WIND SPEED ANALYSIS FOR LAKE OKEECHOBEE

MINGYAN HU

WIND SPEED ANALYSIS FOR LAKE OKEECHOBEE

by

Mingyan Hu

A Thesis Submitted to the Faculty of

The Charles E. Schmidt College of Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

Florida Atlantic University

Boca Raton, Florida

May 2002

To

My husband: Minqing Lu

My daughter: Wendy Lu

My son: Boris Hu Lu

WIND SPEED ANALYSIS FOR LAKE OKEECHOBEE

by

Mingyan Hu

This thesis was prepared under the direction of the candidate's thesis advisor, Dr. Lianfen Qian, Department of Mathematical Sciences, and has been approved by the members of her supervisory committee. It was submitted to the faculty of The Charles E. Schmidt College of Science and was accepted in partial fulfillment of the requirements for the degree of Master of Science.

SUPERVISORY COMMITTEE:

Lianfen Qian

Thesis Advisor

Stephen C. Locke

Chairman, Department of Mathematical Sciences

Dean, Charles E. Schmidt College of Science

Nathan W. Dean

Vice Provost

4/3/02

Date

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Lianfen Qian, who spent a lot of time on advising my research work. I sincerely appreciate her for giving me the opportunity to work with her. Her rich knowledge, wise thoughts and kind guidance encouraged me to complete this research. I would like to thank Dr. Kangren Jin, for providing data set and suggestions. I would like to thank Dr. Jose Correa for agreeing to serve on my thesis committee. Special thanks also goes to Dr. Heinrich Niederhausen and Dr. Frederick Hoffman for reviewing my thesis. I would like to thank my husband Mr. Minqing Lu, for his support and understanding of my work. And I would like to thank my mother, MS. Xiuen Zhang and my mother-in-law, MS. Caixian Li, who helped me to take care of my two kids.

I also thank Meihui Mao, Kehan Gao and all my friends for their kind support.

ABSTRACT

Author: Mingyan Hu

Title: Wind Speed Analysis for Lake Okeechobee

Institution: Florida Atlantic University

Thesis Advisor: Dr. Lianfen Qian

Degree: Master of Science

Year: 2002

In this thesis, we analyze wind speeds collected by South Florida Water Management District at stations L001, L005, L006 and LZ40 in Lake Okeechobee from January 1995 to December 2000. There are many missing values and outliers in this data. To impute the missing values, three different methods are used: Nearby window average imputation, Jones imputation using Kalman filter, and EM algorithm imputation. To detect outliers and remove impacts, we use ARIMA models of time series. Innovational and additive outliers are considered. It turns out that EM algorithm imputation is the best method for our wind speed data set. After imputing missing values, detecting outliers and removing the impacts, we obtain the best models for all four stations. They are all in the form of seasonal ARIMA$(2,0,0) \times (1,0,0)_{24}$ for the hourly wind speed data.

CONTENTS

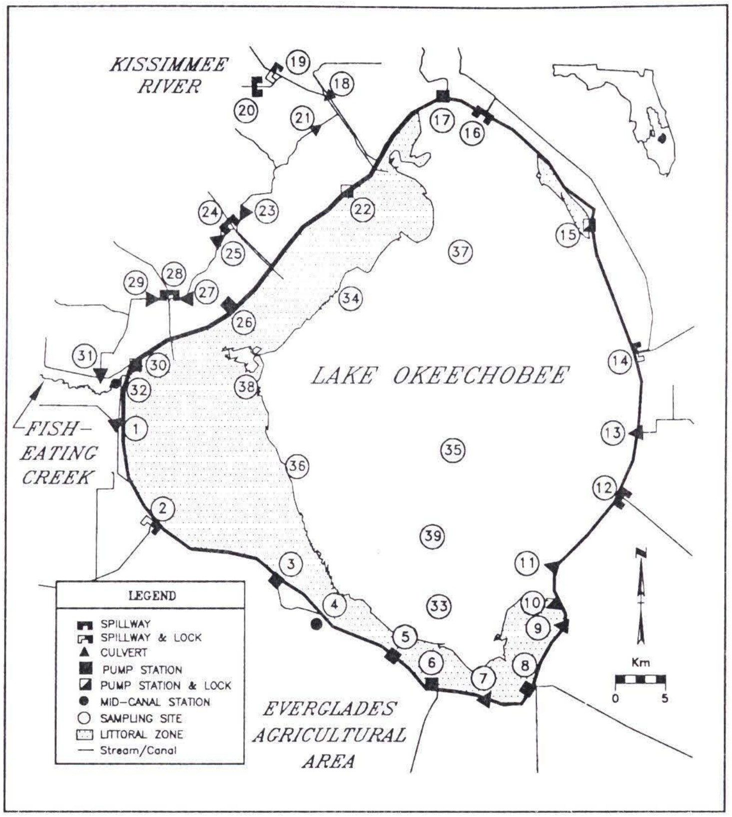
TABLES

FIGURES

Chapter 1

INTRODUCTION

Lake Okeechobee (Figure 1.1) is a natural lake in South Central Florida. Its name comes from two Indian words and means "big water". It is the second largest natural lake in the United States of America and is located at 27 N Latitude and 80 W Longitude. Its surface area is approximately $1730km^2$. It is very shallow, with mean and maximum depths of $2.7m$ and $5.5m$, respectively. A flood control dike built between 1930 and 1960 encircled the natural lake [12]. Currently, the lake has a storage capacity of about 40 billion cubic meters of water. Water levels are regulated according to a schedule developed by the U.S. Army Corps of Engineers. In addition to providing regional flood control, primary uses of the lake include agricultural water supply, drinking water for lakeside cities and towns and a backup water supply for the communities of the lower east coast of Florida. Other uses are commercial and recreational fishing, navigation and wildlife habitat. Lake Okeechobee is also a major component of the Kissimmee-Okeechobee-Everglades hydrologic system, receiving drainage from the Kissimmee River and discharging to the Everglades Agricultural Area [17].



Sample Sites: 16=L001, 38=L005, 39=L006, 35=LZ40

Figure 1.1: Lake Okeechobee and Data Collection Sites

Lake Okeechobee wind speed data are routinely collected by sensors and transcribed from field/laboratory forms to an electronic format. The data set analyzed in this thesis is the wind speeds (miles per hour) collected at stations L001, L005, L006 and LZ40 (corresponding sample sites 16, 38, 39 and 35, respectively in Figure 1.1) from January 1995 to December 2000. From the exploratory data analysis in Chapter 2, we observed that the monthly means of wind speeds are around 8mph in summer, while they are greater than 10mph in all other seasons. The patterns of wind speeds for all four stations are similar. But the monthly means of wind speeds at station L001 is substantially different from those of the other stations in September 1995 and February 1998, and there are more missing values at station L001 than at other stations. In 1995, the monthly means at station L001 are obviously less than those of other stations. This little difference at station L001 may be caused by various reasons, such as location of the station, device failures or bird interruptions. The wind speeds of the four stations are correlated positively. In an attempt to explain the distribution of the data in a three-parameter Weibull distribution, the goodness of fit tests in Table 2.7 show that the distribution does not fit well. A possible improvement may be to use a lognormal, beta or mixed distribution.

Since there are lots of missing values in this data set, we have to impute the missing values before we detect outliers. In this thesis, we use three imputation methods: Nearby window average imputation, Jones imputation using Kalman filter [13] and EM algorithm imputation [19]. In Chapter 3, we introduce the three methods.

The effects of extraneous objects, measuring device failures and human errors may distort the field data. Usually qualified engineers, scientists or technicians identify abnormalities after inspecting the data manually. This manual process is slow, costly and sometimes inconsistent among inspectors [12]. Various methods, such as artificial intelligence [8], neural network [12] and outlier detection in time series models, have been used for detecting abnormal data. In Chapter 3, we also use

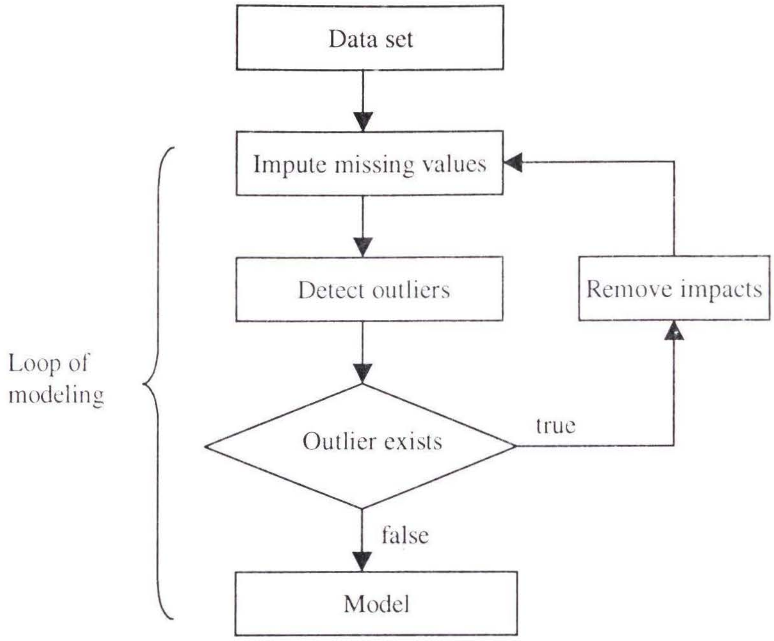


Figure 1.2: Flow Chart of Modeling Process

time series analysis to detect and remove the abnormal data. A common approach to deal with outliers in a time series is to identify the locations, , and the types of outliers and then remove the impacts by using intervention models. Four types of outliers are usually considered: Innovational outlier (IO), additive outlier (AO), level shift (LS) and temporary change (TC) [20]. For the wind speed data, the outliers could be either IO or AO. Hence, only IO and AO are considered in this thesis. We also studied the power of three imputation methods by using a small portion of time series from station L001. Based on the results, we use EM algorithm to impute missing values for the data set used in this thesis.

To get the best model for the wind speed data, the idea is shown in Figure

1.2. After imputing the missing values and removing the impacts of outliers, we can get the best model. This is presented in Chapter 4. Due to the computing problem, the data set used in Chapter 4 is the hourly wind speeds of all four stations from May to August in 2000 only. The best models are seasonal ARIMA$(2,0,0) \times (1,0,0)_{24}$ for all four stations. The term $(2,0,0)$ gives the order of the nonseasonal part of the ARIMA model; the term $(1,0,0)_{24}$ gives the order of the seasonal part. The form of this model is given by

$$(1-\phi_{1,1}B-\phi_{1,2}B^2)(1-\phi_{2,1}B^{24})x_t = \mu+\epsilon_t \quad t=1,\cdots,n,$$

where n is the number of observations in the time series; B is the backshift operator such that $Bx_t = x_{t-1}$; $(1-\phi_{1,1}B-\phi_{1,2}B^2)(1-\phi_{2,1}B^{24})$ is a polynomial of B with all roots outside the unit circle; $\{\epsilon_t\}$ is uncorrelated and identically distributed with mean zero and variance σ^2; The value 24 reflects a daily circle in the hourly wind speed data. Thus, it shows that the wind speed in all stations under study behaves similarly. This suggests that it is not necessary to collect data from all the stations under study.

In an appendix, we include the Matlab codes and SAS programs used for this thesis.

Chapter 2

DATA EXPLORATION

Lake Okeechobee wind speed data are routinely collected by sensors and transcribed from field/laboratory forms to an electronic format. Field data were collected every 15 minutes by the South Florida Water Management District (SFWMD) at a permanent data collection site (Figure 1.1, [11]). Wind speeds (miles per hour) were measured with a Skyvane Wind Sensor Model 2100. Occasionally the effects of extraneous factors such as birds, measuring device failures and human errors, may distort field data [12]. The data set analyzed in this thesis consists of the wind speeds collected at stations L001, L005, L006 and LZ40 (corresponding sample sites 16, 38, 39 and 35, respectively in Figure 1.1) from January 1995 to December 2000.

Table 2.1: Descriptive Statistics for All Stations

Station	N	N Miss	Mean	Std Dev	Min	Max
L001	189408	21024	10.214	5.549	0	72.4
L005	206703	849	10.479	5.404	0	56.26
L006	204162	6270	11.056	5.702	0	49.95
LZ40	203414	7018	11.041	5.722	0	55.68

Table 2.2: Descriptive Statistics for L001

Month	Obs	Miss	Mean	Min	Max	Month	Obs	Miss	Mean	Min	Max
Jan-95	2976	0	8.03	0.44	30.69	Jan-98	2976	0	10.78	0.50	27.18
Feb-95	2688	0	8.16	0.44	30.35	Feb-98	562	2126	17.59	3.53	**40.66**
Mar-95	2976	0	9.44	0.44	27.21	Mar-98	1613	1363	12.31	0.81	25.71
Apr-95	2880	0	9.48	0.44	32.16	Apr-98	2880	0	12.92	0.76	26.90
May-95	2976	0	7.81	0.44	30.78	May-98	2870	106	9.80	0.30	25.80
Jun-95	2839	41	8.98	0.44	31.00	Jun-98	1741	1139	10.88	0.84	26.14
Jul-95	2976	0	7.53	0.44	28.47	Jul-98	1987	989	8.52	0.00	**68.10**
Aug-95	2660	316	9.18	0.44	34.39	Aug-98	1419	1557	8.06	0.00	**58.60**
Sep-95	1578	1302	4.70	0.44	26.01	Sep-98	2612	268	9.91	0.00	39.90
Oct-95	2887	89	10.34	0.44	29.86	Oct-98	2976	0	8.64	0.00	**72.40**
Nov-95	2880	0	10.66	0.44	26.31	Nov-98	2879	1	8.04	0.00	37.45
Dec-95	2976	0	9.51	0.44	25.58	Dec-98	2885	91	9.51	0.40	31.09
Jan-96	2976	0	9.70	0.44	28.64	Jan-99	2976	0	9.61	0.38	30.80
Feb-96	2784	0	9.80	0.44	31.68	Feb-99	2688	0	10.35	0.00	27.45
Mar-96	2976	0	12.34	0.44	32.28	Mar-99	2976	0	10.66	0.65	29.01
Apr-96	2880	0	11.09	0.44	26.74	Apr-99	2880	0	10.47	0.48	34.97
May-96	2915	61	10.55	0.45	31.12	May-99	2976	0	10.26	0.72	29.36
Jun-96	2880	0	9.61	0.46	33.02	Jun-99	2880	0	9.58	0.56	27.64
Jul-96	2976	0	10.14	0.45	30.87	Jul-99	2976	0	8.81	0.83	31.43
Aug-96	2976	0	9.07	0.44	37.12	Aug-99	2974	2	9.40	0.52	28.81
Sep-96	2880	0	9.36	0.44	34.36	Sep-99	2880	0	10.28	0.52	34.00
Oct-96	2976	0	10.66	0.44	29.34	Oct-99	2976	0	12.39	0.55	**55.13**
Nov-96	2880	0	12.21	0.44	27.16	Nov-99	2880	0	11.74	0.52	28.65
Dec-96	1988	988	9.25	0.45	31.68	Dec-99	2976	0	9.76	0.44	24.68
Jan-97	2976	0	9.89	0.50	35.67	Jan-00	2976	0	9.90	0.42	32.69
Feb-97	2688	0	10.79	0.31	27.82	Feb-00	2784	0	9.78	0.42	34.35
Mar-97	2976	0	12.25	0.98	32.98	Mar-00	2976	0	11.57	0.88	29.36
Apr-97	2880	0	13.72	1.08	**63.59**	Apr-00	2879	1	12.35	1.06	25.83
May-97	376	2600	9.67	0.81	24.70	May-00	2975	1	11.40	0.83	25.81
Jun-97	800	2080	11.46	0.74	22.27	Jun-00	2880	0	10.49	0.75	35.24
Jul-97	0	2976	.	.	.	Jul-00	2976	0	10.11	0.70	34.89
Aug-97	1981	995	13.31	0.84	29.53	Aug-00	2976	0	9.68	0.66	31.13
Sep-97	1116	1764	8.93	0.40	21.40	Sep-00	2880	0	9.63	0.57	30.57
Oct-97	2973	3	10.48	0.51	33.95	Oct-00	2976	0	11.81	0.77	29.86
Nov-97	2715	165	10.25	0.64	30.42	Nov-00	2880	0	10.86	0.49	26.55
Dec-97	2976	0	10.60	0.44	31.91	Dec-00	2976	0	11.57	0.66	35.09

"." represents the missing value

2.1 Station L001

Station L001 is in the north of Lake Okeechobee (Figure 1.1). Table 2.2 and Figure 2.1 show that the monthly means are between 8mph and 13mph except September 1995, July 1997 and February 1998. The monthly mean of wind speeds in July 1997 is missing because the wind speeds are all missing. Table 2.2 shows that there are 6 large values of maximum (bold) wind speed (wind speed that is 40mph or

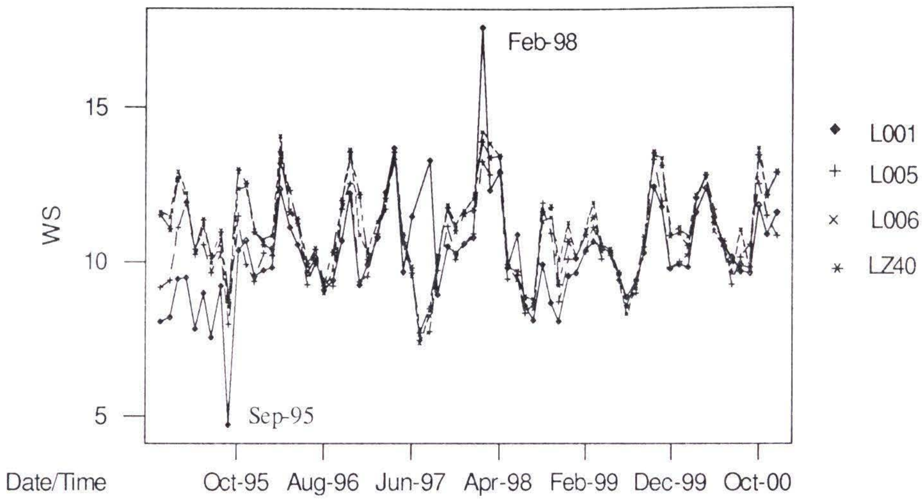


Figure 2.1: Plots of Monthly Mean Values for All Stations

above is considered as a large value). There are 21024 missing values at station L001 (Table 2.1). The number of missing values on September 1995 and February 1998 are 1302 and 2126 (Table 2.2) with missing rates 45.2% and 79%, respectively. It is possible that the large number of missing values caused the unusual monthly means: 4.70mph and 17.59mph for September 1995 and February 1998, respectively. There are more missing values in 1997 and 1998 than in the other years on this station.

2.2 Station L005

Station L005 is in the west of Lake Okeechobee (Figure 1.1). Table 2.3 and Figure 2.1 show that all the monthly means are between 8mph and 14mph. There are 4 large values of maximum wind speeds which are greater than 40mph. There are only 96 observed values on December 2000. The total number of missing values

Table 2.3: Descriptive Statistics for L005

Month	Obs	Miss	Mean	Min	Max	Month	Obs	Miss	Mean	Min	Max
Jan-95	2976	0	9.17	0.44	29.55	Jan-98	2976	0	10.92	0.44	27.84
Feb-95	2688	0	9.37	0.44	35.39	Feb-98	2688	0	13.23	0.44	37.54
Mar-95	2976	0	11.08	0.44	25.87	Mar-98	2976	0	12.80	0.44	29.72
Apr-95	2880	0	11.96	0.44	27.75	Apr-98	2880	0	12.85	0.44	26.54
May-95	2976	0	10.23	0.45	34.33	May-98	2976	0	9.45	0.44	32.58
Jun-95	2880	0	10.53	0.44	30.96	Jun-98	2880	0	9.56	0.44	33.98
Jul-95	2976	0	9.61	0.44	35.37	Jul-98	2976	0	8.29	0.44	32.16
Aug-95	2938	38	10.33	0.44	33.08	Aug-98	2976	0	8.64	0.44	34.11
Sep-95	2880	0	7.92	0.44	27.10	Sep-98	2880	0	11.91	0.44	33.22
Oct-95	2908	68	11.47	0.44	29.00	Oct-98	2976	0	10.92	0.44	25.81
Nov-95	2839	41	9.92	0.44	23.66	Nov-98	2833	47	8.68	0.44	35.93
Dec-95	2929	47	9.35	0.44	26.72	Dec-98	2836	140	10.09	0.00	29.14
Jan-96	2976	0	10.24	0.44	31.88	Jan-99	2976	0	10.07	0.00	25.98
Feb-96	2784	0	10.19	0.44	33.71	Feb-99	2688	0	10.33	0.00	27.04
Mar-96	2976	0	13.21	0.44	31.06	Mar-99	2976	0	11.11	0.00	31.44
Apr-96	2413	467	12.31	0.44	26.09	Apr-99	2880	0	10.10	0.01	32.62
May-96	2976	0	11.24	0.44	31.38	May-99	2976	0	10.16	0.04	35.94
Jun-96	2880	0	9.25	0.44	30.27	Jun-99	2880	0	9.40	0.00	33.78
Jul-96	2976	0	9.92	0.44	32.81	Jul-99	2976	0	8.82	0.01	26.81
Aug-96	2976	0	9.38	0.44	28.30	Aug-99	2976	0	8.94	0.00	32.31
Sep-96	2880	0	9.17	0.44	25.58	Sep-99	2880	0	10.43	0.00	34.62
Oct-96	2976	0	11.71	0.44	28.82	Oct-99	2976	0	13.30	0.00	**45.91**
Nov-96	2880	0	12.56	0.44	29.07	Nov-99	2880	0	11.91	0.00	27.51
Dec-96	2976	0	9.33	0.44	27.30	Dec-99	2976	0	9.79	0.00	22.82
Jan-97	2976	0	9.52	0.44	33.18	Jan-00	2976	0	9.96	0.00	31.62
Feb-97	2688	0	10.93	0.44	25.02	Feb-00	2784	0	10.20	0.00	23.79
Mar-97	2976	0	11.72	0.44	30.76	Mar-00	2976	0	12.02	0.00	28.35
Apr-97	2880	0	13.31	0.44	34.61	Apr-00	2880	0	12.75	0.03	26.73
May-97	2976	0	10.62	0.44	**45.39**	May-00	2975	1	11.65	0.00	25.23
Jun-97	2880	0	9.62	0.44	27.66	Jun-00	2880	0	10.44	0.00	32.79
Jul-97	2976	0	7.44	0.44	31.79	Jul-00	2976	0	9.25	0.00	34.80
Aug-97	2976	0	7.69	0.44	29.21	Aug-00	2976	0	10.13	0.00	26.11
Sep-97	2880	0	9.73	0.44	27.11	Sep-00	2880	0	10.55	0.67	34.90
Oct-97	2976	0	11.17	0.44	27.56	Oct-00	2976	0	12.51	0.00	**52.13**
Nov-97	2880	0	10.09	0.44	24.60	Nov-00	2880	0	11.49	0.00	**56.26**
Dec-97	2976	0	10.57	0.44	30.81	Dec-00	96	0	10.80	3.21	17.09

is 849 (Table 2.1). There are many more missing values (467) on April 1996 than on any other months.

2.3 Station L006

Station L006 is in the south of Lake Okeechobee (Figure 1.1). Table 2.4 shows that the monthly means are between 8mph and 14mph. Figure 2.1 shows that there is no extreme monthly mean value. Table 2.4 shows that four values of

Table 2.4: Descriptive Statistics for L006

Month	Obs	Miss	Mean	Min	Max	Month	Obs	Miss	Mean	Min	Max
Jan-95	2976	0	11.62	0.44	31.74	Jan-98	2976	0	12.03	0.44	31.98
Feb-95	2466	222	11.44	0.44	36.61	Feb-98	2688	0	14.22	0.44	37.89
Mar-95	2760	216	12.93	0.45	32.07	Mar-98	2976	0	13.84	0.44	33.61
Apr-95	2741	139	12.23	0.44	34.59	Apr-98	2880	0	13.42	0.44	29.78
May-95	2976	0	10.27	0.44	30.37	May-98	2976	0	9.97	0.44	36.04
Jun-95	2878	2	11.20	0.44	33.78	Jun-98	2880	0	9.40	0.44	33.12
Jul-95	2976	0	9.94	0.44	32.72	Jul-98	2976	0	8.49	0.44	34.22
Aug-95	2976	0	10.66	0.44	34.83	Aug-98	2976	0	8.44	0.44	34.98
Sep-95	2880	0	8.49	0.44	28.63	Sep-98	2880	0	11.27	0.44	28.13
Oct-95	2976	0	12.37	0.44	27.45	Oct-98	2976	0	11.42	0.44	31.28
Nov-95	2843	37	12.43	0.44	29.15	Nov-98	1782	1098	10.03	0.00	**40.89**
Dec-95	2976	0	10.92	0.44	29.03	Dec-98	906	2070	11.24	0.00	32.72
Jan-96	2976	0	10.70	0.44	35.49	Jan-99	2574	402	10.09	0.00	30.38
Feb-96	2784	0	10.84	0.44	39.36	Feb-99	2688	0	11.03	0.00	31.66
Mar-96	2976	0	14.08	0.44	37.99	Mar-99	2976	0	11.91	0.00	29.24
Apr-96	2880	0	11.66	0.44	27.93	Apr-99	2880	0	10.56	0.00	35.93
May-96	2976	0	11.18	0.45	36.88	May-99	2976	0	10.39	0.00	**40.29**
Jun-96	2880	0	9.83	0.44	36.27	Jun-99	2880	0	9.38	0.00	31.72
Jul-96	2976	0	10.21	0.44	38.34	Jul-99	2976	0	8.27	0.00	30.35
Aug-96	2976	0	8.97	0.44	29.32	Aug-99	2976	0	9.08	0.00	31.51
Sep-96	2880	0	9.72	0.44	25.47	Sep-99	2880	0	10.61	0.00	39.56
Oct-96	2976	0	12.03	0.44	33.43	Oct-99	2976	0	13.50	0.00	49.95
Nov-96	2880	0	13.65	0.44	31.67	Nov-99	2880	0	13.31	0.00	32.22
Dec-96	2976	0	10.85	0.44	31.31	Dec-99	2976	0	10.87	0.00	29.79
Jan-97	2976	0	10.28	0.44	31.64	Jan-00	2976	0	11.12	0.00	32.19
Feb-97	2688	0	11.27	0.44	31.76	Feb-00	2784	0	10.79	0.00	24.79
Mar-97	2976	0	12.00	0.44	31.39	Mar-00	2976	0	12.11	0.00	26.40
Apr-97	2880	0	13.53	0.44	30.88	Apr-00	2880	0	12.79	0.00	31.77
May-97	2976	0	10.85	0.44	26.19	May-00	2973	3	10.96	0.00	25.32
Jun-97	2880	0	9.54	0.44	**41.34**	Jun-00	2880	0	10.48	0.00	36.53
Jul-97	2976	0	7.34	0.44	29.60	Jul-00	2652	324	9.61	0.00	31.75
Aug-97	2976	0	8.21	0.44	33.75	Aug-00	1219	1757	11.00	0.00	**42.00**
Sep-97	2880	0	10.05	0.44	28.04	Sep-00	2880	0	9.81	0.47	30.09
Oct-97	2976	0	11.82	0.44	26.02	Oct-00	2976	0	13.65	0.00	36.12
Nov-97	2880	0	11.21	0.44	32.94	Nov-00	2880	0	12.14	0.32	34.93
Dec-97	2976	0	11.68	0.44	36.34	Dec-00	2976	0	12.86	0.60	36.42

maximum wind speed are over 40mph. There are 6270 missing values at this station (Table 2.1). The number of missing values on November 1998, December 1998 and August 2000 is more than 1000 (Table 2.4).

2.4 Station LZ40

Station LZ40 is in the middle of Lake Okeechobee (Figure 1.1). Table 2.5 shows that the monthly means are between 8mph and 14mph. Figure 2.1 shows

Table 2.5: Descriptive Statistics for LZ40

Month	Obs	Miss	Mean	Min	Max	Month	Obs	Miss	Mean	Min	Max
Jan-95	2976	0	11.50	0.44	33.86	Jan-98	2976	0	11.67	0.50	30.80
Feb-95	2688	0	11.03	0.44	34.67	Feb-98	2688	0	13.88	0.46	37.67
Mar-95	2699	277	12.68	0.44	31.33	Mar-98	2976	0	13.38	0.50	33.64
Apr-95	1815	1065	11.90	0.45	24.87	Apr-98	2880	0	13.43	0.44	29.65
May-95	2976	0	10.40	0.46	25.64	May-98	2976	0	9.90	0.44	38.95
Jun-95	2880	0	11.32	0.45	34.91	Jun-98	2880	0	9.66	0.45	32.36
Jul-95	2976	0	10.17	0.49	33.14	Jul-98	2976	0	8.82	0.45	31.35
Aug-95	2976	0	10.98	0.44	36.67	Aug-98	2976	0	8.75	0.46	33.99
Sep-95	2880	0	8.71	0.46	30.04	Sep-98	2880	0	11.59	0.44	29.58
Oct-95	1730	1246	12.95	0.45	26.60	Oct-98	2530	446	11.75	0.45	30.73
Nov-95	2403	477	12.55	0.44	27.92	Nov-98	2268	612	9.25	0.00	**40.59**
Dec-95	2939	37	10.97	0.44	27.97	Dec-98	2976	0	10.65	0.41	31.39
Jan-96	2976	0	10.56	0.44	32.46	Jan-99	2976	0	10.12	0.00	28.86
Feb-96	2784	0	10.41	0.44	36.86	Feb-99	2688	0	10.89	0.00	30.91
Mar-96	2976	0	13.55	0.45	37.54	Mar-99	2976	0	11.42	0.01	29.33
Apr-96	2880	0	11.62	0.44	27.44	Apr-99	2880	0	10.37	0.03	35.60
May-96	2976	0	11.36	0.45	37.80	May-99	2976	0	10.35	0.10	35.00
Jun-96	2880	0	9.96	0.49	30.74	Jun-99	2880	0	9.64	0.01	29.53
Jul-96	2976	0	10.41	0.45	**40.75**	Jul-99	2976	0	8.53	0.01	29.44
Aug-96	1786	1190	9.21	0.44	30.20	Aug-99	2976	0	9.34	0.01	28.84
Sep-96	2263	617	10.30	0.44	30.08	Sep-99	2880	0	10.76	0.01	37.79
Oct-96	2976	0	11.98	0.47	34.28	Oct-99	2976	0	13.50	0.22	**55.68**
Nov-96	2880	0	13.55	0.44	31.47	Nov-99	2880	0	13.10	0.42	31.54
Dec-96	2121	855	12.19	0.44	30.95	Dec-99	2927	49	10.84	0.00	29.02
Jan-97	2831	145	10.07	0.44	31.90	Jan-00	2976	0	10.91	0.01	32.78
Feb-97	2688	0	10.84	0.44	27.59	Feb-00	2784	0	10.48	0.00	25.05
Mar-97	2976	0	12.06	0.59	31.44	Mar-00	2976	0	12.09	0.03	30.26
Apr-97	2880	0	13.56	0.44	32.61	Apr-00	2880	0	12.78	0.13	30.14
May-97	2976	0	10.86	0.45	**41.42**	May-00	2975	1	11.23	0.02	25.66
Jun-97	2880	0	9.81	0.45	**40.98**	Jun-00	2880	0	10.65	0.12	36.52
Jul-97	2976	0	7.68	0.45	33.52	Jul-00	2976	0	10.00	0.02	32.41
Aug-97	2976	0	8.47	0.44	30.61	Aug-00	2976	0	9.89	0.26	30.02
Sep-97	2880	0	10.15	0.44	33.25	Sep-00	2880	0	9.80	0.00	29.56
Oct-97	2976	0	11.74	0.47	26.95	Oct-00	2976	0	13.41	0.68	34.32
Nov-97	2880	0	11.00	0.44	25.95	Nov-00	2879	1	12.13	0.00	**49.30**
Dec-97	2976	0	11.56	0.44	36.59	Dec-00	2976	0	12.88	0.00	**42.60**

that there is no extreme monthly mean value. There are 7 values of maximum wind speed which are over 40mph (Table 2.5). The total number of missing values is 7018 (Table 2.1). The number of missing values in April and October 1995, and August 1996 is more than 1000 (Table 2.5).

2.5 Using a Weibull Distribution to Describe Wind Speed

The Weibull distribution is usually used to describe wind speeds and study wind power. It is very practical for this application, because the distribution does not allow for negative values and it is easy to appropriately consider the fact that on most days there will be a bit of wind and on some days a lot.

The three-parameter Weibull distribution has probability density function given by

$$f(y) = \frac{c}{\sigma}(\frac{y-\theta}{\sigma})^{c-1}\exp(-(\frac{y-\theta}{\sigma})^c) \quad \text{for } y > \theta,\ c > 0,\ \sigma > 0,$$

where θ is the threshold parameter, σ is the scale parameter and c is the shape parameter [9]. The cumulative distribution function is given by

$$F(y) = 1 - \exp(-(\frac{y-\theta}{\sigma})^c) \quad \text{for } y > \theta.$$

The mean and variance are given by

$$E(y) = \theta + \sigma\Gamma(1 + \frac{1}{c})$$

and

$$Var(y) = \sigma^2\left[\Gamma(1 + \frac{2}{c}) - \Gamma^2(1 + \frac{1}{c})\right],$$

where Γ is the gamma function. The mean wind speed is used to indicate how windy the site is. The shape parameter tells how peaked the distribution is; i.e., if the wind speeds always tend to be very close to a certain value, the distribution will have a high shape parameter value and will be very peaked.

Table 2.6: Exploratory Data Analysis

Station	Mean	SD	Skew	Kurtosis	Q1	Q2	Q3	max
L001	10.214	5.549	0.420	0.731	6.358	10.050	13.740	72.400
L005	10.479	5.404	0.513	0.576	6.577	10.050	14.010	56.260
L006	11.056	5.702	0.611	0.543	6.960	10.540	14.570	49.950
LZ40	11.041	5.722	0.629	0.609	6.831	10.536	14.590	55.680

To check if a Weibull distribution fits a data set well, we use the Anderson-Darling test [18]. The hypotheses of the test are:

H_0: the data follow Weibull distribution,

H_a: the data do not follow Weibull distribution.

The test statistic is given by

$$A^2 = -n - S,$$

where $S = \sum_{i=1}^{n}(\ln F(y_i) + \ln(1 - F(y_{n+1-i})))$, n is sample size, y_i are ordered and F is the cumulative distribution function.

The descriptive statistics for the 15-minute wind speed data for the four stations are shown in Table 2.6. We can see that more than 75% of the wind speed data are below 15mph. The means and standard deviations of Stations L006 and LZ40 are about the same and greater than those of Stations L001 and L005. The histograms in Figure 2.2 show that the distributions are skewed to the right. The maximum likelihood estimations of the parameters of the Weibull distribution are reported in Table 2.7 for all four stations. The p-values of Anderson-Darling

Table 2.7: Weibull Distribution Parameters and Goodness-of-Fit Tests

Station	Threshold (θ)	Scale (σ)	Shape (c)	Mean	SD	Anderson-Darling test statistic	p-value
L001	-1.864	13.614	2.297	10.197	5.568	448.981	<0.001
L005	-0.889	12.822	2.211	10.467	5.424	106.016	<0.001
L006	-0.617	13.170	2.146	11.046	5.723	82.962	<0.001
LZ40	-0.241	12.727	2.061	11.033	5.737	54.885	<0.001

goodness-fit-test are all less than 0.001. This means that a three-parameter Weibull distribution does not fit our wind speed data well. A possible suggestion will be to use a lognormal, beta or mixed distribution.

2.6 Conclusion

Comparing the number of missing values in Table 2.1, there are much more missing values at station L001 than at any other station. Comparing plots of monthly means of wind speeds for all four stations (Figure 2.1), we can see that the patterns of the plots for station L005, L006 and LZ40 are similar. Hence we further check the correlations of wind speeds among these four stations. The Pearson product moment correlation coefficient of two variables is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y},$$

where n is the number of observations, $\bar{x}$ and $\bar{y}$ are sample means of two variables, and s_x and s_y are sample standard deviations of two variables, respectively. Table 2.8 shows Pearson correlations of wind speeds among the stations in 2000. All correlation coefficients are greater than 0.6, and the p-values for the hypothesis

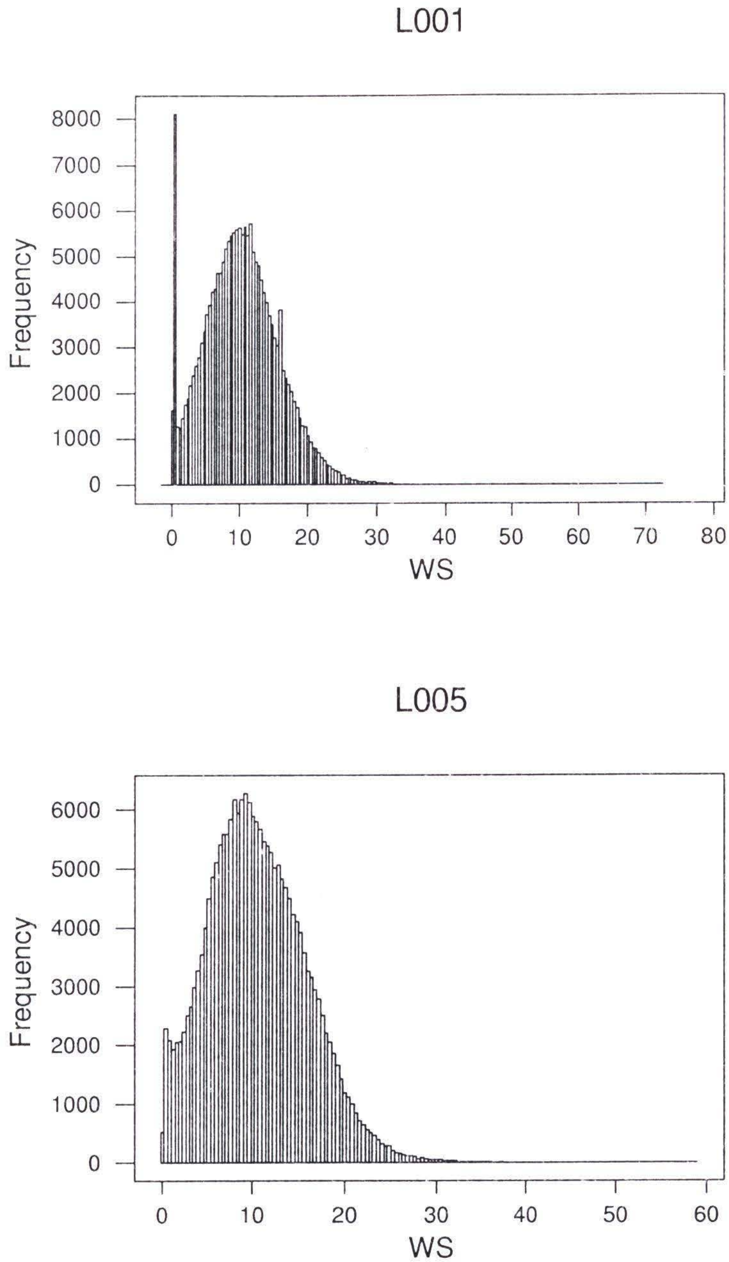


Figure 2.2: Histograms of Wind Speeds

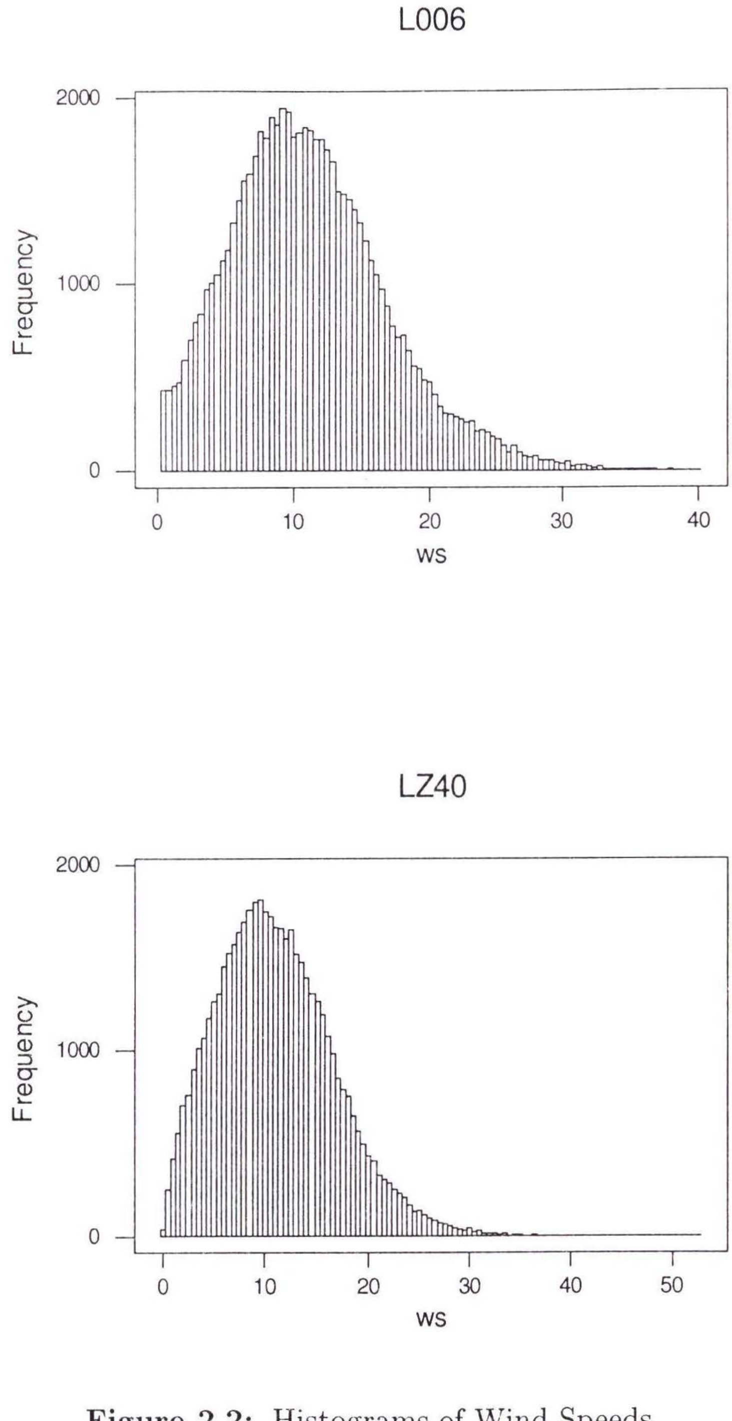


Figure 2.2: Histograms of Wind Speeds

Table 2.8: Correlations among Four Stations

Station	L001	L005	L006
L005	0.668 (<0.001)		
L006	0.748 (<0.001)	0.678 (<0.001)	
LZ40	0.758 (<0.001)	0.697 (<0.001)	0.897 (<0.001)

Note: values in parentheses are p-values

tests of the correlation coefficients being zero are less than 0.001. Therefore, the wind speeds of the four stations are correlated positively.

The monthly means of wind speeds at station L001 is substantially different from those of any other stations on September 1995 and February 1998, and there are more missing values at station L001 than at other stations. In 1995, the monthly means at station L001 are obviously less than at any other station. This little difference at station L001 may be caused by various reasons such as location of the station, measuring device failures or bird interruptions. Further detection is needed. To check the large values of maximum wind speeds (i.e. gerater than 40mph) for all stations, we compared the maximum wind speeds at all stations for the months that have large values. Table 2.9 shows that there are large values on October 1999 at all stations. There was a hurricane named Floyd on October 1999. At station L001, the maximum wind speeds on April 1997, July 1998, August 1998 and October 1998 are obviously much higher than those at the other stations. In Figure 2.3, one can

Table 2.9: Large Values for All Stations

Time	L001	L005	L006	LZ40
07/96	30.87	32.81	38.34	*40.75
04/97	*63.59	34.61	30.88	32.61
05/97	24.7	*45.39	26.19	*41.42
06/97	22.27	27.66	*41.34	*40.98
02/98	*40.66	37.54	37.89	37.67
07/98	*68.1	32.16	34.22	31.35
08/98	*58.6	34.11	34.98	33.99
10/98	*72.4	25.81	31.28	30.73
11/98	37.45	35.93	*40.89	*40.59
05/99	29.36	35.94	*40.29	35
10/99	***55.13**	***45.91**	***49.95**	***55.68**
08/00	31.13	26.11	*42	30.02
10/00	29.86	*52.13	36.12	34.32
11/00	26.55	*56.26	34.93	*49.3
12/00	35.09	17.09	36.42	*42.6

"*" represents that the value is unusually large

see that those four points ($63.59, 68.1, 58.6$ and 72.4) are outlier points, which might be affected by local climate or extraneous factors.

Finaly, we conclude that there are outliers and many missing values in the data sets. The patterns of wind speeds for all four stations are similar and the wind speeds of these four stations are correlated positively. We also observed that for the four stations the monthly means of wind speeds are around 8mph in summer while they are greater than 10mph in all other seasons. The monthly means of wind speeds at station L001 are substantially different from those of the other stations on September 1995, February 1998 and in 1995. There are more missing values at station L001 than at other stations. A three-parameter Weibull distribution does

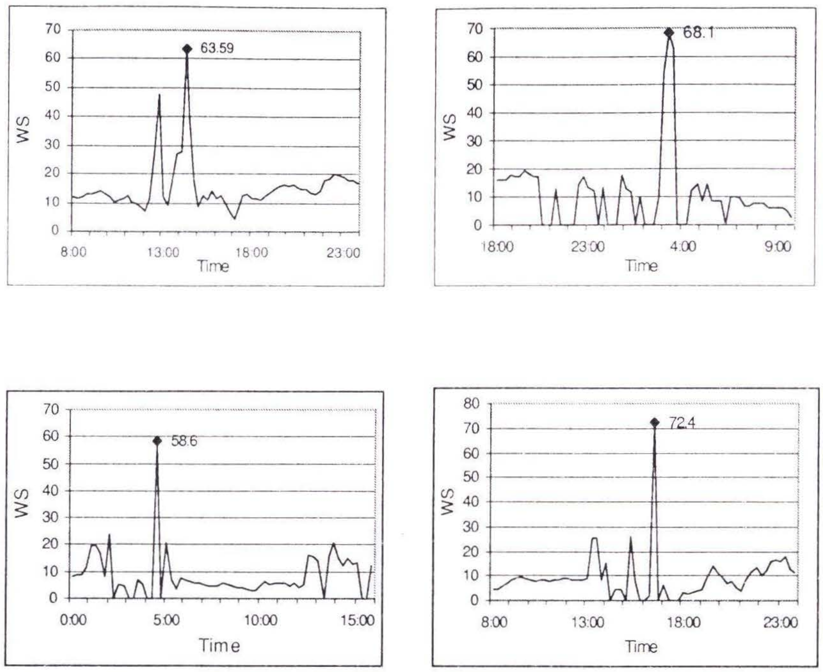


Figure 2.3: Plots of Possible Outliers at Station L001

not fit this data well, as is seen by checking the goodness of fit tests in Table 2.7. A possible improvement may be to use a lognormal, beta or mixed distribution.

Chapter 3

MISSING VALUE IMPUTATION AND OUTLIER DETECTION

As pointed out in Chapter 2, there are many missing values and outliers in the wind speed data. Let x_t be the true time series, y_t be the observed series with missing values and outliers, and z_t be the observed series with outliers after imputation. Thus, we impute the missing values and then detect outliers. In this thesis, three imputation methods are used: Nearby window average imputation, Jones imputation using Kalman filter [13] and EM algorithm imputation [19]. Two types of outliers are considered in this thesis: Innovational outlier (IO) and additive outlier (AO) [20].

3.1 Missing Value Imputation

Nearby window average imputation, Jones imputation using Kalman filter and EM algorithm imputation are used to impute missing values. The idea of the Nearby window method is to use the average value of one value before the missing value begins and one value after the missing value ends to impute the missing values. The other two methods are introduced in the following.

3.1.1 Imputation Using Kalman Filter

Richard H. Jones considered a state-space model using Kalman recursive estimation for time series data with missing values in 1980. Here we only introduce state-space model and Kalman filter (see [13] for details).

3.1.1.1 State-space Model and Kalman Recursive Estimation

A State-space model has two equations: the observation equation and the state equation. Let y_t be an observed time series. Then the observation equation is given by

$$y_t = H\theta_t + \nu_t, \tag{3.1}$$

where H is a $(1\times m)$ vector, θ_t is a $(m\times 1)$ state vector, and ν_t denotes the observation error. The ν_t's are assumed to be uncorrelated and identically distributed with mean zero and variance R. Although the state vector θ_t is unobservable, we can assume that it follows the state equation

$$\theta_t = G\theta_{t-1} + w_t, \tag{3.2}$$

where G is assumed to be a known $(m\times m)$ matrix. The term w_t denotes a vector of deviates, which is white noise with zero mean vector and known variance-covariance matrix Q, and is assumed to be uncorrelated with ν_t.

Assuming that the best unbiased estimator for θ_{t-1} is $\hat{\theta}_{t-1}$ based on our knowledge about the process prior to time $t-1$, the variance-covariance matrix of $\hat{\theta}_{t-1}$ is P_{t-1}. Let $\hat{\theta}_{t|t-1}$ be the one-step ahead forecast of θ_t from time $t-1$, i.e.

$$\hat{\theta}_{t|t-1} = G\hat{\theta}_{t-1}. \tag{3.3}$$

Then the estimation error is

$$\begin{aligned} e_{t|t-1} &= \theta_t - \hat{\theta}_{t|t-1} \\ &= G\theta_{t-1} + w_t - G\hat{\theta}_{t-1} \\ &= G(\theta_{t-1} - \hat{\theta}_{t-1}) + w_t \end{aligned}$$

and the associated error covariance matrix is

$$\begin{aligned} P_{t|t-1} &= E\left[e_{t|t-1}e'_{t|t-1}\right] && (3.4) \\ &= GVar(\hat{\theta}_{t-1})G' + Var(w_t) && (3.5) \\ &= GP_{t-1}G' + Q, && (3.6) \end{aligned}$$

where $e'_{t|t-1}$ is the transposition of $e_{t|t-1}$. If y_t is available, then we may use the observed y_t to improve the estimate of θ_t. Let $\hat{\theta}_t$ be the updated estimate of θ_t satisfying the following equation:

$$\hat{\theta}_t = \hat{\theta}_{t|t-1} + K_t(y_t - H\hat{\theta}_{t|t-1}), \tag{3.7}$$

where K_t is called the Kalman gain [21]. The reason for constructing this $\hat{\theta}_t$ is to minimize the variance of the prediction error $e_t \equiv \theta_t - \hat{\theta}_t$. To derive K_t we use the minimum mean-square error criterion [3]. From (3.1) and (3.7), the error covariance matrix associated with the updated estimate is

$$\begin{aligned}
P_t &\equiv E\left[e_t e_t'\right] = E\left[(\theta_t - \hat{\theta}_t)(\theta_t - \hat{\theta}_t)'\right] \\
&= E\left[(\theta_t - \hat{\theta}_{t|t-1} - K_t(H\theta_t + \nu_t - H\hat{\theta}_{t|t-1}))(\theta_t - \hat{\theta}_{t|t-1} - K_t(H\theta_t + \nu_t - H\hat{\theta}_{t|t-1}))'\right] \\
&= E\left[(e_{t|t-1} - K_t(He_{t|t-1} + \nu_t))(e_{t|t-1} - K_t(He_{t|t-1} + \nu_t))'\right] \\
&= E[e_{t|t-1}e_{t|t-1}' - e_{t|t-1}e_{t|t-1}'H'K_t' - e_{t|t-1}\nu_t'K_t' \\
&\quad -K_tHe_{t|t-1}e_{t|t-1}' + K_tHe_{t|t-1}e_{t|t-1}'H'K_t' + K_tHe_{t|t-1}\nu_t'K_t' \\
&\quad -K_t\nu_te_{t|t-1}' + K_t\nu_te_{t|t-1}'H'K_t' + K_t\nu_t\nu_t'K_t'] \\
&= P_{t|t-1} - P_{t|t-1}H'K_t' - K_tHP_{t|t-1} + K_tHP_{t|t-1}H'K_t' + K_tRK_t'
\end{aligned}$$

Rewrite the error covariance matrix associated with the updated estimate in the form:

$$P_t = P_{t|t-1} - P_{t|t-1}H'K_t' - K_tHP_{t|t-1} + K_t(HP_{t|t-1}H' + R)K_t'. \tag{3.8}$$

Differentiate the trace of P_t with respect to K_t. By the facts that

$$\frac{d\left[trace(AB)\right]}{dA} = B' \text{ (AB must be square)},$$
$$\frac{d\left[trace(ACA')\right]}{dA} = 2AC \text{ (C must be square)}$$

and

$$trace(P_{t|t-1}H'K_t') = trace(K_tHP_{t|t-1}),$$

we have

$$\frac{d(traceP_t)}{dK_t} = -2(HP_{t|t-1})' + 2K_t(HP_{t|t-1}H' + R). \tag{3.9}$$

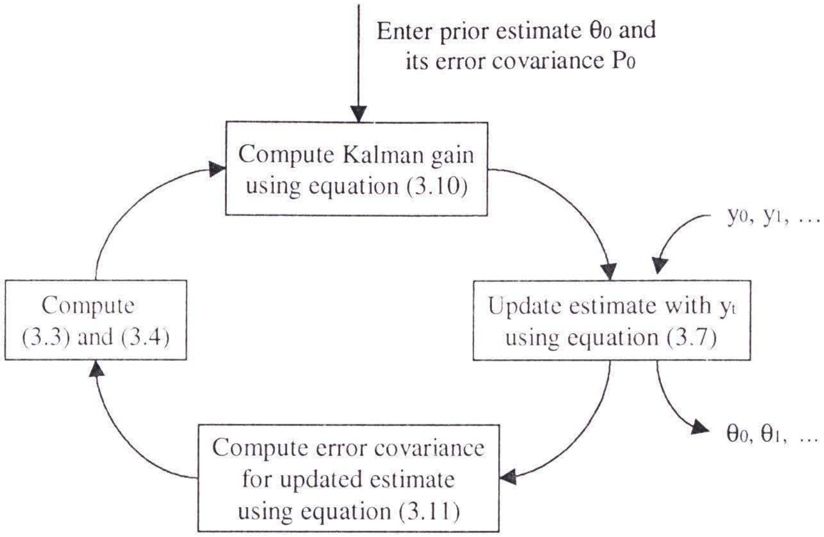


Figure 3.1: Kalman Recursive Process

Setting (3.9) to be zero and solving for K_t, we get

$$K_t = P_{t|t-1}H'(HP_{t|t-1}H' + R)^{-1}. \tag{3.10}$$

From (3.8) and (3.10), we have

$$P_t = P_{t|t-1} - K_tHP_{t|t-1}. \tag{3.11}$$

Equations (3.3), (3.4), (3.7), (3.10) and (3.11) are the Kalman filter recursive equations. The Kalman recursive process is shown in Figure 3.1.

3.1.1.2 State-space Model Representations of ARMA and ARIMA Models

Let x_t be a time series following an autoregressive-moving average (ARMA) model with order (p, q), i.e.,

$$\phi(B)x_t = \psi(B)\epsilon_t \quad t = 1, \cdots, n, \tag{3.12}$$

where n is the number of observations in the time series; B is the backshift operator such that $Bx_t = x_{t-1}$; $\phi(B) = 1-\phi_1 B - \cdots - \phi_p B^p$ and $\psi(B) = 1-\psi_1 B - \cdots - \psi_q B^q$ are polynomials of B with all roots outside the unit circle; $\{\epsilon_t\}$ is white noise with mean zero and variance σ^2. Let y_t still be an observed time series. Define the state vector of this process as

$$\theta_t = \begin{bmatrix} x(t|t) \\ x(t+1|t) \\ \vdots \\ x(t+m-1|t) \end{bmatrix},$$

where $m = \max(p, q+1)$, $x(t|t) = x_t$ and $x(t+1|t)$ is the projection of x_{t+j} on the values of the times series up to time t. Then the observation equation is

$$y_t = [1 \; 0 \; \cdots \; 0]\theta_t + \nu_t, \tag{3.13}$$

where ν_t is the observational error, uncorrelated at different times and uncorrelated with the ϵ's. The mean of ν_t is 0 and its variance is $R = E[\nu_t]^2$. The state equation is

$$\theta_{t+1} = G\theta_t + A\epsilon_{t+1}, \tag{3.14}$$

where

$$G = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ & & \vdots & & \\ \phi_m & \phi_{m-1} & \cdots & \phi_2 & \phi_1 \end{bmatrix};$$

$\phi_i = 0$ for $i > p$; $A = [1, a_2, \cdots, a_m]'$; $a_1 = 1$, $a_j = -\psi_{j-1} + \sum_{i=1}^{j-1} \phi_i a_{j-i}$ for $j > 1$ and $\psi_j = 0$ for $j > q$ (See [13] for details).

For the ARMA model, the likelihood for n observations of the zero mean process is

$$L = \prod_{t=1}^{n} (2\pi V_t)^{-\frac{1}{2}} \exp(-\frac{\tilde{y}_t^2}{2V_t}), \tag{3.15}$$

where $\tilde{y}_t = y_t - x(t|t-1)$ and $V_t = P_{t|t-1} + R$ [13]. Dropping the constant 2π, we get

$$l = -2\ln L = \sum_{t=1}^{n} \left[\frac{\tilde{y}_t^2}{V_t} + \ln V_t \right]. \tag{3.16}$$

From (3.14) we have

$$P_{t|t-1} = GP_{t-1}G' + \sigma^2 AA'.$$

Hence, the variance σ^2 can be removed from the nonlinear estimation problem by dividing R by σ^2. The observational error variance is then replaced by the ratio of the observational error variance to σ^2. In the recursions, since all variances have the same scale factor, $P_{t|t-1}$ and $P_{t|t}$ are replaced by $\sigma^2 P_{t|t-1}$ and $\sigma^2 P_{t|t}$, respectively, and the likelihood becomes

$$l = -2\ln L = \sum_{t=1}^{n}\left[\frac{\tilde{y}_t^2}{\sigma^2 V_t} + \ln(\sigma^2 V_t)\right]. \tag{3.17}$$

Differentiating this with respect to σ^2 and equating it to zero gives

$$\sigma^2 = \frac{1}{n}\sum_{t=1}^{n}\frac{\tilde{y}_t^2}{V_t}. \tag{3.18}$$

Then substituting into (3.17) and dropping the constants gives

$$l = n\ln\sum_{t=1}^{n}\frac{\tilde{y}_t^2}{V_t} + \sum_{t=1}^{n}\ln V_t \tag{3.19}$$

the function to be minimized with respect to the remaining parameters $\phi_1, \cdots, \phi_p$, $\psi_1, \cdots, \psi_q$, d and R.

Jones uses a vector of zeros as initial state vector θ_0, as well as the Akaike method to calculate the initial state covariance matrix $P_0 = P_{0|0}$ (see [13] for details). If an observation y_{t+1} is missing, σ^2 in (3.17) through (3.19) is set to 1 and estimated later. Equations (3.7) and (3.11) are replaced by

$$\hat{\theta}_{t+1|t+1} = \hat{\theta}_{t+1|t},$$

and

$$P_{t+1|t+1} = P_{t+1|t},$$

respectively. The corresponding term in (3.19) for the accumulation of $-2\ln$ likelihood is skipped. If a large block of data is missing, the recursion is equivalent to restarting the recursion at the other end.

Let x_t be a time series following an (ARIMA) model with order (p, d, q), i.e.,

$$\phi(B)\alpha(B)x_t = \psi(B)\epsilon_t, \quad t = 1, \cdots, n, \tag{3.20}$$

where n is the number of observations in the time series; B is the backshift operator such that $Bx_t = x_{t-1}$; $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$; $\psi(B) = 1 - \psi_1 B - \cdots - \psi_q B^q$ are polynomials of B with all roots outside the unit circle; and $\alpha(B) = (1-B)^d$ with all roots of $\alpha(B)$ on the unit circle. Also, $\{\epsilon_t\}$ is white noise with mean zero and variance σ^2. Notice that $\alpha(B)x_t$ satisfy ARMA(p, q). So we can use a state-space representation for the ARMA model to solve the state-space model for ARIMA model.

3.1.2 EM Algorithm

The EM algorithm is a general iterative algorithm for ML estimation in an incomplete data problem [19]. It consists of an Expectation step followed by a Maximization step. The idea is to fill in the missing data X_{miss} based on an initial estimate of the parameter θ, re-estimate θ based on X_{obs} and the filled-in X_{miss}, and iterate until the estimates converge. The specific applications of this idea have appeared in the statistical literature, and go as far back as 1926 [14]. The term EM was introduced by Dempster, Laird and Rubin [7] in 1977. Since then, there have been many new uses of the EM algorithm, as well as further work on its convergence properties, e.g. Wu (1983) [22], Little and Rubin (1987) [14], Schafer (1997) [19]. In any incomplete data problem, the distribution of the complete data X can be factored as

$$f(X|\theta) = f(X_{obs}|\theta) f(X_{miss}|X_{obs}, \theta). \tag{3.21}$$

Let $l(\theta|X) = \ln f(X|\theta)$. The corresponding log-likelihood is

$$l(\theta|X) = l(\theta|X_{obs}) + \ln f(X_{miss}|X_{obs}, \theta). \tag{3.22}$$

Since X_{miss} is unknown, we take the expectation of (3.22) with respect to the distribution $f(X_{miss}|X_{obs}, \theta^t)$, where θ^t is an estimate of the unknown parameter θ. Then we get

$$Q(\theta|\theta^t) = l(\theta|X_{obs}) + H(\theta|\theta^t), \tag{3.23}$$

where

$$Q(\theta|\theta^t) = \int l(\theta|X) f(X_{miss}|X_{obs}, \theta^t) dX_{miss}$$

and

$$H(\theta|\theta^t) = \int [\ln f(X_{miss}|X_{obs}, \theta)]\, f(X_{miss}|X_{obs}, \theta^t) dX_{miss}.$$

Let θ^{t+1} be the value of θ that maximizes $Q(\theta|\theta^t)$; then

$$Q(\theta^{t+1}|\theta^t) \geq Q(\theta^t|\theta^t).$$

By the fact that $\ln x \leq x - 1$, we have

$$\begin{aligned}
H(\theta^t|\theta^t) - H(\theta^{t+1}|\theta^t) &= \int \left[\ln f(X_{miss}|X_{obs}, \theta^t)\right] f(X_{miss}|X_{obs}, \theta^t) dX_{miss} \\
&\quad - \int \left[\ln f(X_{miss}|X_{obs}, \theta^{t+1})\right] f(X_{miss}|X_{obs}, \theta^t) dX_{miss} \\
&= -\int \left[\ln \frac{f(X_{miss}|X_{obs}, \theta^{t+1})}{f(X_{miss}|X_{obs}, \theta^t)}\right] f(X_{miss}|X_{obs}, \theta^t) dX_{miss} \\
&\geq -\int \left[\frac{f(X_{miss}|X_{obs}, \theta^{t+1})}{f(X_{miss}|X_{obs}, \theta^t)} - 1\right] f(X_{miss}|X_{obs}, \theta^t) dX_{miss} \\
&= -\int \left[f(X_{miss}|X_{obs}, \theta^{t+1}) - f(X_{miss}|X_{obs}, \theta^t)\right] dX_{miss} \\
&= 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
l(\theta^{t+1}|X_{obs}) - l(\theta^t|X_{obs}) &= Q(\theta^{t+1}|\theta^t) - H(\theta^{t+1}|\theta^t) - (Q(\theta^t|\theta^t) - H(\theta^t|\theta^t)) \\
&= Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t) + H(\theta^t|\theta^t) - H(\theta^{t+1}|\theta^t) \\
&\geq 0.
\end{aligned}$$

That is,

$$l(\theta^{t+1}|X_{obs}) \geq l(\theta^t|X_{obs}).$$

Thus maximizing $l(\theta|X_{obs})$ is sufficed to maximizing $Q(\theta|\theta^t)$. One iteration of the EM algorithm includes two steps:

1. E-step: the function $Q(\theta|\theta^t)$ is calculated by taking the expectation of $l(\theta|X)$ with the distribution $f(X_{miss}|X_{obs}, \theta^t)$.

2. M-step: the parameter θ is found by maximizing $Q(\theta|\theta^t)$.

The two steps are iterated until the iterations converge. In SAS, the EM algorithm by Schafer [19] is used in the MI procedure. Let the parameter $\theta = (\mu, \Sigma)$. For multivariate normal data, suppose there are G groups with distinct missing patterns. Then the observed-data log-likelihood can be expressed as

$$l(\theta|X_{obs}) = \sum_{g=1}^{G} l_g(\theta|X_{obs}),$$

where $l_g(\theta|X_{obs})$ is the observed-data log-likelihood from the gth group, and

$$l_g(\theta|X_{obs}) = -\frac{n_g}{2}\ln|\Sigma_g| - \frac{1}{2}\sum_{ig}\left[(\mathbf{x}_{ig} - \mu_g)'\Sigma_g^{-1}(\mathbf{x}_{ig} - \mu_g)\right],$$

where n_g is the number of observations in the gth group, the summation is over observations in the gth group, $\mathbf{x}_{ig}$ is a vector of observed values of $\mathbf{x}_g$ variables, μ_g is the corresponding mean vector, and Σ_g is the associated covariance matrix. The initial values for the first iteration are the sample means and sample variances from the observed data. The E-step uses the standard sweep operator [14] on the covariance matrix of the observations to calculate the conditional expectation and variance of missing values. Suppose that A is a $(p \times p)$ symmetric matrix with elements a_{ij}. The standard sweep operator $SWP[k]$ operates on A by replacing it with another $(p \times p)$ symmetric matrix B, where the elements of B are given by

$$\begin{aligned} b_{kk} &= -\frac{1}{a_{kk}}; \\ b_{jk} &= b_{kj} = \frac{a_{jk}}{a_{kk}} \quad \text{for } k \neq j; \\ b_{jl} &= b_{lj} = a_{jl} - \frac{a_{jk}a_{kl}}{a_{kk}} \quad \text{for } k \neq j \text{ and } k \neq l. \end{aligned}$$

Let $B = SWP[k]A$. For example, assume x_t is a time series following the model:

$$(1 - \phi B)x_t = \mu + \epsilon_t \quad \text{for } t = 1, \cdots, n, \tag{3.24}$$

where $|\phi| < 1$, $\{\epsilon_t\}$ is white noise with mean zero and variance σ^2. Let $\theta = (\mu, \phi, \sigma)$. The ML estimate is $\hat{\theta} = (\hat{\mu}, \hat{\phi}, \hat{\sigma})$. Hence the variance and covariance of missing values can be estimated by $\hat{\theta}$. Suppose that x_j is missing, and that x_{j-1} and x_{j+1} are present. The covariance matrix of x_{j-1}, x_j and x_{j+1} is

$$A = \frac{\sigma^2}{1 - \phi^2} \begin{bmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{bmatrix}.$$

Sweeping on $var(x_{j-1})$, i.e. row and column 1, we get

$$A_{j-1} \equiv SWP[1]A = \begin{bmatrix} -\frac{1-\phi^2}{\sigma^2} & \phi & \phi^2 \\ \phi & \sigma^2 & \sigma^2\phi \\ \phi^2 & \sigma^2\phi & \sigma^2(1+\phi^4) \end{bmatrix}.$$

Then sweeping on $var(x_{j+1})$, i.e. row and column 3,

$$SWP[3]A_{j-1} = \frac{1}{1+\phi^2} \begin{bmatrix} -\frac{1}{\sigma^2} & \phi & -\frac{\phi^2}{\sigma^2} \\ \phi & \sigma^2 & \phi \\ -\frac{\phi^2}{\sigma^2} & \phi & -\frac{1}{\sigma^2} \end{bmatrix}. \tag{3.25}$$

From (3.25), we get

$$Var(x_j | x_{j-1}, x_{j+1}, \theta) = \frac{\sigma^2}{1+\phi^2}$$

and

$$\begin{aligned}
E(x_j|x_{j-1}, x_{j+1}, \theta) &= \mu + \frac{\phi}{1+\phi^2}(x_{j-1} - \mu) + \frac{\phi}{1+\phi^2}(x_{j+1} - \mu) \\
&= \mu(1 - \frac{2\phi}{1+\phi^2}) + \frac{\phi}{1+\phi^2}(x_{j-1} + x_{j+1}).
\end{aligned}$$

3.2 Outlier Detection

The effects of extraneous objects, device failure and human errors may distort the field data. Usually qualified engineers, scientists or technicians identify abnormalities after inspecting the data manually. This manual process is slow, costly, and sometimes inconsistent among inspectors. Various methods, such as artificial intelligence [8], neural networks [12] and outlier detection in time series models, have been used for detecting abnormal values in data. In this thesis, we use time series analysis to detect and remove the abnormal data.

The effect of an outlier could be either a short-term transient effect or a long-term change. With short-term effects, one or more outliers may be visible in the time series plot and these can create problems for handling non-stationary with standard time series methods. Thus detecting and removing outliers becomes important in modeling. Four types of outliers are usually considered: innovational outlier (IO), additive outlier (AO), level shift (LS) and temporary change (TC) [20]. An IO represents an extraordinary shock at a time point influencing a sequence of points. An AO causes an immediate and one-shot effect on the observed series. A LS produces an abrupt and permanent step change in the series. A TC causes an initial effect at a time point, and this effect dies out gradually over time. Since any effect on wind speed is short-term, only IO and AO are considered in this thesis. The approach to deal with outliers here is using intervention models to identify the locations and the types of outliers, and to remove the impacts of outliers.

3.2.1 Estimates of Outlier Impacts and Hypothesis Testing

Let x_t be a time series following an autoregressive-integrated-moving average (ARIMA) model with order (p, d, q); that is,

$$\phi(B)\alpha(B)x_t = \psi(B)\epsilon_t, \quad t = 1, \cdots, n, \tag{3.26}$$

where n is the number of observations in the time series; B is the backshift operator such that $Bx_t = x_{t-1}$; $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\psi(B) = 1 - \psi_1 B - \cdots - \psi_q B^q$ are polynomials of B with all roots outside the unit circle; $\alpha(B) = (1 - B)^d$ with all roots of $\alpha(B)$ on the unit circle; and $\{\epsilon_t\}$ are independent and identically normal distributed with mean zero and variance σ^2. We consider the estimation problem when both the location and the dynamic pattern of an outlier are not known. The approach is to classify an outlier impact into two types: IO and AO.

If the location and the dynamic pattern of an event are known, then the models [1] are:

$$\begin{aligned} IO: \quad & z_t = \frac{\psi(B)}{\phi(B)\alpha(B)}(\epsilon_t + \omega\zeta_t^{(T)}) \quad \text{and} \\ AO: \quad & z_t = \frac{\psi(B)}{\phi(B)\alpha(B)}\epsilon_t + \omega\zeta_t^{(T)}, \end{aligned} \tag{3.27}$$

where B, $\phi(B)$, $\psi(B)$, $\alpha(B)$ and $\{\epsilon_t\}$ are the same as in model (3.12), ω is the impact of the possibly unknown outlier at T, and

$$\zeta_t^T = \begin{cases} 1 & \text{for } t = T, \\ 0 & \text{otherwise} \end{cases}$$

and indicates the time of occurrence of the outlier impact. Here T is the possibly unknown location of the outlier. Then (3.27) can be written in the form

$$\begin{aligned} IO: &\quad z_t = x_t + \tfrac{\psi(B)}{\phi(B)\alpha(B)}\omega\zeta_t^{(T)}; \\ AO: &\quad z_t = x_t + \omega\zeta_t^{(T)}. \end{aligned} \tag{3.28}$$

The effect of an IO is more intricate than the effects of other types of outliers. An IO represents an extraordinary shock at time point T influencing $z_T, z_{T+1}, \cdots$, through the dynamic system described by $\frac{\psi(B)}{\phi(B)\alpha(B)}$. To examine the effects of outliers on the estimated residuals in model (3.12), we assume that the time series parameters are known and the series is observed from $t = -J$ to $t = n$, where J is an integer larger than $p+d+q$, and $1 \leq T \leq n$. Let $\pi(B) = \frac{\phi(B)\alpha(B)}{\psi(B)} = 1 - \pi_1 B - \pi_2 B^2 - \cdots$. Because the zeros of $\psi(B)$ are all outside the unit circle, the weights π_j's for j beyond J would in practice become essentially equal to zero with J of moderate size. We use the outlier contaminated data $\{z_t\}$ for model (3.12) to get the estimated residual $\hat{e}_t = \hat{\pi}(B)z_t$ for $t = 1, \cdots, n$. For our two types of outliers, from (3.27) we have

$$\begin{aligned} IO: &\quad \hat{e}_t = \omega\zeta_t^{(T)} + \epsilon_t \text{ and} \\ AO: &\quad \hat{e}_t = \omega\hat{\pi}(B)\zeta_t^{(T)} + \epsilon_t, \end{aligned} \tag{3.29}$$

where $\hat{\pi}(B) = \pi_{\hat{\psi}}(B)$ and $\hat{\psi}$ is MLE of ψ in (3.12) [2]. From the theory of least squares, the estimators of the impact ω in these two models are

$$\begin{aligned} IO: &\quad \hat{\omega}_{IO} = \hat{e}_T \quad \text{and} \\ AO: &\quad \hat{\omega}_{AO} = \hat{\rho}^2 \hat{\pi}(F)\hat{e}_T = \hat{\rho}^2 \hat{\pi}(F)\hat{\pi}(B) z_T, \end{aligned} \tag{3.30}$$

where $\hat{\rho}^2 = (1 + \hat{\pi}_1^2 + \hat{\pi}_2^2 + \cdots + \hat{\pi}_{n-T}^2)^{-1}$ and F is the forward-shift operator. Let H_0 be the null hypothesis that $\omega = 0$ at T, H_{IO} be the alternative hypothesis that an IO exists at T, and H_{AO} be the alternative hypothesis that an AO exists at T. From (3.12) and (3.28), the variances of the estimators for the impacts under H_0 are the following:

$$\begin{aligned} IO: &\quad var(\hat{\omega}_{IO}) = \sigma^2; \\ AO: &\quad var(\hat{\omega}_{AO}) = \rho^2 \sigma^2. \end{aligned}$$

Noticing $\mathbf{E}\hat{\omega}_{IO} = \mathbf{E}\hat{\omega}_{AO} = 0$ ($\mathbf{E}$ means expectation under H_0), hence the results can be used to construct test statistics for testing the existence of an outlier. Thus the likelihood ratio tests are:

$$\begin{aligned} H_0 \;\; vs \;\; H_{IO}: &\quad \hat{\lambda}_{IO,T} = \tfrac{\hat{\omega}_{IO}}{\hat{\sigma}} \quad \text{and} \\ H_0 \;\; vs \;\; H_{AO}: &\quad \hat{\lambda}_{AO,T} = \tfrac{\hat{\omega}_{AO}}{\hat{\rho}\hat{\sigma}}, \end{aligned} \tag{3.31}$$

where $\hat{\sigma} = 1.483 \times median\{|\hat{e}_t - \tilde{e}|\}$, and $\tilde{e}$ is the median of the estimated residuals [6]. The standardized statistics of the outlier effects $\hat{\lambda}_{IO,T}$ and $\hat{\lambda}_{AO,T}$ in (3.31) asymptotically have a standard normal distribution [4].

To locate an IO or AO, the following decision rules are used:

$$IO: \quad \hat{\eta}_{IO} = \max_{1 \leq T \leq n} |\hat{\lambda}_{IO,T}| > c \tag{3.32}$$

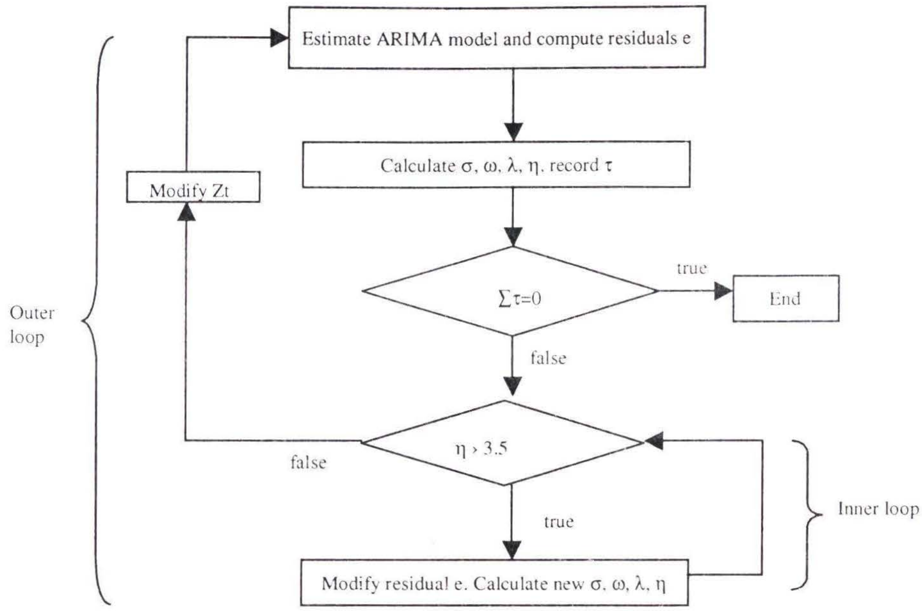


Figure 3.2: Flow Chart for the Procedure of Outlier Detection

$$AO: \quad \hat{\eta}_{AO} = \max_{1 \leq T \leq n} |\hat{\lambda}_{AO,T}| > c, \tag{3.33}$$

where c is some suitably chosen positive constant. In practice, it is recommended to use $c = 3.0$ for high sensitivity, $c = 3.5$ for medium sensitivity, and $c = 4.0$ for low sensitivity in the outlier-detecting procedure when the length of the series is less than 200 [4]. In this thesis $c = 3.5$ is chosen to detect the outliers at any suspected point T. The possible outlier is classified as an IO if $|\hat{\lambda}_{IO,T}| > |\hat{\lambda}_{AO,T}|$, else it is classified as an AO.

3.2.2 Outlier Detection Algorithm

The procedure for detecting outliers is described as follows (the flow chart is shown in Figure 3.2):

1. Estimate ARIMA model (3.26) using $\{z_t\}$ and compute the residuals $\hat{e}_t$ to get $\hat{\omega}_{IO}$ and $\hat{\omega}_{AO}$.

2. Find the median of the residuals $\tilde{e}$, and use $\hat{\sigma} = 1.483 \times median\{|\hat{e}_t - \tilde{e}|\}$ as the estimate of σ. Compute $\hat{\lambda}_{IO,t}$ and $\hat{\lambda}_{AO,t}$ for $t = 1, \cdots, n$. Let $\eta_t = max\{|\hat{\lambda}_{IO,t}|, |\hat{\lambda}_{AO,t}|\}$ for $t = 1, \cdots, n$. Record the location $\tau_t = t$ if $\eta_t > 3.5$, else $\tau_t = 0$. If $\sum_{i=1}^{n} \tau_i = 0$, stop. If $\eta = \max_t \eta_t = |\hat{\lambda}_{IO,T}| > 3.5$, then there is the possibility of an IO at T. The impact ω is estimated by $\hat{\omega}_{IO}$ in (3.30). If $\eta = \max_t \eta_t = |\hat{\lambda}_{AO,T}| > 3.5$, then there is the possibility of an AO at T. The impact ω is estimated by $\hat{\omega}_{AO}$ in (3.30).

3. For the point T in step 2, the new residual for IO is set to

$$\breve{e}_t = \begin{cases} 0, & \text{for } t = T; \\ \hat{e}_t, & \text{else.} \end{cases}$$

The new residuals adjusting for AO are

$$\breve{e}_t = \begin{cases} \hat{e}_t, & \text{for } t < T; \\ \hat{e}_t - \hat{\omega}_{AO}\hat{\pi}(B)\zeta_t^{(T)} & \text{for } t \geq T. \end{cases}$$

A new estimate $\breve{\sigma}$ is computed from the modified residuals. Recompute $\hat{\omega}_{IO}, \hat{\omega}_{AO}, \hat{\lambda}_{IO,t}$ and $\hat{\lambda}_{AO,t}$ based on the same initial estimates of the time series parameters, but using the modified residuals $\breve{e}_t$'s and the estimate $\breve{\sigma}$.

4. Repeat steps 2 and 3 until no further outlier candidates can be identified, that is, $\sum_{i=1}^{n} \tau_i = 0$.

5. Suppose that the k time points $T_1, \cdots, T_k$ are detected as IO's or AO's. Treat these times as known, and estimate the outlier parameters $\omega_1, \omega_2, \cdots, \omega_k$ and the time series parameters simultaneously, using models of the form

$$z_t = \sum_{i=1}^{k} \omega_i L_i(B) \zeta_t^{(T_i)} + \frac{\psi(B)}{\phi(B)\alpha(B)} \epsilon_t, \tag{3.34}$$

where

$$L_i(B) = \begin{cases} 1 & \text{for an AO at } t = T_i, \\ \frac{\psi(B)}{\phi(B)\alpha(B)} & \text{for an IO at } t = T_i. \end{cases}$$

6. Repeat step 1 to 5 until no further new outlier is detected.

3.2.3 Outlier Detection with Missing Values

Before detecting outliers, we first impute missing values. In this section, three imputation methods are used: nearby window average imputation, Jones imputation using Kalman filter [13] and EM algorithm imputation [19]. We study the power of these three imputation methods by using a small portion of time series from station L001. Three data sets are used. Data set A (True) is the hourly wind speed data of January 1996 without missing values. Data sets B and D are constructed from the data set A with missing values by deleting some observations and then imputing these missing values using the EM and Jones imputation, respectively. The locations of missing values are listed in Table 3.1. Data set C is constructed from the data set A with missing values by deleting some observations and imputing these missing values using nearby window average imputation. The average value is the average

Table 3.1: Location of Missing Values

Beginning	End	Number of missing
03JAN96:12	03JAN96:13	2
19JAN96:04	19JAN96:06	3
23JAN96:13	23JAN96:17	5
25JAN96:16	26JAN96:05	14

of one value before the missing value begins and one value after the missing value ends. To fit the models for the data, we use the Time Series Forecasting System in SAS. The system can generate the best model by using 12 criteria such as Mean square error, R-square, Akaike Information criterion (AIC), and Schwarz Bayesian Information Criterion (SBC). Here we use AIC and SBC. For ARIMA models, AIC and SBC are computed as follows:

$$AIC : -2\ln(L) + 2k \text{ and}$$
$$SBC : -2\ln(L) + k\ln(n),$$

where L is the likelihood function, k is the number of free parameters and n is the number of residuals that can be computed for the time series. For the exponential models, AIC and SBC are computed as follows:

$$AIC : n\ln(\frac{SSE}{n}) + 2k \text{ and}$$
$$SBC : n\ln(\frac{SSE}{n}) + k\ln(n),$$

where $SSE = \sum_{t=0}^{n}(y_t - \hat{y}_t)^2$, and $\hat{y}_t$ is the one-step predicted value for the series. The smaller the values of AIC and SBC are, the better the model is. By comparing the values of AIC and SBC for several possible models for these three data sets, we choose single exponential smoothing models. The single exponential smoothing model in SAS is defined as follows [16]: Let x_t be a time series observation at period t. The single exponential smoothing operation is

$$s_t = \alpha x_t + (1 - \alpha)s_{t-1} \tag{3.35}$$

and

$$\hat{x}_{t+1} = s_t, \tag{3.36}$$

where s_t is the smoothed value at period t, α is the smoothing constant $(0 < \alpha < 1)$, and F_{t+1} is the forecast for x_{t+1}. Thus (3.36) can be rewritten as

$$\begin{aligned} \hat{x}_{t+1} &= \alpha x_t + (1 - \alpha)\hat{x}_t && (3.37) \\ &= \alpha[x_t + (1 - \alpha)x_{t-1} + (1 - \alpha)^2 x_{t-2} + \cdots] && (3.38) \end{aligned}$$

Theorem The single exponential smoothing model is equivalent to the ARIMA(0,1,1) model [5].

Proof:

Let x_t be a time series following ARIMA(0,1,1) model, that is

$$(1 - B)x_t = (1 - \psi B)\epsilon_t \quad t = 1, \cdots, n, \tag{3.39}$$

where n, B, ψ and ϵ_t are the same as model (3.12). Rewrite (3.39) as:

$$x_t - x_{t-1} = \epsilon_t - \psi\epsilon_{t-1} \tag{3.40}$$

or

$$\epsilon_t = x_t - x_{t-1} + \psi\epsilon_{t-1}. \tag{3.41}$$

Then we have

$$x_t = x_{t-1} + \epsilon_t - \psi\epsilon_{t-1} = x_{t-1} - \psi\epsilon_{t-1} + \epsilon_t. \tag{3.42}$$

Therefore, the one-step-ahead forecast for x_{n+1} based on $x_1, \cdots, x_n$ is

$$\hat{x}_{n+1} = x_n - \psi\epsilon_n. \tag{3.43}$$

From (3.41) and (3.43), we have

$$\begin{aligned}
\hat{x}_{n+1} &= x_n - \psi(x_n - x_{n-1} + \psi\epsilon_{n-1}) \\
&= x_n - \psi(x_n - \hat{x}_n - \psi\epsilon_{n-1} + \psi\epsilon_{n-1}) \\
&= x_n - \psi(x_n - \hat{x}_n) \\
&= x_n - \psi x_n + \psi\hat{x}_n \\
&= (1-\psi)x_n + \psi\hat{x}_n.
\end{aligned}$$

Setting $\alpha = 1 - \psi$, the above equation is the same as (3.35).

Let $\epsilon_t = x_t - \hat{x}_t$ for all t. Then $\hat{x}_t = x_t - \epsilon_t$. From (3.37), we have

$$\begin{aligned}
x_{t+1} - \epsilon_{t+1} &= \alpha x_t + (1-\alpha)(x_t - \epsilon_t) \\
x_{t+1} - \epsilon_{t+1} &= \alpha x_t + x_t - \alpha x_t - (1-\alpha)\epsilon_t \\
x_{t+1} - x_t &= \epsilon_{t+1} - (1-\alpha)\epsilon_t \\
(1-B)x_{t+1} &= (1-(1-\alpha)B)\epsilon_{t+1}.
\end{aligned}$$

Table 3.2: Summary of Outlier Detection

Time	Data A Impact	Data A Type	Data B Impact	Data B Type	Data C Impact	Data C Type	Data D Impact	Data D Type
501	15.13	IO	15.02	IO	15.11	IO	15.13	IO
581	12.22	IO	12.17	IO	12.21	IO	12.22	IO
131	-11.77	IO	-11.75	IO	-11.77	IO	-11.77	IO
20	9.08	AO	8.98	AO	9.06	AO	9.09	AO
275	8.96	AO	8.90	AO	8.95	AO	8.96	AO
159	11.08	IO	11.12	IO	11.08	IO	11.07	IO
60	10.36	IO	8.28	IO	*	*	*	*
65	10.05	IO	10.31	IO	10.26	IO	10.26	IO
50	-9.16	IO	-9.02	IO	-9.13	IO	-9.17	IO
278	9.12	IO	9.06	IO	9.11	IO	9.13	IO
714	8.99	IO	8.91	IO	8.98	IO	9.00	IO
177	-8.49	IO	-8.53	IO	-8.50	IO	-8.49	IO
641	6.72	AO	6.69	AO	6.71	AO	6.72	AO
17	-8.04	IO	-8.13	IO	-8.06	IO	-8.04	IO
649	7.95	IO	7.97	IO	7.95	IO	7.95	IO
354	7.66	IO	7.58	IO	7.64	IO	7.66	IO
738	-6.10	AO	-6.13	AO	-6.11	AO	-6.10	AO
201	6.09	AO	6.11	AO	6.09	AO	6.09	AO
379	-6.07	AO	-6.04	AO	-6.07	AO	-6.07	AO
658	-5.84	AO	-5.84	AO	-5.84	AO	-5.84	AO
650	7.18	IO	7.40	IO	7.22	IO	7.16	IO
503	10.39	IO	10.06	IO	10.37	IO	10.46	IO
19	-9.59	IO	-9.43	IO	-9.54	IO	-9.58	IO
52	-8.99	IO	-8.64	IO	-8.90	IO	-8.99	IO
583	8.87	IO	9.56	IO	9.61	IO	9.63	IO
67	8.23	IO	8.26	IO	8.39	IO	8.43	IO
444	-5.65	AO	-5.65	AO	-5.68	AO	-5.68	AO
206	7.34	IO	7.31	IO	6.99	IO	7.02	IO
309	7.22	IO	7.18	IO	6.89	IO	6.92	IO
69	7.92	AO	7.84	AO	7.84	IO	7.90	IO
505	8.74	IO	8.33	IO	8.63	IO	8.75	IO
55	-8.01	IO	-7.87	IO	-6.67	IO	-6.68	IO
21	-7.85	IO	-7.67	IO	-7.73	IO	-7.78	IO
490	5.44	AO	5.47	AO	5.41	AO	5.42	AO
582	7.05	IO	7.34	IO	7.08	IO	7.00	IO
716	7.03	IO	*	*	7.00	IO	7.06	IO
585	7.00	IO	8.28	IO	8.31	IO	8.52	IO
229	5.33	AO	*	*	5.64	AO	5.62	AO
620	5.25	AO	*	*	5.24	AO	5.24	AO
592	*	*	-10.04	IO	*	*	*	*
436	*	*	9.03	IO	*	*	*	*
594	*	*	-11.09	IO	*	*	*	*
596	*	*	-9.27	IO	*	*	*	*

"*" : Outlier is not detected.

Table 3.2: Summary of Outlier Detection

	Data A		Data B		Data C		Data D	
Time	Impact	Type	Impact	Type	Impact	Type	Impact	Type
70	*	*	*	*	-7.00	IO	-7.01	IO
204	*	*	*	*	-6.98	IO	-6.96	IO
427	*	*	*	*	6.79	IO	6.78	IO
296	*	*	*	*	5.41	AO	5.39	AO
187	*	*	*	*	-6.70	IO	-6.68	IO
72	*	*	*	*	-10.04	IO	-10.06	IO
189	*	*	*	*	-8.49	IO	-8.52	IO
337	*	*	*	*	-5.11	AO	-5.11	AO
250	*	*	*	*	-6.57	IO	-6.62	IO
191	*	*	*	*	-10.24	IO	-10.20	IO
74	*	*	*	*	-9.08	IO	-9.81	IO
252	*	*	*	*	-7.54	IO	-7.06	IO
487	*	*	*	*	-5.04	AO	-4.96	AO
718	*	*	*	*	6.47	IO	6.63	IO
652	*	*	*	*	4.92	AO	4.97	AO
89	*	*	*	*	-4.89	AO	-4.88	AO
123	*	*	*	*	-6.42	IO	-6.44	IO
429	*	*	*	*	6.41	IO	6.41	IO
687	*	*	*	*	6.39	IO	6.41	IO
259	*	*	*	*	-4.84	AO	-4.86	AO
383	*	*	*	*	-4.83	AO	-4.79	AO
395	*	*	*	*	4.82	AO	4.89	AO
344	*	*	*	*	6.14	IO	6.13	IO
506	*	*	*	*	4.66	AO	*	*
193	*	*	*	*	-9.32	IO	-9.80	IO
507	*	*	*	*	7.79	IO	*	*
76	*	*	*	*	-6.46	IO	-7.55	IO
587	*	*	*	*	6.31	IO	7.44	IO
509	*	*	*	*	10.77	IO	*	*
589	*	*	*	*	7.07	IO	8.15	IO
511	*	*	*	*	9.05	IO	*	*
591	*	*	*	*	6.60	AO	9.76	IO
310	*	*	*	*	5.96	IO	5.91	IO
584	*	*	*	*	*	*	5.87	IO
139	*	*	*	*	*	*	5.86	IO
266	*	*	*	*	*	*	5.86	IO
692	*	*	*	*	*	*	-5.84	IO
546	*	*	*	*	*	*	6.35	IO
191	*	*	*	*	*	*	-10.20	IO
74	*	*	*	*	*	*	-9.81	IO
718	*	*	*	*	*	*	6.63	IO
141	*	*	*	*	*	*	6.47	IO
143	*	*	*	*	*	*	8.43	IO
71	*	*	*	*	*	*	7.19	IO
78	*	*	*	*	*	*	-6.33	IO
268	*	*	*	*	*	*	5.89	IO
23	*	*	*	*	*	*	-5.72	IO
179	*	*	*	*	*	*	-5.69	IO
593	*	*	*	*	*	*	8.53	IO
595	*	*	*	*	*	*	7.64	IO
597	*	*	*	*	*	*	6.85	IO
606	*	*	*	*	*	*	-8.24	IO
599	*	*	*	*	*	*	6.13	IO
608	*	*	*	*	*	*	-6.89	IO
71	*	*	*	*	*	*	-6.07	IO
161	*	*	*	*	*	*	6.04	IO
307	*	*	*	*	*	*	-5.96	IO

"*" : Outlier is not detected.

Table 3.3: Classification Matrix

Type	Data B			Data C			Data D			Total
	AO	IO	*	AO	IO	*	AO	IO	*	
AO	10	0	2	11	1	0	11	1	0	12
IO	0	26	1	0	26	1	0	26	1	27
*	0	4	57	10	23	28	8	45	8	61
Total	10	30	60	21	50	29	19	72	9	100

Setting $\psi = 1 - \alpha$, the above equation is the same as (3.39). $\square$

Thus, the ARIMA(0,1,1) model is used for outlier detection. Table 3.2 lists the locations, impacts and types of outliers that are detected in the four data sets. Given data A are true, from Table 3.3, the overall correct rates of outliers (IO,AO and over detected outliers labeled by *) of data B, C and D are $\frac{10+26+57}{100} = 93\%$, $\frac{11+26+29}{100} = 66\%$ and $\frac{11+26+8}{100} = 45\%$, respectively, while the correct rates of outliers (IO, AO) of data B, C and D are $\frac{10+26}{12+27} = 92\%$, $\frac{11+1+26}{12+27} = 97\%$ and $\frac{11+1+26}{12+27} = 97\%$, respectively; the correct rates of IO outliers of all three methods are $\frac{26}{27} = 96\%$; the correct rates of AO outliers of data B is $\frac{10}{12} = 83\%$; the correct rates of AO outliers of data C and D are $\frac{11}{12} = 92\%$.

By comparing the correction rates, we see that the best result is the EM algorithm. We also know that when the data are not fully observed, the EM algorithm is a general technique for finding maximum-liklihood estimates for parametric models [19]. Hence we use the EM algorithm to fill the missing values and then detect outliers for the data set in this thesis.

Chapter 4

MODELING

To get the best model for the wind speed data, we use EM algorithm to impute missing values and the method introduced in Chapter 3 to detect outliers and remove impacts of outliers. Let x_t be the true time series, y_t be the observed series with missing values and outliers, and z_t be the observed series with outliers after imputation. The idea is the following (Figure 4.1):

Step 1: Impute the missing values in y_t using the EM algorithm. The data set we get then is z_t. SAS code is in Appendix A.1.

Step 2: Detect outliers and remove the impacts of outliers in z_t. The data set we get then is x_t.

Step 3: Let $y_t^{'}$ be x_t but with the same missing values as y_t. Re-do steps 1 and 2. If there exist outliers in step 2, finish step 2 and do step 3. Otherwise, fit the best models for z_t.

The data set used in this chapter is the hourly wind speeds of all four stations from May to August in 2000. To fit the models for the data, we still use Time Series Forecasting System in SAS. AIC and SBC are used as information criterions. During

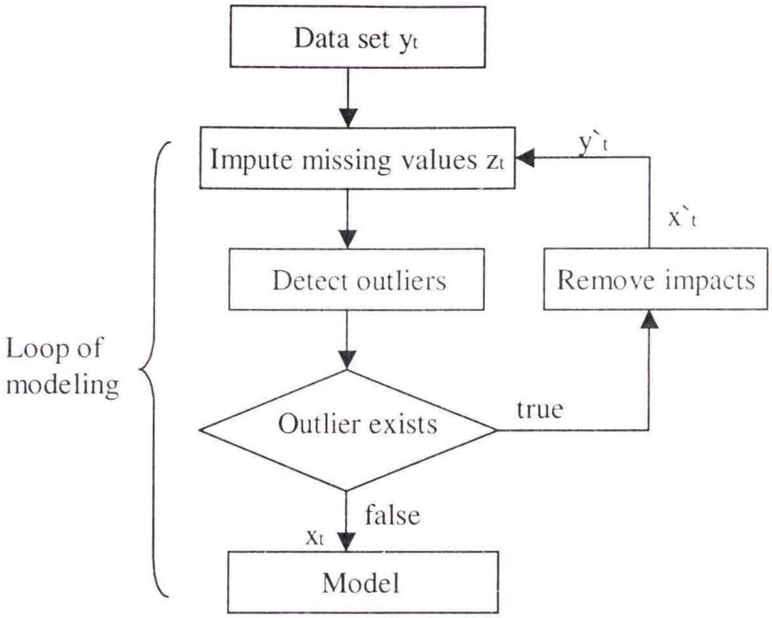


Figure 4.1: Flow Chart of Modeling Process

the process of imputing missing values, detecting outliers and removing impacts of outliers, we get the fitted models are seasonal ARIMA models. In SAS, the seasonal ARIMA model is denoted by ARIMA$(p, d, q) \times (P, D, Q)_s$. The term (p, d, q) gives the order of the nonseasonal part of the ARIMA model; the term $(P, D, Q)_s$ gives the order of the seasonal part. The value of s is the number of observations in a seasonal cycle such as 12 for monthly series. The fitted models are ARIMA$(2, 0, 0) \times (1, 0, 0)_{24}$ of the form

$$(1 - \phi_{1,1}B - \phi_{1,2}B^2)(1 - \phi_{2,1}B^{24})x_t = \mu + \epsilon_t.$$

For convenience of outlier detection stage, we de-mean before fitting the models. Table 4.1 reports the summary of outer loops in outlier detection stage in the first

Table 4.1: Outlier Detection Report

Outer	L001			L005			L006			LZ40		
Loop	mean	AO	IO	mean	AO	IO	mean	AO	IO	mean	AO	IO
1	10.407	47	57	10.373	40	49	10.181	44	45	10.431	47	55
2	10.114	21	16	10.120	21	31	9.884	27	24	10.039	31	24
3	10.114	2	6	10.058	12	18	9.773	4	8	9.946	7	7
4	10.121	1	1	10.034	8	8	9.725	2	5	9.931	2	6
5	10.120	0	0	10.007	2	4	9.694	1	0	9.934	2	2
6				10.007	0	4	9.691	2	1	9.938	2	2
7				10.006	0	3	9.687	1	0	9.942	1	0
8				10.002	1	2	9.688	1	0	9.944	0	0
9				9.993	1	1	9.690	0	0			
10				9.991	0	0						

loop of modeling. In the second loop of modeling, no outlier is detected in Station L001, L005 and LZ40. Hence we go on to model for Station L006 until no outlier is detected in the locations of observed values. Finally, after imputing missing values, and detecting and removing impacts of outliers, we get the following best models for the hourly wind speeds of stations L001, L005, L006 and LZ40 from May to August in 2000:

$$\text{L001: } (1 - 0.895B + 0.097B^2)(1 - 0.156B^{24})x_t = 10.144 + \epsilon_t.$$

$$\text{L005: } (1 - 0.924B + 0.100B^2)(1 - 0.207B^{24})x_t = 10.014 + \epsilon_t.$$

$$\text{L006: } (1 - 0.878B + 0.050B^2)(1 - 0.240B^{24})x_t = 9.659 + \epsilon_t.$$

$$\text{LZ40: } (1 - 0.988B + 0.146B^2)(1 - 0.225B^{24})x_t = 9.991 + \epsilon_t.$$

The parameter estimates and goodness of fit tests are shown in Table 4.2. We can see that all the parameter estimates are significant. To check the white-noise assumption, we draw the histograms for residuals. The histograms in Figure 4.2 are about normal. This means that the assumptions for residuals of the four models are

Table 4.2: Parameter Estimates and Good-fitness Tests

		L001	L005	L006	LZ40
intercept	estimate	10.144	10.014	9.659	9.991
	T	43.735	36.996	31.227	33.624
	p-value	< 0.001	< 0.001	< 0.001	< 0.001
$\phi_{1,1}$	estimate	0.895	0.924	0.878	0.988
	T	48.728	50.308	47.311	53.862
	p-value	< 0.001	< 0.001	< 0.001	< 0.001
$\phi_{1,2}$	estimate	-0.097	-0.100	-0.050	-0.146
	T	-5.290	-5.463	-2.685	-7.990
	p-value	< 0.001	< 0.001	< 0.007	< 0.001
$\phi_{2,1}$	estimate	0.156	0.207	0.240	0.225
	T	8.517	11.388	13.236	12.375
	p-value	< 0.001	< 0.001	< 0.001	< 0.001
AIC		4545.242	4291.438	4650.076	4075.418
SBC		4569.203	4315.399	4674.036	4099.379

valid. From the four models, we can conclude that the wind speeds in these four stations have the similar patterns. This conclusion is the same as the one we get in Chapter 2. The first plot in Figure 4.3 is the plot of wind speeds vs time for station L006 from August 14 to 23, 2000. We can see that there is a large block of missing values. The second plot is the plots of wind speed for station L001, L005, LZ40 and imputation wind speeds of L006 at the same time. Again we can see that the plots have similiar patterns. This means that EM algorithm is a very good method to impute missing values for our wind speed data set. We also can see that there is a daily cycle in wind speed data from the models.

Through analyzing of Lake Okeechobee wind speed data, we can conclude that the wind speeds of the four stations we study have similar patterns and a daily

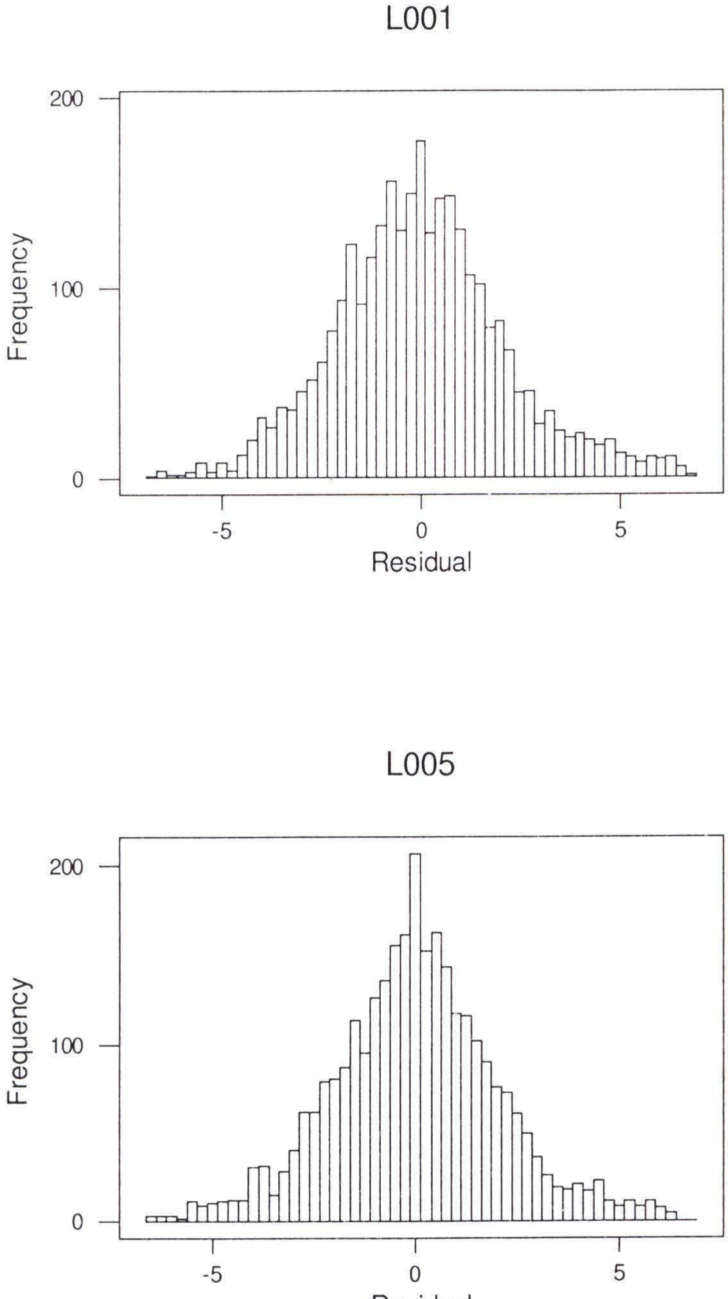


Figure 4.2: Histograms of Residuals

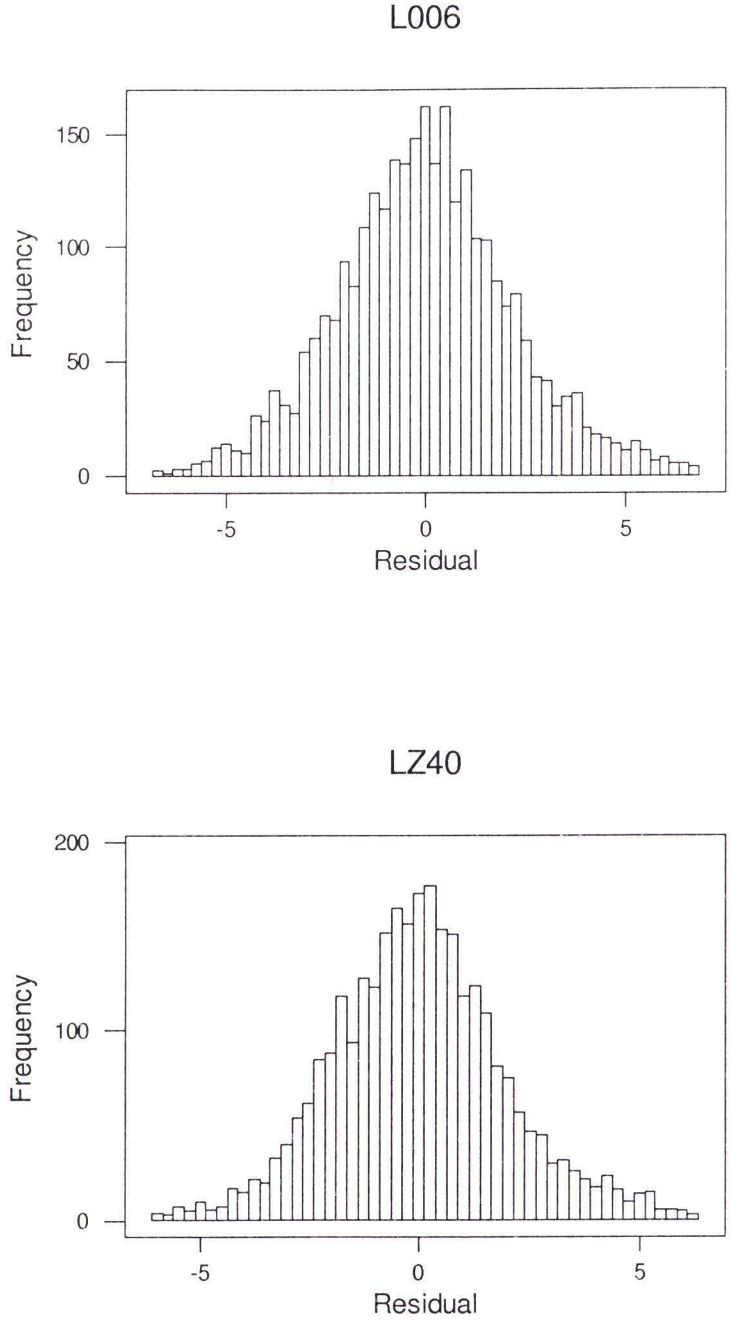


Figure 4.2: Histograms of Residuals

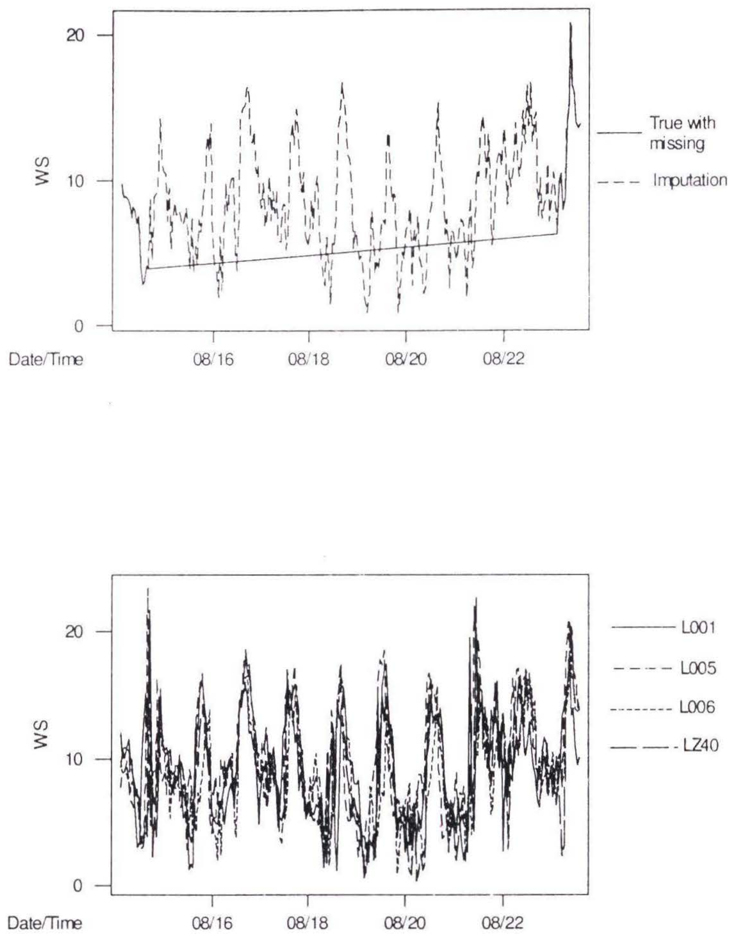


Figure 4.3: Plots of Wind Speeds

cycle. For the data we study in this thesis, the best method to impute missing values is the EM algorithm and the best fitted model is the seasonal ARIMA$(2,0,0) \times (1,0,0)_{24}$. The fact that the wind speeds of the four stations have similar patterns and models shows that the wind speed in all stations under study behave in a similar way. Furthermore the method of outlier detection using intervention models in time series models and the EM algorithm to impute missing values are more effective than the manual process of inspecting abnormal values and filling missing values in the data set.

Chapter 5

CONCLUSIONS

In this thesis, we analyzed wind speeds at four stations in Lake Okeechobee. There are lots of missing values and outliers in the data. The patterns of wind speeds for all four stations are similar and have a daily cycle. But the monthly means of wind speeds at station L001 are substantially different from those of the other stations in February 1998 and in 1995. This little difference at station L001 may be caused by various reasons such as location of the station, device failures or bird interruptions. Further study is needed. The wind speeds of the stations are correlated positively. A three-parameter Weibull distribution does not fit the data well. The EM algorithm is good for imputing missing values of the data. The method of outlier detection seems more effective than the manual process of inspecting abnormal values and filling missing values in the data set. In a future study, we may consider using a lognormal, beta or mixed distribution to fit the data. We also need combine the computer programs only using SAS.

Appendix

COMPUTER PROGRAM

A.1 SAS Program

```
/*********************************************************/
/* Title: EM imputation                                  */
/* Input: hourly wind speeds of 4 stations from 05/00 to 08/00  */
/* Output: wind speeds after EM imputation               */
/*********************************************************/
option ls=70 ps=750 nodate nonumber;

data miss00;
infile 'c:\data00.prn';
input year month day hour ws1 ws5 ws6 ws40;
datetime=dhms(mdy(month, day, year),hour,0,0);
format datetime datetime10.;
drop year month day hour;
run;

proc mi data=miss00 out=a;
var ws1 ws5 ws6 ws40;
run;
```

A.2 Matlab Program

MatLab code for outlier detection of ARIMA$(2,0,0) \times (1,0,0)$:

```
%File re.m: Detect outlier, compute impact
%Input file: re.txt is residuals
%Ouput file: impact.txt is impacts, postions and types of outliers clear;
hu=1;
j=1;
while hu==1;
dataset=load('c:\re.txt');    %input residuals
resi=dataset(:,1);
[n, m] =size(resi);
```

```
m0=median(resi);
m1=median(abs(resi-m0));
sigma=1.483*m1;
phi1=0.8948;    %φ1,1
phi2=-0.0971;   %φ1,2
phi=0.1564;     %φ2,1
%compute λ for IO
for t = 1 : n;
lambda_io(t)=resi(t)/sigma;
end;
%compute ω, λ for AO
for t=1:(n-26);
p(t)=1+phi1∧2+phi2∧2+phi∧2+(phi1*phi)∧2+(phi2*phi)∧2;
pp(t)=resi(t)-phi1*resi(t+1)-phi2*resi(t+2)
−phi*resi(t+24)+phi1*phi*resi(t+25)+phi2*phi*resi(t+26);
end;
for t=n-25;
p(t)=1+phi1∧2+phi2∧2+phi∧2+(phi1*phi)∧2;
pp(t)=resi(t)-phi1*resi(t+1)-phi2*resi(t+2)-phi*resi(t+24)+phi1*phi*resi(t+25);
end;
for t=n-24;
p(t)=1+phi1∧2+phi2∧2+phi∧2;
pp(t)=resi(t)-phi1*resi(t+1)-phi2*resi(t+2)-phi*resi(t+24);
end;
for t=(n-23):(n-2);
p(t)=1+phi1∧2+phi2∧2;
pp(t)=resi(t)-phi1*resi(t+1)-phi2*resi(t+2);
end;
for t=n-1;
p(t)=1+phi1∧2;
pp(t)=resi(t)-phi1*resi(t+1);
end;
for t=n;
p(t)=1;
pp(t)=resi(t);
end;
for t=1:n;
w_ao(t)=pp(t)/p(t);
lamda_ao(t)=w_ao(t)/(sqrt(1/p(t))*sigma);
%check if IO exist
if abs(lamda_io(t))>= 3.5 k_io(t)=t;
else k_io(t)=0;
end;
%check if AO exist
if abs(lamda_ao(t))>= 3.5 k_ao(t)=t;
else k_ao(t)=0;
end;
%decide outlier type: 0 for AO, 1 for IO
```

```
if abs(lamda_ao(t))> abs(lamda_io(t)) diff(t)=0;
eta(t)=abs(lamda_ao(t)); tau(t)=k_ao(t);
w(t)=w_ao(t);
else diff(t)=1;
eta(t)= abs(lamda_io(t)); tau(t)=k_io(t);
w(t)=resi(t);
end;
end;
ita=max(eta);
k=1;
for t=1:n
if eta(t)==ita & tau(t)> 0 k=t;
impa(j)=w(k);
loc(j)=tau(k);
d(j)=diff(k);
break;
else k=0;
end;
end;
if diff(k)==1 resi(k)=0
else
if k==n;
resi(k)=resi(k)-w(k);
elseif k==n-1;
resi(k)=resi(k)-w(k);
resi(k+1)=resi(k+1)+w(k)*phi1;
elseif k < n - 1 & k> n-24;
resi(k)=resi(k)-w(k);
resi(k+1)=resi(k+1)+w(k)*phi1;
resi(k+2)=resi(k+2)+w(k)*phi2;
elseif k == n-24;
resi(k)=resi(k)-w(k);
resi(k+1)=resi(k+1)+w(k)*phi1;
resi(k+2)=resi(k+2)+w(k)*phi2;
resi(k+24)=resi(k+24)+w(k)*phi;
elseif k == n-25;
resi(k)=resi(k)-w(k);
resi(k+1)=resi(k+1)+w(k)*phi1;
resi(k+2)=resi(k+2)+w(k)*phi2;
resi(k+24)=resi(k+24)+w(k)*phi;
resi(k+25)=resi(k+25)-w(k)*phi1*phi;
elseif k < n-25;
resi(k)=resi(k)-w(k);
resi(k+1)=resi(k+1)+w(k)*phi1;
resi(k+2)=resi(k+2)+w(k)*phi2;
resi(k+24)=resi(k+24)+w(k)*phi;
resi(k+25)=resi(k+25)-w(k)*phi1*phi;
```

```
resi(k+26)=resi(k+26)-w(k)*phi2*phi;
end;
end;
re=[resi];
fid=fopen('re.txt','w');
fprintf(fid,'%10.4f\n',re);
fclose(fid);
if sum(loc)==0
break;
else xy=[impa;loc;d];
j=j+1;
fid=fopen('impact.txt','w');
fprintf(fid,'%10.4f %4.0f %4.0f\n',xy);
fclose(fid);
end
end
```

```
%File ws.m: remove impact of outlier.
%input: impact.txt(impact, location, type of outlier) ws.txt (wind speed)
%output:wsa.txt(wind speed after removing impacts of outliers)
phi1=0.9878;
phi2=-.146;
phi=0.2253;
dataset=load('c:\ws.txt');
ws=dataset(:,1);
dataset=load('c:\impact.txt');
impact=dataset(:,:);
[m, n] =size(impact);     %m is row, n is column
w=impact(:,1);
loca=impact(:,2);
d=impact(:,3);    % 0 for ao, 1 for io
for t=1:m
if d(t)==0
ws(loca(t))=ws(loca(t))-w(t);
else ws(loca(t))=ws(loca(t))-w(t);
ws(loca(t)+1)=ws(loca(t)+1)-phi1*w(t);
end
end
speed=[ws];
fid=fopen('ws.txt','w');
fprintf(fid,'%10.4f\n',speed);
fclose(fid);
```

BIBLIOGRAPHY

[1] G. E. P. Box and G. C. Tiao. Intervention Analysis with Applications to Aconomic and Environmental Problems. *Journal of the American Statistical Association*, 70:70–79, 1975.

[2] G. E. P. Box and G. C. Tiao. *Time Series Analysis, Forecasting and Control.* Holden-Day, San Francisco, 1976.

[3] R. G. Brown and P. Y. C Hwang. *Introduction to Random Signals and Applied Kalman Filtering.* John Wiley & Sons, New York, 1997.

[4] L. H. Chang, G. C. Tiao and C. Chen. Estimation of Time Series Parameters in the Presence of Outliers. *Technometrics*, 30(2):193–204, 1988.

[5] C. Chatfield. *The Analysis of Time Series: An Introduction.* Chapman & Hall, New York, 1996.

[6] C. Chen and L. M. Liu. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 88(421):284–297, 1993.

[7] A. P. Dempster, N. M. Laird and D. B. Rubin Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm (with discussion). *Journal of Royal Statistical Society Series B*, 39:1–38, 1977.

[8] K. Y. Huang. Fuzzy Functional-link Net for Seismic Trace. *Proc. IEEE, Int. Conf. Neural Networds.*, 3:1650–1653, 1994.

[9] L. E. Holmedal, D. Myrhaug and H. Rue. Seabed Shear Stresses under Irregular Waves Plus Current from Monte Carlo Simulations of Parameterized Models. *Coastal Engineering*, 39:123–147, 2000.

[10] R. T. James, V. H. Smith and B. L. Jones. Historical trends in the Lake Okeechobee ecosystem, III. water quality. Arch. Hydrobiol./Suppl. (Monographische Beitrage), 1995.

[11] K. R. Jin and K. H. Wang. Wind Generated Waves in Lake Okeechobee. *Journal of the American Water Resources Association*, 34(5):1099–1108, 1998.

[12] K. R. Jin, D. C. Chen, L. Fang, H. C. Chen and J. Martin. Data Quality Control For Lake Temperature by Neural Network. *Journal of Lake and Reservoir Management*, 15(4):272–284, 1999.

[13] R. H. Jones. Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations. *Technometrics*, 22(3):389–395, 1980.

[14] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, New York, 1987.

[15] J. M. Otero and V. Floris. Lake Okeechobbee vegualtion schedule simulation - South Florida regional routing model. *South Florida Water Management District, West Palm Beach, FL* 1994.

[16] SAS Institute Inc.. *SAS/ETS User's Guide.* SAS Institute Inc., Cary NC, 1995.

[17] South Florida Water Management District. About Lake Okeechobee. *http://www.sfwmd.gov/org*, 2001.

[18] M. A. Stephens. EDF Statistics for Goodness of Fit and Some Comparisons. *Journal of the American Statistical Association*, 69:730–737, 1974.

[19] J. L. Schafer. *Analysis of Imcomplete Multivariate Data.* Chapman & Hall, Boca Raton, 1999.

[20] R. S. Tsay. Outliers, Level Shifts, and Variance Changes in Time Series. *Journal of Forecasting*, 7(1):1–20, 1988.

[21] G. Welch and G. Bishop. An Introduction to the Kalman Filter. *http://www.cs.unc.edu/~welch*, 1995.

[22] C. F. J. Wu. On the Convergence Properties of the EM algorithm. *Annals of Statistics*, 47:635–646, 1983.