

**SPATIAL AND TEMPORAL MODELLING OF  
WATER ACIDITY IN TURKEY LAKES WATERSHED**

**Spatial and Temporal Modelling of  
Water Acidity in Turkey Lakes Watershed**

By  
Jing Lin, M.A.

A Project  
Submitted to the School of Graduate Studies  
in Partial Fulfilment of the Requirements  
for the Degree  
Master of Science

McMaster University

© Copyright by Jing Lin, May, 2005

MASTER OF SCIENCE (2005)  
(Statistics)

McMaster University  
Hamilton, Ontario

TITLE: Spatial and Temporal Modelling for  
Turkey Lakes Watershed Acidity

AUTHOR: Jing Lin , M.A.  
(Shandong University, P. R. China)

SUPERVISOR: Professor Abdel H. El-Shaarawi

NUMBER OF PAGES: x, 74

# Abstract

Acid rain continues to be a major environmental problem. Canada has been monitoring indicators of acid rain in various ecosystems since the 1970s. This project focuses on the analysis of a selected subset of data generated by the Turkey Lakes Watershed (TLW) monitoring program from 1980 to 1997. TLW consists of a series of connected lakes where 6 monitoring stations are strategically located to measure the input from an upper stream lake into a down stream lake. Segment regression models with AR(1) errors and unknown point of change are used to summarize the data. Relative likelihood based methods are applied to estimate the point of change. For pH, all the regression parameters except autocorrelation have been found to change significantly between the model segments. This was not the case for  $SO_4^{2-}$  where a single model was found to be adequate. In addition pH has been found to have a moderate increasing trend and pronounced seasonality while  $SO_4^{2-}$  showed a dramatic decreasing trend but little seasonality. Multivariate dimension reduction methods are used to provide an overall graphical summary of the changes in TLW water system. We also report the result of applying segment regression for the analysis of first two principal components in selected stations. The results show that the efforts of the Canadian and US governments to reduce the emission of  $SO_2$  have been successful in

controlling the acid rain problem in Eastern Canada. The project ends with suggestions for various extensions of the present work.

**Key Words:** Acid rain; TLW; Likelihood method; Change point; Segment regression; Auto-correlation; Principal components.

# Acknowledgements

I would like to sincerely thank my supervisor, Dr. Abdel H. El-Shaarawi for his great guidance, support, encouragement and patience throughout the entire process of this project.

I would like to thank Dr. Peter Macdonald and Dr. Roman Viveros-Aguilera for serving in my examination committee and providing me with valuable advice and help over the years.

Thanks go to my professors and staff in the Department of Mathematics and Statistics at McMaster University for their valuable support.

Last but not least I express my appreciation to Dr. Dean S. Jeffries, of Environment Canada, for providing the data and some of the background information.

Thanks also go to all my friends for their friendship and help during the last two years, to my daughter for her support.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Data Structure and Variables . . . . .	5
<b>2</b>	<b>Graphical Description of the Data</b>	<b>8</b>
2.1	The Graphical Display of pH Value . . . . .	9
2.2	The Graphical Display of $SO_4^{2-}$ . . . . .	14
<b>3</b>	<b>Multivariate Data Reduction</b>	<b>22</b>
3.1	Matrix Factorization . . . . .	22
3.2	Matrix Approximation . . . . .	24
3.3	Biplots . . . . .	27
<b>4</b>	<b>Regression Models with Change Point</b>	<b>33</b>
4.1	Likelihood Methods . . . . .	34

4.1.1	The Likelihood under the of Independent Assumption . . . . .	36
4.1.2	The Likelihood under the AR(1) Assumption . . . . .	37
4.2	Results . . . . .	41
4.2.1	Modelling the Changes in pH . . . . .	41
4.2.2	Modelling the Changes of $SO_4^{2-}$ . . . . .	49
4.2.3	Change Point of the First Two Principal Components . . . . .	51
<b>5</b>	<b>Conclusion and Future work</b>	<b>58</b>
5.1	Conclusion . . . . .	58
5.2	Future Work . . . . .	59
<b>A</b>	<b>R functions</b>	<b>62</b>
A.1	Monthly mean . . . . .	62
A.2	Change point under the independent assumption . . . . .	63
A.3	Likelihood function under AR(1) assumption . . . . .	65
A.4	Result analysis . . . . .	68
<b>B</b>	<b>R command</b>	<b>71</b>



# List of Tables

1.1	Example of the data set. . . . .	7
4.1	Change of pH under assumption of independence. . . . .	42
4.2	Change of pH under assumption of AR(1). . . . .	45
4.3	The estimates of $\sigma$ in two segments. . . . .	45
4.4	$\hat{\phi}_1, \hat{\phi}_2$ and their std for the AR(1) model. . . . .	48
4.5	Coefficients of fitting the trend of $SO_4^{2-}$ . . . . .	50
4.6	Change of the first two principal components in S0 and S5. . .	52

# List of Figures

1.1	Map of the TLW with location of the observation stations. . . . .	3
2.1	Monthly mean of pH (Jan.– April). . . . .	10
2.2	Monthly mean of pH (May– Aug.). . . . .	11
2.3	Monthly mean of pH (Sept.– Dec.). . . . .	12
2.4	Seasonality of pH. . . . .	13
2.5	pH difference among stations. . . . .	15
2.6	Monthly mean of $SO_4^{2-}$ (Jan.– April). . . . .	17
2.7	Monthly mean of $SO_4^{2-}$ (May– Aug.). . . . .	18
2.8	Monthly mean of $SO_4^{2-}$ (Sept.– Dec.). . . . .	19
2.9	Seasonality of $SO_4^{2-}$ . . . . .	20
2.10	$SO_4^{2-}$ difference among stations. . . . .	21
3.1	Biplot 1 (Jan.– April). . . . .	28
3.2	Biplot 2 (May – Aug.). . . . .	29

3.3	Biplot 3 (Sept. – Dec.). . . . .	30
3.4	Boxplot of angles and distances between pH and the given variable. . .	31
4.1	The change point of pH in station S0 under assumption of independence.	43
4.2	The change point of pH in station S5 under assumption of independence.	44
4.3	The change point of pH in station S0 under assumption of AR(1). . . .	46
4.4	The change point of pH in station S5 under assumption of AR(1). . . .	47
4.5	The trend of $SO_4^{2-}$ . . . . .	50
4.6	Boxplots of $SO_4^{2-}$ before and after cutting off the extreme values. . . .	51
4.7	The change point of first principal component of station S0. . . . .	53
4.8	The change point of second principal component of station S0. . . . .	54
4.9	The change point of first principal component of station S5. . . . .	55
4.10	The change point of second principal component of station S5. . . . .	56

# Chapter 1

## Introduction

### 1.1 Background

Acid rain has been a hot topic for over thirty years. It has direct and indirect impact on natural environment and human health. It kills aquatic life, trees, crops and other vegetation, damages buildings and monuments, corrodes copper and lead piping, damages man-made things such as automobiles, reduces soil fertility and can cause toxic metals to leach into underground drinking water sources. In early 1970, several cases of surface water acidification and loss of fish population were reported in Scandinavia, Canada and the United States of America. Acidification has been shown to be related to fisheries losses (Beamish and Harvey, 1972; Wales and Beggs, 1986; Smith and Underwood, 1986). As a result, environmental agencies in these countries became concerned about the potential impact of the long range transport of air pollutants (LRTAP) and made funds available to study various aspects of acid rain problem.

In Eastern Canada, the area south of latitude 52° is generally considered sensitive

to elevated atmospheric deposition (Kelso, 1986). The distribution of lake sizes in the Sault Ste. Marie district is considered representative of this area, and the lakes in the Turkey Lakes Watershed (TLW) are among the type considered most at risk (Jeffries, 1988). At that time, several other study sites already existed, covering a variety of terrain types. TLW was chosen from some 100 potential candidates to fill the gap in the overall Canadian research programs investigating the effects of LRTAP on shield terrain. Another reason for selecting TLW was to utilize expertise in a truly integrated watershed study that was already available in Sault Ste. Marie at the Great Lakes Forest Centre (Canadian Forestry Service, Department of Natural Resources) and the Great Lakes Laboratory for Fisheries and Aquatic Science (Department of Fisheries and Oceans).

The TLW study was established in 1980 to define the impact of acidic deposition on undeveloped aquatic and terrestrial terrain. It was organized as a joint initiative of Canadian Forest Service, Environment Canada and Department of Fisheries and Oceans. The National Water Research Institute (NWRI) of Environment Canada is one of the members in the study team.

The TLW (Figure 1.1) is located 50 km north of Sault Ste. Marie, Ontario, on the Canadian Shield near the northern margin of the Great Lakes–St. Lawrence forest region. It is 10.5 km<sup>2</sup> in area and contains a chain of four lakes. There are few emission sources within 100 km of the TLW, so most of the atmospheric pollutants that reach it are transported from sources hundreds or thousands of kilometers away.

Acid rain refers to all types of precipitation—rain, snow, sleet, hail, fog—that is acidic in nature. The sources of the acid deposition are mainly two kinds of air pollutants: sulfur dioxide ( $SO_2$ ) and nitrogen oxides (include  $NO_2$  and  $NO_3$ , denoted as  $NO_x$ ).

These chemicals—produced by burning of fossil fuels, smelting of ore, burning of coal, and processing of natural gas—can be carried long distances by the wind before dissolving in precipitation and being deposited on the earth's surface. The acid deposition is an international issue. North America is a huge contributor to the world's air pollution and acid rain.

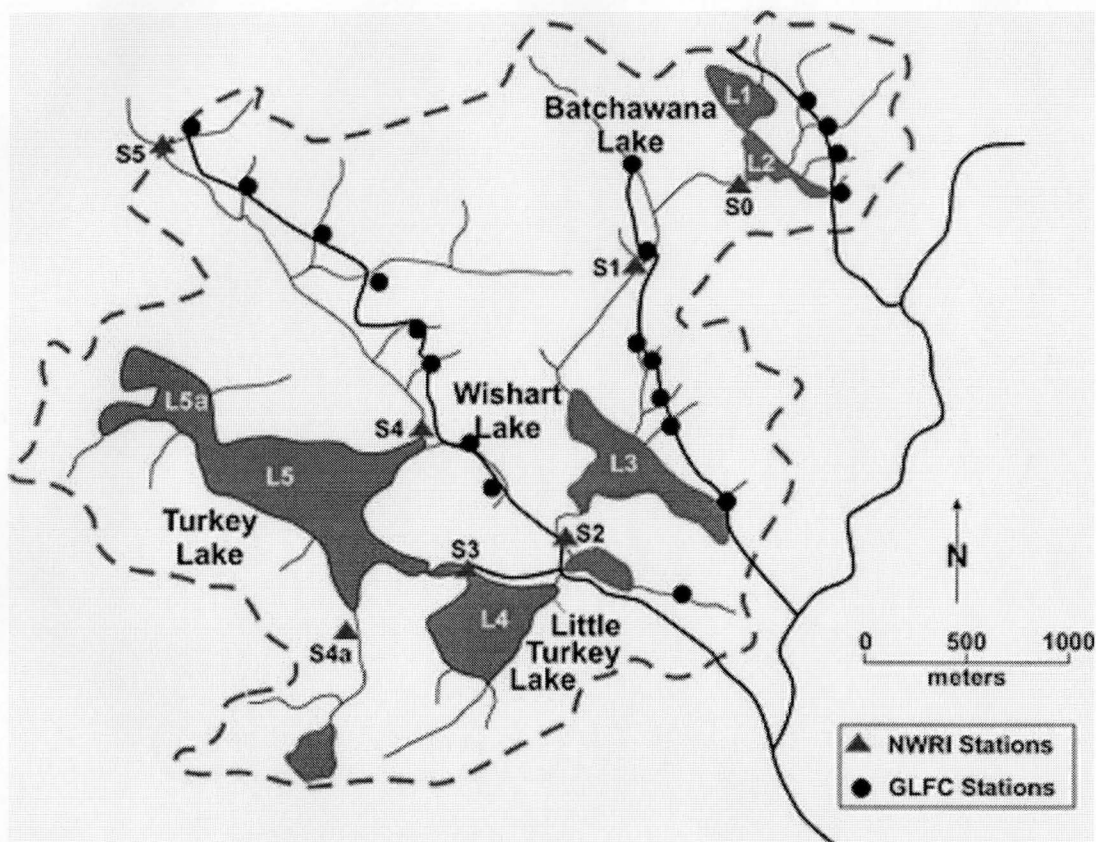


Figure 1.1: Map of the TLW with location of the observation stations.

Acidity is measured by the pH value, the logarithm of the concentration of the free hydrogen ions  $H^+$  (electrically charged atoms) in water. The pH scale is logarithmic

where pH value changes 1 unit as the concentration of free hydrogen ions changes 10 times. Because normally carbon dioxide exists in the atmosphere and it reacts with water to form carbonic acid, so “pure” rain is acidic with pH 5.6-5.7.

A selected subset of the data collected on Turkey Lake Watershed Study during 1880 to 1997 is used in this project. During the study period, there were certain interventions. In 1985 the governments of Canada and the seven eastern provinces joined forces to take action on reducing sulfur dioxide, the major contributor to acid rain. They launched a program to cut sulfur dioxide emissions in the eastern provinces in half by 1994. In 1990, the U.S. launched action to reduce emissions of sulfur dioxide by amending its “Clean Air Act”. Do these interventions have significant effect on the stream water acidity in TLW? So in this investigation we concentrate mainly on the analysis of the pH and  $SO_4^{2-}$  data. pH is a response variable and  $SO_4^{2-}$  is one of explanatory variables. The objective of this project is to evaluate within- and between-year variability, seasonal cycles and multi-year trends of the stream water acidity in TLW, to identify the change of the acidity of the water by using suitable statistical methods.

We begin by performing exploratory analysis using graphical methods for displaying univariate and multivariate data using monthly mean values. This was then followed by concentrating on the detailed modelling of pH and  $SO_4^{2-}$ . The results show clear spatial and temporal patterns in the two variables: pH value and the concentration of  $SO_4^{2-}$  are increasing as the water travels from the upstream lake to a downstream lake. Temporally, pH is increasing while  $SO_4^{2-}$  is decreasing.

The result shows that pH has modestly increased and  $SO_4^{2-}$  has clearly decreased after those interventions took place.

A brief description of the data set is in the next section. Chapter 2 presents a graphical display of pH and  $SO_4^{2-}$  data. Chapter 3 is concerned with the multivariate variable reduction of the data and uses biplots to show the relations among the variables. Chapter 4 presents the details of the modelling process and the results. Conclusions and suggestions for future work are given in Chapter 5.

## 1.2 Data Structure and Variables

The map of the TLW is shown on Figure 1.1 which indicates the sampling locations of stations ( $S_0, S_1, \dots, S_5$ ). Batchawana Lake is the headwater lake of TLW and is separated into two basins. The outflow stream draining Batchawana Lake South (and subsequent portions of the watershed) is called Norberg Creek and goes through a rapid change in elevation prior to entering Wishart Lake. Then water flows from Wishart Lake to Little Turkey Lake and finally Turkey Lake. The outflow from the TLW enters the Bachawana River and then passes on to Lake Superior.

Samples were collected at the outflow of each lake. The water chemistries (major ions, nutrients, DOC) were measured and recorded. Station  $S_0$  is at the outflow of the first lake (Batchwana Lake) and the others are along the main drainage channel,  $S_5$  is at the lowest stream. Sampling frequency was approximately weekly from 1980 to 1997. We analyzed 6 data sets from 6 stations.

Table 1.1 shows a sample of the data sets which include the following variables: Time (year, month and date), pH value, the concentration of cations (including calcium ( $Ca^{2+}$ , mg/L), magnesium ( $Mg^{2+}$ , mg/L), sodium ( $Na^+$ , mg/L), potassium ( $K^+$ , mg/L)), ammonia ( $NH_4^+$ , mg/L-N), alkalinity (Alk, meq/L), sulfate anion ( $SO_4^{2-}$ ,



mg/L), nitrite and nitrate ( $NO_3^- + NO_2^- = NO_x^-$ , mg/L-N), chlorine ( $Cl^-$ , mg/L). Dissolved Organic Carbon (DOC) is included in the data set but not considered in this work, because a lot of its values were recorded as missing. An empty cell in the table indicates a missing measurement. In these 6 data sets between 15% - 20% of the observations contain one or more missing measurement(s). The issue of missing data does not have a big impact on the univariate analysis since the analysis is based on the monthly mean value of the measurements. In the original data there are some obvious mistakes, such as some of the pH values are out of range, we just treat them as missing. Before the analysis was conducted, the data was checked to make sure all the values are reliable.

Table 1.1: **Example of the data set.**

Y/M/D	pH	$Ca^{2+}$	$Mg^{2+}$	$Na^+$	$K^+$	$NH_4^+$	Alk	$SO_4^{2-}$	$NO_x^-$	$Cl^-$
		mg/L	mg/L	mg/L	mg/L	mg/L-N	meq/L	mg/L	mg/L-N	mg/L
80/09/29	6.14	2.35	0.40	0.32	0.14	0.025	0.051	5.95		0.48
80/10/08	6.19	3.09	0.48	0.37	0.20	0.028	0.062	5.51		0.48
80/10/16	6.26	3.50	0.50	0.49	0.15	0.037	0.070	5.95	0.077	
80/10/22	6.36	3.33	0.45	0.54	0.22		0.072	6.42	0.032	0.40
80/10/29	6.36	3.22	0.44	0.51	0.24		0.066	6.25		0.47
80/11/13	6.11	3.36	0.47	0.44	0.17	0.067	0.062	6.75		0.43
80/11/19	5.87	3.17	0.49	0.58	0.08	0.042		6.28		0.37
80/12/03	5.79	2.99	0.57	0.41	0.17	0.060	0.067	6.59		0.36
80/12/22	5.73	2.48	0.47	0.53	0.22	0.060	0.043	5.98		0.38
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
97/12/22	6.28	2.744	0.413	0.483	0.195	0.0474	0.0567	5.07	0.11	0.28

## Chapter 2

# Graphical Description of the Data

In the study of water acidity, pH value is the variable of primary interest. Thus we choose it as the most important variable to investigate. In addition,  $SO_4^{2-}$  is a very important variable, because it is directly related to the emission of  $SO_2$ , which is considered to be a big contributor to water acidification. This chapter focuses on the visual inspection of the data. Since pH values and  $SO_4^{2-}$  are recorded during the whole period of study as time series, the objective is to decompose each series into three components: long-term trend, seasonality and error terms. That is

$$data = trend + seasonality + error \tag{2.1}$$

We consider a variety of graphical displays of the data with the objective of identifying the patterns of seasonality, trend and differences between sampling locations.

## 2.1 The Graphical Display of pH Value

Figures 2.1 to 2.5 show various types of graphical display of the pH data. In all figures the symbols used to identify the stations are as follows:

$$\begin{array}{l} \circ \text{ station } S0 \quad \triangle \text{ station } S1 \quad + \text{ station } S2 \\ \times \text{ station } S3 \quad \diamond \text{ station } S4 \quad \nabla \text{ station } S5 \end{array}$$

Figures 2.1 - Figure 2.3 show the yearly trends for each month and stations. It is apparent from these figures that the pH level increases as the station elevation level decreases with S0 having the lowest level and S5 the highest level. In winter and early spring months, the differences between the curves decreases as we move downstream. Thus in this period of time, the difference between S0 and S1 is the largest. But in summer and fall months, the levels of the 6 stations roughly form three groups: ( S0 and S1), S2, ( S3, S4 and S5). As for the time trend there is a general increase in the level for the later years. In addition winter months show the strongest trend, while the weakest change occurred in the summer months. Around 1985, there appears to be a deep pH depression and this seems to be the case with  $SO_4^{2-}$  as well.

Figure 2.4 shows the seasonal cycles of the pH change. The patterns are similar in different stations: pH values progressively decrease from January to reach a minimum in April, rise to reach a maximum in the mid summer and then decline to the winter level. The variation in April is much bigger than other months. The boxplots appear to be symmetric, so the median level and the upper and lower quartiles provide an

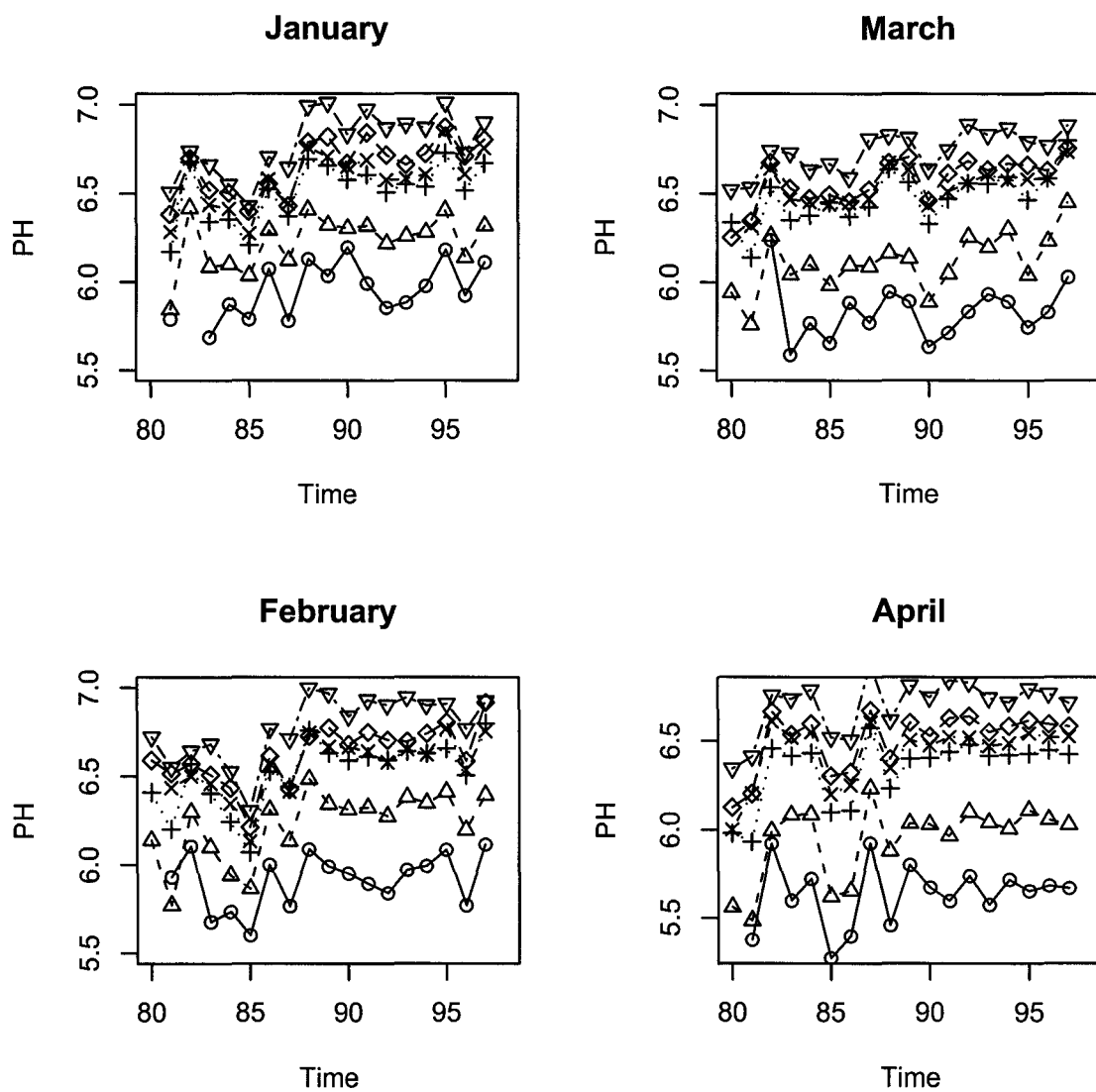


Figure 2.1: Monthly mean of pH (Jan.- April).

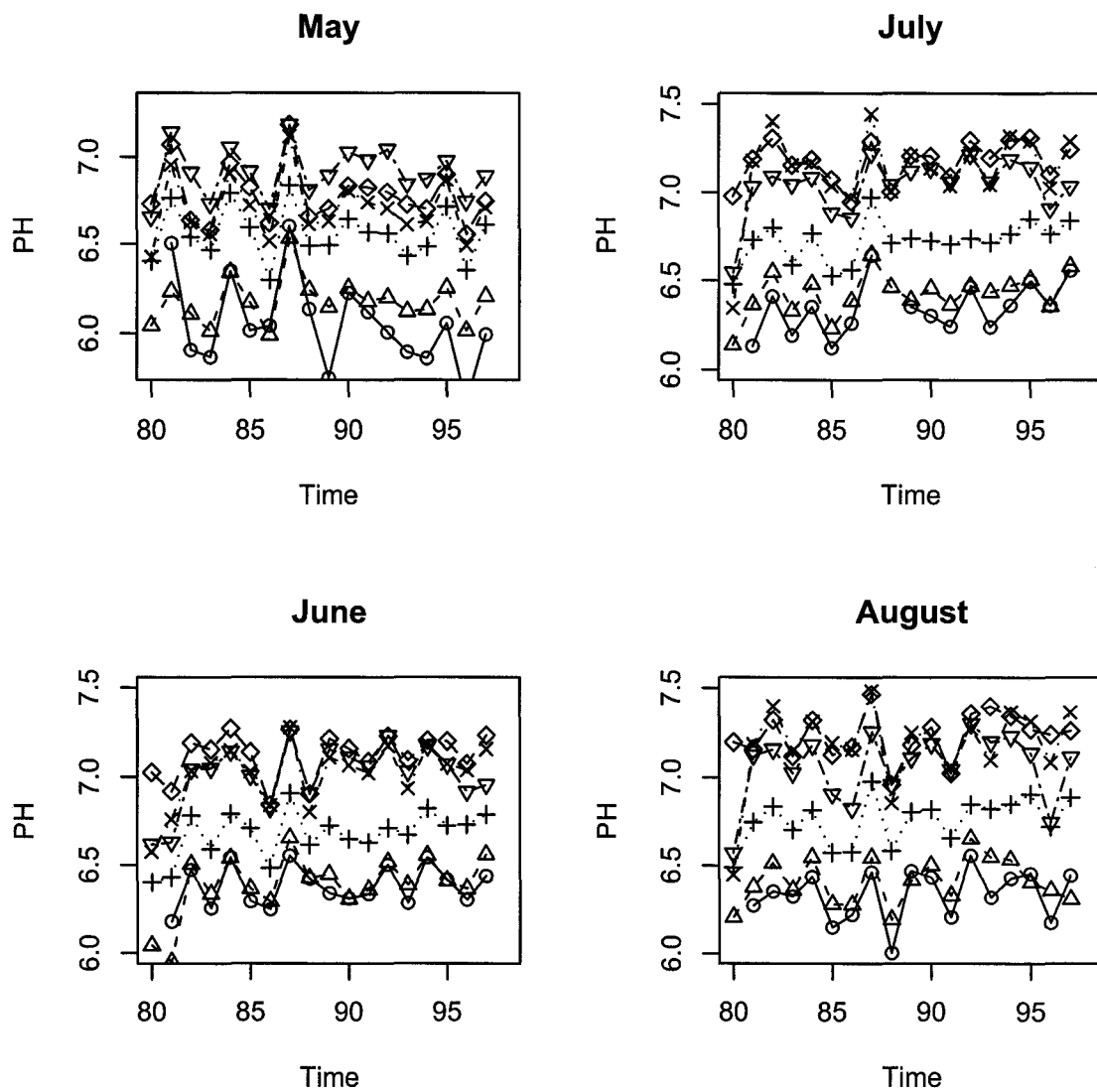


Figure 2.2: Monthly mean of pH (May- Aug.).

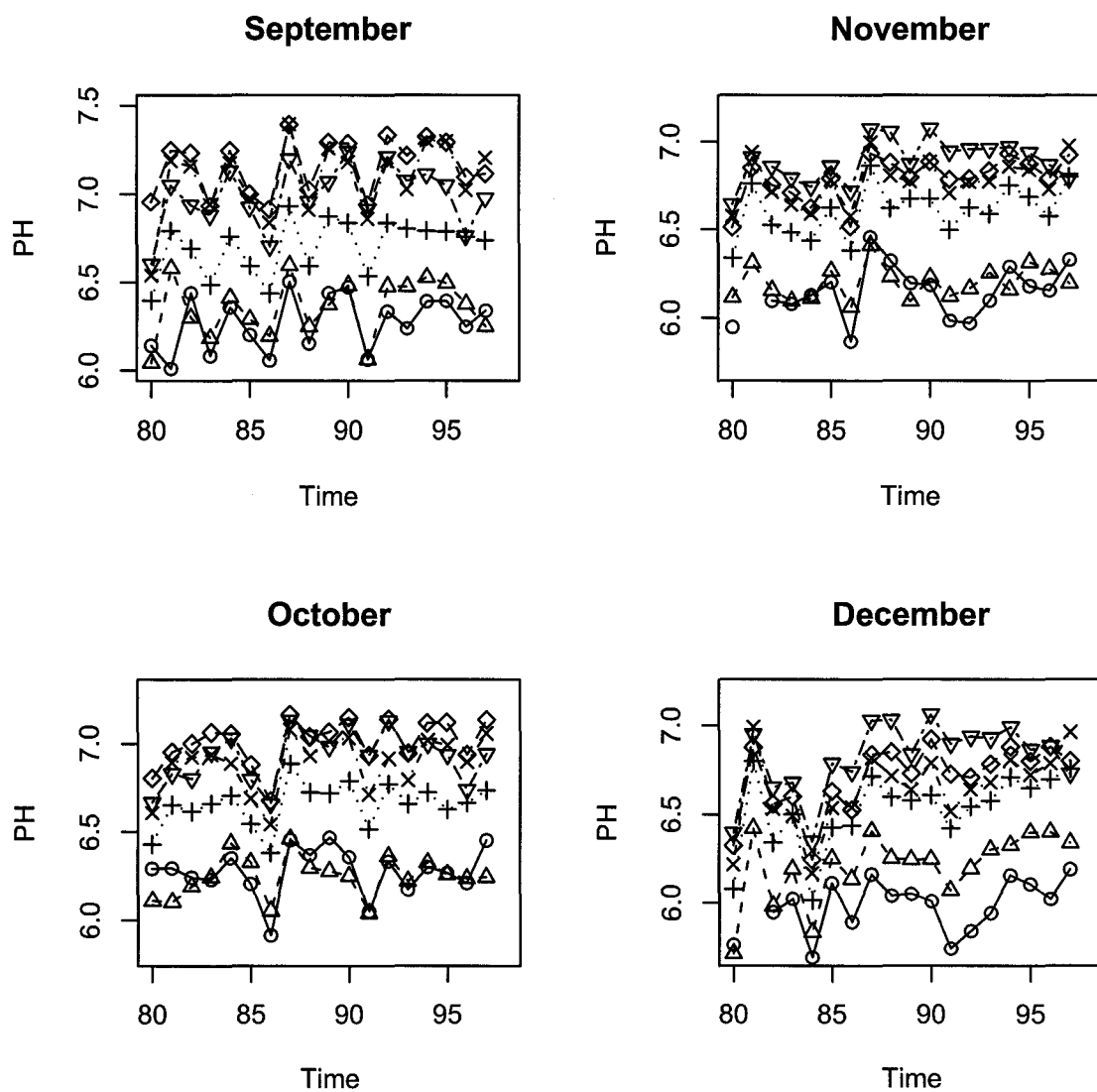


Figure 2.3: Monthly mean of pH (Sept.– Dec.).

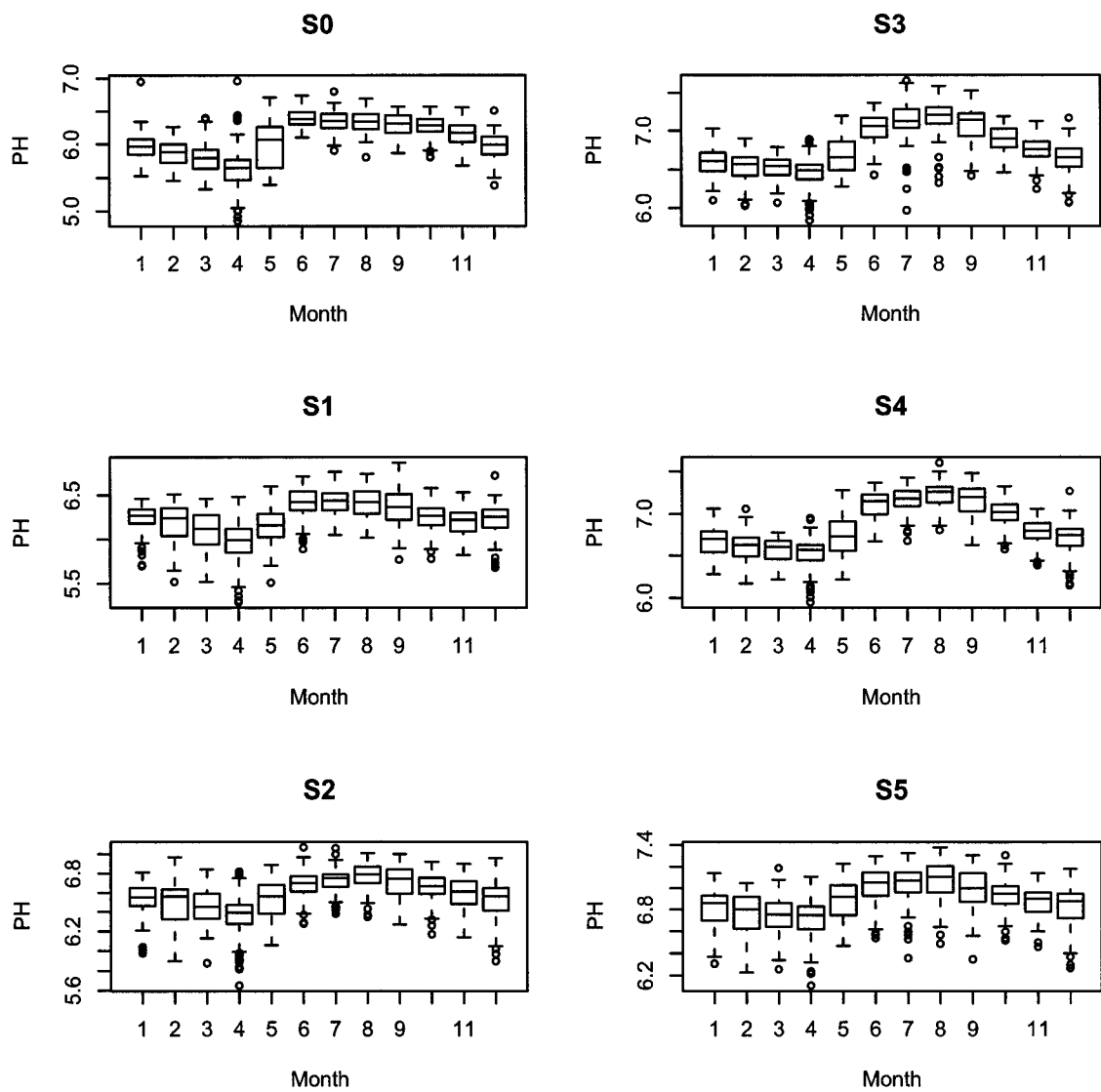


Figure 2.4: Seasonality of pH.



adequate summary statistics of the data. This is the bases for Figure 2.5 which shows at glance the pattern of changes in pH in each month and each station.

The decrease of pH in early spring is caused by the snowmelt. Studies (Semkin and Jeffries, 1986) have shown that up to 50% of the chemicals are released in the first 30% of the snowmelt, as a result, the water from the snowmelt is much more acidic than at other times of the year.

## 2.2 The Graphical Display of $SO_4^{2-}$

Figures 2.6 to 2.10 are graphical displays of the variable  $SO_4^{2-}$ .

The time trend patterns appear to be consistent at all stations and month. From its starting point  $SO_4^{2-}$  concentration declined and reached a minimum around 1985. This was followed by steep rise reaching a maximum around 1990, then declined sharply afterwards to reach minimum in 1997. The strength of this pattern appears to vary seasonally. It is stronger during winter and early spring months and is weaker during the remainder of the year. In addition more variability with the appearance of many extreme values occurs during the summer and fall months. There is also a consistent pattern for the stations with S0 showing the lowest  $SO_4^{2-}$  concentration while S5 showing the highest. Thus the concentration of  $SO_4^{2-}$  increases as the water moves from the higher level to the lower level elevation. The difference between the station levels is not constant from month to month. The separation between station patterns is more pronounced during the winter months and early spring season. This may be due to the melting ice which rapidly changes the concentration in the lower lakes due to the supply of water derived from the melting ice. In addition the fluctuation is not

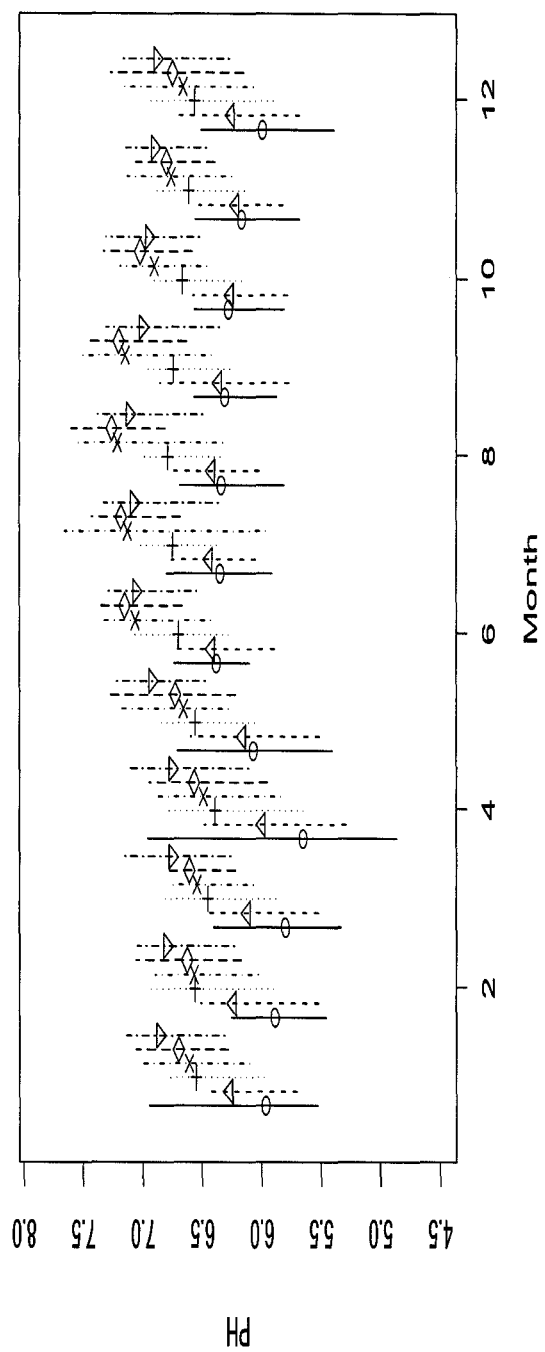


Figure 2.5: pH difference among stations.

the same for different stations with S1 showing very high values particularly during some years. This is unlikely due to analytical problem, since they are confined only to this station. It may be instead due to other sources of  $SO_4^{2-}$  that are confined to this lake during that period.

Figures 2.9 shows the pattern of seasonal cycle which appears to be generally similar for the six stations. It is high and nearly constant at its station level in the winter. A large drop in the level occurs in May and then a slow increase during the summer and fall to reach the winter maximum. The range of  $SO_4^{2-}$  concentration is narrower in the winter and wider during the spring and summer. The shape of the distribution tends to be more symmetric in the winter than the rest of the year. As expected based on the above the variation at station S1 is more pronounced.

Further insight about the pattern is shown in Figure 2.10. This compares the main feature of the distribution of  $SO_4^{2-}$  by station and month. The pattern agrees with the outcome reached from the previous figures.

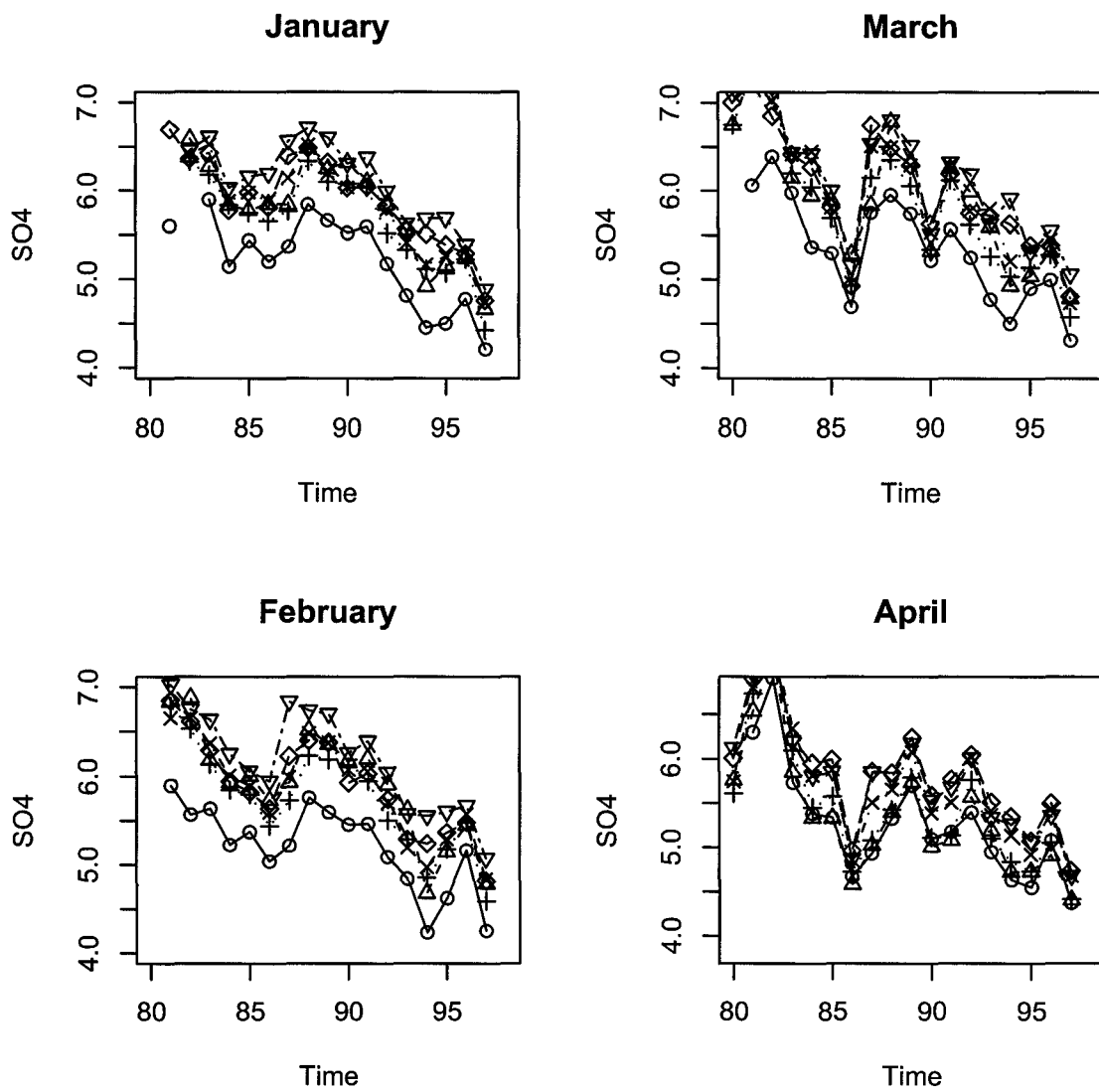


Figure 2.6: Monthly mean of  $SO_4^{2-}$  (Jan.- April).

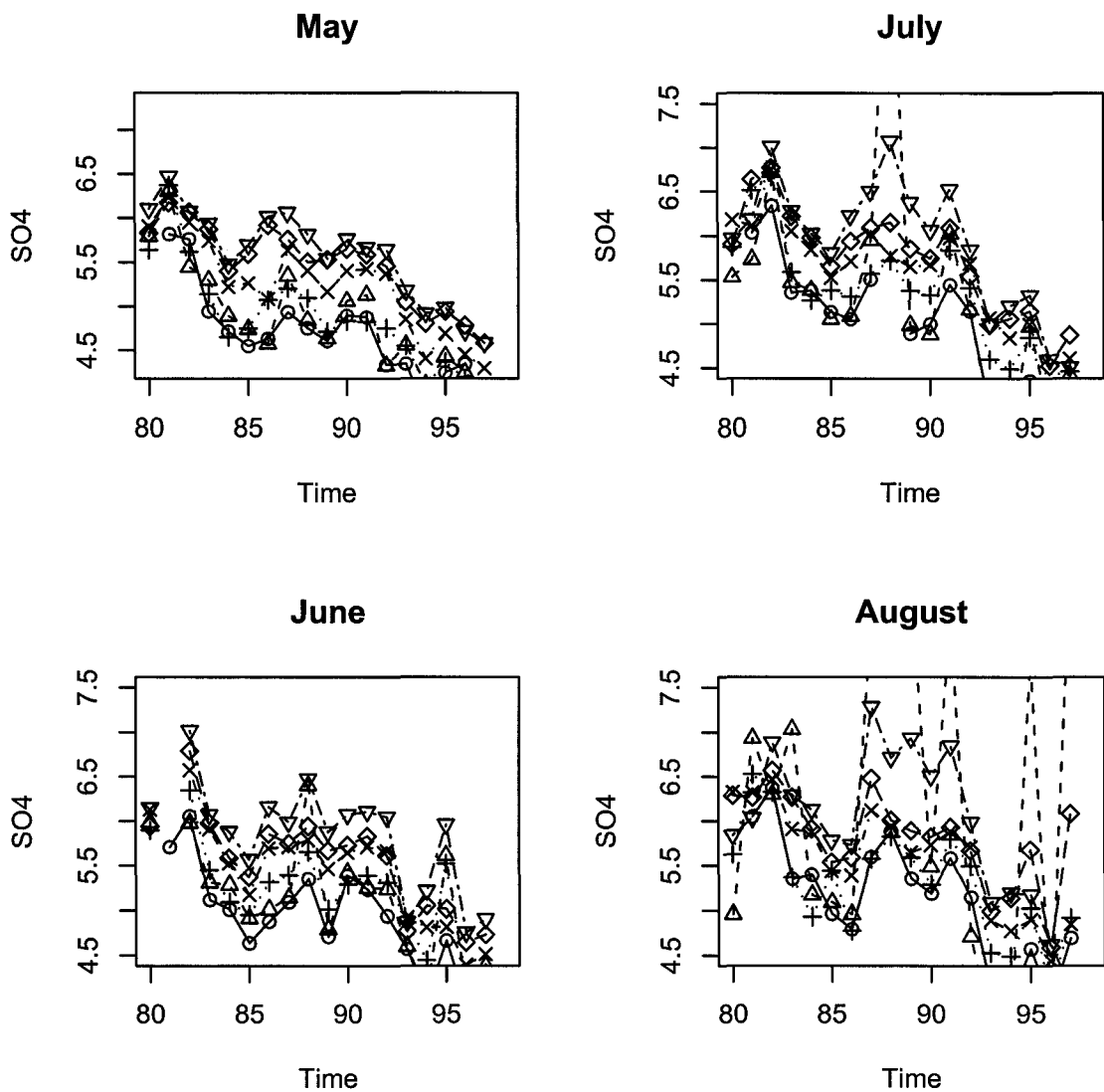


Figure 2.7: Monthly mean of  $SO_4^{2-}$  (May– Aug.).

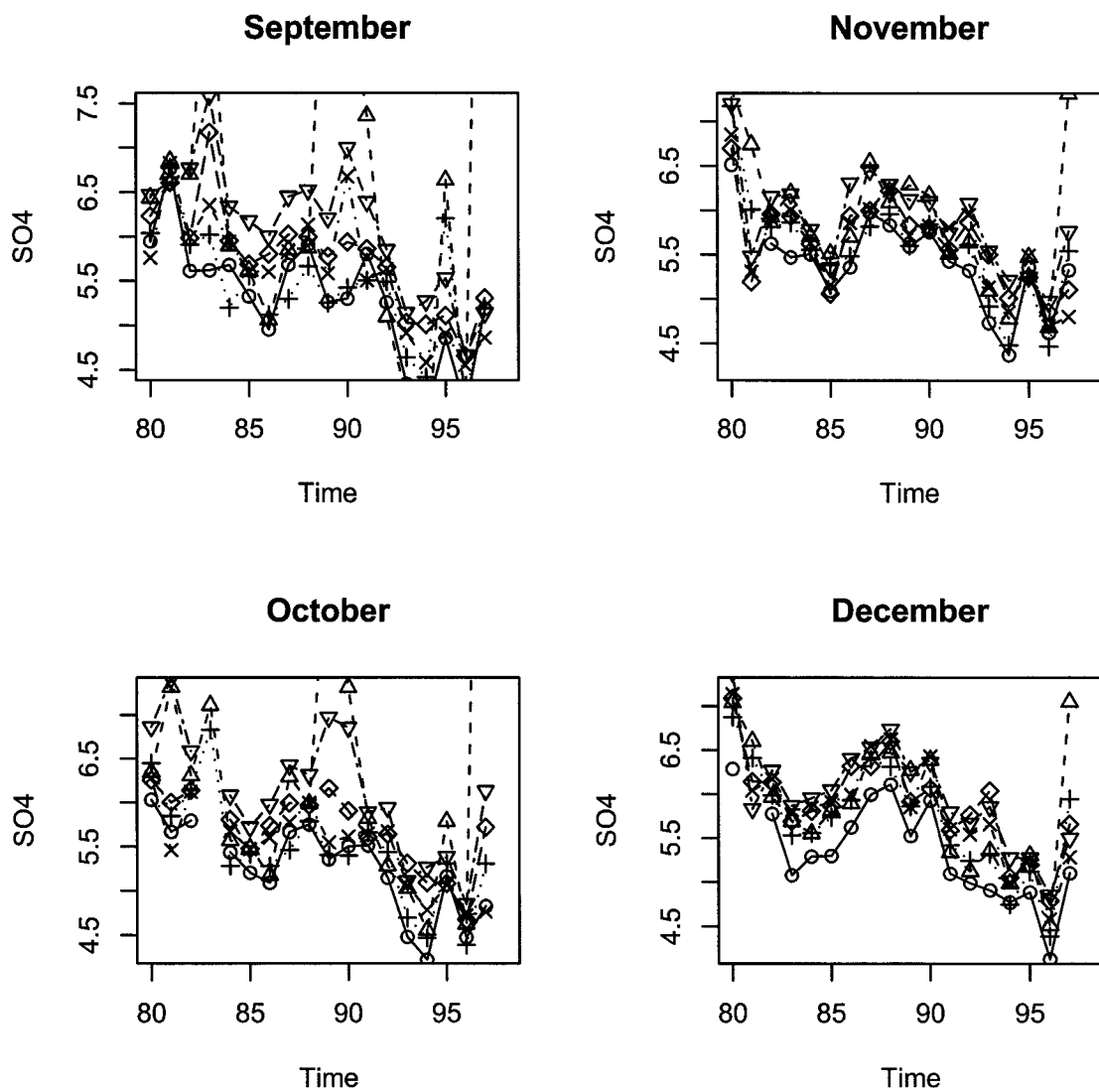


Figure 2.8: Monthly mean of  $SO_4^{2-}$  (Sept.– Dec.).

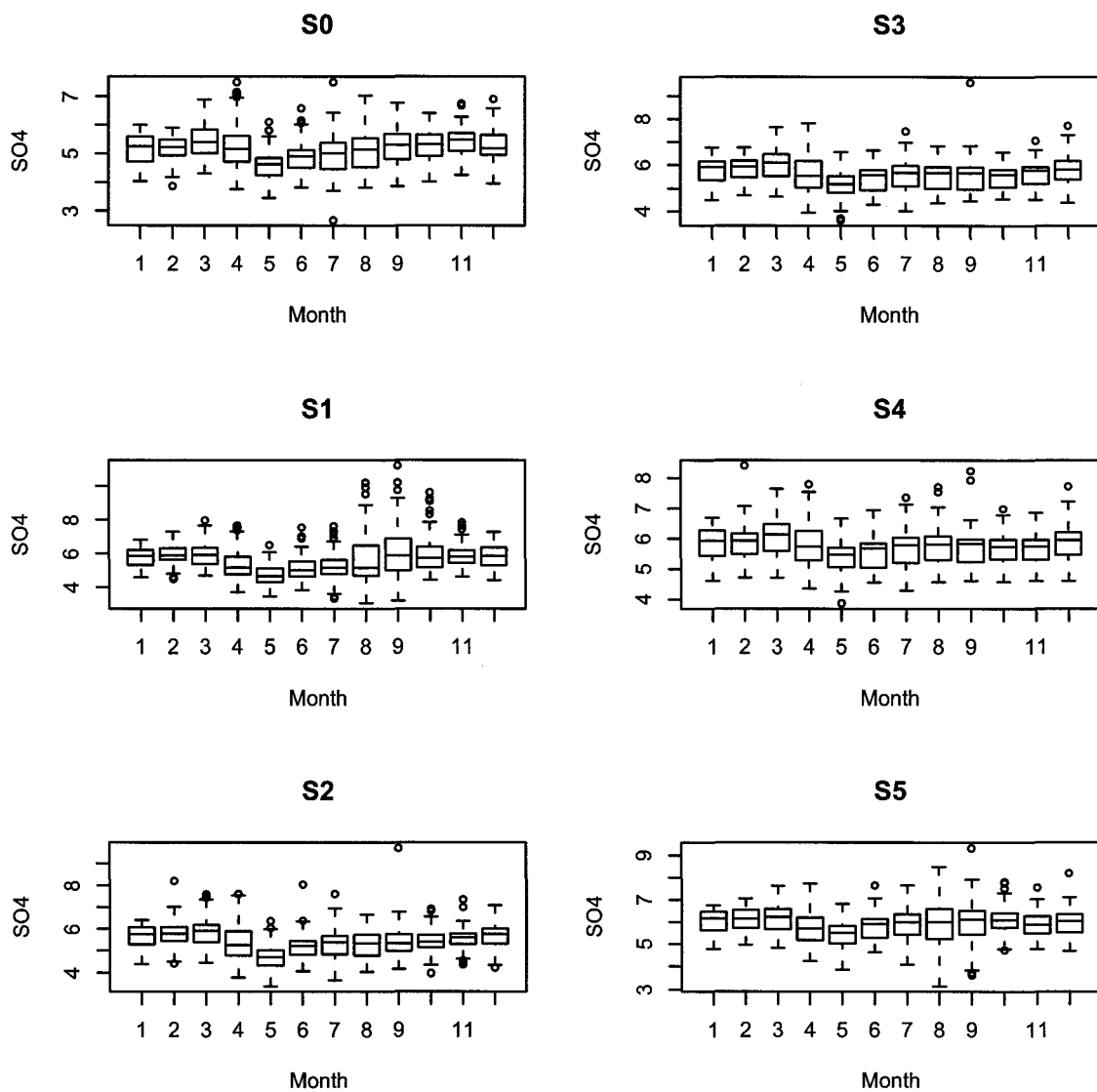


Figure 2.9: Seasonality of  $SO_4^{2-}$ .

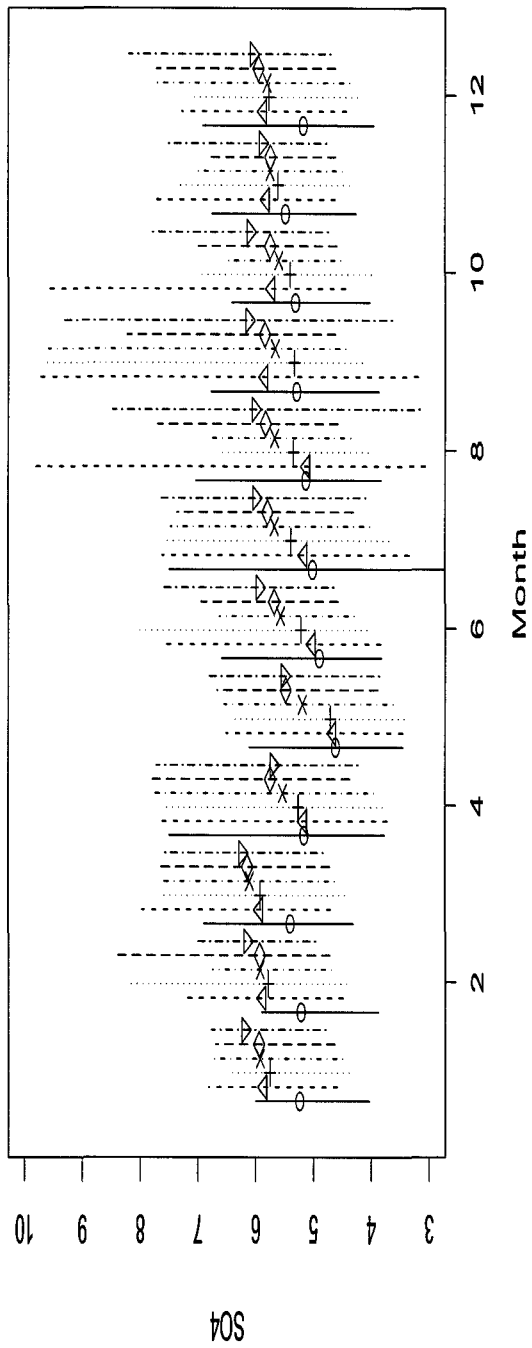


Figure 2.10:  $SO_4^{2-}$  difference among stations.



# Chapter 3

## Multivariate Data Reduction

In environmental studies, scientists typically collect observations on a large number of highly correlated variables. To obtain a concise summary of the data, it is desirable to focus on small set of functions of variables that account for most of the information. The most common method is to find linear combination of the variables that explain most of the variation. Principal components and associated singular value decomposition is a popular method and can be easily implemented in S-plus or R computer packages. Here we summarize the mathematical bases of these method and give their applications to TLW data.

### 3.1 Matrix Factorization

Generally, any  $n \times m$  matrix  $\mathbf{Y}$  of rank  $r$  can always be expressed as the product of two matrices  $\mathbf{G}$  and  $\mathbf{H}'$  of rank  $r$  (Rao, 1965). That is

$$\mathbf{Y}=\mathbf{GH}', \tag{3.1}$$

where  $\mathbf{G}$  is an  $n \times r$  matrix,  $\mathbf{H}$  a  $m \times r$  matrix and  $\mathbf{H}'$  is the transpose of  $\mathbf{H}$ , both of them are of rank  $r$ . The factorization also can be expressed as the inner product

$$y_{ij} = \mathbf{g}'_i \mathbf{h}_j \quad (3.2)$$

$y_{ij}$  is the  $i^{th}$  row and  $j^{th}$  column element of matrix  $\mathbf{Y}$

$\mathbf{g}'_i$  is the  $i^{th}$  row of matrix  $\mathbf{G}$

$\mathbf{h}_j$  is the  $j^{th}$  row of matrix  $\mathbf{H}$

We can understand the meaning of the factorization as follow: Assign the vectors  $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n$  to each of the  $n$  rows of  $\mathbf{Y}$  and the vectors  $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m$  to each of  $m$  column of  $\mathbf{Y}$ . The matrix  $\mathbf{Y}$  then is represented by those  $m + n$  vectors in  $r$ -space. In this sense the  $\mathbf{g}$ 's may be regarded as the "row effect" and the  $\mathbf{h}$ 's may be regarded as the "column effect". Gabriel (1971) introduced the biplot by applying it to a matrix of rank two. When the matrix  $\mathbf{Y}$  is of rank two  $\mathbf{g}$ 's and  $\mathbf{h}$ 's are all vectors of order two and thus may be plotted in a plane. The elements of  $\mathbf{Y}$  are represented by the inner products of the corresponding row effect and column effect vectors. This biplot provides a visual appraisal of the structure of the matrix. It represents the rank 2 matrix exactly, to the accuracy of plotting.

The factorization of the matrix is not unique since the same representation is obtained when  $\mathbf{G}$  is replaced by  $\mathbf{G}^* = \mathbf{G}\Delta^{-1}$  and  $\mathbf{H}$  by  $\mathbf{H}^* = \mathbf{H}\Delta'$ , where  $\Delta$  is a square matrix of order and rank  $r$ . In order to put it in practical use, some constraints need to be placed on the vectors  $\mathbf{g}$ 's and/or  $\mathbf{h}$ 's.

## 3.2 Matrix Approximation

For a matrix of rank higher than two, it is not possible to represent it exactly by a biplot. These are the cases we encounter most often in multivariate analysis practice. Fortunately, if the matrix  $\mathbf{Y}$  can be satisfactorily approximated by  $\mathbf{Y}^{(2)}$ , a rank two matrix, then the biplot of  $\mathbf{Y}^{(2)}$  may serve as a good approximation for the visual inspection of matrix  $\mathbf{Y}$ . The next step is to find the best approximate  $\mathbf{Y}^{(2)}$  matrix for  $\mathbf{Y}$ . Suppose  $\mathbf{Y}$  is an arbitrary real  $n \times m$  matrix of rank  $r$ , then  $\mathbf{Y}$  can be expressed as the sum of  $r$  matrices of rank 1 in a variety of ways. Among them the most useful one is the singular value decomposition, or SVD.

$$\mathbf{Y} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1' + \sigma_2 \mathbf{u}_2 \mathbf{v}_2' + \cdots + \sigma_r \mathbf{u}_r \mathbf{v}_r' \quad (3.3)$$

where the  $\sigma$ 's are real, positive numbers and  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$ . In matrix notations this is expressed as

$$\mathbf{Y} = \mathbf{U} \Sigma \mathbf{V}' \quad (3.4)$$

where  $\mathbf{U}$  is an  $n \times k$  orthonormal matrix, its column vectors are the eigenvectors of  $n \times n$  matrix  $\mathbf{Y}\mathbf{Y}'$ ,

$\mathbf{V}'$  is a  $k \times m$  orthonormal matrix with the eigenvectors of  $m \times m$  matrix  $\mathbf{Y}'\mathbf{Y}$  as its row vectors.

$\Sigma$  is a diagonal matrix with the elements  $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ .  $\sigma$ 's are the square roots of the positive eigenvalues of either  $\mathbf{Y}\mathbf{Y}'$  or  $\mathbf{Y}'\mathbf{Y}$ .

We wish to approximate  $\mathbf{Y}$  as closely as possible by another  $n \times m$  matrix  $\mathbf{X}$  of smaller rank  $k$  ( $r > k$ ). There might be various definitions of what is meant by “as

closely as possible”, but in statistical content, it would be in the least square sense of minimizing the sum of squared discrepancies between the original elements  $y_{ij}$  of  $\mathbf{Y}$  and their fitted counterparts  $x_{ij}$  of matrix  $\mathbf{X}$

$$S = \sum_{i=1}^n \sum_{j=1}^m (y_{ij} - x_{ij})^2 \quad (3.5)$$

Then the best-fitting matrix  $\mathbf{X}$  is the singular value decomposition of  $\mathbf{Y}$  but with all  $\sigma'_i$ s for  $i > k$  set to zero. That is  $\mathbf{X}$  is same as the sum of the first  $k$  terms on the right-hand side of ( 3.3).

Furthermore, the minimum value of  $S$  achieved by the singular value decomposition is

$$S_{min} = \sigma_{k+1}^2 + \sigma_{k+2}^2 + \cdots + \sigma_r^2 \quad (3.6)$$

In other words, the approximation matrix  $\mathbf{X}$  of rank  $k$  represents the maximum proportion of the total variation of matrix  $\mathbf{Y}$ , the fraction is used as a goodness of fit measure criterion.

$$\rho^{(k)} = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2} \quad (3.7)$$

There is a direct connection between SVD and principal components analysis(Good, 1969). Based on the above result, if we want to approximately represent a matrix  $\mathbf{Y}$  (of rank  $r$ ) by a matrix  $\mathbf{X}$  of rank 2, then  $\mathbf{X}$  is given by

$$X = \sum_{i=1}^2 \sigma_i \mathbf{u}'_i \mathbf{v}_i \quad (3.8)$$

with  $\mathbf{u}_i = (u_{i1}, u_{i2}, \dots, u_{in})$  and  $\mathbf{v}_i = (v_{i1}, v_{i2}, \dots, v_{im})$  for  $i = 1, 2$  i.e.:

$$X = \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \end{pmatrix} \quad (3.9)$$

When we try to factorize the matrix  $\mathbf{X}$  in order to obtain a bipolt, the three simple expressions come out intuitively:

$$X = \begin{pmatrix} u_{11}\sqrt{\sigma_1} & u_{21}\sqrt{\sigma_2} \\ u_{12}\sqrt{\sigma_1} & u_{22}\sqrt{\sigma_2} \\ \vdots & \vdots \\ u_{1n}\sqrt{\sigma_1} & u_{2n}\sqrt{\sigma_2} \end{pmatrix} \begin{pmatrix} v_{11}\sqrt{\sigma_1} & v_{12}\sqrt{\sigma_1} & \dots & v_{1m}\sqrt{\sigma_1} \\ v_{21}\sqrt{\sigma_2} & v_{22}\sqrt{\sigma_2} & \dots & v_{2m}\sqrt{\sigma_2} \end{pmatrix} \quad (3.10)$$

$$= \begin{pmatrix} u_{11}\sigma_1 & u_{21}\sigma_2 \\ u_{12}\sigma_1 & u_{22}\sigma_2 \\ \vdots & \vdots \\ u_{1n}\sigma_1 & u_{2n}\sigma_2 \end{pmatrix} \begin{pmatrix} v_{11} & v_{12} & \dots & v_{1m} \\ v_{21} & v_{22} & \dots & v_{2m} \end{pmatrix} \quad (3.11)$$

$$= \begin{pmatrix} u_{11} & u_{21} \\ u_{12} & u_{22} \\ \vdots & \vdots \\ u_{1n} & u_{2n} \end{pmatrix} \begin{pmatrix} v_{11}\sigma_1 & v_{12}\sigma_1 & \dots & v_{1m}\sigma_1 \\ v_{21}\sigma_2 & v_{22}\sigma_2 & \dots & v_{2m}\sigma_2 \end{pmatrix} \quad (3.12)$$

They are corresponding to the different weights assigned to the rows and the columns effect. These three versions of the choices of  $\mathbf{G}$  and  $\mathbf{H}$  in (3.1) cover the

most useful cases (Krzanowski, 1988). They correspond respectively to a general factorization where the emphasis is placed neither on rows nor columns, and two special factorizations placing emphasis on rows and columns in turn.

### 3.3 Biplots

Biplot is a very efficient method to summarize the information available in a multivariate data set. It generates a visual summary that shows both the relationship between variables and between cases. In our case, the original data sets have over 1000 rows (cases) each. It is not possible to plot the biplot directly, not even the monthly means, since these are 200 rows long. It is hard to separate one point from the other. Considering that some of the variables have yearly cycles, the biplots are produced by month. Because missing values are spread over the variables, station S0, S1 and S3 do not have sufficient monthly mean values to produce all 12 months' biplots. Figure 3.1– Figure 3.3 give the biplots for station S2 data. In this plots the variables are presented by arrows (vectors) and the years by points. The smaller the angle between two vectors, the larger is similarity. Orthogonal vectors indicate lack of association between the variables. Angles larger than  $90^\circ$  indicate negative association.

From these biplots we can see that the variables have different contribution to the first two principal components in different months. Generally, if two variable have similar contributions to the first two principal components, they will appear in the biplot as close (or overlapped) arrows.

To investigate the relationship, we calculate the angles between pH and each of the other 9 variables. The distance between two vectors is also a useful measure of the

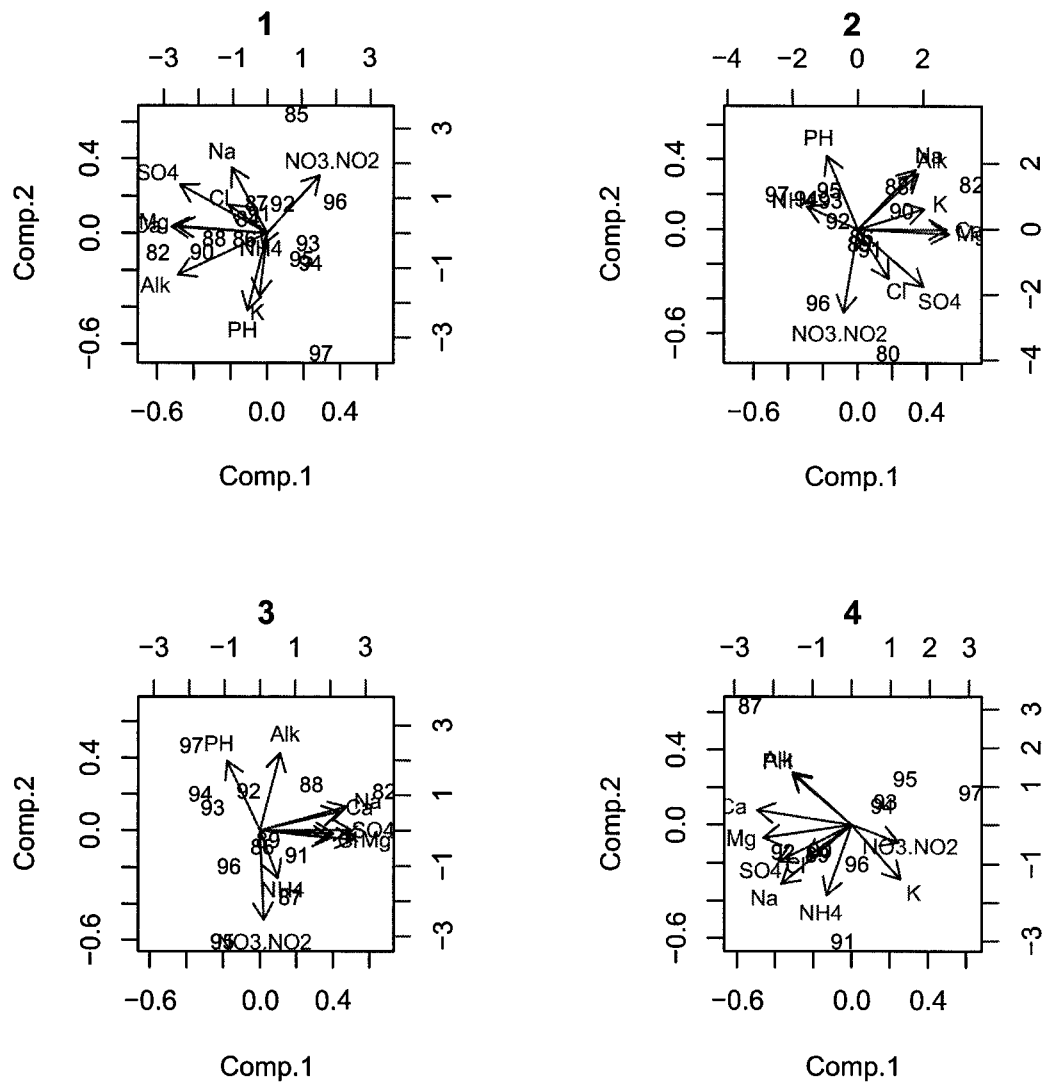


Figure 3.1: Biplot 1 (Jan.– April).

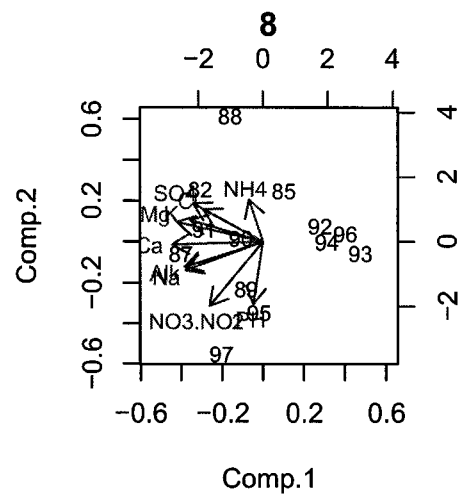
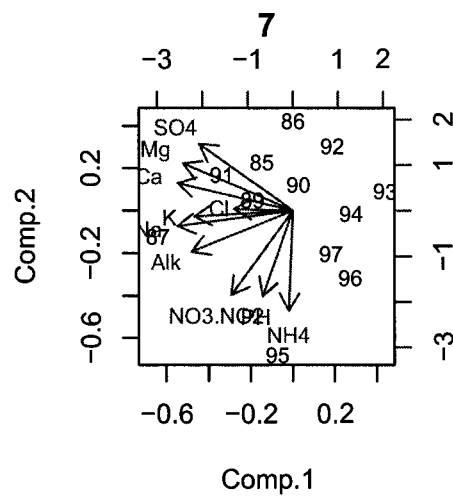
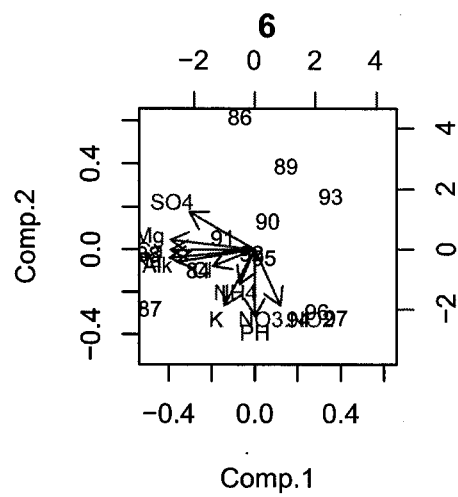
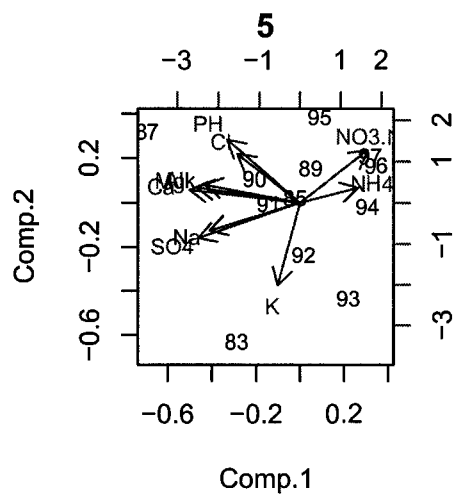


Figure 3.2: Biplot 2 (May – Aug.).



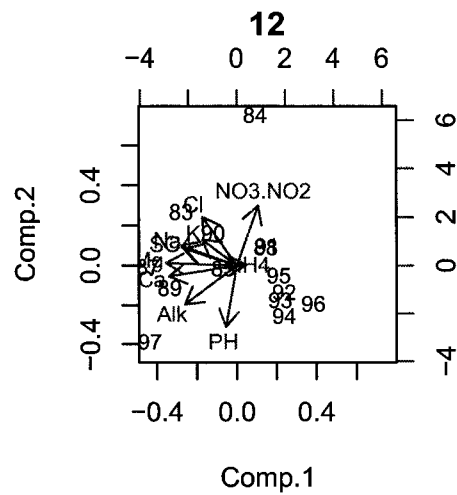
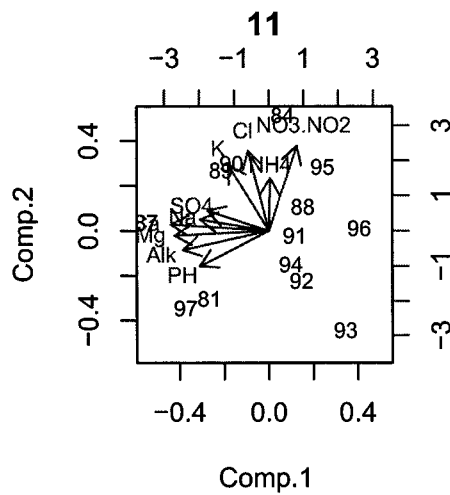
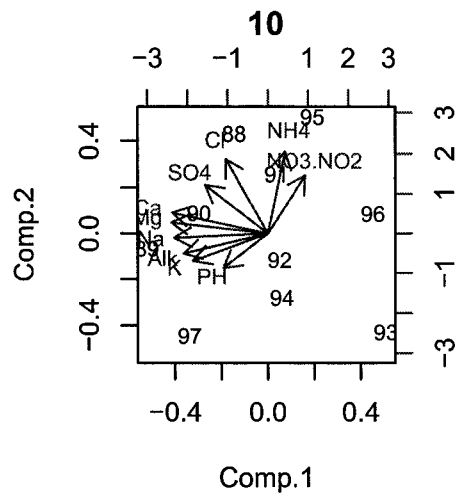
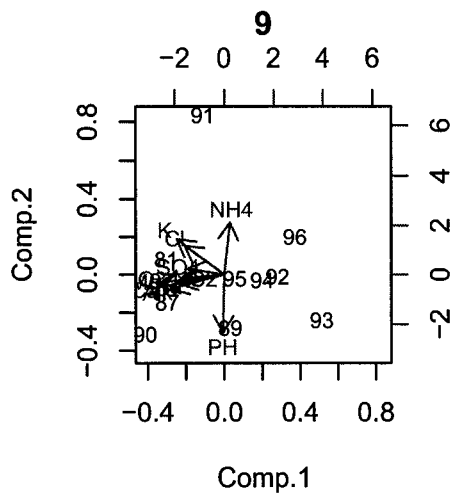


Figure 3.3: Biplot 3 (Sept. – Dec.).

relationship between two variable since it combines the information of the directions and the magnitude of the difference between two variables. As mentioned above, the angle and distance between pH and a given variable change from one month to another. Figure 3.4 presents boxplots of the angles and distances.

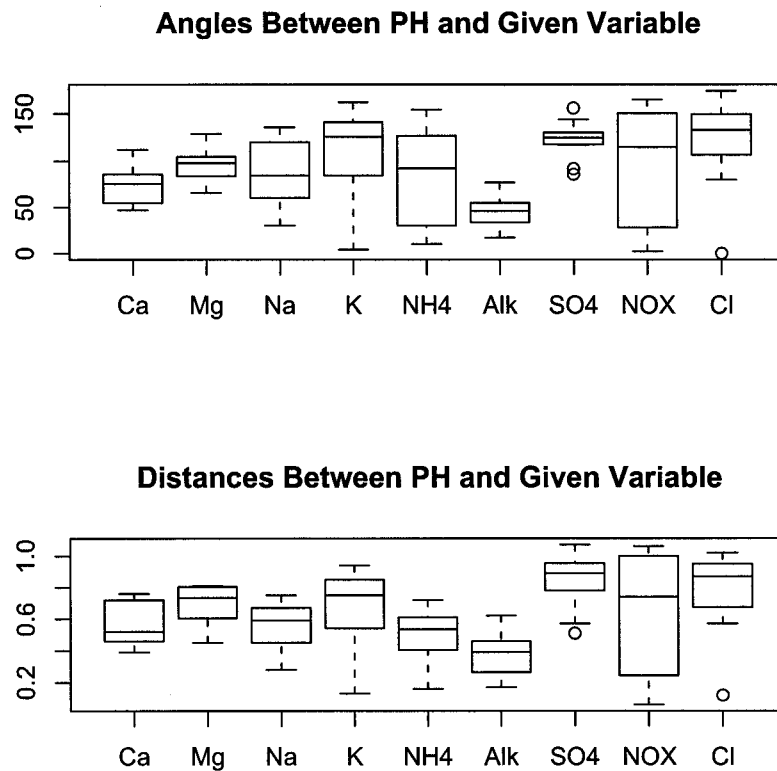


Figure 3.4: Boxplot of angles and distances between pH and the given variable.

Overall, Alkalinity (Alk) is the closest one to pH, the average angle is about 40° and the variation is the smallest. This angle is close to zero in April and reaches its maximum in June. The seasonal variation of Alk angle suggests maximum association

with pH in April and minimum association in June. Also the distance between Alk and pH is small. Another interesting variable is  $SO_4^{2-}$ , the average angle between  $SO_4^{2-}$  and pH is not the biggest one but it exceeds  $90^\circ$ . While the variation of the angles is much smaller than that of other variables. In addition, the distance between pH and  $SO_4^{2-}$  is the largest. This puts it in the position that have a constant negative correlation with pH as expected.  $NO_x$  has the largest variable angle and distance. The median angles of Ca, Mg, Na are close to  $90^\circ$  indicating little association with pH. In contrast, negative associations are found with  $NO_x$  and with Cl.

Combining the information from the biplot and the boxplot, we can see that of all the variable Ca, Mg, Na, and Alk have average angle smaller than  $90^\circ$  and  $NH_4$ ,  $SO_4^{2-}$ ,  $NO_x$  and Cl have average angle bigger than  $90^\circ$ . When comparing this with the correlation of pH with these variables, we find them agree with each other very well.

It should also be noted from the biplots that points representing the later years are clustered together. That means the Turkey Lakes chemistry has changed over the years.

## Chapter 4

# Regression Models with Change Point

As shown earlier, pH values vary temporally and spatially. Specifically an increasing time trend and a strong periodic seasonality have been observed. Furthermore, upper lakes show higher acidity (lower pH ) than lower lakes. Here the focus is on capturing the trend and seasonal structure by developing regression models with regime changes. Likelihood based methods are used to make inferences about the model's parameters including the point at which the regression function changes (change point). It should be mentioned that there is a vast literature dealing the detection of the point change in regression models. Some are parametric, others are non-parametric or semi-parametric. These are summarized in the Encyclopedia of Environmetrics (2001, vol.1).

Here we use the approach presented in Esterby and El-Shaarawi (1981) which uses likelihood method for making inferences about the change point and regression parameter under the assumption of normal error. Specifically we apply this approach

to develop models for the changes in pH,  $SO_4^{2-}$  and the first two principal components.

The two variables are selected due to their importance for acid rain problem.

## 4.1 Likelihood Methods

Let  $y_t$  be the observed value of response variable at time T, where T represents the Julian day which is defined as:

$$T = (year - 1980) + \frac{(month - 1)}{12} + \frac{date}{365}$$

and let  $y_t$  be expressed as:

$$\begin{aligned} y_t &= \mu_{1t} + e_{1t} && (for \ t = 1, 2, \dots, k) \\ &= \alpha_0 + \alpha_1 T_t + \dots + \alpha_p T_t^{p-1} + \sum_{j=1}^{m_1} (\beta_{j1} \sin(2\pi j T_t) + \beta_{j2} \cos(2\pi j T_t)) + e_{1t} \\ y_t &= \mu_{2t} + e_{2t} && (for \ t = k + 1, k + 2, \dots, n) \\ &= \gamma_0 + \gamma_1 T_t + \dots + \gamma_q T_t^{q-1} + \sum_{j=1}^{m_2} (\lambda_{j1} \sin(2\pi j T_t) + \lambda_{j2} \cos(2\pi j T_t)) + e_{2t}, \end{aligned} \quad (4.1)$$

a combination of a polynomial and periodic functions in T and an error term. The point  $k$  is known as the change point at which the regression model changes its course.

This model can be expressed in the concise form:

$$y_t = I(t \leq k)(\mu_{1t} + e_{1t}) + \{1 - I(t > k)\}(\mu_{2t} + e_{2t}), \quad (4.2)$$

where the indicator function  $I(t \leq k) = 1$  if  $t \leq k$  and 0 otherwise. The error terms are assumed to be an autoregressive error process  $AR(p)$  of order  $p$ . We shall consider here only two special cases.  $p = 0$  and 1 for both regression segments. that is

$$e_{it} = \phi_i e_{it-1} + a_{it} \quad \text{where} \quad a_{it} \sim N(0, \sigma_i^2) \quad (4.3)$$

Note that

1.  $p = 0$  corresponds to the case of independence while  $p = 1$  corresponds to AR(1);
2. the error term parameters in segment 1 and 2 are different

letting

$$\Theta'_1 = (\alpha_0, \dots, \alpha_p, \beta_{11}, \beta_{12}, \dots, \beta_{m_11}, \beta_{m_12}) \text{ and}$$

$$\Theta'_2 = (\gamma_0, \dots, \gamma_q, \lambda_{11}, \lambda_{12}, \dots, \lambda_{m_21}, \lambda_{m_22})$$

the above model can be written in matrix form as:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_k \\ y_{k+1} \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{pmatrix} \begin{pmatrix} \Theta_1 \\ \Theta_2 \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} \quad (4.4)$$

Where  $\mathbf{A}_1$  and  $\mathbf{A}_2$  are the design matrices associated with the vectors  $\mathbf{Y}_1 = (y_1, y_2, \dots, y_k)'$  and  $\mathbf{Y}_2 = (y_{k+1}, y_{k+2}, \dots, y_n)'$  of observations. It is assumed that the matrices  $\mathbf{A}_1$  and  $\mathbf{A}_2$  are of full ranks.

### 4.1.1 The Likelihood under the of Independent Assumption

First, we assume that the observations are independent from each other and follow the normal distribution with constant variances in the two segments respectively. That is

$$\begin{aligned} e_i &\sim N(0, \sigma_1^2) & (\text{for } i = 1, 2, \dots, k) \\ e_i &\sim N(0, \sigma_2^2) & (\text{for } i = k+1, k+2, \dots, n) \end{aligned} \quad (4.5)$$

Let  $k_i$  denote the number of observations in segment  $i$  (for  $i=1$  and  $2$ ), we have  $k_2 = n - k_1$ . If the numbers of coefficients (parameters) in the regression are  $p_i$ , we have then :

The coefficients of regression are estimated by

$$\hat{\Theta}_i = (\mathbf{A}_i' \mathbf{A}_i)^{-1} \mathbf{A}_i' \mathbf{Y}_i \quad (4.6)$$

The estimation of the mean values (fitted values) are :

$$\hat{\mu}_i = \mathbf{A}_i \hat{\Theta}_i \quad (4.7)$$

The variances are estimated as

$$\hat{\sigma}_i^2 = \frac{1}{k_i - p_i} \mathbf{Y}_i' (I - \mathbf{A}_i (\mathbf{A}_i' \mathbf{A}_i)^{-1} \mathbf{A}_i') \mathbf{Y}_i \quad (4.8)$$

Plugging these estimates in the joint density function of the observations we obtain the profile likelihood function for the change point  $k$  ( $=k_1$ ) as

$$L(k) = L(k_1)L(k_2) \propto \hat{\sigma}_1^{-k_1} \hat{\sigma}_2^{-k_2} \quad (4.9)$$

Maximum likelihood estimate of  $k$  is  $\hat{k}$ , the value which maximizes the above function. This is obtained numerically through the calculation of the likelihood function

for all the possible value of  $k$ . On the other hand, inferences about  $k$  is conducted based on the relative likelihood function in the manner discussed in Kalbfleish, (1985).

$$ReL = \frac{L(k)}{\sup\{L(k)\}} \quad (4.10)$$

The  $ReL$  varies between zero and 1 with the value of  $\hat{k}$  corresponds to  $ReL = 1$  which is the most plausible value. Relative likelihood intervals can be used to summarize the information about  $k$ . For example, the values of  $k$  such that  $ReL \geq .5$  are highly plausible while those for which  $ReL \leq .1$  are implausible. So the relative likelihood function may be used to rank the parameter values according to their plausibilities. So that by plotting the relative likelihood against  $k$ , we can find out what is the plausible value of  $k$  corresponding to the position of the maximum likelihood  $L(\hat{k})$ , that is where the change takes place.

#### 4.1.2 The Likelihood under the AR(1) Assumption

Since the data are taken as a sequence of time ordered observations at each of the six sampling stations, it is then expected that they are likely to be autocorrelated which needs to be taken into account in model development.

Under the assumption that in each segment the observations are autocorrelated with different autocorrelation coefficient and ignoring the correlation between observations from different segments, the model for the  $i^{th}$  ( $i = 1$  and  $2$ ) segment is

$$\mathbf{Y}_i = \mathbf{X}_i\Theta_i + \mathbf{u}_i \quad (4.11)$$

Now the error terms are not independent. Instead, they are successive variables



from two stationary processes with the variance-covariance matrices:

$$Var(\mathbf{Y}_i) = Var(\mathbf{u}_i) = \Sigma_i$$

Estimates of regression coefficients are obtained using the generalized least squares method.

$$\hat{\Theta}_i = (\mathbf{A}_i' \Sigma_i^{-1} \mathbf{A}_i)^{-1} \mathbf{A}_i \Sigma_i^{-1} \mathbf{Y}_i \quad (4.12)$$

and the joint density of  $\mathbf{Y}_i$  is

$$f(\mathbf{Y}_i) = \frac{1}{(2\pi)^{\frac{k_i}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{Y}_i - \mathbf{A}_i \Theta_i)' \Sigma_i^{-1} (\mathbf{Y}_i - \mathbf{A}_i \Theta_i)}$$

For the AR(1) process, we have

$$u_{it} = \phi_i u_{it-1} + a_{it} \quad (for \ t = 0, \pm 1, \pm 2, \dots) \quad (4.13)$$

where  $|\phi_i| < 1$  to ensure that the autoregressive process is stationary and

$$a_{it} \sim N(0, \sigma_i^2)$$

From the definition of stationarity, we have

$$u_{it} \sim N(0, \frac{\sigma_i^2}{1 - \phi_i^2}) \quad (for \ t = 1, 2, \dots, k_i) \quad (4.14)$$

The variance and covariance matrix of  $\mathbf{u}_i$  (or equivalently  $\mathbf{Y}_i$ ) are

$$Var(\mathbf{Y}_i) = Var(\mathbf{u}_i) = \frac{\sigma_i^2}{1 - \phi_i^2} \begin{pmatrix} 1 & \phi_i & \phi_i^2 & \dots & \phi_i^{k_i-1} \\ \phi_i & 1 & \phi_i & \dots & \phi_i^{k_i-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_i^{k_i-1} & \phi_i^{k_i-2} & \phi_i^{k_i-3} & \dots & 1 \end{pmatrix}$$

The density of  $(u_1, u_2, \dots, u_{k_i})$  is

$$f(u_{i1}, u_{i2}, \dots, u_{ik_i}) = \frac{1}{(2\pi)^{\frac{k_i}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2} \mathbf{u}_i' \Sigma_i^{-1} \mathbf{u}_i}$$

In the case all the  $u$ 's ( or  $y$ 's) are equally spaced we can calculate the determinant and the inverse of the variance-covariance matrix  $\Sigma_i$  :

$$\Sigma_i^{-1} = \left( \frac{1 - \phi_i^2}{\sigma_i^2} \right)^{k_i} \begin{pmatrix} 1 & -\phi_i & 0 & \cdots & 0 & 0 & 0 \\ -\phi_i & 1 + \phi_i^2 & -\phi_i & \cdots & 0 & 0 & 0 \\ 0 & -\phi_i & 1 + \phi_i^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 + \phi_i^2 & -\phi_i & 0 \\ 0 & 0 & 0 & \cdots & -\phi_i & 1 + \phi_i^2 & -\phi_i \\ 0 & 0 & 0 & \cdots & 0 & -\phi_i & 1 \end{pmatrix}$$

and

$$|\Sigma_i| = \frac{\sigma_i^{2k_i}}{1 - \phi_i^2} \quad (4.15)$$

By plugging these expressions in to likelihood function, we obtain the density, equivalently, the likelihood function of  $\mathbf{u}$ :

$$L_i(u_1, u_2, \dots, u_{k_i}) = (2\pi\sigma_i^2)^{-\frac{k_i}{2}} (1 - \phi_i^2)^{\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma_i^2} \left\{ (1 - \phi_i^2)u_{i1}^2 + \sum_{t=2}^{k_i} (u_{it} - \phi_i u_{it-1})^2 \right\} \right\}$$

For any give  $k_1, k_2 (=n - k_1)$  and  $\phi_1, \phi_2$  the segment regression coefficients are first estimated by using generalized least squares, they are functions of  $k_1, \phi_1$  and  $k_2, \phi_2$ . Consequently, residuals are expressed as functions of  $k_i$  and  $\phi_i$ . The variance  $\sigma_i^2$  can then be estimated as  $\hat{\sigma}_i^2$

$$\hat{\sigma}_i^2 = \frac{1}{k_i} \sum_{t=1}^n (u_{it} - \phi_i u_{it-1})^2 \quad (4.16)$$

Substituting these estimates into the likelihood function, we obtain a profile likelihood function for  $k_i$  and  $\phi_i$ . For a given  $k_i$ , the maximum likelihood estimate of  $\phi_i$  can be obtained by equating the first order derivative of the profile likelihood with respect to  $\phi_i$  to zero and solving the equation for  $\phi_i$ . This leads to a cubic equation and the solutions could be obtained numerically as suggested in El-Shaarawi and Esterby (1982). In practice, the possible value of  $\phi_i$  is limited to  $(-1, 1)$ . So we tried to set  $\phi_i$  and  $k_i$  to their possible values and calculate the value of likelihood function. Then the relative likelihood is tabulated and used to rank the plausibility of the parameters.

In the ideal situation, the observations are equally spaced in terms of time. In our case, even though we used the mean value of each month to make them approximately equally spaced, still have some missing values in certain month. So it is difficult to get an explicit expression for  $|\Sigma_i|$  and  $\Sigma_i^{-1}$ . In this case a simple modification of the likelihood is used to deal with missing values. The simplest approach is to apply the EM Algorithm which involves replacing the missing values by their conditional expectation and then maximization in an iterative sequence.

Another assumption we have made is that the observations in two segments (before and after the occurrence of the change) are independent from each other, but in each segment the observations are dependent. The autocorrelations  $\phi_i$  are assumed different in two segments. This allows us to calculate the likelihood as

$$L(k, \phi_1, \phi_2 | Y) = L_1(k_1, \phi_1 | Y_1) L_2(k_2, \phi_2 | Y_2) \quad (4.17)$$

or the log-likelihood function

$$l = l_1 + l_2 \quad (4.18)$$

Calculation consideration:

1. For a possible value of  $k_1$ , the likelihoods  $L(k_1, \phi_1|Y_1)$  and  $L(k_i, \phi_2|Y_2)$  for each of the two segments with all possible values of  $\phi_i$  are computed ;
2. Determine  $\phi_1$  and  $\phi_2$  that maximize  $L_1$  and  $L_2$  respectively;
3. The overall likelihood for  $k$  is then computed as  $L(k) = L_1(k, \phi_1^{max}|Y_1)L_2(n - k, \phi_2^{max}|Y_2)$ ;
4. For all possible value of  $k_1$  (  $2 \leq k_1 \leq n - 1$ ), repeat step 1-3 to form a sequence  $\{L(k)\}$ ;
5. Select the MLE of the change point  $k$  by identifying the value of  $k_1$  which maximizes the over all likelihood.

## 4.2 Results

The results of the application of the above methods to the data sets from stations S0 and S5 are discussed below.

### 4.2.1 Modelling the Changes in pH

The monthly mean of pH data for each station are modeled first assuming independence using ( 4.1). Since the assumption of independence was not supported by the data the

Table 4.1: Change of pH under assumption of independence.

St. No.	Change Time	$\log(L_{max}^{IN})$	$\hat{\alpha}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\gamma}_0$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$
S0	June 87	52.796	6.057	0.147	0.262	6.108	0.211	0.186
S1	Dec. 87	87.191	6.194	0.085	0.138	6.294	0.077	0.090
S2	July 87	127.075	6.525	0.051	0.174	6.637	0.107	0.067
S3	Aug. 87	90.187	6.760	0.158	0.389	6.833	0.186	0.232
S4	July 87	106.156	6.799	0.147	0.366	6.911	0.192	0.228
S5	Feb. 87	128.606	6.805	0.062	0.199	6.969	0.081	0.102

model was then modified to take account of serial dependence. As the model assumes changes in the regression regime, inferences need to be made on regression parameters as well as the point at which regression has changed. Table 4.1 lists the estimates of the parameters and the change points for all the stations. In addition the log maximized likelihood is also given. It is interesting to note that the ML estimates of the change point are consistent for all stations, since all occur in 1987 but in different month. The consistency of the signs and approximate magnitudes of the regression parameters.

Figures 4.1 and 4.2 present the profile relative likelihood function and the fitted regression models for stations S0 and S5. The *ReL* for S5 has a sharp peak indicated a precise estimate of the change point. This is contrast with that of S0 where double peaks are indicated, suggesting that change could have occurred over a wider time period. It is interesting however to note that the first peak at S0 has nearly occurred at the same time as that of S5. The fitted model seems to present an adequate representation of the main features of the data patterns. Figure 4.1 and Figure 4.2 also show that the changes are not only in the levels (trend) but also in the variation

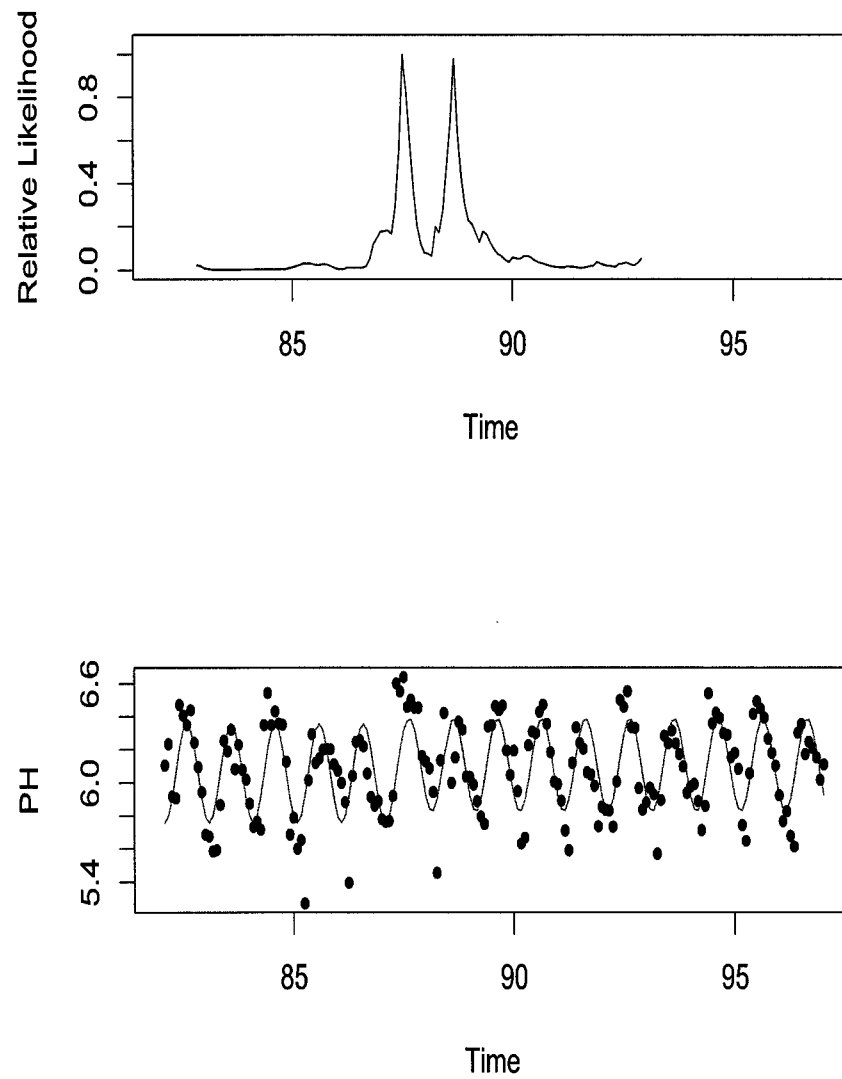


Figure 4.1: The change point of pH in station S0 under assumption of independence.

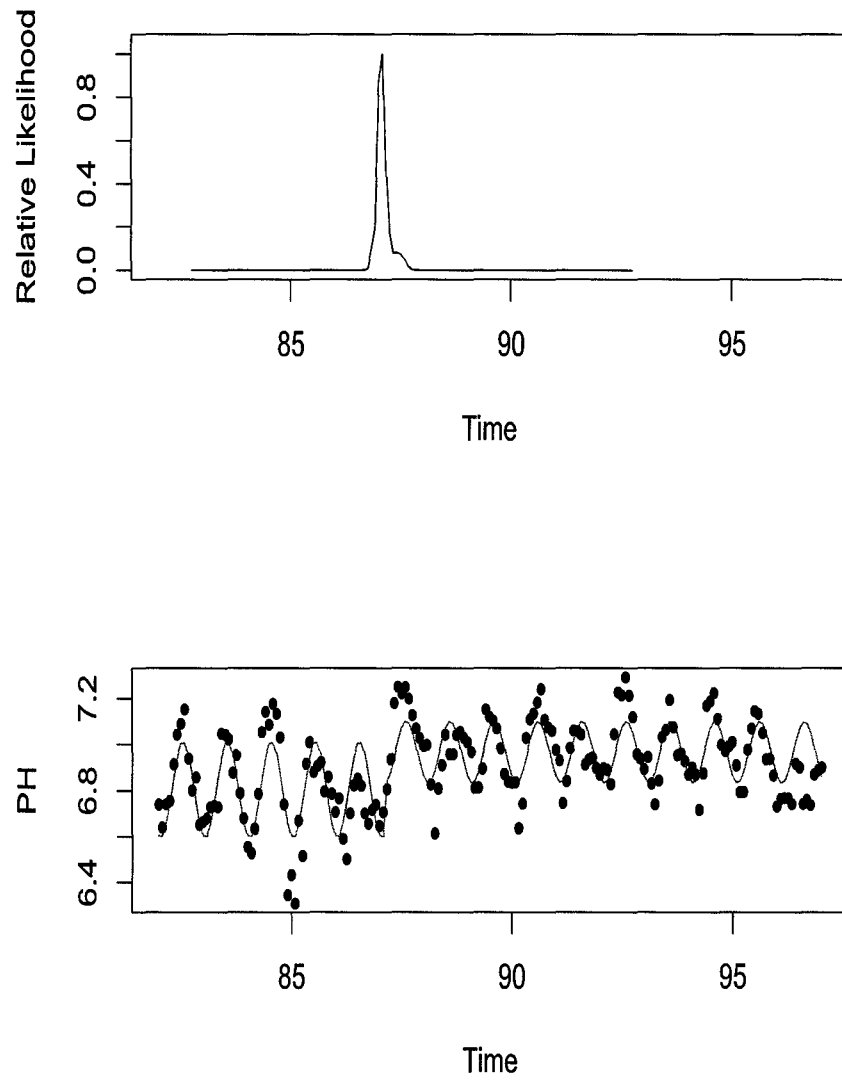


Figure 4.2: The change point of pH in station S5 under assumption of independence.

Table 4.2: **Change of pH under assumption of AR(1).**

St.No.	Change Time	$\log(L_{max}^{AR(1)})$	$\hat{\alpha}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\gamma}_0$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$
S0	July 88	150.8	6.076	0.163	0.227	6.114	0.210	0.198
S1	March 88	189.7	6.251	0.079	0.131	6.279	0.070	0.081
S2	March 88	236.7	6.570	0.074	0.150	6.636	0.102	0.074
S3	April 87	209.5	6.741	0.161	0.350	6.862	0.182	0.247
S4	April 87	261.1	6.787	0.158	0.339	6.924	0.182	0.238
S5	June 85	272.2	6.832	0.074	0.264	6.928	0.075	0.102

Table 4.3: **The estimates of  $\sigma$  in two segments.**

St.No.	S0	S1	S2	S3	S4	S5
$\hat{\sigma}_1$	0.193	0.146	0.130	0.136	0.130	0.101
$\hat{\sigma}_2$	0.149	0.128	0.093	0.117	0.112	0.084

within the yearly cycles.

The results of fitting model 4.1 are summarized in Table 4.2 assuming that the errors are correlated according to an AR(1) process. It appears that the maximum of log-likelihood function is substantially higher than that obtained under the assumption of dependence. According to the likelihood ratio test

$$2[\log(L_{max}^{AR(1)}) - \log(L_{max}^{IN})] \gg \chi_{0.05,2}^2 \quad (4.19)$$

that the model under the AR(1) assumption is a significant improvement over from the model with independent error process.

Figure 4.3 and Figure 4.4 are the profile relative likelihood function and the fitted



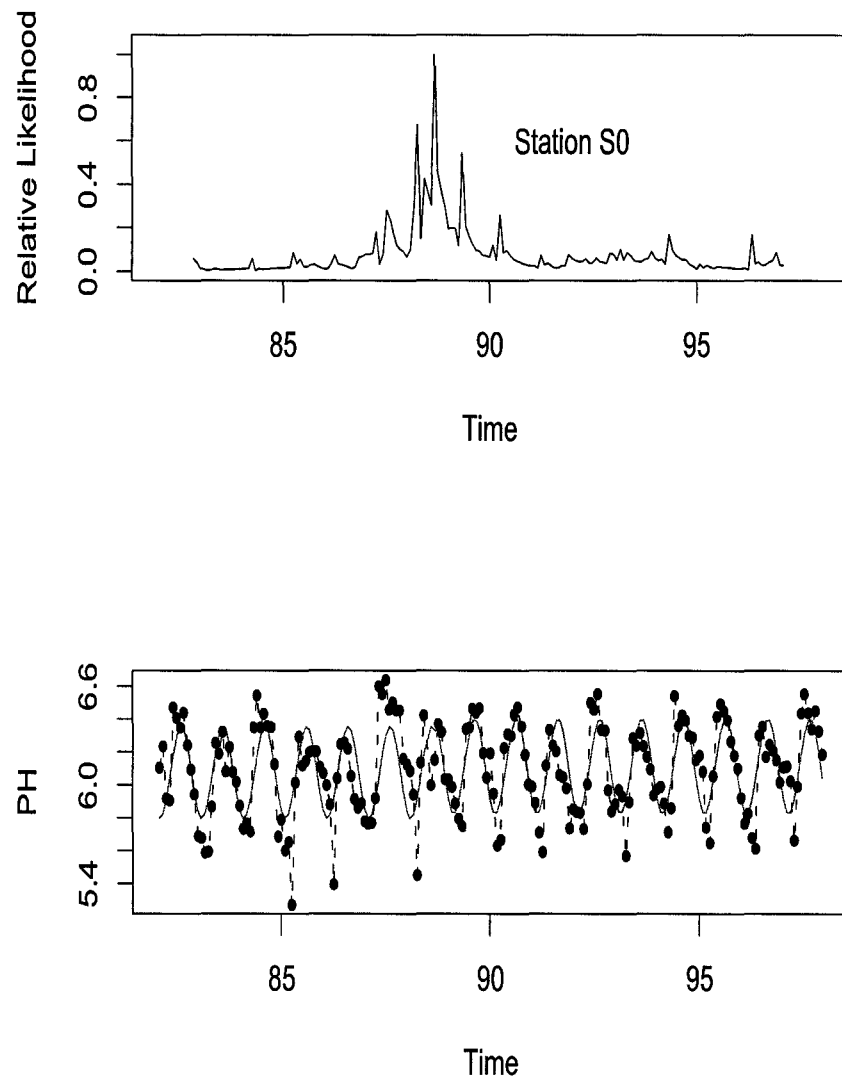


Figure 4.3: The change point of pH in station S0 under assumption of AR(1).

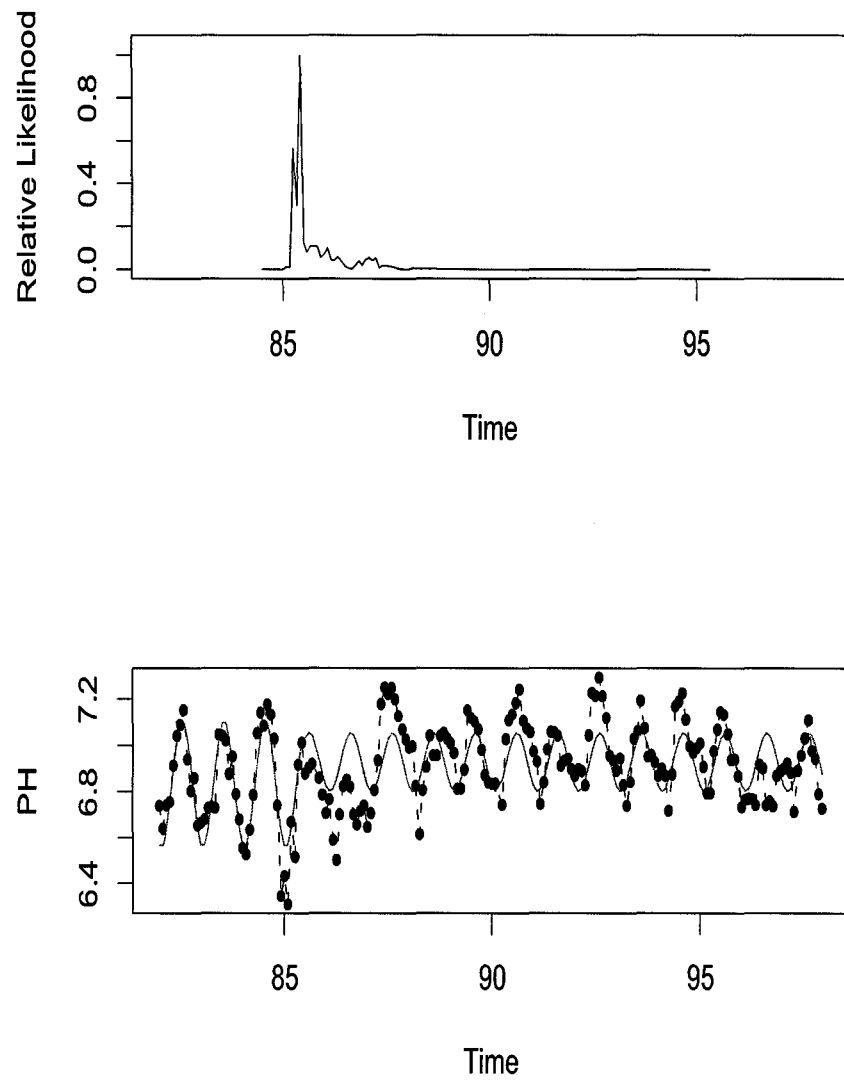


Figure 4.4: The change point of pH in station S5 under assumption of AR(1).

Table 4.4:  $\hat{\phi}_1$ ,  $\hat{\phi}_2$  and their std for the AR(1) model.

St.No.	$\hat{\phi}_1$	$\sigma_{\hat{\phi}_1}$	$\hat{\phi}_2$	$\sigma_{\hat{\phi}_2}$	$z_i$
S0	0.45	0.115	0.35	0.098	0.66
S1	0.60	0.120	0.35	0.095	1.64
S2	0.70	0.121	0.45	0.095	1.63
S3	0.45	0.129	0.60	0.090	-0.95
S4	0.45	0.129	0.50	0.091	-0.32
S5	0.30	0.164	0.75	0.081	-2.46

regression models for stations S1 and S5 under AR(1) assumption. In station S0, the value of likelihood functions change in a narrow range during the period of 88-89. So, the peaks in Figure 4.3 indicate that the change at this station could be some where during that time period. For station S5, it seems the extremely variable data around the summer months of 86 has a major impact on the likelihood function.

Table 4.3 presents the estimates of the  $\sigma$ 's before and after the changes taking place: segment 1 and segment 2. The results consistently show that the square root of the SSE is smaller in segment 2 than in segment 1.

The estimates of autoregressive coefficients of two segments for 6 stations are listed in Table 4.4 along with their standard deviations. It shows that they are significantly different from zero. The quantity  $z_i$  is

$$z_i = \frac{\hat{\phi}_1 - \hat{\phi}_2}{\sqrt{\frac{1}{k_1 - p_1} + \frac{1}{k_2 - p_2}}}$$

is the test statistic for testing the equality of  $\phi_1$  and  $\phi_2$ . Under the null hypotheses

of  $\phi_1 = \phi_2$ , it is asymptotically distributed as  $N(0,1)$ . The test indicate significant evidence against the hypotheses (5 % level) except for the lowest station.

### 4.2.2 Modelling the Changes of $SO_4^{2-}$ .

As we have seen in Chapter 2, in all 6 stations the concentration of  $SO_4^{2-}$  decreased dramatically with time and the yearly cycle is not as clear as that for the pH value. The model is different from the one used for pH where segment regression is used. Here one regression will describe the data fairly well, so the first part of general model 4.1 is used in the fitting. For the monthly mean values of  $SO_4^{2-}$ , coefficients of the polynomial terms are significant up to the quadratic term. i.e.  $\Theta_1 = (\alpha_0, \alpha_1, \alpha_2, \beta_{11}, \beta_{12})$ .

In the process of fitting different autoregressive processes we have used the AIC criterion to select the ARMA order. The results show that the AR(1) has the smallest AIC, so it is kept in the final model. The fit results are given in Figure 4.5. And the coefficients are listed in Table 4.5.

From Table 4.5 and Fig. 4.5, we can see that: There is a clear gradient in the concentration of  $SO_4^{2-}$ , lower downstream lakes have higher  $SO_4^{2-}$  concentration. The seasonality in downstream lakes become weaker. In all 6 stations the  $SO_4^{2-}$  decreased in the study period, however the decreasing rates vary from station to station and change with time. In recent years the downstream stations have decreased the most.

Station S1 is an exception in the sense that the coefficients of linear and quadratic terms have opposite sign as those of other stations and the amplitude of periodic function describing its yearly cycle is much bigger. The reason is that there are several very extreme observations ( the box plot of the data show a long heavy tail in the high

Table 4.5: Coefficients of fitting the trend of  $SO_4^{2-}$ .

St.No.	$\alpha_0$	$\alpha_1$	$\alpha_2$	$\beta_{11}$	$\beta_{12}$	$\phi$
S0	5.656	-0.064	0.002	-0.137	0.184	0.687
S1	6.060	0.047	-0.021	-0.3999	0.286	0.667
S2	6.137	-0.194	0.020	-0.082	0.296	0.615
S3	6.346	-0.195	0.024	-0.028	0.211	0.566
S4	6.325	-0.117	0.011	-0.031	0.119	0.588
S5	6.494	-0.097	0.012	-0.139	0.096	0.592

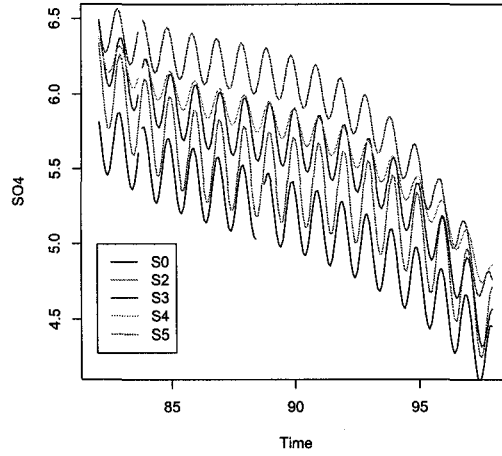


Figure 4.5: The trend of  $SO_4^{2-}$ .

end), even after removing all data exceeding 10 ( see the before and after boxplot in Figure 4.6). It seems there was serious problem with the quality of the data. Fitting of S1 is not included in Figure 4.5.

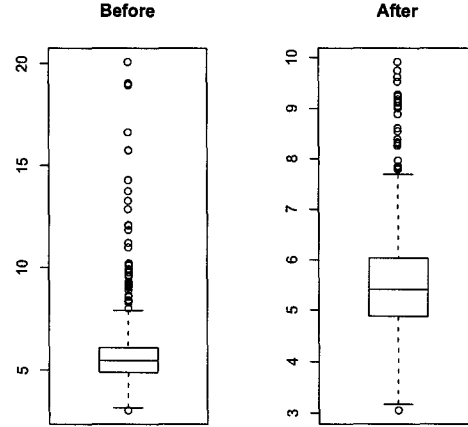


Figure 4.6: Boxplots of  $SO_4^{2-}$  before and after cutting off the extreme values.

### 4.2.3 Change Point of the First Two Principal Components

In the principal components analysis, the first principal component  $Y_{(1)}$  is a combination of the original variables which explain the biggest proportion of the variation in the data. The mathematical details ( Krzanowski, 1988) also show that the combination coefficients is the eigenvector corresponding to the largest eigenvalue of the variance-covariance matrix. These are the first column vector of the matrix  $V'$  in equation ( 3.4) (denoted as  $\mathbf{v}'_1$ ), so that

Table 4.6: Change of the first two principal components in S0 and S5.

St. No.	Components	Change Time	$\hat{\phi}_1$	$\hat{\sigma}_1$	$\hat{\alpha}_0$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	likelihood
			$\hat{\phi}_2$	$\hat{\sigma}_2$	$\hat{\gamma}_0$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$	
S0	1st	April 92	0.60	1.064	0.963	0.637	0.260	-187.2930
			0.50	0.864	1.873	0.575	0.257	
	2nd	April 85	0.25	1.102	0.744	1.224	1.797	-189.7873
			0.30	0.980	-0.140	1.198	1.160	
S5	1st	March 88	0.65	1.087	-0.538	1.036	0.992	-252.9772
			0.55	1.602	0.269	1.548	1.303	
	2nd	April 92	0.45	0.773	-0.405	0.403	0.620	-134.8241
			0.50	0.694	0.981	0.545	0.526	

$$Y_{(1)} = Y\mathbf{v}'_1 \quad (4.20)$$

Similarly, the second principal component is

$$Y_{(2)} = Y\mathbf{v}'_2 \quad (4.21)$$

In our data sets, the first two components explain about 65 % of the variation for both station S0 and S5. Each of them contains more information of the data set than a single variable. So in this section, we present the result of modelling the first two components of S0 and S5. The results are given in Figures 4.7 4.10 and Table 4.6.

The change pattern of the first two components in S1 different from that of S5. It indicates that the data structure are different for different stations. More investigation

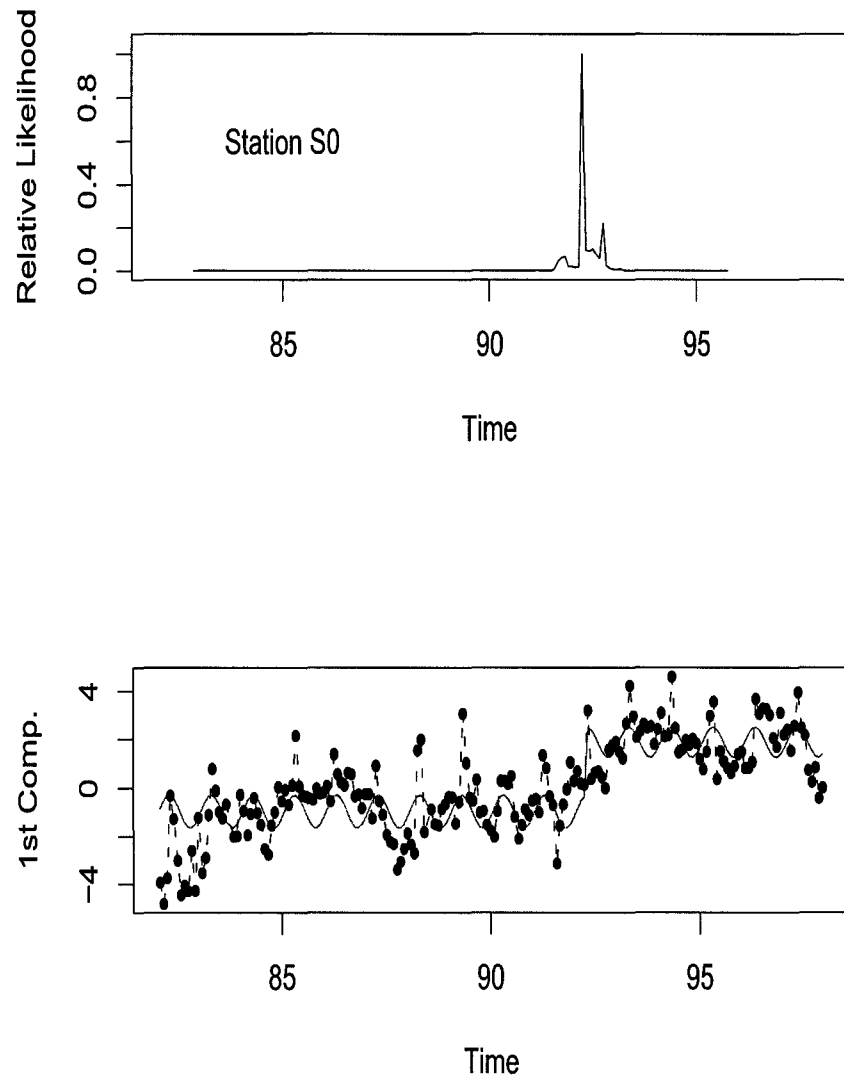


Figure 4.7: The change point of first principal component of station S0.



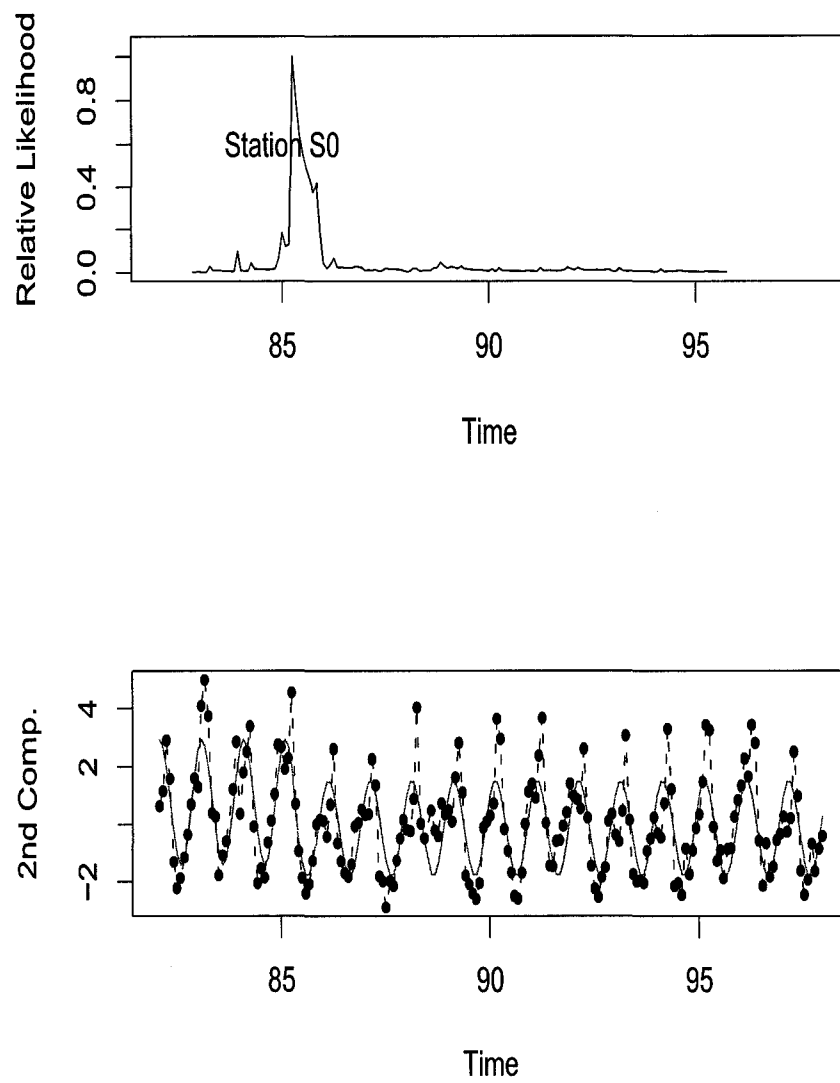


Figure 4.8: The change point of second principal component of station S0.

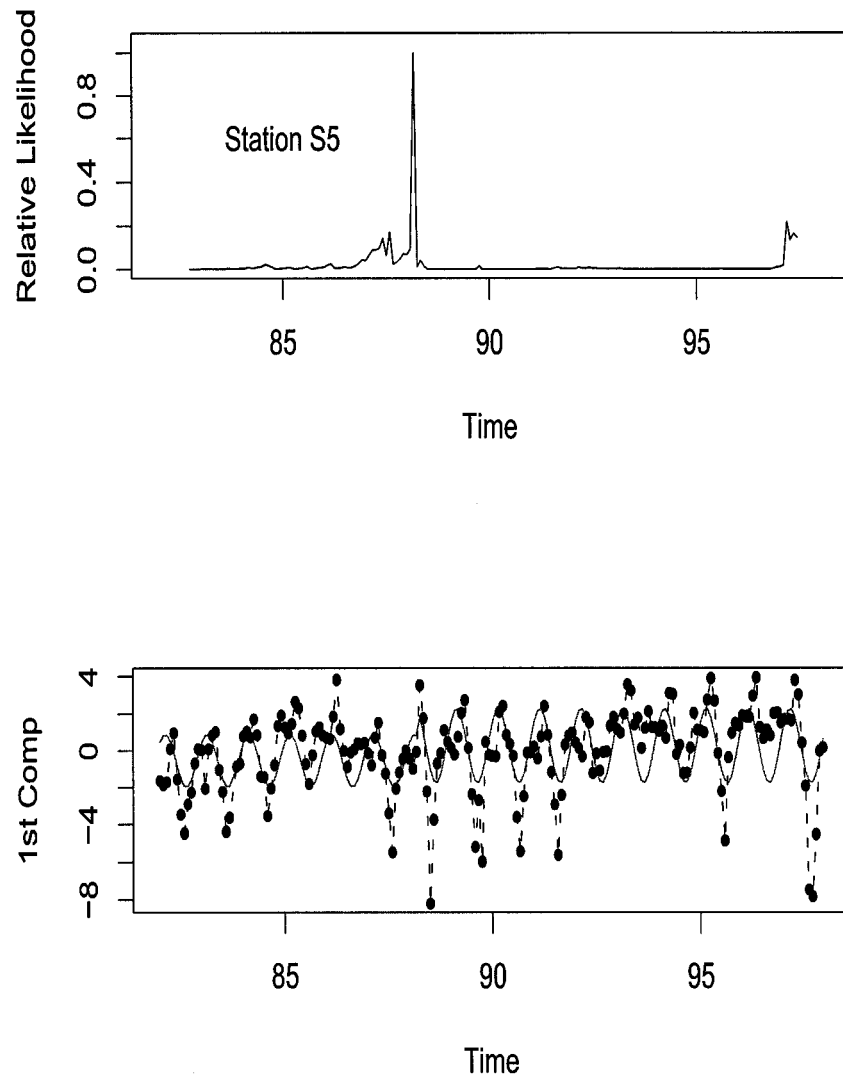


Figure 4.9: The change point of first principal component of station S5.

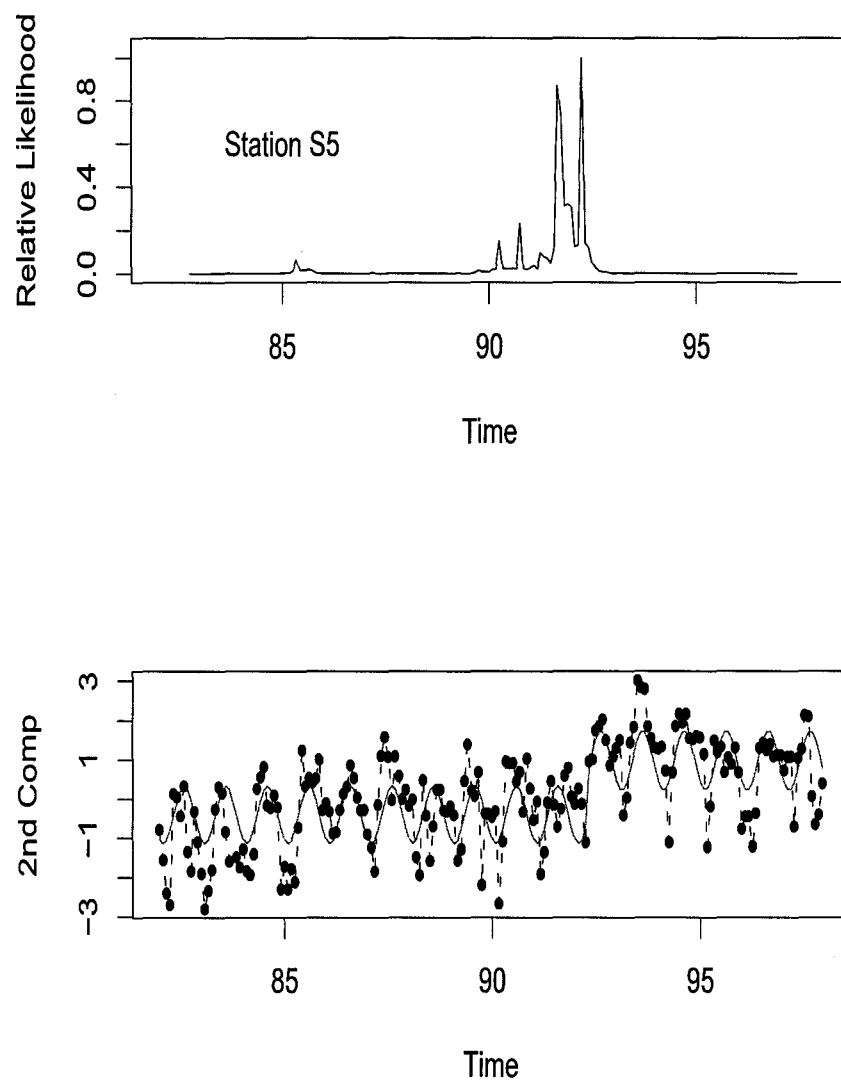


Figure 4.10: The change point of second principal component of station S5.

about variables other than pH and  $SO_4^{2-}$  is needed in order to explain the change of principal components.

The upper panels of Figures 4.7 and 4.9 show the plots of the relative likelihood functions for the change point of the first principal components at stations S0 and S5 respectively. While the corresponding lower panels show the data time plots along with the fitted models. Figures 4.8 and 4.10 show the plots based on the second principal component. The estimated model parameters and the maximized likelihoods are given in Table 4.6 for S0 and S5. The change point of the regression model differed for two components within the same stations and also between the stations. Inspection of the time plots show that time trend was dominating the first component for S0 while seasonality was the dominant for S5. This resolves somehow the discrepancy between the two stations for it is important to compare the pattern of component 1 at S0 with that of component 2 at S5. This shows that the change occurred in April 92 for both stations (Table 4.6). There is a clear indication that changes in the seasonality pattern at S0 (represented by component 2) occurred in 1985. This is earlier than that at S5 (1988 for component 1). The table shows that the value of the autoregressive parameters hardly changed between the two segments of the regression models for S0 and S5. It is interesting to note that the variability was lower in segment 2 in comparison with segment 1 of the regression models of S0 which is not the case at S5.

# Chapter 5

## Conclusion and Future work

### 5.1 Conclusion

In this project, inferences about changes in the water chemistry of TWL are made using regression models, change point, biplots and principle component analysis. Detailed modelling was presented for pH and  $SO_4^{2-}$  because of their importance in the characterization and control of anthropogenic acid rain. It has been concluded that pH has increased and  $SO_4^{2-}$  has decreased over the years. The pH trend is less pronounced than that of  $SO_4^{2-}$  and is well represented by a segmented regression model with the point at which the regression changed is also considered as an unknown parameter to be estimated from the data. In addition, serial correlation was found to be significant and this led us to include an AR(1) error term in the model. The analysis did not reject the constancy of the AR(1) parameter for the two segments of the regression model, however the variance was different in a consistent pattern. We presented the modelling of the two principal components to provide summary for the overall changes

in the chemistry of TWL despite the fact more subject matter is needed to interpret the findings.

## 5.2 Future Work

The topic of acid rain will remain important since more and more countries and people are realizing its potential effect. The TLW study will continue to generate more data and will require a more thorough analysis. Issues that need to be addressed in the future are:

1. Include more data in the analysis so that we will be able to detect more than one change point and study other single variable intensively.
2. Apply other methods including non-parametric and semi-parametric to the data set, for example, cumulative sum and recursive residuals, smoothing and compare the results.
3. Build a general model to connect the different stations together to account for the spatial feature of the data sites.
4. Multivariate extension of the regression model will be valuable for the integration of the information for the entire TLW system.

# Bibliography

- [1] Beamish, R. J., and Harvey, H. H. (1972). Acidification of the La Cloche Mountain Lakes, Ontario, and resulting fish mortalities. *J. Fish. Res. Board Can.* **29**, 1131-1143.
- [2] El-Shaarawi, A. H. and Esterby, S. R. (1982). *Inference About the Point of Change in a Regression Model with a Stationary Error Process*, Elsevier Scientific Publishing Company, Amsterdam, The Netherlands.
- [3] Esterby, S. R. and El-Shaarawi, A. H. (1981). Inference about the point of Change in a Regression Model. *Appl. Statist.* **30**, 277-285.
- [4] Gabriel, K. R. (1971). The biplot graphic display of the matrices with application to principal component analysis. *Biometrika.* **58**, 453-467.
- [5] Good, I. J. (1969). Some applications of singular value decomposition of a matrix. *Technometrics*, **11**, 823-31.
- [6] Jandhyale, V. K., Fotopoulos, S. B. and El-Shaarawi, A. H. Change-point method,(2001). *Encyclopedia of Envirometrics* Vol.1, 324-332. New York: Wiley.

- [7] Jeffries, D. S., Kelso, J. R. M. and Morrison, I. K., (1988). Physical, chemical, and biological characteristics of the Turkey Lakes Watershed, Central Ontario. *Can. J. Fish. Aquat. Sci.* **45**, 3-13.
- [8] Kalbfleish, J. G.(1985) *Probability and Statistical Inference volume 2: Statistical Inference*. Springer-Verlag: New York.
- [9] Kelso, J. R. M., Minns, C. K., Lipsit, J. H. and Jeffries, D. S.. (1986) Headwater lake chemistry during the spring freshet in north-central Ontario. *Wat. Air Soil Pollut.* **29**, 245-259.
- [10] Krzanowski, W. J. ( 1988). *Principals of Multivariate Analysis*. Clarendon press. Oxford.
- [11] Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: Wiley.
- [12] Semkin, R. G. and Jeffries, D. S. (1986). Storage and release of major ionic contaminants from the snowpack in the Turkey Lakes Watershed, *Water, Air and Soil Pollution* **31**, 215-221.
- [13] Smith, D. L. and Underwood, J. K. (1986) Fish species distribution and water chemistry in Nova Scotia Lakes, *Air, Soil Pollution (Historical Archive)*, **30**, 489 - 495
- [14] Wales, D. L and Beggs, G. L.(1986) Fish species distribution in relation to lake acidity in Ontario, *Water, Air, Soil Pollution (Historical Archive)*, **30**, 601 - 609.



# Appendix A

## R functions

### A.1 Monthly mean

```
mean.monthly<-function(Dat,St.No)
{
  mean.monthly=matrix(NA,ncol=14,nrow=216)
  no.row=0
  mean.monthly[,1]=rep(St.No,216)
  mean.monthly[,2]=rep(seq(1,12),18)
  for (i in 80:97)
  for (j in 1:12)
  {
    no.row=no.row+1
    mean.monthly[no.row,3]=i+(j-1)/12
    for (k in 4:14)
    { flag=((Dat$Yr==i)&(Dat$Mon==j))
```

```

        cell<-Dat[flag,k]

        if(length(na.omit(cell)!=0))

            mean.monthly[no.row,k]<-mean(na.omit(cell))

    }

}

dimnames(mean.monthly)[[2]]=c("St.No","Mon","Time",
+dimnames(Dat)[[2]][-(1:3)])

data.frame(mean.monthly)

}

```

## A.2 Change point under the independent assumption

```

Chglike=function (X = mean.mon.s0, name.var = "PH") {

    if (name.var == "PH")

        Dat = data.frame(X$Time, X$PH)

    if (name.var == "SO4")

        Dat = data.frame(X$Time, X$SO4)

    Dat.new <- na.omit(Dat)

    mark = ((Dat.new$X1 >= 82) & (Dat.new$X1 <= 97))

    Dat.ana <- Dat.new[mark, ]

    n <- length(Dat.ana[[2]])

    lk <- rep(NA, n)

    for (k in 10:(n - 10)) {

```

```

Dat.1 <- Dat.ana[1:k, ]
Dat.2 <- Dat.ana[(k + 1):n, ]
r1.lm <- lm(X2 ~ sin(2 * pi * X1) + cos(2 * pi * X1),
           data = Dat.1)
r2.lm <- lm(X2 ~ sin(2 * pi * X1) + cos(2 * pi * X1),
           data = Dat.2)
Dat.1 <- cbind(Dat.1, r1.lm$fitted)
Dat.2 <- cbind(Dat.2, r2.lm$fitted)
Dat.com <- rbind(as.matrix(Dat.1), as.matrix(Dat.2))
sigma.1 <- sum((Dat.com[1:k, 3] - Dat.com[1:k, 2])^2)/(k -
              3)
sigma.2 <- sum((Dat.com[(k + 1):n, 3] - Dat.com[(k +
              1):n, 2])^2)/(n - k - 3)
lk[k] <- sum(log(dnorm(Dat.com[1:k, 2], Dat.com[1:k,
              3], sqrt(sigma.1)))) + sum(log(dnorm(Dat.com[(k +
              1):n, 2], Dat.com[(k + 1):n, 3], sqrt(sigma.2))))
}

par(mfrow = c(2, 1))
plot(c(82, 97), c(0, 1.05), type = "n", xlab = "Time",
     ylab = "Relative Likelihood" )
lines(Dat.ana$X1[10:130], exp(lk[10:130])/exp(max(lk[10:130])))
point.chang = which.max(lk[10:130]) + 9
k <- point.chang
Dat.1 <- Dat.ana[1:k, ]

```

```

Dat.2 <- Dat.ana[(k + 1):n, ]

r1.lm <- lm(X2 ~ sin(2 * pi * X1) + cos(2 * pi * X1), data = Dat.1)
r2.lm <- lm(X2 ~ sin(2 * pi * X1) + cos(2 * pi * X1), data = Dat.2)

Dat.1 <- cbind(Dat.1, r1.lm$fitted)
Dat.2 <- cbind(Dat.2, r2.lm$fitted)

Dat.com <- rbind(as.matrix(Dat.1), as.matrix(Dat.2))

plot(c(82, 97), c(min(Dat.com[, 2]), max(Dat.com[, 2])),
     type = "n", xlab = "Time", ylab = paste(name.var))
points(Dat.com[, 1], Dat.com[, 2], pch=20)
lines(Dat.com[, 1], Dat.com[, 3])
lines(smooth.spline(Dat.com[1:k, 1], Dat.com[1:k, 3], df = 5),
      lty = 3)
lines(smooth.spline(Dat.com[(k + 1):n, 1], Dat.com[(k + 1):n,
      3], df = 5), lty = 3)

coef.value <- round(1000 * c(Dat.com[k, 1], max(lk[10:150]),
      as.numeric(r1.lm$coef), as.numeric(r2.lm$coef)))/1000

coef.value
}

```

### A.3 Likelihood function under AR(1) assumption

```

# miss- row number of missing data

# X--n1 by 2 matrix (time, observation), after deleting missing value

cor.mat<-function(n,phi) {power<-matrix(NA,ncol=n,nrow=n)

```

```

for (i in 1:n)
for (j in 1:n)\oint
power[i,j]=abs(i-j)
matrix(as.matrix(outer(phi,power,"^")),ncol=n,nrow=n)
} function (phi1, phi2, u0 = 0, X = na.omit(mean.mon.s0[-(1:25),
3:4]), k, miss = miss0)
{
n1 = length(X[[2]])
y1 = X[1:k, ]
y2 = X[(k + 1):n1, ]
if (k < miss[1]) {
cor.mat.1 = cor.mat(k, phi1)
cor.mat.2 = cor.mat((n - k), phi2)[-(miss - k), -(miss -
k)]
}
if ((k >= miss[1]) && (k < miss[length(miss)])) {
m = 0
for (i in 1:length(miss)) if (k >= miss[i])
m = m + 1
miss.new1 = miss[1:m]
miss.new2 = miss[-(1:m)] - k
cor.old1 = cor.mat(k + length(miss.new1), phi1)
cor.old2 = cor.mat((n1 - k + length(miss.new2)), phi2)
cor.mat.1 = cor.old1[-miss.new1, -miss.new1]

```

```

        cor.mat.2 = cor.old2[-miss.new2, -miss.new2]
    }

    if (k >= miss[length(miss)]) {
        cor.mat.1 = cor.mat((k + length(miss)), phi1)[-miss,
            -miss]
        cor.mat.2 = cor.mat((n1 - k), phi2)
    }

    A = cbind(rep(1, n1), sin(2 * pi * X[, 1]), cos(2 * pi *
        X[, 1]))
    A1 = A[1:k, ]
    A2 = A[(k + 1):n1, ]
    theta1 = solve(t(A1) %*% solve(cor.mat.1) %*% A1) %*% t(A1) %*%
        solve(cor.mat.1) %*% y1[, 2]
    theta2 = solve(t(A2) %*% solve(cor.mat.2) %*% A2) %*% t(A2) %*%
        solve(cor.mat.2) %*% y2[, 2]
    u1 = y1[, 2] - A1 %*% theta1
    u2 = y2[, 2] - A2 %*% theta2
    sigma.sq1 = as.numeric(((1 - phi1^2) * t(u1) %*% solve(cor.mat.1) %*%
        u1/k)
    sigma.sq2 = as.numeric(((1 - phi2^2) * t(u2) %*% solve(cor.mat.2) %*%
        u2/(n1 - k))
    like1 = -log(det(sigma.sq1/(1 - phi1^2) * cor.mat.1))/2 -
        t(u1) %*% solve((sigma.sq1/(1 - phi1^2)) * cor.mat.1) %*%
        u1

```

```

like2 = -log(det(sigma.sq2/(1 - phi2^2) * cor.mat.2))/2 -
        t(u2) %*% solve((sigma.sq2/(1 - phi2^2)) * cor.mat.2) %*%u2
c(as.numeric(like1), as.numeric(like2))
}

```

## A.4 Result analysis

```

LikeAr1Res=function(phi1,phi2,u0=0,X,k,miss) {
  n1=length(X[[2]])
  n=n1+length(miss)
  y1=X[1:k,]
  y2=X[(k+1):n1,]
  #generate the correlation matrix according to the missing value
  if (k<miss[1])
  { cor.mat.1=cor.mat(k,phi1)
    cor.mat.2=cor.mat((n-k),phi2)[-(miss-k),-(miss-k)]
  }
  if ((k>=miss[1]) && (k<miss[length(miss)]))
  { m=0
    for (i in 1:length(miss))
      if (k>=miss[i]) m=m+1
    miss.new1=miss[1:m]
    miss.new2=miss[-(1:m)]-k
  }
}

```

```

cor.old1=cor.mat(k+length(miss.new1),phi1)
cor.old2=cor.mat((n1-k+length(miss.new2)),phi2)

cor.mat.1=cor.old1[-miss.new1,-miss.new1]
cor.mat.2=cor.old2[-miss.new2,-miss.new2]
}

if (k>=miss[length(miss)])
{ cor.mat.1=cor.mat((k+length(miss)),phi1)[-miss,-miss]
  cor.mat.2=cor.mat((n1-k),phi2)
}

A=cbind(rep(1,n1),sin(2*pi*X[,1]),cos(2*pi*X[,1]))
A1=A[1:k,]
A2=A[(k+1):n1,]
theta1=solve(t(A1)%*%solve(cor.mat.1)%*%A1)%*%t(A1)%*%solve(cor.mat.1)%*%y1[,2]
theta2=solve(t(A2)%*%solve(cor.mat.2)%*%A2)%*%t(A2)%*%solve(cor.mat.2)%*%y2[,2]
u1=y1[,2]-A1%*%theta1
u2=y2[,2]-A2%*%theta2
sigma.sq1=as.numeric((1-phi1^2)*t(u1)%*%solve(cor.mat.1)%*%u1/k)
sigma.sq2=as.numeric((1-phi2^2)*t(u2)%*%solve(cor.mat.2)%*%u2/(n1-k))
like1=-log(det(sigma.sq1/(1-phi1^2)*cor.mat.1))/2-t(u1)
      %*%solve((sigma.sq1/(1-phi1^2))*cor.mat.1)%*%u1
like2=-log(det(sigma.sq2/(1-phi2^2)*cor.mat.2))/2-t(u2)

```



```

%%solve((sigma.sq2/(1-phi2^2))*cor.mat.2)%%u2

plot(X$Time,X$PH,xlab="Time",ylab="PH",pch=20)
lines(X$Time,X$PH,lty=2)
lines(X$Time,c(A1*%%theta1,A2*%%theta2),col="red")

list(ChangeTime=t.change,Phi1=phi1,Phi2=phi2,Likelihood=like1+like2,
      Sigma1=sigma.sq1,Sigma2=sigma.sq2,coefficients=cbind(theta1,theta2))
}

```

# Appendix B

## R command

```
source("e:\\program/fns.R.txt")
s0=read.csv("e:\\ProjectData/s0.csv",sep=",",header=T)
s1=read.csv("e:\\ProjectData/s1.csv",sep=",",header=T)
s2=read.csv("e:\\ProjectData/s2.csv",sep=",",header=T)
s3=read.csv("e:\\ProjectData/s3.csv",sep=",",header=T)
s4=read.csv("e:\\ProjectData/s4.csv",sep=",",header=T)
s5=read.csv("e:\\ProjectData/s5.csv",sep=",",header=T)
mean.mon.s0<-mean.monthly(s0,0)
mean.mon.s1<-mean.monthly(s1,1)
mean.mon.s2<-mean.monthly(s2,2)
mean.mon.s3<-mean.monthly(s3,3)
mean.mon.s4<-mean.monthly(s4,4)
mean.mon.s5<-mean.monthly(s5,5)
m0<-descrip(s0,11)
```

```

m1<-descrip(s1,11)
m2<-descrip(s2,11)
m3<-descrip(s3,11)
m4<-descrip(s4,11)
m5<-descrip(s5,11)

plot(c(.5,12.5),c(3,10),type="n",xlab="Month",ylab="S04")
for (i in 1:12)
{points((m0[i,1]-.5+d),m0[i,3],pch=1)
  lines(rep(m0[i,1]-.5+d,2),m0[i,c(2,4)],lty=1)
  points((m1[i,1]-.5+2*d),m1[i,3],pch=2)
  lines(rep(m1[i,1]-.5+2*d,2),m1[i,c(2,4)],lty=2)
  points((m2[i,1]-.5+3*d),m2[i,3],pch=3)
  lines(rep(m2[i,1]-.5+3*d,2),m2[i,c(2,4)],lty=3)
  points((m3[i,1]-.5+4*d),m3[i,3],pch=4)
  lines(rep(m3[i,1]-.5+4*d,2),m3[i,c(2,4)],lty=4)
  points((m4[i,1]-.5+5*d),m4[i,3],pch=5)
  lines(rep(m4[i,1]-.5+5*d,2),m4[i,c(2,4)],lty=5)
  points((m5[i,1]-.5+6*d),m5[i,3],pch=6)
  lines(rep(m5[i,1]-.5+6*d,2),m5[i,c(2,4)],lty=6)
}

X1=na.omit(mean.month.s0[-(1:25),2:3])
n<-length(X1[,1])
X<-na.omit(X1)

#only use the data from 82 to 95

```

```

flag=((X[,1]>=2)&(X[,1]<=17))
X=X[flag,]
n1=length(X[,1])
pi<-seq(-.9,.9,by=.05)
#calculate the likelihood corresponding to
k and phi1 and phi2
lk1<-lk2<-matrix(NA,ncol=n1,nrow=length(phi)) #
for (j in 5:(n1-5))
  for (i in 1:length(phi))
{ lk1[i,j]<-like1(phi[i],phi2=0,u0=0,X,j)[1]
  lk2[i,j]<-like1(phi1=0,phi[i],u0,X,j)[2] }

write(lk1,"c:\\\\like1PHst01.txt")
write(lk2,"c:\\\\like2PHst01.txt")

lk1<-read("E:\\\\like1PHst01")
lk2<-read("E:\\\\like2PHst01")

like.max1<-apply(lk1,2,max)
like.max2<-apply(lk2,2,max)

X=na.omit(mean.mon.s0[-(1:25),3:4])
max.l=(like.max2+like.max1)[10:180]

```

```

ex<-max(max.l)
plot(c(82,98),c(0,1.05),type="n",xlab="Time",ylab="Relative Likelihood")
lines(X$Time[10:180],exp(max.l)/exp(ex))

K=n.change=which.max(max.l)+9
t.change=X$Time[n.change]
phi1.max=phi[which.max(lk1[,n.change])]
phi2.max=phi[which.max(lk2[,n.change])]

LikeAr1Res(phi1.max,phi2.max,u0=0,X,K,miss)

```