} \\ Y_{t-p} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}$$

This suggests a simple F-test for Granger Causality using the univariate regression for each

$$X_t = C_1 + \sum_{i=1}^p \Phi_{11,i} X_{t-i} + \sum_{i=1}^p \Phi_{12,i} Y_{t-i} + \epsilon_{1t}$$

$$Y_t = C_2 + \sum_{i=1}^p \Phi_{21,i} X_{t-i} + \sum_{i=1}^p \Phi_{22,i} Y_{t-i} + \epsilon_{2t}$$

where

Ho:  $\Phi_{12,i} = 0$  for  $i = 1, \dots, p$  ( $Y_t$  does not Granger cause  $X_t$ )

Ha:  $\Phi_{12,i} \neq 0$  for some  $i = 1, \dots, p$  ( $Y_t$  does Granger cause  $X_t$ )

The Test Statistic is given as

$$\hat{F} = \frac{(RSS_r - RSS_{ur})/p}{RSS_{ur}/(n - 2p - 1)} \sim F(p, n - 2p - 1)$$

In a bivariate VAR(P) model

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + \begin{bmatrix} \Phi_{xx}(B) & \Phi_{xy}(B) \\ \Phi_{yx}(B) & \Phi_{yy}(B) \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{xt} \\ \epsilon_{yt} \end{bmatrix}$$

The variables  $\{X_t\}$  are said to cause  $Y_t$ , but  $Y_t$  does not cause  $X_t$  if  $\Phi_{xy}(B) = 0$ , but  $\Phi_{yx}(B) \neq 0$ . This implies that future values of the process  $X_t$  are influenced by its own past values and not by the past of  $Y_t$ , where as future values of  $Y_t$  are influenced by both past values of  $X_t$  and  $Y_t$ . If the future  $X_t$  values are not influenced by the past values of  $Y_t$ , then it can be better to model  $X_t$  separately from  $Y_t$ .

#### 4.2.6 Defining the Order of a VAR Model

In the first step it is assumed that all the series in the VAR model have equal lags lengths. To determine the number of lags that should be included, criterion functions can be utilized the same manner as in the univariate case [12].

The underlying assumption with AIC and other related criteria is that the residuals follow a multivariate normal distribution, i.e.

$$\epsilon_t \sim N_2(0, \Sigma_\epsilon)$$

In the original forms AIC and BIC are defined as

$$AIC = -2\text{Log}L + 2s$$

$$BIC = -2\text{Log}L + s\text{Log}n$$

where "L" stands for the *Likelihood function*, and "s" denotes the number of estimated parameters [3][8]. The best fitting model is one that minimizes the criterion function.

For example in a bivariate VAR(p) model, there are  $s = 2(1 + 2p) + 2(2 + 1)/2$  estimated parameters.

When building the VAR models, we will be using ordinary least squares method for estimation. For multivariate case, the model checking and forecasting are similar to the univariate case and will not be discussed here.

### 4.3 Fitting Multivariate Models to Monthly Data

Let  $Y_t^*$  be the monthly series of Electrical Load and  $X_t^*$  be the monthly series of Temperature plotted from April, 1992 to December, 2002, as shown in Figure 4.1.

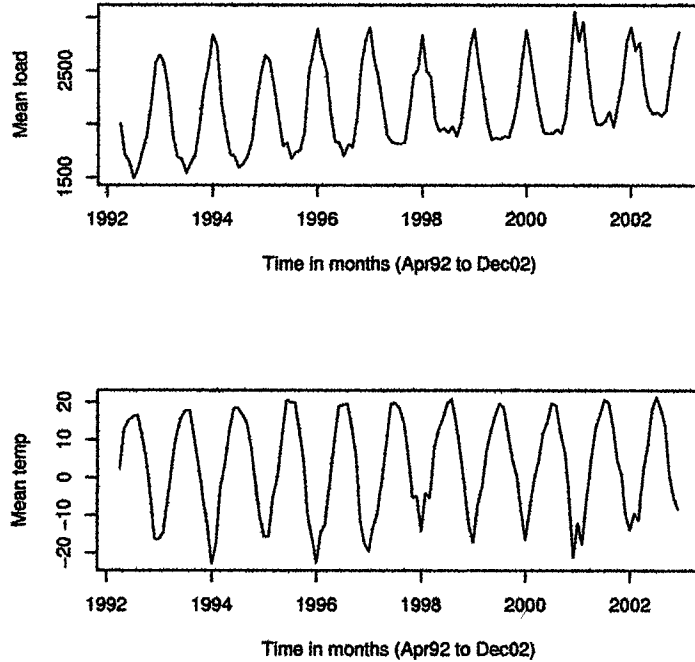


Figure 4.1: Time series plot of the monthly mean load and temperature

We see the two series follow a similar seasonal pattern with temperature exhibiting an inverse relationship compared to load. A general upward trend is clearly seen in the demand for electricity or load as time increases while temperature remains constant. Beyond that, it can be seen that the members of the series tend to move together and there is an interaction between them. As shown in section 3.2.1, there is a need for differencing the data for both time series. Taking the difference of lag 12 for both series, we see in Figure 4.2 the two series are now stationary with mean approximately zero and the variance is fairly constant with the exceptions of a few irregular spikes.

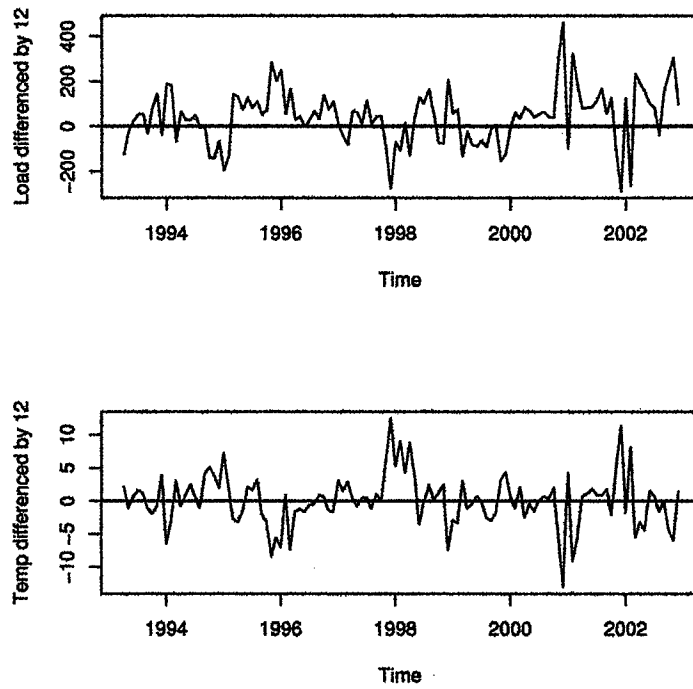


Figure 4.2: Time series plot of the differenced monthly mean load and temperature

For notational purposes, we will now denote the differenced series as

$$X_t = (1 - B^{12})X_t^* \quad (4.4)$$

$$Y_t = (1 - B^{12})Y_t^* \quad (4.5)$$

$$\mathbf{Z}_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \quad (4.6)$$

where  $X_t$  is the differenced series for temperature and  $Y_t$  is the differenced series for load. Table 4.1 displays the sample correlation matrices along with their indicator symbols for the differenced series of load and temperature from lags 1 through 12 and Figure 4.3 provides the sample ACF, PACF and CCF plots for the differenced series of load and temperature. .

Schematic Representation of Cross Correlations													
variable/lag	0	1	2	3	4	5	6	7	8	9	10	11	12
Load	+-	+-	+-	..	..	..	..	..	..	..	..	..	+-
Temp	-+	-+	-+	..	..	..	..	..	..	..	..	..	+-

Schematic Representation of Partial Autoregression													
variable/lag	0	1	2	3	4	5	6	7	8	9	10	11	12
Load	+	..	..	..	..	..	..	..	..	..	..	..	..
Temp	..	..	..	..	..	..	..	..	..	..	..	..	..

+ is  $> 2 \times \text{std error}$ , - is  $< 2 \times \text{std error}$ , . is between, \* is NA

Table 4.1: Sample correlation matrices for the differenced series  $Z_t$

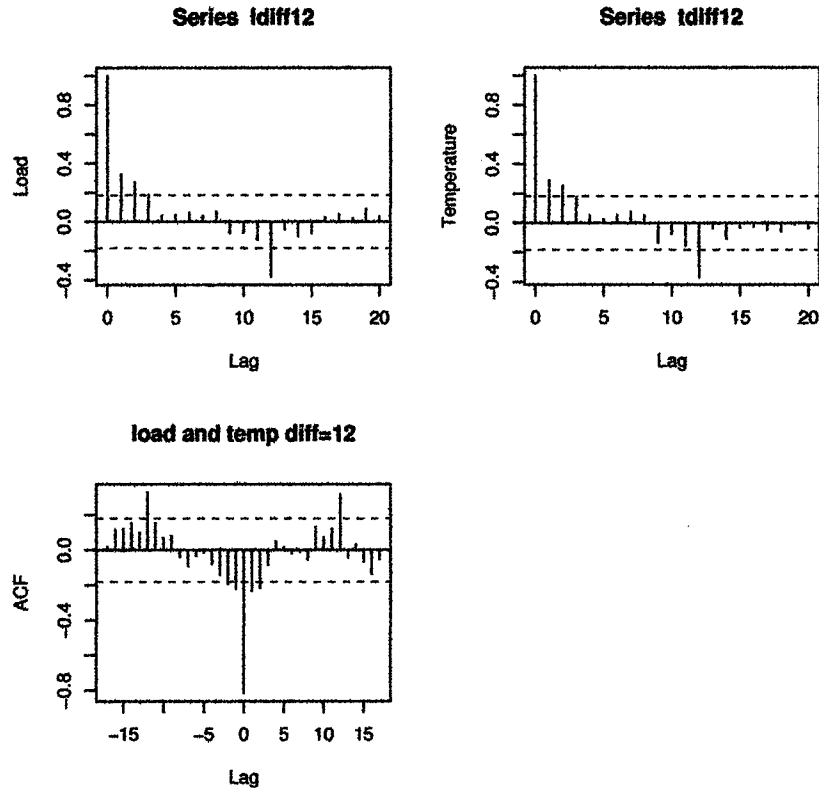


Figure 4.3: Sample auto and cross correlation of differenced data  $Z_t$

Since the sheer volume of information given in Table 4.1 makes the arrays of coefficients difficult to assess, for simplicity, we denote codes to the values as follows:

- + denotes a value greater than 2 estimated standard errors
- denotes a value less than -2 estimated standard errors
- denotes a value within 2 estimated standard errors

These correlations show that the ACFs and CCF decay fairly quickly to zero after the second lag with a seasonal spike at lag 12. We also see that the partial correlations drop quickly to zero at lag one which indicates a reasonable initial model is VAR(2).



Table 4.2 provides a summary of results for the preliminary VARMA(p,q) determination procedure, based on the least squares regression estimation method using the AIC method with values given by AIC(p,q) from the MINIC method in SAS.

Minimum Information Criterion									
lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6	MA 7	MA 8
AR 0	11.579	11.60	11.538	11.422	11.407	11.406	11.368	11.351	11.391
AR 1	11.089	11.099	11.131	11.129	11.128	11.116	11.184	11.205	11.325
AR 2	11.048	11.114	11.138	11.153	11.178	11.168	11.180	11.172	11.233
AR 3	<b>11.002</b>	11.076	11.137	11.172	11.215	11.187	11.228	11.239	11.147
AR 4	11.027	11.113	11.170	11.209	11.265	11.194	11.265	11.231	10.994
AR 5	11.052	11.098	11.153	11.215	11.213	11.257	11.326	11.214	11.014
AR 6	11.107	11.179	11.235	11.297	11.292	11.339	11.398	11.258	11.087

Table 4.2: Minic Information criterion using AIC method to identify possible VARMA model

The AIC, as well as BIC and other criterion applied to these linear least squares estimation results all suggest that a VAR(3) might be essentially the ideal model to use. The VARMA(4,8) could be used as well since the AIC is the smallest, but the model will contain too many parameters. Hence, we find that a vector AR(3) model might be acceptable.

Once we have an initial model, we use the CAUSAL statement from PROC VARMAX to compute the Granger-Causality test for the VAR(3) model shown in Table 4.3

Granger-Causality Wald Test			
Test	DF	Chi-Square	P-value
1	3	0.13	0.9879
Test1: Group1 Variables			Temp
Group2 Variables			Load

Table 4.3: Granger-Causality test for exogeneous variables

The CAUSAL statement fits the VAR(3) model using the variables in two groups considering them as dependent variables. The CAUSAL statement fits the VAR(3) model using the variables load and temperature. The output in Table 4.3 shows that you cannot reject that temperature is influenced by itself and not by load at the 0.05 significance level for Test 1. Since the test of hypothesis fails to reject the null hypothesis, the variable temperature in group 1 is considered as an independent variable or exogeneous.

The autoregressive model with an exogeneous variable is called the ARX(p,s) model. The form of the ARX(p,s) model can be written as

$$Y_t = \delta + \sum_{i=1}^p \Phi_i Y_{t-i} + \sum_{i=0}^s \Theta_i^* X_{t-i} + \epsilon_t$$

where

$\delta$  is a constant

$X_t$  is the exogeneous variable temperature

$Y_t$  is the dependent variable load

The AIC first suggested to use a VAR(3) model as a possible model , we will use it in the VARX(p,s) model with p=3 and s=3 as our initial model. The estimates and schematic representation of the estimates are given in Table 4.4.

Schematic Representation of Parameter Estimates								
Variable/lag	C	XL0	XL1	XL2	XL3	AR1	AR2	AR3
Load	+	-	+	.	+	+	.	+

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
Load	Const1	17.563	7.586	2.32	0.0225	1
	<i>XL0</i>	-25.884	1.577	-16.40	0.0001	temp(t)
	<i>XL1</i>	8.603	2.917	2.95	0.0039	temp(t-1)
	<i>XL2</i>	1.001	3.026	0.33	0.7414	temp(t-2)
	<i>XL3</i>	8.069	2.869	2.81	0.0059	temp(t-3)
	<i>AR1</i>	0.337	0.093	3.60	0.0005	load(t-1)
	<i>AR2</i>	0.091	0.098	0.93	0.3568	load(t-2)
	<i>AR3</i>	0.225	0.093	2.41	0.0175	load(t-3)

Table 4.4: Parameter Estimates for ARX(3,3)

The coefficient estimates of  $XL2$  ( $\Theta_2^*$ ) and  $AR2$  ( $\Phi_2$ ) in the model were not significant and could be omitted from the model. We will first drop only  $XL2$  from the model since its estimate was less than 1.5 times its estimated standard errors and the  $AR2$  estimate was not. The estimates and schematic representation are given in Table 4.5.

Schematic Representation of Parameter Estimates							
Variable/lag	C	XL0	XL1	XL3	AR1	AR2	AR3
Load	+	-	+	+	+	.	+

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
Load	Const1	17.971	7.454	2.41	0.0176	1
	<i>XL0</i>	-25.832	1.563	-16.52	0.0001	temp(t)
	<i>XL1</i>	8.900	2.764	3.22	0.0017	temp(t-1)
	<i>XL3</i>	8.384	2.695	3.11	0.0024	temp(t-3)
	<i>AR1</i>	0.346	0.088	3.89	0.0002	load(t-1)
	<i>AR2</i>	0.063	0.052	1.20	0.2309	load(t-2)
	<i>AR3</i>	0.234	0.088	2.65	0.0094	load(t-3)

Table 4.5: Parameter Estimates for ARX(3,3) with XL2 dropped

Our estimates in Table 4.5 seem to improve slightly with XL2 coefficient dropped. We now examine the residuals  $\hat{\epsilon}_t$  to check how our model fits. The residual correlations and schematic representations were obtained and presented in Table 4.6.

Cross-Correlation of Residuals							
lag	0	1	2	3	4	5	6
Load	1.0	-0.009	0.049	0.076	-0.113	0.009	0.044
lag	7	8	9	10	11	12	
Load	-0.044	0.052	0.047	0.004	0.064	-0.343	

Schematic Representation of Cross-Correlation of Residuals													
Variable/lag	0	1	2	3	4	5	6	7	8	9	10	11	12
Load	+	.	.	.	.	.	.	.	.	.	.	.	-

Table 4.6: Residuals diagnostics for ARX(3,3)

One notable feature of these residual correlations is the marginally significant correlation at lag 4. There are no seasonal structure that exists except the large negative value at lag 12. On the Portmanteau test for cross-correlation of residuals displayed in Table 4.7, the large spike at lag 12 marginally influences the P-values at the lags 12 or larger, creating some inadequacy of the model.

Portmanteau Test for Cross-Correlation of Residuals								
up to lag	4	5	6	7	8	9	10	11
Pr > ChiSq	0.113	0.284	0.431	0.559	0.650	0.730	0.824	0.844
up to lag	12	13	14	15	16	17	18	19
Pr > ChiSq	0.023	0.035	0.054	0.047	0.055	0.058	0.068	0.062
up to lag	20	21	22	23	24			
Pr > ChiSq	0.077	0.043	0.049	0.061	0.081			

Table 4.7: Check for white noise in the residuals

Therefore we consider a coefficient term at lag 12 for the exogeneous variable temp and a coefficient at lag 4 for the AR section of the model.

Using PROC VARMAX [7] we obtain the coefficient estimates with the added terms displayed in Table 4.8. Most of the parameter estimates can be seen in the

Schematic Representation of Parameter Estimates									
Variable/lag	C	XL0	XL1	XL3	XL12	AR1	AR2	AR3	AR4
Load	+	-	+	+	.	+	.	+	.

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
Load	Const1	17.913	7.660	2.34	0.0214	1
	XL0	-25.823	1.712	-15.08	0.0001	temp(t)
	XL1	9.512	2.865	3.32	0.0013	temp(t-1)
	XL3	8.762	2.819	3.11	0.0025	temp(t-3)
	XL12	0.410	1.697	0.24	0.8096	temp(t-12)
	AR1	0.359	0.091	3.94	0.0002	load(t-1)
	AR2	0.101	0.054	1.88	0.0635	load(t-2)
	AR3	0.261	0.095	2.75	0.0071	load(t-3)
	AR4	-0.096	0.052	-1.84	0.0691	load(t-4)

Table 4.8: Parameter Estimates for ARX(4,12) model

table to be significant, with the exception of the XL coefficient at lag 12. The P-values of AR2 and AR4 are slightly larger than 0.05, but removing them causes the model to become inadequate. Since they contribute to fitting the model with respect to the other parameter coefficients, we will retain the two parameters in the model. The P-value of the parameter of XL12 is considered large and the estimated value of XL12 was less than 1.5 of the estimated standard error. Omitting this parameter only induces the model to become inadequate in the diagnosis stage and thus will be retained.

Since most of the parameter estimates from Table 4.8 are significant, this suggests that the model is significant and now need to guard against model misspecification. Therefore, a detailed diagnostic analysis of the residuals given in Table 4.9 is necessary for the model to be adequate.

Schematic Representation of Cross-Correlation of Residuals															
Variable/lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Load	+	.	.	.	.	.	.	.	.	.	.	.	.	-	.

Portmanteau Test for Cross-Correlation of Residuals								
up to lag	5	6	7	8	9	10	11	12
Pr > ChiSq	0.292	0.529	0.655	0.708	0.576	0.697	0.793	0.073
up to lag	13	14	15	16	17	18	19	20
Pr > ChiSq	0.104	0.144	0.192	0.239	0.212	0.246	0.255	0.195
up to lag	21	22	23	24				
Pr > ChiSq	0.127	0.143	0.180	0.194				

Table 4.9: Residuals diagnostics for ARX(4,12)

The P-values of the Chi-squares in Table 4.9 have values larger than the 0.05 level of significance, indicating that the residuals in the model are uncorrelated and adequate. Our estimated model becomes

$$\begin{aligned}
 Y_t &= \delta + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3} + \Phi_4 Y_{t-4} \\
 &\quad + \Theta_0^* X_t + \Theta_1^* X_{t-1} + \Theta_3^* X_{t-3} + \Theta_{12}^* X_{t-12} + \epsilon_t \\
 &= 17.913 + 0.359 Y_{t-1} + 0.101 Y_{t-2} + 0.261 Y_{t-3} - 0.096 Y_{t-4} \\
 &\quad - 25.823 X_t + 9.512 X_{t-1} + 8.762 X_{t-3} + 0.410 X_{t-12} + \epsilon_t
 \end{aligned}$$

where

$$\begin{aligned}
 X_t &= (1 - B^{12}) X_t^* \\
 Y_t &= (1 - B^{12}) Y_t^*
 \end{aligned}$$

Now that we have an adequate model we are ready to forecast. Figure 4.4 displays the prediction error plot of our fitted model predicted values against the actual data values. This is similar to a residual plot of predicted against actual data values.

The dark shaded area in the prediction error plot are the boundaries for one standard error and the lightly shaded area contains the boundaries for 2 standard errors.

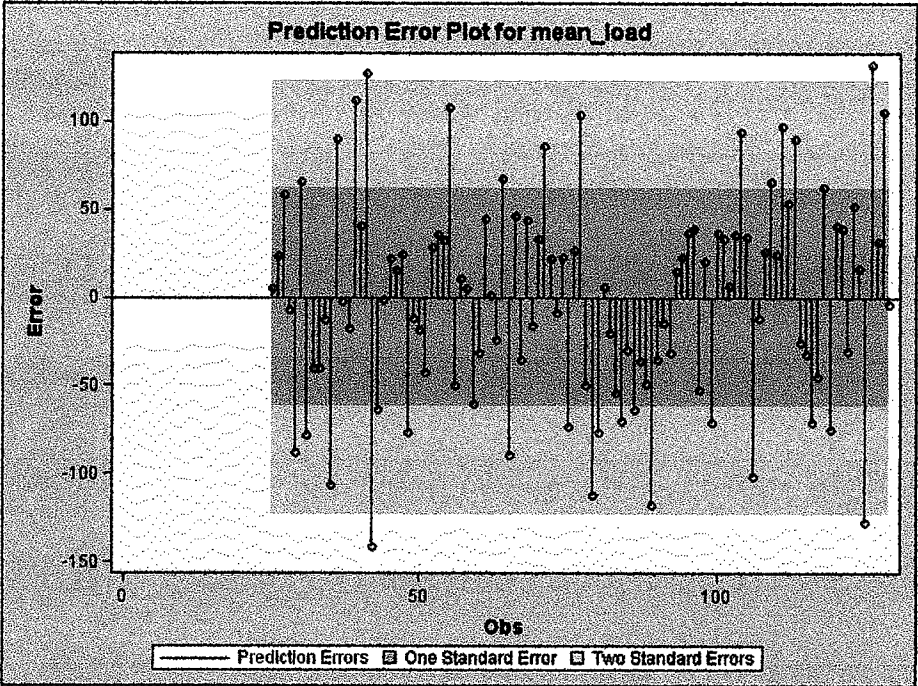


Figure 4.4: Prediction error plot of an ARX(4,12) model

Our forecasted values lies mainly within the two standard errors of the actual data with four of the predicted values lying barely past two standard errors. This shows that our model fits the data reasonably well. The forecasted values and plot are given in Table 4.10 and Figure 4.5.



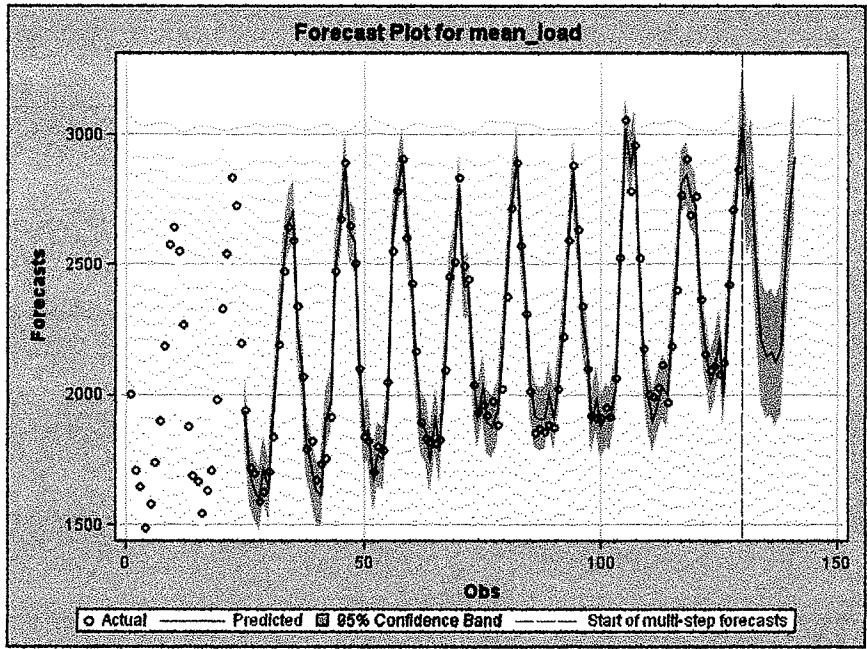


Figure 4.5: Forecast plot with forecasted monthly values for 2003

Time	Load	Forecast	Res	Std Err	95% LCI	95% UCI
Mar02	2764.78	2720.87	43.910	61.667	2600.01	2841.74
Apr02	2370.65	2330.30	40.346	61.667	2209.44	2451.17
May02	2159.34	2190.09	-30.746	61.667	2069.22	2310.95
Jun02	2096.61	2044.62	51.996	61.667	1923.75	2165.48
Jul02	2109.82	2093.61	16.206	61.667	1972.75	2214.48
Aug02	2080.15	2207.11	-127.199	61.667	2086.48	2328.21
Sep02	2126.87	1995.19	131.684	61.667	1874.32	2116.05
Oct02	2423.81	2389.87	33.947	61.667	2269.00	2510.73
Nov02	2711.05	2607.11	103.941	61.667	2486.25	2727.98
Dec03	2865.72	2873.69	-7.964	61.667	2752.82	2994.55
Jan03	.	3044.19	.	116.319	2816.20	3272.17
Feb03	.	2773.94	.	118.426	2541.82	3006.05
Mar03	.	2822.79	.	119.284	2589.00	3056.58
Apr03	.	2438.04	.	121.591	2199.73	2676.36
May03	.	2207.94	.	121.985	1968.85	2447.02
Jun03	.	2148.61	.	122.217	1909.06	2388.15
Jul03	.	2162.33	.	122.588	1922.06	2402.60
Aug03	.	2127.32	.	125.389	1881.56	2373.08
Sep03	.	2175.93	.	126.625	1927.74	2424.11
Oct03	.	2470.58	.	129.410	2216.94	2724.22
Nov03	.	2755.31	.	129.923	2500.66	3009.95
Dec03	.	2912.86	.	130.230	2657.61	3168.11

Table 4.10: Forecasted values from March 2002 to December 2003

The forecasted values from Table 4.10 are fairly close to the actual data values with August and September of 2002 actual values outside the 95% confidence intervals. This can be seen from the prediction error plot shown in Figure 4.4. Overall, the model is adequate for forecasting load, but this is only one of the reasonable models that we could possibly use.

Another model that was reasonable is the ARX(3,12) model, which was given as

$$\begin{aligned}
 Y_t &= \delta + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \Phi_3 Y_{t-3} \\
 &\quad + \Theta_0^* X_t + \Theta_1^* X_{t-1} + \Theta_3^* X_{t-3} + \Theta_{12}^* X_{t-12} + \epsilon_t \\
 &= 16.074 + 0.356 Y_{t-1} + 0.085 Y_{t-2} + 0.233 Y_{t-3} \\
 &\quad - 25.594 X_t + 9.544 X_{t-1} + 8.561 X_{t-3} + 0.816 X_{t-12} + \epsilon_t
 \end{aligned}$$

The forecasts and prediction error of the alternative model are given in Figure 4.6 and Table 4.11.

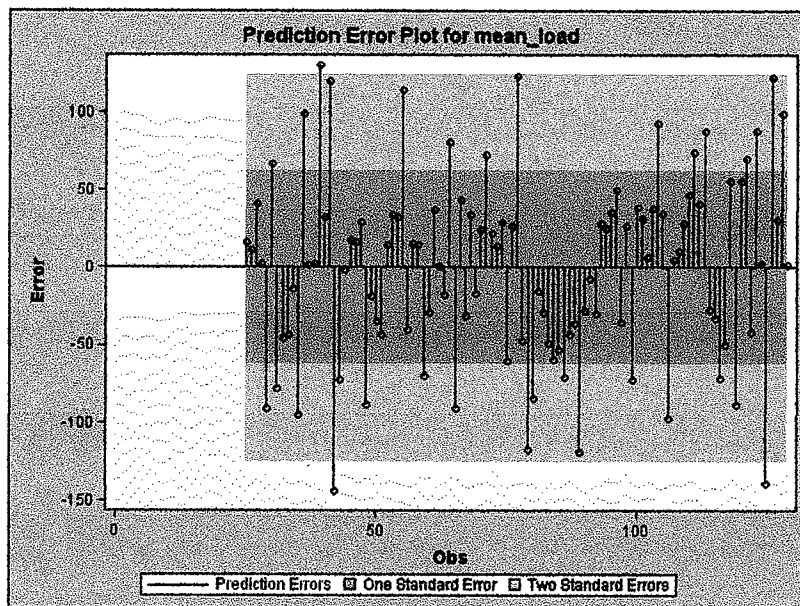


Figure 4.6: Prediction error plot of an ARX(3,12) model

The forecasted values for the alternative model fits the actual data values reasonably well. This can be seen from the prediction error plot from Figure 4.6 with approximately 5% of the values lies outside the two standard errors. Just like the first model, it also had a spike at the 12th lag on the residual cross-correlation, but it neglected to affect the P-values on the portmanteau test of the residuals. This implies that perhaps there was an anomaly that occurred at lag 12 which explains why there was a spike for both models at the 12th lag for the residuals.

Lastly, even though the alternative model has less estimated parameters, it also has a larger variance than the first model, but they both forecasted the load data reasonably well. Therefore, either model would have been acceptable for this data set.

Time	Load	Forecast	Res	Std Err	95% LCI	95% UCI
Mar02	2764.78	2709.01	50.768	62.419	2586.68	2831.35
Apr02	2370.65	2300.15	70.497	62.419	2177.81	2422.49
May02	2159.34	2201.18	-41.842	62.419	2078.84	2323.52
Jun02	2096.61	2008.94	87.679	62.419	1886.60	2131.27
Jul02	2109.82	2107.21	2.607	62.419	1984.87	2229.55
Aug02	2080.15	2218.49	-138.340	62.419	2096.15	2340.83
Sep02	2126.87	2004.60	122.271	62.419	1882.26	2126.94
Oct02	2423.81	2393.38	30.433	62.419	2271.04	2515.72
Nov02	2711.05	2611.91	99.140	62.419	2489.58	2734.25
Dec02	2865.72	2863.54	2.184	62.419	2741.20	2985.37
Jan03	.	3042.15	.	115.421	2815.93	3268.37
Feb03	.	2793.18	.	117.549	2562.79	3023.57
Mar03	.	2857.09	.	118.296	2625.24	3088.95
Apr03	.	2450.11	.	120.183	2214.56	2685.67
May03	.	2227.46	.	120.982	1990.34	2464.58
Jun03	.	2165.36	.	121.382	1927.45	2403.26
Jul03	.	2173.74	.	121.842	1934.93	2412.54
Aug03	.	2138.38	.	125.427	1892.55	2384.22
Sep03	.	2184.35	.	126.868	1935.69	2433.01
Oct03	.	2475.93	.	129.539	2222.04	2729.82
Nov03	.	2758.65	.	130.769	2502.35	3014.95
Dec03	.	2917.22	.	131.419	2659.64	3174.79

Table 4.11: Forecasted values from March 2002 to December 2003 for the alternative model

## 4.4 Fitting Multivariate Models to Daily data

In this section, we will extend our discussion to daily vector autoregressive (VAR) models for load and temperature taken from 01Jan99 to 31Dec02. In Figure 4.7, we see the original time plots of load and temperature.

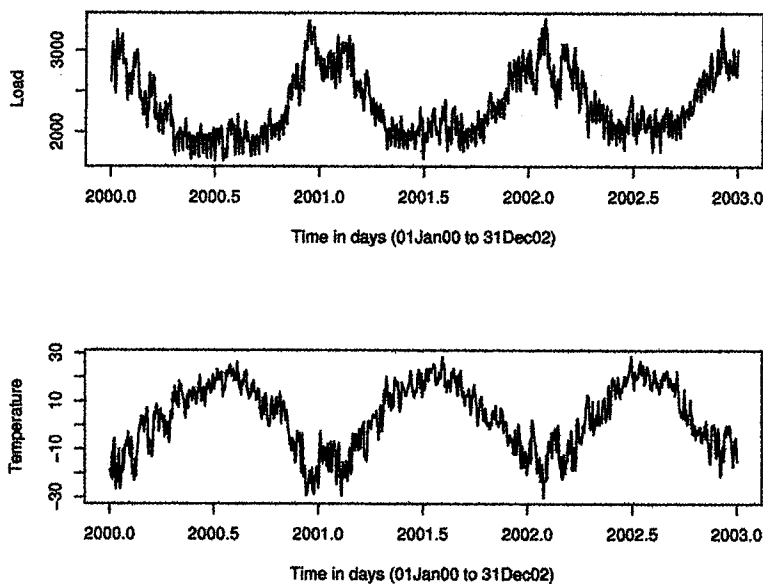


Figure 4.7: Daily time plot of the original series

Just like the monthly case, the load and temperature have a similar inverse relationship. There is a distinguishing seasonal pattern occurring every year for both load and temperature series, as well as another possible seasonal pattern occurring inside due to the systematic fluctuations along the yearly pattern. To check if another seasonal pattern exists, let's look at the behaviour of load broken down by the day of the week shown in Figure 4.8.

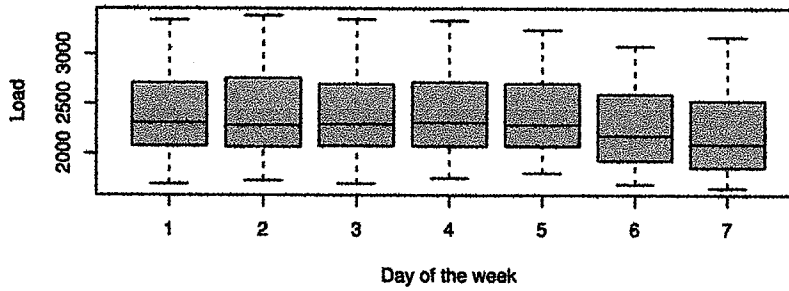


Figure 4.8: Behaviour of load throughout the week

From Figure 4.8, we notice the averages of the load from Monday to Friday appears to be higher than the weekend. This could possibly be due to the fact that there are less people working on the weekend and also commercial buildings require more energy on light fixtures, heating cooling, etc than housing. This means that we may need to adjust our model by breaking apart our daily data into forecasted weekdays and weekends separately.

#### 4.4.1 Fitting Multivariate Models to Weekday Daily Data

Lets consider the weekdays of the daily data from 01Jan00 to 31Dec02 shown in Figure 4.9.

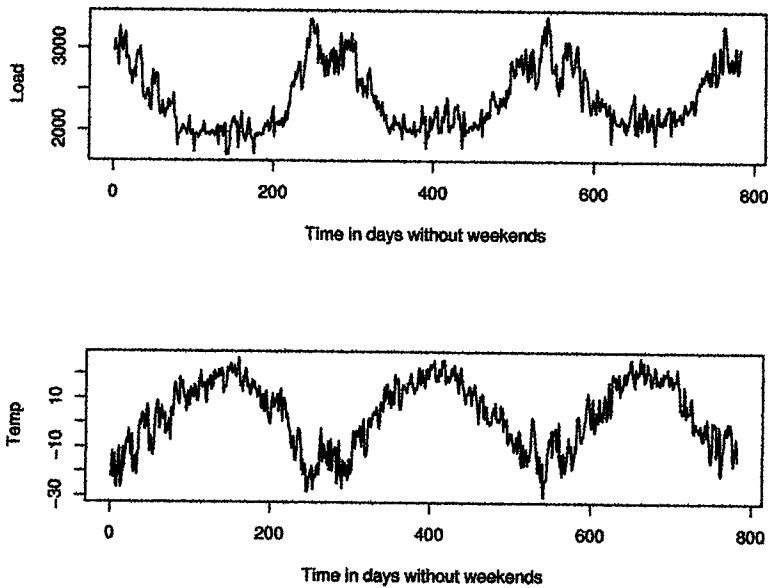


Figure 4.9: Time plots of load and temp without weekends

With the exclusion of weekends, we see that there is far less variability in both plots of load and temp than when weekends are included. We see this more with load than we do with temperature. Figure 4.10 shows the ACF of load and temp without weekends.

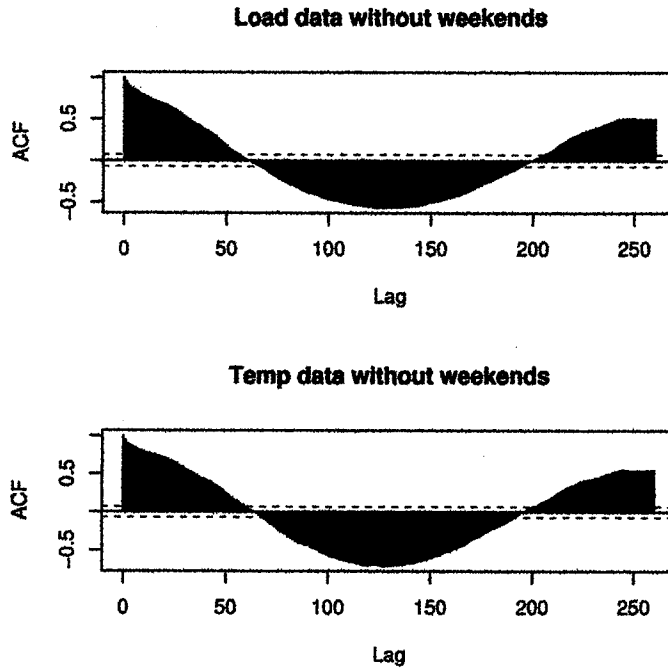


Figure 4.10: ACF plots of load and temp without weekends

We see the autocorrelations decays very slowly which indicates a nonstationary process. The weekly seasonal pattern has now disappeared and only the yearly seasonal exists. This indicates that we can now take a seasonal difference of 260 to make the data stationary and we won't have to worry about the weekly seasonal pattern. Taking a seasonal difference of 260 in Figure 4.11, the plots now appear stationary with mean now close to zero and seasonality eliminated.



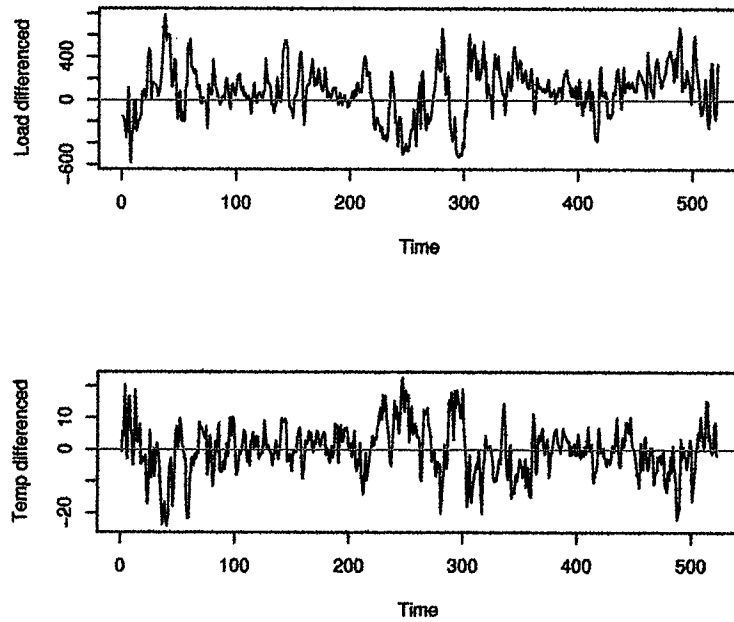


Figure 4.11: Time plots of seasonally differenced load and temp without weekends

For notational purposes, we will now denote the seasonally differenced series as

$$X_t = (1 - B^{260})X_t^* \quad (4.7)$$

$$Y_t = (1 - B^{260})Y_t^* \quad (4.8)$$

$$Z_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \quad (4.9)$$

Where  $X_t$  is the differenced series for temperature and  $Y_t$  is the differenced series for load. Figure 4.12 displays the ACF and PACF of the seasonally differenced series  $Z_t$ .

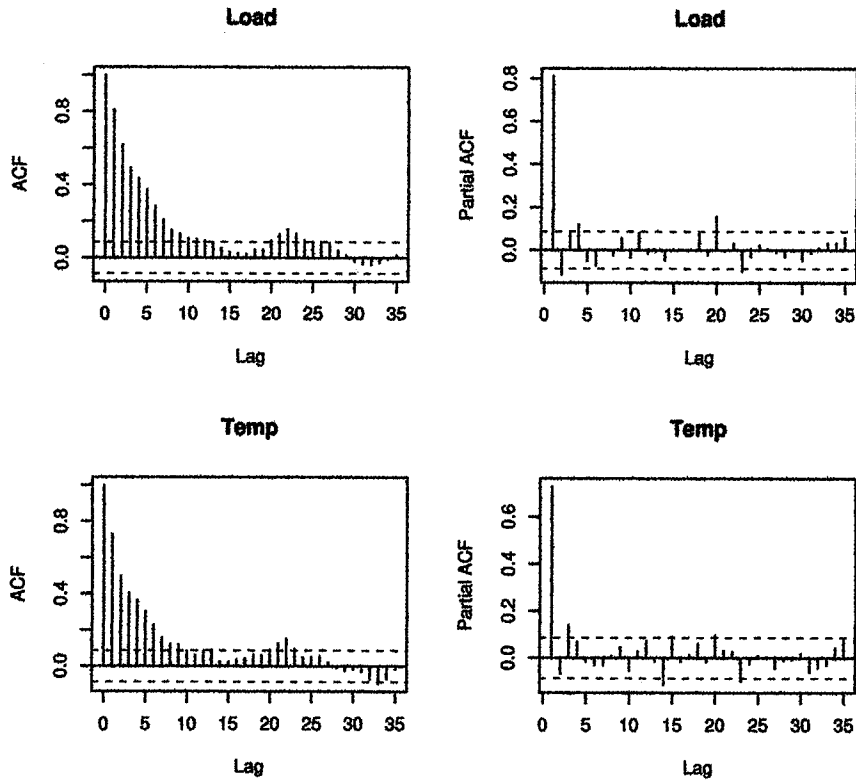


Figure 4.12: Sample ACF and PACF of differenced data  $Z_t$  without weekends

These correlations on the ACFs for both load and temperature decay fairly quickly to zero after lag 10 with another spike at the 22nd lag. The partial correlations on the otherhand, drop to zero quickly after lag 2 and 3. This indicates that a reasonable initial model is possibly a VAR(3) model. To determine a possible initial model, Table 4.12 provides summary of results for the initial VARMA(p,q) model based on the AIC method obtained from SAS.

Minimum Information Criterion							
	Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR	0	14.282	13.705	13.422	13.318	13.214	13.070
AR	1	12.776	12.786	12.788	12.786	12.790	12.788
AR	2	12.760	12.794	12.797	12.798	12.794	12.793
AR	3	12.764	12.791	12.792	12.800	12.796	12.801
AR	4	<b>12.757</b>	12.783	12.788	12.802	12.800	12.811
AR	5	12.767	12.788	12.789	12.797	12.810	12.824
AR	6	<b>12.754</b>	12.772	12.779	12.788	12.801	12.799
AR	7	12.767	12.779	12.784	12.796	12.794	12.795
AR	8	12.779	12.795	12.800	12.812	12.810	12.813
AR	9	12.787	12.803	12.818	12.830	12.828	12.831
AR	10	12.801	12.817	12.833	12.848	12.847	12.850
AR	11	12.795	12.811	12.826	12.842	12.858	12.863
AR	12	12.804	12.819	12.835	12.851	12.866	12.882

Table 4.12: Minic Information criterion using AIC method to identify possible VARMA model

The AIC results suggest that a VAR(4) or VAR(6) might be an ideal model to use. Since the AIC values for both are very close, we will attempt to use a VAR(6) as a preliminary model. Once we have a initial model, we use the CAUSAL statement from PROC VARMAX to compute the Granger-Causality test for the VAR(6) model shown in Table 4.13.

Granger-Causality Wald Test			
Test	DF	Chi-Square	P-value
1	6	15.37	0.0089
Test1: Group1 Variables			Temp
Group2 Variables			Load

Table 4.13: Granger-Causality test for exogeneous variables

The CAUSAL statement fits the VAR(6) model using the variables in the two groups and considering them as dependent variables. The CAUSAL statement fits the VAR(6) model using the variables load and temperature. At the 0.05 level of significance for Test 1, the P-value is very small from Table 4.13. This show that you cannot reject at the 0.05 level of significance that weekday daily temperature is influenced by itself and by load as well. Since the test of hypothesis rejects the null hypothesis, the variable temperature in group 1 is considered a dependent variable. Later from the Granger-Causality test, we will see that for the weekend daily temperature, it is influenced by itself but not by load at the 0.05 significance level. This suggest that we don't consider the variable temperature as exogeneous and still retain the VAR model.

Without going into as much detail, our fifth parameter coefficient in the VAR(6) model was not significant and was omitted. The estimates and schematic representation of our estimates for the remaining VAR(6) model is now given in Table 4.14.

Schematic Representation of Parameter Estimates						
Variable/lag	C	AR1	AR2	AR3	AR4	AR6
Load	+	+-	..	..	+	..
Temp	.	-+	--	..	..	..

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
load	CONST1	28.654	7.378	3.88	0.0001	1
	AR1 <sub>11</sub>	0.835	0.050	16.61	0.0001	load(t-1)
	AR1 <sub>12</sub>	-3.556	1.279	-2.78	0.0056	temp(t-1)
	AR2 <sub>11</sub>	-0.123	0.063	-1.94	0.0535	load(t-2)
	AR2 <sub>12</sub>	2.189	1.524	1.44	0.1517	temp(t-2)
	AR3 <sub>11</sub>	-0.092	0.063	-1.45	0.1473	load(t-3)
	AR3 <sub>12</sub>	-1.978	1.509	-1.31	0.1905	temp(t-3)
	AR4 <sub>11</sub>	0.138	0.053	2.57	0.0105	load(t-4)
	AR4 <sub>12</sub>	-1.364	1.306	-1.04	0.2970	temp(t-4)
	AR6 <sub>11</sub>	-0.085	0.039	-2.17	0.0304	load(t-6)
	AR6 <sub>12</sub>	-0.481	1.080	-0.45	0.6560	temp(t-6)
	temp	CONST2	0.500	0.291	1.72	0.0867
AR1 <sub>21</sub>		-0.005	0.001	-2.76	0.0061	load(t-1)
AR1 <sub>22</sub>		0.714	0.050	14.13	0.0001	temp(t-1)
AR2 <sub>21</sub>		0.001	0.002	0.28	0.7818	load(t-2)
AR2 <sub>22</sub>		-0.195	0.060	-3.24	0.0013	temp(t-2)
AR3 <sub>21</sub>		0.001	0.002	0.35	0.7274	load(t-3)
AR3 <sub>22</sub>		0.097	0.059	1.64	0.1026	temp(t-3)
AR4 <sub>21</sub>		-0.003	0.002	-1.46	0.1458	load(t-4)
AR4 <sub>22</sub>		0.026	0.051	0.52	0.6019	temp(t-4)
AR6 <sub>21</sub>		0.001	0.001	0.47	0.6357	load(t-6)
AR6 <sub>22</sub>		-0.042	0.042	-1.00	0.3177	temp(t-6)

Table 4.14: Parameter Estimates for VAR(6)

In the last Table, the second column gives the parameter name  $ARL_{ij}$ , which indicates the (ij)th element of the lag P autoregressive coefficient. It is apparent that

not all of the (ij)th elements of the lag P autoregressive coefficient are significant. If we retain the model as it is, the model is adequate in the Portmanteau test for cross-correlation of residuals displayed in Table 4.15.

Schematic Representation of Cross-Correlation of Residuals																			
Var/lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Load	+-	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
Temp	+-	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Var/lag	19	20	21	22	23	24
Load	..	+-	..	+-	..	..
Temp	..	..	..	+-	..	..

Portmanteau Test for Cross-Correlation of Residuals								
up to lag	7	8	9	10	11	12	13	14
Pr > ChiSq	0.456	0.669	0.552	0.670	0.514	0.548	0.330	0.351
up to lag	15	16	17	18	19	20	21	22
Pr > ChiSq	0.252	0.403	0.301	0.298	0.141	0.121	0.148	0.033
up to lag	23	24	25	26	27	28		
Pr > ChiSq	0.032	0.046	0.077	0.101	0.114	0.166		

Table 4.15: Residuals diagnostics for VAR(6)

We see that the majority of the P-values are large except for lag 22 and its neighboring lags 21 and 23 which are less than 0.05. It doesn't appear to affect the other lags beyond lag 22 so we can say that the spike was caused largely by chance. The large spikes at lag 22 could also be due to the fact that if we are just looking at weekdays, there is roughly 22 days in a month. Weather can change drastically in a month, which in turn can affect the electrical load. Overall, the model is adequate and can be used to forecast.

Instead of using the VAR(6) model to forecast, let's first reduce the model by restricting some of the (ij) elements of the Pth lag autoregressive coefficients to zero that are not significant. After eliminating the off-diagonal terms from the LS estimate of  $\Phi_6$ , the second column terms of  $\Phi_4$ , and the (2,1) position of  $\Phi_2$ , which are found to be clearly nonsignificant, and reestimating the simplified model, we arrive at the estimated model presented in Table 4.16.

Schematic Representation of Parameter Estimates						
Variable/lag	C	AR1	AR2	AR3	AR4	AR6
Load	+	+·	··	··	+*	·*
Temp	+	+·	*·	+·	·*	*·

+ is > 2×std error, - is < 2×std error, · is between, \* is NA

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
load	CONST1	26.011	6.973	3.73	0.0002	1
	AR1 <sub>11</sub>	0.836	0.047	17.41	0.0001	load(t-1)
	AR1 <sub>12</sub>	-3.640	1.269	-2.87	0.0043	temp(t-1)
	AR2 <sub>11</sub>	-0.116	0.056	-2.07	0.0390	load(t-2)
	AR2 <sub>12</sub>	2.470	1.487	1.66	0.0973	temp(t-2)
	AR3 <sub>11</sub>	-0.107	0.061	-1.75	0.0804	load(t-3)
	AR3 <sub>12</sub>	-2.906	1.270	-2.29	0.0226	temp(t-3)
	AR4 <sub>11</sub>	0.161	0.047	3.37	0.0008	load(t-4)
	AR4 <sub>12</sub>	0.000	0.000			temp(t-4)
	AR6 <sub>11</sub>	-0.070	0.029	-2.37	0.0184	load(t-6)
	AR6 <sub>12</sub>	0.000	0.000			temp(t-6)
	temp	CONST2	0.579	0.275	2.10	0.0358
AR1 <sub>21</sub>		-0.005	0.001	-3.30	0.0010	load(t-1)
AR1 <sub>22</sub>		0.718	0.049	14.55	0.0001	temp(t-1)
AR2 <sub>21</sub>		0.000	0.000			load(t-2)
AR2 <sub>22</sub>		-0.206	0.055	-3.72	0.0002	temp(t-2)
AR3 <sub>21</sub>		0.001	0.002	0.69	0.4932	load(t-3)
AR3 <sub>22</sub>		0.118	0.049	2.36	0.0188	temp(t-3)
AR4 <sub>21</sub>		-0.003	0.001	-1.92	0.0553	load(t-4)
AR4 <sub>22</sub>		0.000	0.000			temp(t-4)
AR6 <sub>21</sub>		0.000	0.000			load(t-6)
AR6 <sub>22</sub>		-0.059	0.031	-1.86	0.0628	temp(t-6)

Table 4.16: Parameter Estimates for a simplified VAR(6) model



The parameter estimates of a simplified VAR(6) model appear to have more terms significant when some of the terms were restricted, then when using the unrestricted VAR(6) model. Some of the estimates in the model are insignificant, but removing them causes the model to become inadequate. Since most of the parameter estimates from Table 4.17 are significant, this suggests that the model overall is significant, and we need to check for model adequacy in order to protect against model misspecification. Therefore, a detailed diagnostic analysis of the residuals is given in Table 4.17 is necessary for the model to be adequate.

Schematic Representation of Cross-Correlation of Residuals																			
Var/lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Load	+	..	..	..	..	..	..	..	..	..	..	..	..	..	..	+	..	..	..
Temp	-+	..	..	..	..	..	..	..	..	..	..	..	..	+	..	..	..	..	..

Var/lag	19	20	21	22	23	24
Load	..	..	..	+	..	..
Temp	..	..	..	+	..	..

Portmanteau Test for Cross-Correlation of Residuals								
up to lag	7	8	9	10	11	12	13	14
Pr > ChiSq	0.204	0.437	0.379	0.522	0.392	0.431	0.246	0.274
up to lag	15	16	17	18	19	20	21	22
Pr > ChiSq	0.190	0.325	0.256	0.251	0.105	0.097	0.122	0.028
up to lag	23	24	25	26	27	28		
Pr > ChiSq	0.027	0.041	0.069	0.088	0.099	0.147		

Table 4.17: Residuals diagnostics for a simplified VAR(6)

Examination of the residuals from this simplified fitted VAR(6) model gives no indication of inadequacy of the model, except for lag 21, 22 and 23 which are not significant at the 0.05 significance level. The spikes at lag 22 doesn't affect the lags afterwards so this model is accepted as an adequate representation of the bivariate series.

The fitted model for the simplified VAR(6) is given as

$$\begin{aligned}
 Z_t &= C + \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \Phi_3 Z_{t-3} + \Phi_4 Z_{t-4} + \Phi_6 Z_{t-6} + \epsilon_t \\
 \begin{bmatrix} Y_t \\ X_t \end{bmatrix} &= \begin{bmatrix} 26.0111 \\ 0.5793 \end{bmatrix} + \begin{bmatrix} 0.8356 & -3.64 \\ -0.0052 & 0.7175 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.1161 & 2.4703 \\ 0 & -0.2062 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} + \begin{bmatrix} -0.107 & -2.9062 \\ 0.0014 & 0.1176 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} \\
 &+ \begin{bmatrix} 0.161 & 0 \\ -0.0034 & 0 \end{bmatrix} \begin{bmatrix} Y_{t-4} \\ X_{t-4} \end{bmatrix} + \begin{bmatrix} -0.0703 & 0 \\ 0 & -0.0587 \end{bmatrix} \begin{bmatrix} Y_{t-6} \\ X_{t-6} \end{bmatrix} \\
 &+ \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{xt} \end{bmatrix}
 \end{aligned}$$

Where

$$X_t = (1 - B^{260})X_t^*$$

$$Y_t = (1 - B^{260})Y_t^*$$

$$\mathbf{Z}_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix}$$

The Forecasted values and prediction errors for both load and temperature are presented in Figure 4.13, Figure 4.14 and Table 4.18. We can clearly see that the forecasted values of the fitted model performs reasonably well against the actual values for both load and temperature with approximately 95% of the values lying within 2 standard errors. The standard errors for load and temperature were 130.928 and 5.167 respectively.

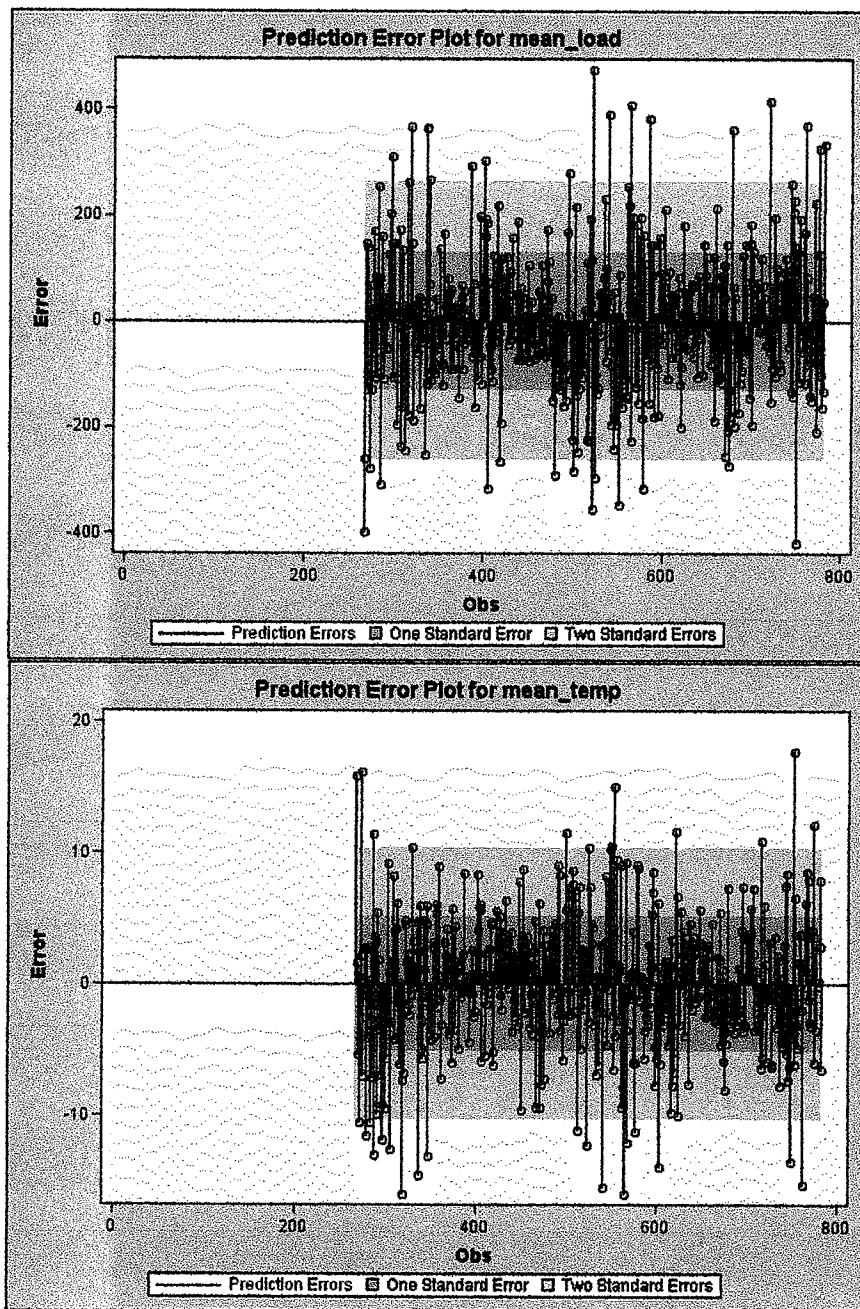


Figure 4.13: Prediction error plot of a re-estimated VAR(6) model

Time	Load	Fore1	95% LCI	95% UCI	Temp	Fore2	95% LCI	95% LCI
20Dec02	2798.25	2754.65	2498.03	3011.26	-5.375	0.555	-9.574	10.684
23Dec02	3019.29	2692.51	2435.89	2949.12	-11.754	-14.88	-25.008	-4.751
24Dec02	2989.92	2863.22	2606.60	3119.83	-17.162	-13.683	-23.812	-3.554
25Dec02	2755.17	2856.56	2599.94	3113.17	-13.220	-13.108	-23.237	-2.979
26Dec02	2710.29	2872.14	2615.53	3128.76	-11.212	-11.38	-21.509	-1.251
27Dec02	2737.21	2868.33	2611.71	3124.94	-5.862	-13.756	-23.885	-3.628
30Dec02	2917.13	2880.50	2623.88	3137.11	-8.591	-11.511	-21.64	-1.382
31Dec02	3005.71	2670.93	2414.32	2927.55	-15.134	-8.664	-18.793	1.465
01Jan03	.	3304.96	3048.35	3561.57	.	-18.701	-28.83	-8.572
02Jan03	.	3112.64	2765.20	3460.09	.	-13.624	-26.526	-0.721
03Jan03	.	2917.67	2529.98	3305.35	.	-9.452	-23.177	4.274
06Jan03	.	2852.27	2449.12	3255.42	.	-4.91	-19.041	9.222
07Jan03	.	2693.95	2276.64	3111.25	.	-0.255	-14.774	14.264

Table 4.18: Forecasted values for daily weekday values of load and temp

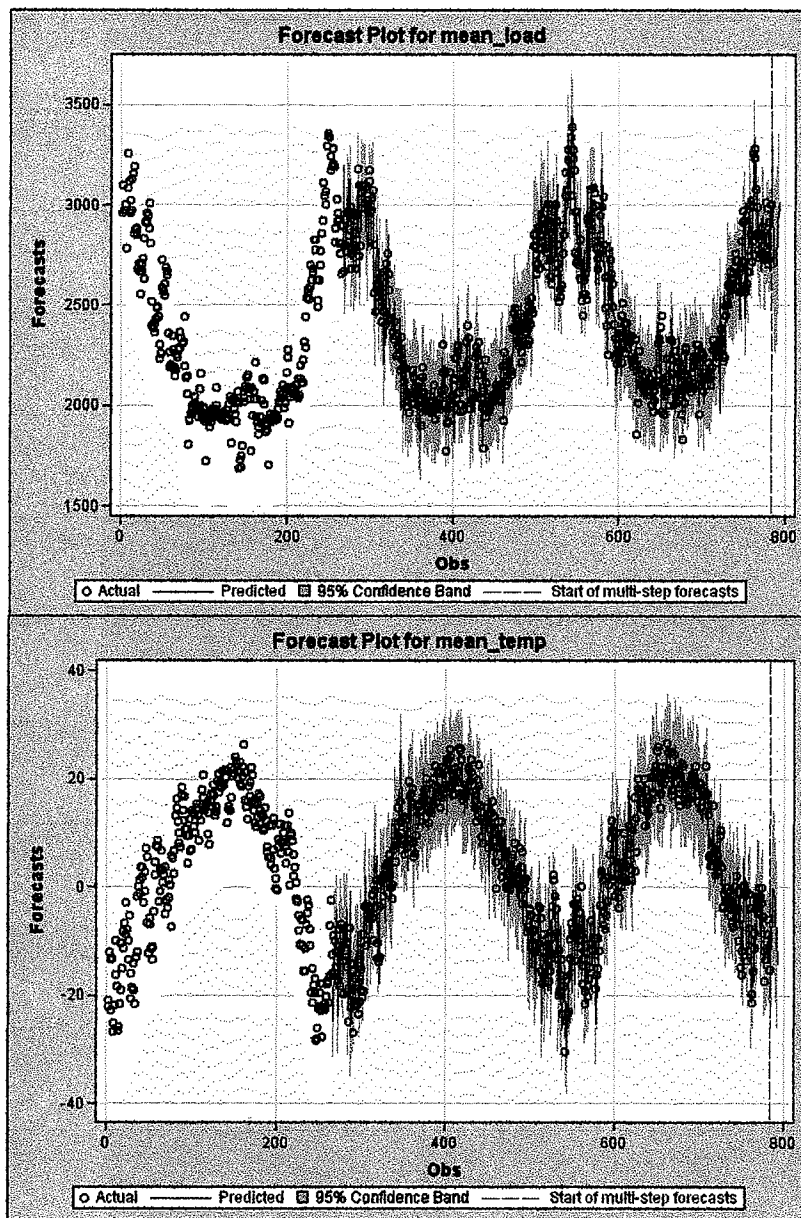


Figure 4.14: Forecast plot of a simplified VAR(6) model

Alternatively we could have taken the VAR(6) model without any restrictions on the (ij) parameter coefficients. If we look at the prediction error plot of load and temperature for the unrestricted VAR(6) model presented in Figure 4.15, we can clearly see that the prediction error plot showed us the same pattern that we noticed with the restricted VAR(6) model.

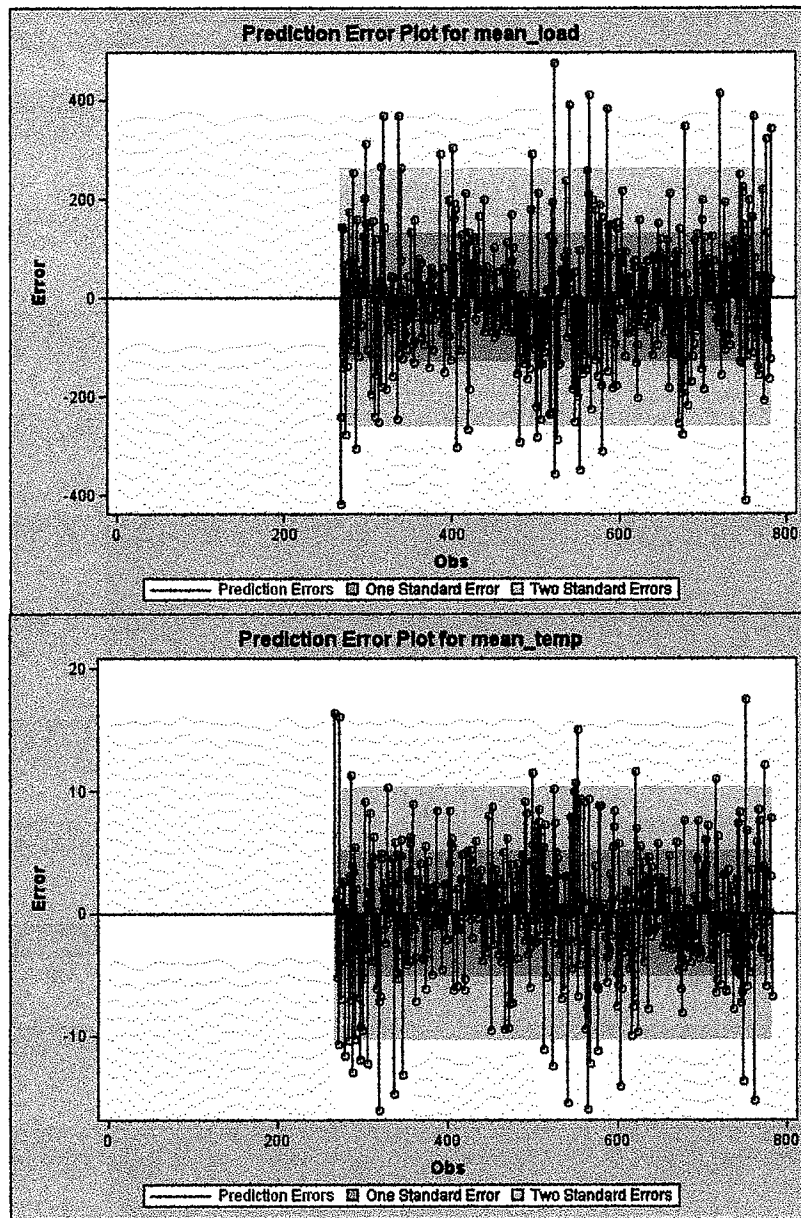


Figure 4.15: Prediction error plot of a unrestricted VAR(6) model

The histogram of the residuals for the error plots from the unrestricted VAR(6) model are shown in Figure 4.16.

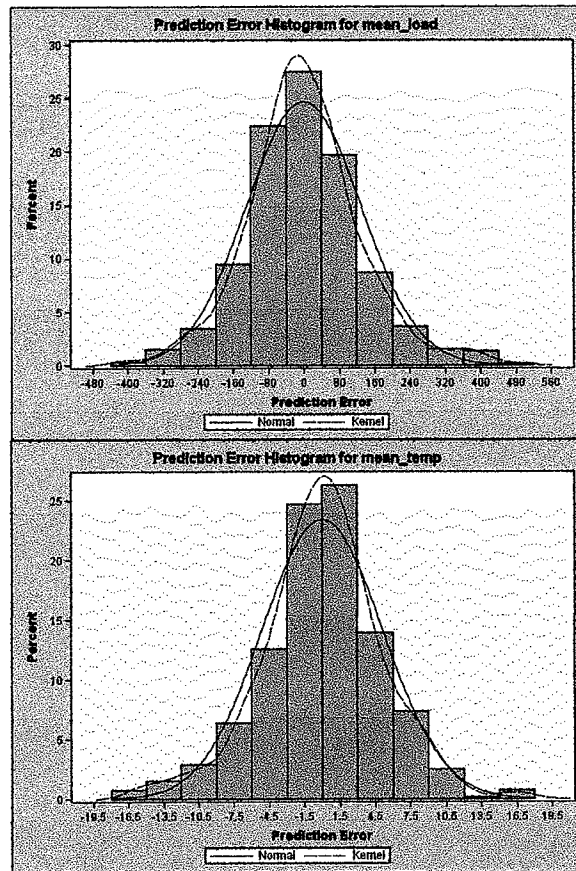


Figure 4.16: Prediction error plot for the unrestricted VAR(6) model

The assumption of normal errors seems reasonable. This implies that there are no inadequacies of the given model. The forecast values for both models in Figure 4.13 and 4.15 are fairly similar with the standard errors for an unrestricted VAR(6) model being slightly smaller. This implies that taking either model would be sufficient for forecasting weekday data.

#### 4.4.2 Fitting Multivariate Models to Weekend Daily Data

Now that we have obtained a model for the weekday case, we need to fit a multivariate model to the weekend daily data. We consider daily data from 01Jan00 to 31Dec02 comprised of the weekend data only shown in Figure 4.17.

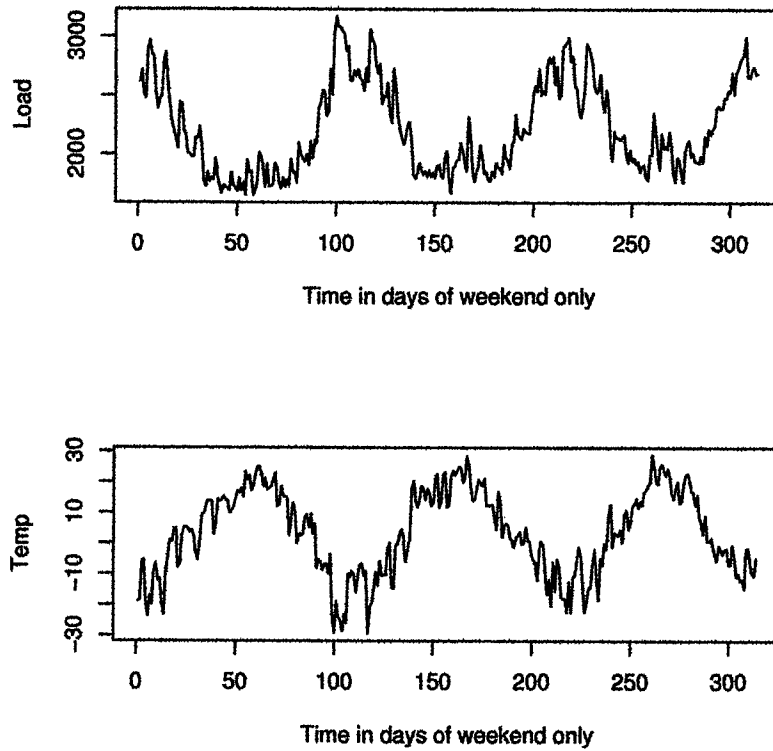


Figure 4.17: Time plots of the daily weekend data for load and temp

There is an obvious seasonal pattern occurring within each year, with high load values occurring in winter and low values in summer, vice versa for temperature. If we take a look at the ACF's of load and temperature in Figure 4.18, we can see the data is not stationary. We could either take a seasonal difference of 104 to remove the seasonal pattern or we could take a nonseasonal first difference since the values of load have an upward trend.



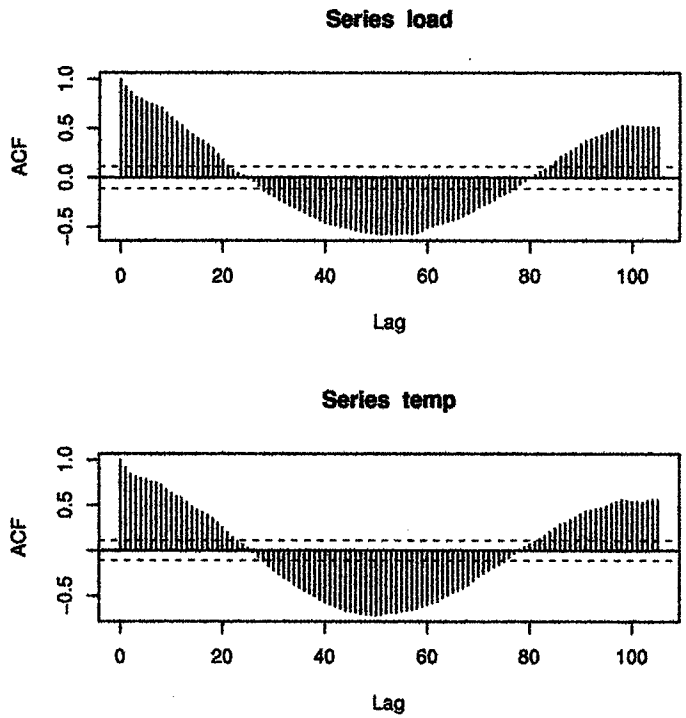


Figure 4.18: ACF plots of the daily weekend data for load and temp

Without showing any detail, since the ACF and PACF of the first difference appears to be more stationary than the seasonal difference for load and temperature, we will use the nonseasonal difference of the first order. The time plots of the differenced series are given in Figure 4.19.

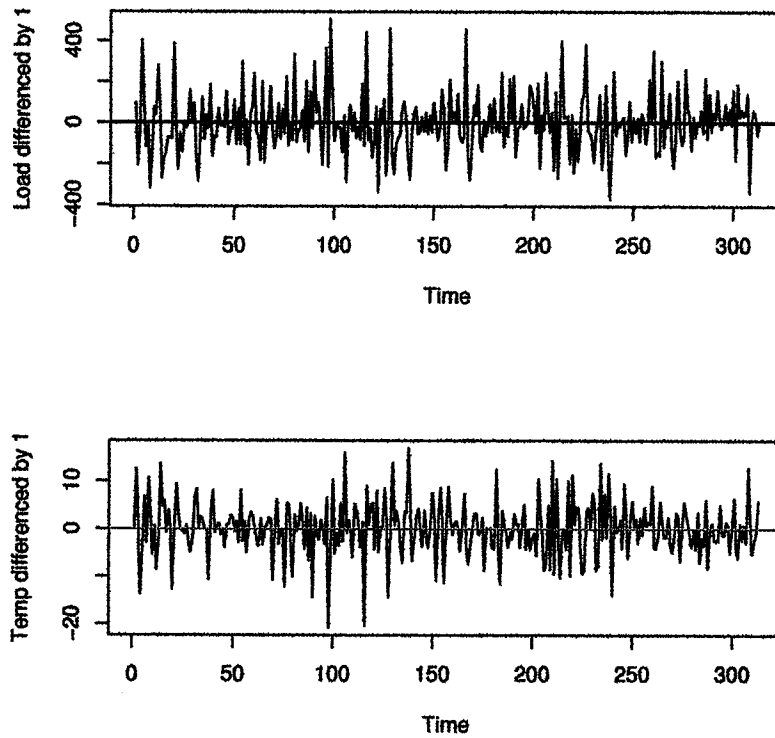


Figure 4.19: Time plots of the first difference daily weekend data.

After taking the first difference, the data now becomes stationary with mean approximately zero, and variance constant. The differenced series are now denoted as

$$X_t = (1 - B)X_t^* \quad (4.10)$$

$$Y_t = (1 - B)Y_t^* \quad (4.11)$$

$$\mathbf{Z}_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \quad (4.12)$$

We also notice that after taking the first difference, the seasonality has disappeared. This also can be seen in the ACF's and PACF's presented in Figure 4.20.

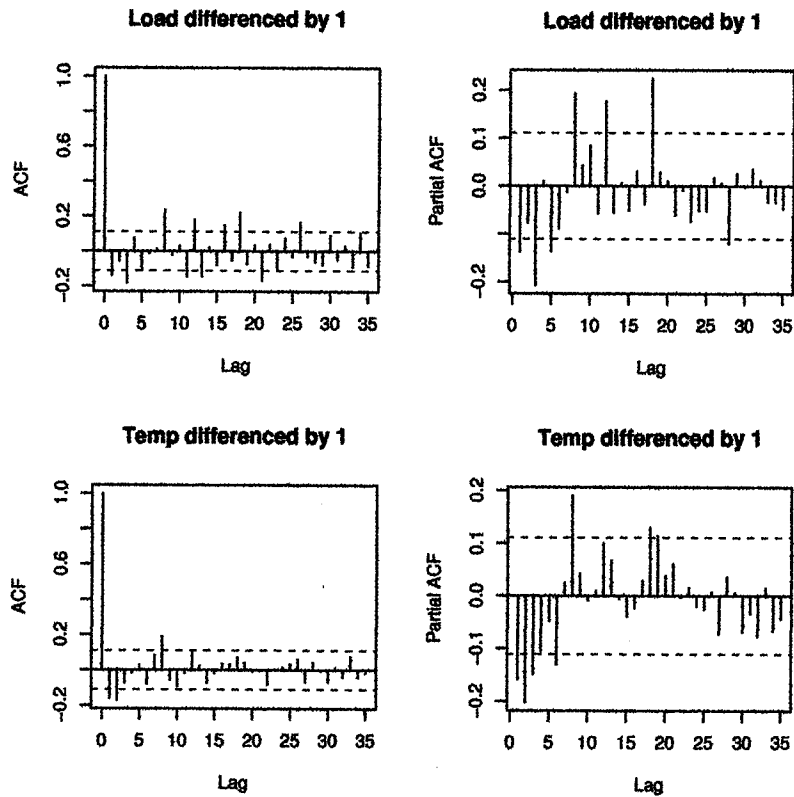


Figure 4.20: ACF plots of the daily weekend data  $Z_t$

When looking at Figure 4.20, the ACFs for both series decay to zero quite quickly after the 2nd or 3rd lag. The ACF plots also show that seasonal portion of the series has disappeared when taking the first difference. There are a few large spikes on both series that occur on the ACF, which makes the model identification a little more difficult to use. To determine a possible initial model, Table 4.19 provides a summary of results for the initial VARMA(p,q) model based on the AIC method obtained in SAS.

Minimum Information Criterion							
	Lag	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5
AR	0	13.056	13.026	12.972	12.959	12.935	12.955
AR	1	13.025	13.017	12.965	12.925	12.924	12.931
AR	2	12.961	12.972	12.924	12.862	12.867	12.881
AR	3	12.827	12.843	12.838	12.853	12.827	12.842
AR	4	12.824	12.849	12.842	12.857	12.835	12.855
AR	5	12.826	12.855	12.842	12.847	12.818	12.821
AR	6	12.811	12.824	12.807	12.800	12.800	12.798
AR	7	12.813	12.800	12.785	12.794	12.803	12.806
AR	8	<b>12.760</b>	12.775	12.776	12.797	12.789	12.803
AR	9	12.781	12.807	12.808	12.829	12.821	12.835
AR	10	12.783	12.809	12.836	12.857	12.850	12.864
AR	11	12.806	12.832	12.859	12.885	12.878	12.892
AR	12	12.787	12.814	12.841	12.867	12.894	12.908

Table 4.19: Minic Information criterion using AIC method to identify possible VARMA model

The AIC results suggest that a VAR(8) might be an ideal model to use. Now that we have an initial model, we use the CAUSAL statement from PROC VARMAX to compute the Granger-Causality test given in Table 4.20.

Granger-Causality Wald Test			
Test	DF	Chi-Square	P-value
1	8	10.69	0.2201
Test1: Group1 Variables			Temp
Group2 Variables			Load

Table 4.20: Granger-Causality test for exogeneous variables

At the 0.05 level of significance, the P-value is fairly large from Table 4.20. The output shows the weekend daily temperature is influenced by itself but not by load at the 0.05 significance level. Since the test of hypothesis fails to reject the null hypothesis, the variable temperature in group 1 is considered an exogeneous variable. Although this is true for the weekend daily data, temperature was not exogeneous for the weekday daily data. This suggests that we will use a VAR(P) model instead of the ARX(p,s) model. The estimates and schematic representation of the estimates for a VAR(8) model is given in Table 4.21.

Schematic Representation of Parameter Estimates								
Variable/lag	AR1	AR2	AR3	AR4	AR5	AR6	AR7	AR8
Load	--	--	--	--	--	--	--	+
Temp	--	--	--	--	--	--	--	+

Model Parameter Estimates						
Equation	Parameter	Estimate	Std Error	t-value	Pr >  t	Variable
load	AR1 <sub>11</sub>	-0.173	0.069	-2.49	0.0132	load(t-1)
	AR1 <sub>12</sub>	-0.942	1.821	-0.52	0.6053	temp(t-1)
	AR2 <sub>11</sub>	-0.160	0.071	-2.26	0.0245	load(t-2)
	AR2 <sub>12</sub>	-2.142	1.913	-1.12	0.2639	temp(t-2)
	AR3 <sub>11</sub>	-0.243	0.074	-3.29	0.0011	load(t-3)
	AR3 <sub>12</sub>	-2.759	2.051	-1.35	0.1796	temp(t-3)
	AR4 <sub>11</sub>	0.021	0.077	0.27	0.7856	load(t-4)
	AR4 <sub>12</sub>	1.566	2.142	0.73	0.4653	temp(t-4)
	AR5 <sub>11</sub>	-0.127	0.076	-1.66	0.0986	load(t-5)
	AR5 <sub>12</sub>	-0.905	2.143	-0.42	0.6730	temp(t-5)
	AR6 <sub>11</sub>	-0.033	0.074	-0.45	0.6503	load(t-6)
	AR6 <sub>12</sub>	1.898	2.052	0.93	0.3556	temp(t-6)
	AR7 <sub>11</sub>	-0.012	0.070	-0.17	0.8638	load(t-7)
	AR7 <sub>12</sub>	-2.945	1.901	-1.55	0.1226	temp(t-7)
	AR8 <sub>11</sub>	0.213	0.069	3.10	0.0021	load(t-8)
	AR8 <sub>12</sub>	1.418	1.816	0.78	0.4354	temp(t-8)
temp	AR1 <sub>21</sub>	-0.002	0.002	-0.87	0.3828	load(t-1)
	AR1 <sub>22</sub>	-0.304	0.069	-4.36	0.0001	temp(t-1)
	AR2 <sub>21</sub>	-0.006	0.002	-2.55	0.0114	load(t-2)
	AR2 <sub>22</sub>	-0.346	0.073	-4.73	0.0001	temp(t-2)
	AR3 <sub>21</sub>	-0.005	0.002	-1.95	0.0520	load(t-3)
	AR3 <sub>22</sub>	-0.273	0.078	-3.48	0.0006	temp(t-3)
	AR4 <sub>21</sub>	-0.001	0.002	-0.35	0.7296	load(t-4)
	AR4 <sub>22</sub>	-0.152	0.082	-1.86	0.0644	temp(t-4)
	AR5 <sub>21</sub>	-0.000	0.002	-0.26	0.7931	load(t-5)
	AR5 <sub>22</sub>	-0.059	0.082	-0.73	0.4667	temp(t-5)
	AR6 <sub>21</sub>	-0.0001	0.002	-0.02	0.9844	load(t-6)
	AR6 <sub>22</sub>	-0.078	0.078	-1.00	0.3195	temp(t-6)
	AR7 <sub>21</sub>	-0.002	0.002	-0.91	0.3633	load(t-7)
	AR7 <sub>22</sub>	0.052	0.072	0.72	0.4732	temp(t-7)
	AR8 <sub>21</sub>	-0.001	0.002	-0.59	0.5542	load(t-8)
	AR8 <sub>22</sub>	0.176	0.069	2.53	0.0120	temp(t-8)

Table 4.21: Parameter Estimates for a VAR(8) model

It is apparent that not all of the (ij)th elements of the lag P autoregressive coefficient are significant, especially lags 4 to 7 for the autoregressive coefficient. Even though those estimated terms were not significant, they contribute to the model being adequate. We check this through the Portmanteau test of the residuals displayed in Table 4.22.

Schematic Representation of Cross-Correlation of Residuals																			
Var/lag	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Load	+-	..	..	..	..	..	..	..	..	..	..	..	+-	..	..	..	..	..	+-
Temp	-+	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	-+

Var/lag	19	20	21	22	23	24
Load	..	..	..	..	..	..
Temp	..	..	..	..	..	..

Portmanteau Test for Cross-Correlation of Residuals								
up to lag	9	10	11	12	13	14	15	16
Pr > ChiSq	0.134	0.110	0.307	0.076	0.157	0.217	0.323	0.444
up to lag	17	18	19	20	21	22	23	24
Pr > ChiSq	0.509	0.144	0.160	0.194	0.214	0.294	0.296	0.258
up to lag	25	26	27	28				
Pr > ChiSq	0.151	0.138	0.114	0.065				

Table 4.22: Residuals diagnostics for VAR(6)

From the given table, we see that all of the P-values are greater than the 0.05 level of significance. This implies that our model is adequate and is ready to forecast. The fitted model for a VAR(8) is denoted as

$$\begin{aligned}
 Z_t &= \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \Phi_3 Z_{t-3} + \Phi_4 Z_{t-4} + \Phi_5 Z_{t-5} \\
 &+ \Phi_6 Z_{t-6} + \Phi_7 Z_{t-7} + \Phi_8 Z_{t-8} + \epsilon_t \\
 \begin{bmatrix} Y_t \\ X_t \end{bmatrix} &= \begin{bmatrix} -0.1793 & -0.9422 \\ -0.0023 & -0.3047 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.1608 & -2.142 \\ -0.007 & -0.347 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.2439 & -2.76 \\ -0.0056 & -0.2739 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} 0.0211 & 1.5664 \\ -0.001 & -0.1526 \end{bmatrix} \begin{bmatrix} Y_{t-4} \\ X_{t-4} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.1275 & -0.9055 \\ -0.0008 & -0.0599 \end{bmatrix} \begin{bmatrix} Y_{t-5} \\ X_{t-5} \end{bmatrix} + \begin{bmatrix} -0.0336 & 1.8987 \\ -0.0001 & -0.0785 \end{bmatrix} \begin{bmatrix} Y_{t-6} \\ X_{t-6} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.0122 & -2.9453 \\ -0.0025 & 0.0524 \end{bmatrix} \begin{bmatrix} Y_{t-7} \\ X_{t-7} \end{bmatrix} + \begin{bmatrix} 0.2138 & 1.4184 \\ -0.0016 & 0.1762 \end{bmatrix} \begin{bmatrix} Y_{t-8} \\ X_{t-8} \end{bmatrix} \\
 &+ \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{xt} \end{bmatrix}
 \end{aligned}$$

Where

$$X_t = (1 - B)X_t^*$$

$$Y_t = (1 - B)Y_t^*$$

$$Z_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix}$$



The daily weekend forecasted values and prediction errors for load and temperature are given Figures 4.21, 4.22 and Table 4.23.

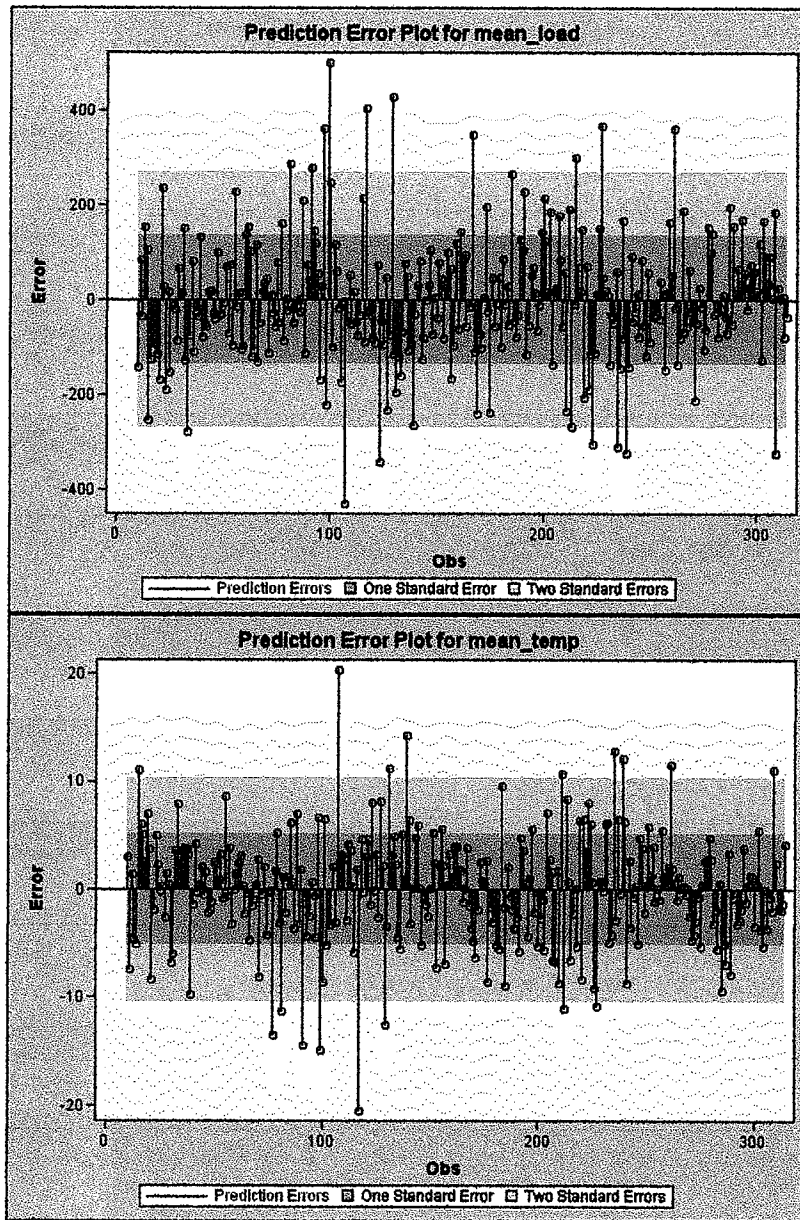


Figure 4.21: Prediction error plot of a VAR(8) model

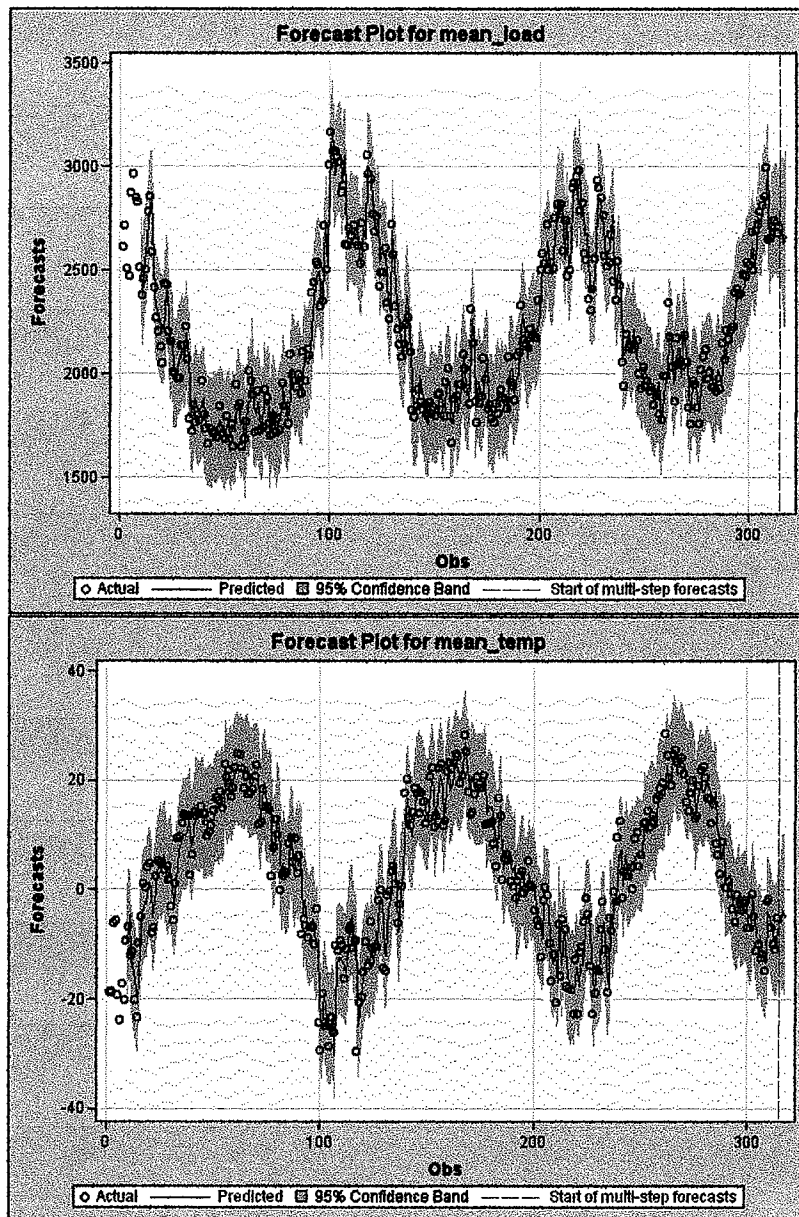


Figure 4.22: Forecasts plot of a VAR(8) model

Time	Load	Fore1	95% LCI	95% UCI	Temp	Fore2	95% LCI	95% LCI
15Dec02	2657.63	2630.04	2365.44	2894.65	-1.52	-4.01	-14.16	6.15
21Dec02	2714.00	2711.01	2446.40	2975.62	-6.86	-5.16	-15.31	4.99
22Dec02	2743.38	2736.39	2471.78	3001.00	-9.83	-7.94	-18.09	2.21
28Dec02	2683.63	2760.80	2496.20	3025.41	-10.68	-9.39	-19.54	0.77
29Dec02	2685.63	2721.06	2456.46	2985.67	-5.01	-9.21	-19.36	0.95
05Jan03	.	2737.00	2472.39	3001.61	.	-6.09	-16.24	4.06
06Jan03	.	2720.42	2373.62	3067.21	.	-5.88	-18.46	7.70
12Jan03	.	2628.19	2227.34	3029.04	.	-3.89	-17.83	10.05
13Jan03	.	2662.08	2229.30	3094.87	.	-5.09	-20.09	9.90

Table 4.23: Forecasted values for daily weekend values of load and temp

We see the forecasted values for the fitted model perform reasonably well with approximately 95% of the values lying within 2 standard errors. Since both the histograms of the residuals shown in Figure 4.23 have symmetrical shapes, the error terms seems to follow a normal distribution. Therefore, there are no inadequacies in our model. The forecast plots are then given in Figure 4.22.

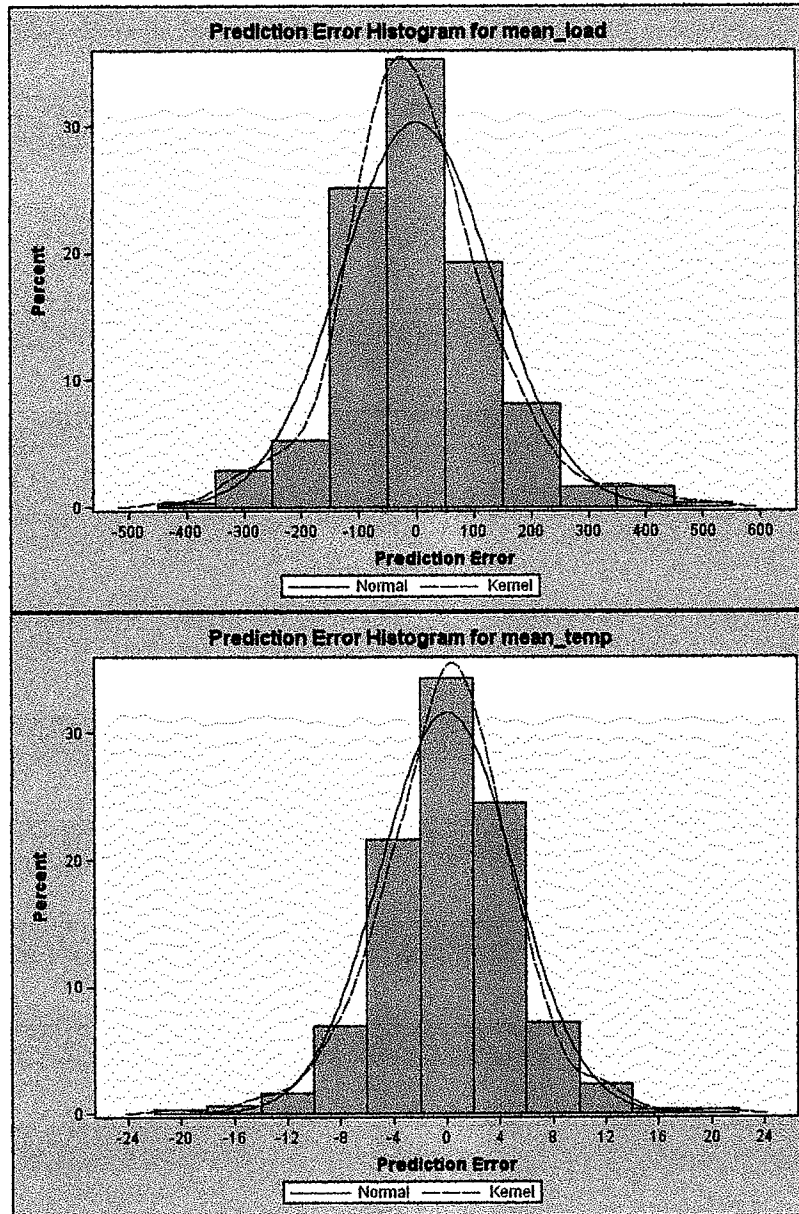


Figure 4.23: Forecasts plot of a VAR(8) model

Alternatively, we could restrict some of the (ij) parameter coefficients in the VAR(8) model. The restricted model is also found to be adequate and is given as

$$\begin{aligned}
 Z_t &= \Phi_1 Z_{t-1} + \Phi_2 Z_{t-2} + \Phi_3 Z_{t-3} + \Phi_4 Z_{t-4} + \Phi_5 Z_{t-5} \\
 &+ \Phi_6 Z_{t-6} + \Phi_7 Z_{t-7} + \Phi_8 Z_{t-8} + \epsilon_t \\
 \begin{bmatrix} Y_t \\ X_t \end{bmatrix} &= \begin{bmatrix} -0.1817 & -1.1362 \\ -0.0019 & -0.2941 \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ X_{t-1} \end{bmatrix} + \begin{bmatrix} -0.1699 & -2.3851 \\ -0.0064 & -0.3337 \end{bmatrix} \begin{bmatrix} Y_{t-2} \\ X_{t-2} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.2531 & -2.9722 \\ -0.0051 & -0.263 \end{bmatrix} \begin{bmatrix} Y_{t-3} \\ X_{t-3} \end{bmatrix} + \begin{bmatrix} 0 & 1.2084 \\ 0 & 1.6092 \end{bmatrix} \begin{bmatrix} Y_{t-4} \\ X_{t-4} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.1409 & -1.2352 \\ 0 & -0.0413 \end{bmatrix} \begin{bmatrix} Y_{t-5} \\ X_{t-5} \end{bmatrix} + \begin{bmatrix} -0.0361 & 1.713 \\ 0 & -0.0693 \end{bmatrix} \begin{bmatrix} Y_{t-6} \\ X_{t-6} \end{bmatrix} \\
 &+ \begin{bmatrix} -0.0225 & -3.1475 \\ -0.0019 & 0.0637 \end{bmatrix} \begin{bmatrix} Y_{t-7} \\ X_{t-7} \end{bmatrix} + \begin{bmatrix} 0.1905 & 1.0908 \\ 0 & 0.1980 \end{bmatrix} \begin{bmatrix} Y_{t-8} \\ X_{t-8} \end{bmatrix} \\
 &+ \begin{bmatrix} \epsilon_{yt} \\ \epsilon_{xt} \end{bmatrix}
 \end{aligned}$$

Now if we look at the prediction error plot shown in Figure 4.24, we see that forecast for the restricted VAR(8) model is essentially the same as the unrestricted VAR(8) model with minor differences in the forecast error. Therefore, either model would be acceptable to use.

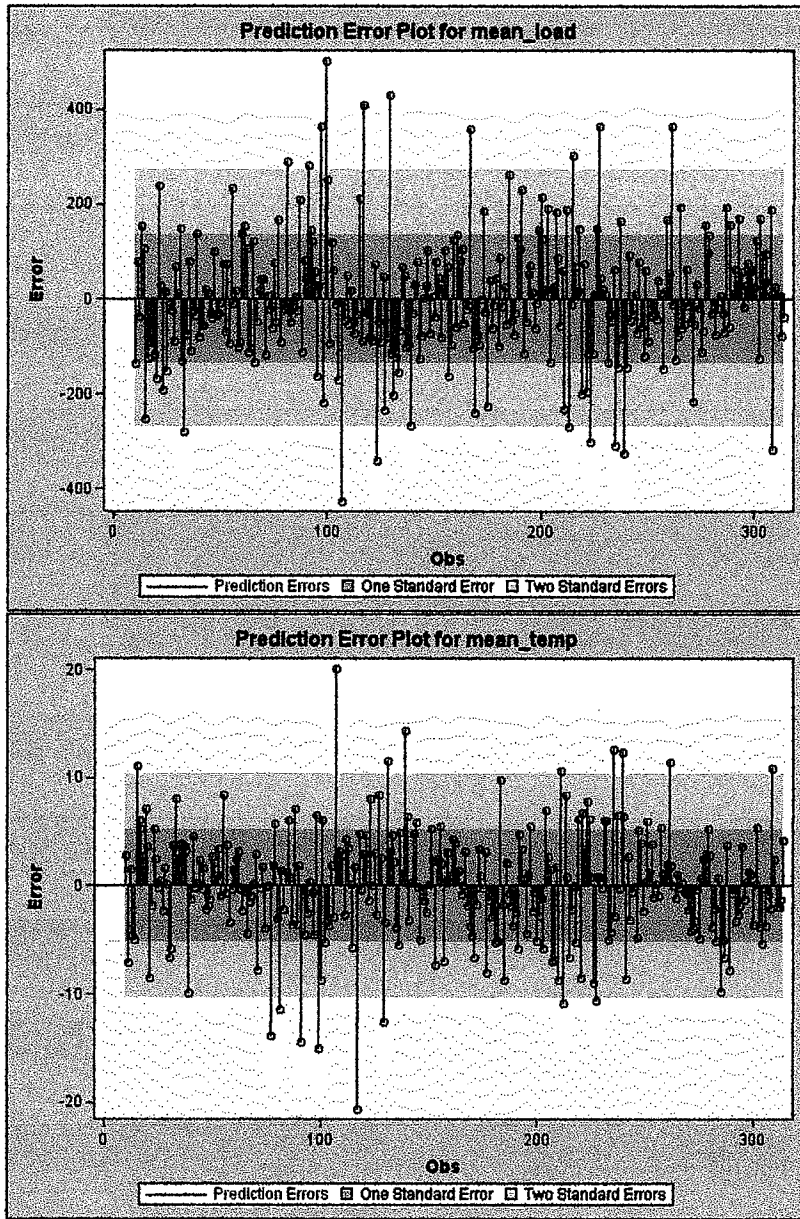


Figure 4.24: Prediction error plot of a VAR(8) model

Theoretically, multivariate models should produce more accurate forecasts than univariate models because multivariate forecasts are based on more information than just the past values of the series being forecasted. In most cases, univariate models usually produce better forecasts than multivariate models. Table 4.24 compares the daily temperature forecasts from the univariate and multivariate models to the actual temperature values.

Day	Date	Actual Temp	Forecasted	
			Multivariate Temp	Univariate Temp
Fri	Dec 20, 2002	-5.375	0.56	<b>-3.3</b>
Sat	Dec 21, 2002	-6.86	<b>-5.16</b>	-4.97
Sun	Dec 22, 2002	-9.83	<b>-7.94</b>	-6.17
Mon	Dec 23, 2002	-11.754	-14.88	<b>-8.67</b>
Tue	Dec 24, 2002	-17.162	<b>-13.68</b>	-9.93
Wed	Dec 25, 2002	-13.22	-13.11	<b>-15.08</b>
Thu	Dec 26, 2002	-11.212	<b>-11.38</b>	-10.78
Fri	Dec 27, 2002	-5.862	-13.75	<b>-10.56</b>
Sat	Dec 28, 2002	-10.68	<b>-9.39</b>	-5.8
Sun	Dec 29, 2002	-5.01	<b>9.21</b>	-11.4
Mon	Dec 30, 2002	-8.59	<b>-11.51</b>	-4.96
Tue	Dec 31, 2002	-15.13	-8.66	<b>-9.82</b>

Table 4.24: Temperature forecasts from multivariate vs. univariate

The numbers that are in bold are the forecasted values that were closer to the actual temperature values. From the above table, the daily forecast values for temperature from the multivariate model tend to exceed the forecasts from the univariate model. Therefore, our multivariate model is the more appropriate model to use for forecasting.

## 4.5 Fitting Multivariate Models to Hourly Data

In the last section, we shown the model for daily data needed to be partitioned into two parts because of the differences of load between the weekdays and weekends. Now lets examine the time plot of hourly data taken across the one week from 01Dec02 to 07Dec02 given in Figure 4.25.

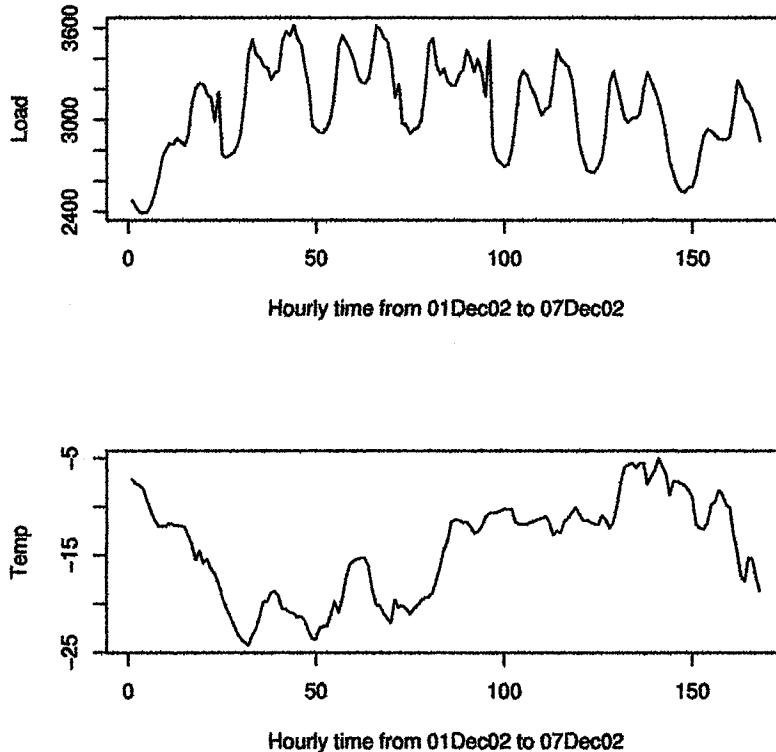


Figure 4.25: Time plot of load and temp across one week by the hour

On the load time plot, the series follows the same pattern as the plot of load behaviour discussed in Chapter 1. The load has a bimodal peak for every 24 hour intervals, while temperature doesn't appear to have any obvious pattern occurring. If we took the average of the hourly temperature behaviour for each month separately shown in Chapter 1, then we have a pattern similar to that of the load. The other problem we also have is that since load is lower on the weekends for the daily series, this applies to the hourly data as well. What does this mean? Can we find a multi-



variate model for the hourly series? If we try to model the series, there will be too many factors affecting the time series like we have seen in the daily case. If we want to forecast load for the hourly time interval, it would be much easier to use a univariate ARIMA model than a multivariate model due to simplicity. The multivariate model would become too complex and partitioning the model any further will only complicate our model even more.

## 4.6 Conclusion

In this section we have expanded the univariate series into a multivariate case involving two time series data: load and temperature. We looked at the monthly case and found that load and temperature have the same pattern to one another, only one is inverse to the other. We also examined the multivariate models for the daily data series. We noticed that electrical load throughout the week is much lower on the weekends than it is on the weekdays. Therefore, there were some complications that existed for transforming a nonstationary series into a stationary one. To alleviate this, we partitioned the daily series into weekend and weekday data, and modeled each separately. Lastly we looked at hourly data for the multivariate series. We have seen that for a given week, load has a consistent pattern occurring throughout the day for each day, while temp had no pattern evident. This suggests that there was no relationship that exists between load and temperature at the hourly intervals and forecasting was not required.

# Chapter 5

## Conclusion

This study has developed time series models for forecasting temperature data and its extremes. In Chapter 1 we discussed that the main goal of this practicum was to model temperature and its probability distribution for extreme occurrences. We introduced the univariate Box-Jenkins ARIMA procedure that could be used to model the temperature data at different time intervals. We also used the ARIMA procedure to model the temperature extremes, maximum and minimum. Before introducing the models for the temperature, we looked at the summary statistics of past historical temperature data. This information helped us understand how the average temperature and its extremes behave at different time intervals. We did see that load and temperature relationship are fairly consistent throughout the day and that the temperature data follows a normal distribution. The temperature data appears to be roughly consistent through time, meaning the ARIMA model should be appropriate for fitting the data.

We did see that the ARIMA models gave us fairly accurate results. Although we saw that using ARIMA models for the smaller time units such as day or hour, we didn't need to use the whole temperature data to achieve good forecasting results. Instead, we needed only to use a small fraction of the temperature data to achieve optimal results. The approximations were fairly similar to the maximum likelihood

results, but we didn't see any problem with the least squares approach. We also examined the multivariate models for the daily data series. We noticed that electrical load throughout the week is much lower on the weekends than it is on the weekdays. Therefore, there were some complications that existed for transforming a nonstationary series into a stationary one. To alleviate this, we partitioned the daily series into weekend and weekday data, and modeled each separately. Lastly we looked at hourly data for the multivariate series. We have seen that for a given week, load has a consistent pattern occurring throughout the day for each day of the week, while temp had no pattern evident. This suggests that there was no relationship that exists between load and temperature at the hourly intervals and therefore forecasting was not required.

One of the problems that exists with the VAR models is the number of parameters to be estimated may be very large. Due to the lack of parsimony, this could possibly lead to serious problems when the model is to be used for forecasting. To alleviate this, imposing restrictions to reduce the number of parameters helps improve their estimability. Another approach to overcome overparameterization is to use the Bayesian VAR model which imposes a prior on the AR parameters. By choosing the appropriate priors, you may be able to get more accurate forecasts using a BVAR model rather than using an unconstrained VAR model.

If we wanted to look into the model further, ARIMA or Vector ARIMA can be converted into state space models (and vice versa), in which the current and relevant past behaviour is included in the current state of the system [5]. Load and weather can then be updated using Kalman Filtering. With state space formulations, it is possible to make new forecast based on results from the previous hour, rather than recomputing the effect of the same past behaviour in several past hours. Alternatively, we could also take several possible models that are adequate for forecasting and make an inference of those models by taking an average of the forecasts to improve the forecasting capability. We might need to also incorporate other weather variables

such as wind and humidity. They might have some influence on the temperature and load relationship as well. Hopefully, the results of this project will give Manitoba Hydro a better understanding of temperature behaviour and its influences on load-temp forecasts. Any impurities on the model forecasts can result in a huge loss of revenues for Manitoba Hydro.

# Appendix A

## The Data

In our discussions we have used data collected by Manitoba Hydro since the beginning of January, 1953 until the end of December, 2002. Only some of the variables were of importance to us, but Table A.1 shows a subset of the data containing the variables of interest.

DATE	HOUR	TEMP (°C)
12AUG79	20	14.4
12AUG79	21	12.8
12AUG79	22	11.7
12AUG79	23	11.3
12AUG79	24	11.5
13AUG79	1	10.3
13AUG79	2	10
13AUG79	3	9.4
13AUG79	4	8.9
13AUG79	5	8.1

Table A.1: Partial data set collected by Manitoba Hydro

DATE is obviously the time the data was taken. HOUR is the hour of the day recorded in standard hour units. TEMP is the average Winnipeg temperature taken over the corresponding hour in degrees Celcius. To have a better understanding of what this means, lets take the first row in Table A.1. The average temperature between 8pm and 9pm on August 12, 1979 was 14.4°C.

## A.1 Monthly and Daily Data

In part of our discussion for Chapter 3, we analyzed the temperature for monthly data. Here is a subset of the data created in SAS shown in Table A.2, containing some of the extra variables that were of importance to us. This was created from the original data collected by Manitoba Hydro. A portion of the code is given in Appendix B.

YEAR	MONTH	MEAN TEMP(°C)	MAX TEMP(°C)	MIN TEMP(°C)
1965	1	-21.13	-2.2	-35.0
1965	2	-17.91	2.8	-33.9
1965	3	-13.39	0.6	-29.4
1965	4	3.32	23.9	-7.2
1965	5	10.54	29.4	-6.1
1965	6	16.94	27.2	2.2
1965	7	18.19	28.9	5.0
1965	8	17.60	35.0	1.7
1965	9	7.96	21.1	-6.7
1965	10	7.00	21.7	-6.1

Table A.2: Partial data set of monthly temp

The DATE variable from section A.1 is now divided into 2 variables called YEAR

and MONTH. MEANTEMP is the average temperature taken for that specific year and month. For example, in the first row of Table A.2, the average temperature for the 31 days in January, as well as the 24 hours in each day is  $-21.13^{\circ}\text{C}$ . MAXTEMP is the maximum temperature taken for that year and month and MINTEMP is the minimum temperature for that year and month.

The daily temperature is shown in Table A.3 with the variables the same as previously discussed.

DATE	MEAN TEMP( $^{\circ}\text{C}$ )	MAX TEMP( $^{\circ}\text{C}$ )	MIN TEMP( $^{\circ}\text{C}$ )
30MAY02	19.03	27.2	12.8
31MAY02	18.99	26.8	10.7
01JUN02	14.56	21.2	6.1
02JUN02	11.50	17.9	2.4
03JUN02	14.55	19.4	11.1
04JUN02	14.63	21.4	8.0
05JUN02	16.49	25.4	5.2
06JUN02	20.09	29.0	13.5
07JUN02	18.02	23.6	12.0
08JUN02	13.77	19.0	5.7

Table A.3: Partial data set of daily temperature

# Appendix B

## SAS and R code

In this appendix we introduce some of the SAS and R code used to fit an ARIMA model for the temperature data.

### B.1 PROC ARIMA code for ARIMA models in SAS

In chapter 3 we fit an ARIMA model for the monthly, daily and hourly temperature data using SAS. The following code was used to perform the three stages of ARIMA modelling.

```
libname hydro 'c:\scott_output';
data one; /*using the data hspot23012003 in a library called hydro */
  set hydro.hspot23012003 (drop=tz -- onoffday);
  where date>'31dec52'd and date<='31dec02'd;
run;

data two;
  set one;
```



```

date=date;
day=day(date);      /*just looking at the day values*/
month=month(date);  /*month values*/
year=year(date);    /*year*/
run;
proc means data=two noprint nway maxdec=2;
  class year month;
  var temp;
  output out = extremes
         mean = mean_temp
         max = max_temp
         min = min_temp;
run;
proc arima data=extremes;      ***for the min temp;
  identify var=min_temp(12);
  estimate noconstant p=(1,28) q=(12) method=cls;
  forecast lead=12 out=fore;
run;

```

As you can see from using PROC MEANS, we have a statement that creates the mean, max and min temperature for the monthly time intervals. By using the output statement we created a data set called "extremes", which has the mean temperature for each month and year for the 50 years as well as the extremes, maximum and minimum. To generate the mean, max and min for daily temperature, change the "year and month" in the class statement to "date". There are many options that can be used for PROC ARIMA. See the SAS user guide [7] for more details on the options available.

## B.2 Using the ESACF method in SAS

In chapter 2 and 3 we discussed the extended sample autocorrelation function (ESACF) for finding a tentative model in the identification stage. The analysis is based on "data extremes" created in the previous section.

```
proc arima data=extremes;          ***for the min temp;
  identify var=min_temp(12) minic esacf;
run;
```

## B.3 Summary Statistics in SAS

In Chapter 1 we presented the monthly summary statistics for the 50 years through the use of boxplots. The following code was used to generate them.

```
*** Sort data by BY variables ***;
proc sort data=SASUSER.monthtemp out=WORK._stsrt_0;
  by MONTH;
run;
*** Box plot ***;
goptions ftext=SWISS ctext=BLACK htext=1 cells;
symbol1 c=BLACK h=1 cells;
proc boxplot data=WORK._stsrt_0 ;
  plot (M_TEMP
        )*MONTH
        / caxis = BLACK
          cframe =
          ctext = BLACK
          cboxes = BLACK
          cboxfill = BLUE
```

```

        idcolor = BLUE
        idsymbol = SQUARE
        boxstyle = SKELETAL
        waxis = 1
        name = 'BOX'
        description = "Box Plots of M_TEMP by MONTH"
        npanel = 15
;
run;
symbol1;
goptions ftext= ctext= htext=;
axis1;
quit;

```

## B.4 PROC VARMAX Code for Multivariate Time Series

In Chapter 4 we looked at cross correlations between the load and temperature using SAS. The following code was used to perform cross correlations between the two time series as well as performing possible VAR models.

```

proc varmax data=extremes;
  model mean_load mean_temp /noint p=(1,2,3) method=ls dify(12)
    minic=(type=aic p=12 q=12) lagmax=24 print=(corry parcoef);
  causal group1=(mean_temp) group2=(mean_load);
  output out=foremonth1 lead=12;
run;

```

The CAUSAL statement in the above code was used to determine Granger-Causality between load and temperature. If load is dependent on temperature then the following code is used where load is the dependent variable and temperature is the independent (exogeneous) variable.

```
proc varmax data=extremes;
  model mean_load = mean_temp / p=(1,2,3,4) dify=(12) difx=(12)
      xlag=(1,3,12) lagmax=24;
  causal group1=(mean_temp) group2=(mean_load);
  output out=foremonth lead=12;
run;
```

## B.5 The ACF, PACF and CCF plots in R

In chapter 3 we used ACF and PACF plots to help aid in the identification stage in ARIMA modelling. We first will look at the ACF and PACF plots for the original data. The following code will perform these plots.

```
ex<- read.table("c:/scott_output/extremes2.txt",header=T)
> attach(ex)
> par(mfrow=c(2,2))
> acf(meantp)
> pacf(meantp)
```

We will now secondly look at the ACF and PACF plots of the differenced data and find an ARIMA model from the original data. The following code was used.

```
> mean.diff12 <- diff(meantp,lag=12)
> acf(mean.diff12)
> pacf(mean.diff12)
> model1 <- arima(meantp, order=c(1,0,0),
```

```
+seasonal=list(order=c(0,1,0), period=12))
```

Lastly, we now look at the CCF plots of load and temperature. The following code is the cross correlation between the two time series of the original data.

```
ex<- read.table("c:/scott_output/extremes2.txt",header=T)
> attach(ex)
> par(mfrow=c(2,2))
> ccf(meantp,meanload) ##produces the cross correlation
                        btw temp and load
```

## B.6 Spline Smoothing in R

In Chapter 1 we discussed the behaviour of average temperature throughout the day for each month. The following code was used to fit spline smoothing for nonlinear regression models on the month of January and February. The rest are similiar as the code below.

```
tempbyhour <- read.table("c:/scott_output/tempvarybyhour.txt",header=T)
attach(tempbyhour)

month.sub1 <-tempbyhour[mon==1,]
month.sub2 <-tempbyhour[mon==2,]
month.sub3 <-tempbyhour[mon==3,]
month.sub4 <-tempbyhour[mon==4,]
month.sub5 <-tempbyhour[mon==5,]
month.sub6 <-tempbyhour[mon==6,]
month.sub7 <-tempbyhour[mon==7,]
month.sub8 <-tempbyhour[mon==8,]
month.sub9 <-tempbyhour[mon==9,]
```

```

month.sub10 <-tempbyhour[mon==10,]
month.sub11 <-tempbyhour[mon==11,]
month.sub12 <-tempbyhour[mon==12,]

par(mfrow=c(3,2))
plot(month.sub1[,2],month.sub1[,4],xlab="hour of day",ylab="Mean temp",
      +main="Mean temp for January by hour")
fit <- smooth.spline(month.sub1[,2],month.sub1[,4],spar=0.3)
lines(fit$x, fit$y, col="blue", lwd=2) #plot smooth spline fit
plot(month.sub2[,2],month.sub2[,4],xlab="hour of day",ylab="Mean temp",
      +main="Mean temp for February by hour")
fit <- smooth.spline(month.sub2[,2],month.sub2[,4],spar=0.3)
lines(fit$x, fit$y, col="blue", lwd=2) #plot smooth spline fit

```

- [10] Silva, M (2002). "The Use of Piecewise Linear Models to Predict Hydroelectric Load for Manitoba Hydro", M.Sc Practicum, University of Manitoba, Department of Statistics.
- [11] Tsay, R. S. and G. C. Tiao (1984). Consistent Estimates of Autoregressive Parameters and Extended Sample Autocorrelation Function for Stationary and Nonstationary ARMA Models. *Journal of the American Statistical Association*, 79, 84-96.
- [12] Wei, William W. S. (1990). *Time Series Analysis: Univariate and Multivariate Methods*. Redwood City, California: Addison - Wesley.
- [13] <http://www.nps.navy.mil/Courses/msa847/sasdoc/sashtml/ets/chap7/sect21.htm>
- [14] <http://www.uwm.edu/IMT/Computing/sasdoc8/sashtml/ets/chap7/sect2.htm>