APPLIED STATISTICAL METHODS ON PRICE DATA AND ANALYZING THE EFFECTS OF ENVIRONMENTAL AND CHEMICAL FACTORS FOR APHANIZOMENON FLOS-AQUAE DENSITY

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistics

in the

College of Graduate Studies

University of Idaho

by

Yuanyang Yu

May 2014

Major Professor: Stephen S. Lee, Ph.D.

Authorization to Submit Thesis

This thesis of Yuanyang Yu, submitted for the degree of Master of Science with a major in Statistics and titled "Applied statistical methods on price data and analyzing the effects of environmental and chemical factors for *Aphanizomenon flos-aquae* density" has been reviewed in final form. Permission, as indicated by the signatures and dates given below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____ Date _____
Stephen S. Lee, Ph.D.

Committee Members _____ Date _____
Christopher Williams, Ph.D.

_____ Date _____
Frank Wilhelm, Ph.D.

Department Administrator _____ Date _____
Christopher Williams, Ph.D.

Discipline's College Dean _____ Date _____
Paul Joyce, Ph.D.

Final Approval and Acceptance

Dean of the College of Graduate Studies _____ Date _____
Jie Chen, Ph.D.

Abstract

This work is divided into two research projects: (i) Comparison of the price reports for lumber products between *Random Lengths* and *Crows'* in North American softwood market; (ii) Analyzing the effects of environmental and chemical factors for *Aphanizomenon flos-aquae* density in Willow Creek, Oregon.

When compared the price reports for lumber products, the parametric tests for different means were not proper methods. This was due to the time dependent of the data which the observations were conducted weekly over 20 years. The paired comparison of price differences changing over time showed how much difference between *Random Lengths* and *Crows'* price reports. Based on the results of autocorrelation function and binomial test, it showed that the prices between *Random Lengths* and *Crows'* had a mean difference of less than $1 and standard error around 0.15 for the 12 different products of which most southern yellow pine and some other products; these differences were stationary time series. Therefore, this was the average difference between *Random Lengths* and *Crows'*. For Douglas Fir Green 8' 2X4 Std&Btr-US and Douglas Fir Green 8' 2X6 #2&Btr-US, the difference means were $26 and $9 and the standard error were 0.027 and 0.026, respectively, and they were strong non-stationary time series. Thus, the difference between the price reports and the real market prices should be mentioned when study the lumber product prices in North America by using third-party price data.

Analysis of *Aphanizomenon flos-aquae* density in the Willow Creek, Oregon area was different from other studies targeting of blue-green algae. Modeling a single species rather than the whole biomass was more difficult because the dominant species strongly influenced the model. If *Aphanizomenon flos-aquae* was not the dominant species, the accuracy of the model would be reduced. The reactive phosphorus may give a negative effect on *Aphanizomenon flos-aquae* density, but positive effect on whole biomass density. When modelling the *Aphanizomenon flos-aquae* density, it was necessary to select the observations which *Aphanizomenon flos-aquae* density higher than 200 counts per liter. If it was lower than 200 counts per liter, it turned to poor model-fit. Moreover, lower density of

Aphanizomenon flos-aquae will not cause toxin problems of public health. Model with Expectation Maximization (EM) algorithm showed that nitrite, reactive phosphorus, pH, and temperature had highly significant effects on *Aphanizomenon flos-aquae* density, DO however, was less significant.

Acknowledgements

I am greatly indebted to my major advisor, Dr. Stephen S. Lee, for his support and guidance throughout my education and research. His guidance, encouragement, dedication, and trust helped with this thesis. Dr. Lee has not only been a great advisor, but also a mentor helping me with a vision of my future.

I am grateful to my committee members, Dr. Chris Williams and Dr. Frank Wilhelm for their time and contributions on this project.

I would like to thank Dr. Steven Shook in the Department of Business for his support and help on this project.

I would also like to thank all the colleagues in the Department of Statistical Science, and wish to thank friends and colleagues whom supported and contributed in many different ways in this research. I am so blessed to be able to work and learn with these brilliant people.

Also, I would especially thank my wife, Jing Dai who has been helped me for years. She is a successful model of mine and the one who made me here today. Finally, I wish to thank my parents, Shanbao Yu and Xiuzhen Zhu, and my brother Yuanqiang Yu, for their non-stop care, love, and support.

Table of Contents

List of Tables

List of Figures

Chapter 1 Comparison of the price reports for lumber products between *Random Lengths* and *Crows'* in North American softwood market

1.1 Abstract

Random Lengths and *Crows'* lumber products price report has been the replacement of real marketing price for a long time for both business design and forest products study. From paired comparison, the average difference between *Random Lengths* and *Crows'* lumber products price was less than $1 with standard error around 0.15 and was stationary over time. However, Douglas fir Green 8' 2X4 Std&Btr-US and Douglas fir Green 8' 2X6 #2&Btr-US had average difference $26 and $10 and they were strongly non-stationary time series. The result revealed that the difference between *Random Lengths* and *Crows'* lumber products price report of these two products may have a lager error difference than the real market price.

1.2 Introduction

The United States consumes nearly half of the world's softwood lumber consumption. According to *Random Lengths* (1993-2007), the United States consumed 142 million cubic meters (60 billion board feet) of softwood lumber in 2005. This amount accounts for 43% of the total 325 million cubic meter lumber consumed worldwide in the same year (FAO, 2009). In the U.S., softwood lumber is primarily used in home building and remodeling (U.S. Bureau of Census, 2009b).

In 2002, production of lumber in the United States amounted to 47.4 billion board feet. Southern yellow pine production amounted to 16.2 billion board feet which took one third of the total production. Meanwhile, Douglas fir production amounted to 9,180 million board feet. It is one of the main forest products in Western lumber regions (U.S. Bureau of Census,

2002). Therefore, the studies on the weekly price report for these two products significantly represented the situation of the whole picture.

As the softwood products market is large in North America, the price acts as a sensitive economic signal. It will affect different parts of the market, includes manufactures, dealers, and householders. In the market, consumers assume to purchase a product based on value (i.e., utility); the price informs consumers to reduce the risk and uncertainty of the quality of goods (Shook, 1999). The price also has an important role on the decision-making process because it induces both consumers and producers to revise their choice, particularly when the price is greater than real value. The manufacturers' decision on the quantity of production is thus based on the future price expecting (Luca, 2014). Also, price affects the speculative and postponement behaviors for both buyers and sellers. Therefore, accurate price information reduces economic waste and increases the efficiency of capital flows; in this way the price-quality is a serious concern.

Price of natural resource products is usually reported by external sources which includes broadcast product list/price data (fax, email, web), requests for quotes (RFQs with a ceiling price listed), futures market data, and third-party publications. The external price is not accurate representation of the value of the products, because they are not generated within the consumer's mind and producers' conditions, however, they are still important signals for the market (Mayhew & Winer, 1992, Zeithaml, 1988, Kopalle & Lindsey-Mullikin, 2003).

External price is a well-established concept in marketing. It has been found to have a significant impact on consumer purchase decisions. The external price has a significant curvi-linear effect on subjects' initial price expectations. The relative impact of external price and actual price on consumers' final price expectations increase up to a point and then start decreasing (Kopalle & Lindsey-Mullikin, 2003). Thus, external price is important to marketing even it is not equal to the value or accurate price in marketing of the products.

The average price of wood products differs from other natural resources in the North American forest products market. For example, the commodity futures exchange price of

petroleum oil at New York Mercantile Exchange (NYMEX), or the commodity futures exchange price of copper at Commodity Exchange, Inc (COMEX) can be reliable sources when looking for average market price for petroleum oil and copper. However, due to the small business size, it is hard to find such information to make educated business decisions for local dealers. The local dealers usually use third-party publications from marketing research companies to decide their price. Several publications give weekly price reports for North American forest products. The *Random Lengths* and *Crows'* are mostly used to helps forest products dealers to make accurate market decisions. Also, their published price data is often used for North American forest products research, such as identifying the inter-market relationships of forest products in the Pacific Northwest (Yin, 2004), according to lumber trade restrictions in North America: application of a spatial equilibrium model (Stennes, 2005), but do these third-party price reports give unbiased prices? Do any of these price reporters give more accurate price report than the others? In this work, the *Random Lengths* and *Crows'* price reports will be compared to find out if their price reports are unbiased.

1.3 Data

Random Lengths and *Crows'* both use phone survey methodology to prepare their weekly price report. Both claim that they conduct hundreds of telephone interviews each week in gathering information. Both also collect information from both buyers and sellers to reduce the risk of being "gamed". The "gamed" means the dealers will try to increase the value of their inventory by giving a higher price report than real exchange price to the phone survey. *Random Lengths* was founded in 1944 and began to be a price guide in 1958. RISI publishes *Crows'* wood products' prices and *Crows'* lumber price report has published for over 80 years. *Crows'* report includes wholesale pricing, detailed market reports, daily and weekly news updates, and market forecasts designed to give reader a complete and accurate understanding of each week's market.

The objective of this paper was to examine the third party price reports of North American softwood products. Because both *Random Lengths* and *Crows'* refuse to expose any detailed information about their surveys and interviews. Some type of survey errors, such as the

biased coverage of samples and lack of randomness of respondents' selection may occur. Due to lack of information of the sampling methods from *Random Lengths*, it's difficult to evaluate how well these weekly price reports represent the real market price. Some of the potential errors could not be examined by any third parties. However, *Random Lengths* and *Crows'* both claim their weekly wood products price reports containing independent and unbiased information for North American forest products market. This fact offers an opportunity to examine the compatibility of the two independent data sets, and find out if they are both unbiased. Table 1.1 shows the price reports and codes of 12 forest products which *Crows'* and *Random Lengths* both published in their reports.

Table1.1 Product data coding in report from *Random Lengths* and *Crows'*

Products	Random Lengths Code	Crows' Code
Douglas Fir Green 8' 2X4 Std&Btr-US	LAMM	CROW-04
Douglas Fir Green 8' 2X6 #2&Btr-US	LAMN	CROW-05
Douglas Fir Green Random Tally 8'-20' 2X6 #2&Btr-Portland	LAAM	CROW-11
Douglas Fir Green Random Tally 8'-20' 2X8 #2&Btr-Portland	LAAN	CROW-13
Douglas Fir KD Random Tally 8'-20' 2X10 #2&Btr-US	LADD	CROW-15
Douglas Fir KD Random Tally 8'-20' 2X12 #2&Btr-US	LADE	CROW-16
Fir/Larch KD Random Tally 8'-20' 2X6 #3-US	LADZ	CROW-12
SYP KD 8' 2X4 #1-West	LALP	CROW-25
SYP KD 8' 2X4 #1-Central	LAVB	CROW-24
SYP KD 8' 2X4 #1-East	LAMB	CROW-23
SYP KD 8' 2X4 #2-East	LAMD	CROW26
Hem/Fir KD 8' 2X4 Std&Btr-US West Coast	LAVL	CROW-20

1.4 Method

Random Lengths and *Crows'* published forest products price weekly as well as 12 different products' prices. The compatibility of the same product was checked to find out if there was any problem for their price reports. For each product, two price values were formed into two time series. The major objective was to observe how much difference existed between these

two time series. In order to observe the difference with a clear view, a new single time series was created by using *Crows'* minus *Random Lengths* price. Thus, a price difference time series with 816 data points for each product was generated.

1.4.1. Time series summary statistics

As the common step to analyze time series data, we check the statistics summary of the datasets. The summary includes time series plot, basic statistics with mean and standard error and box plot. What we expect is there is no price difference between *Crows'* and *Random Lengths* price report, which indicate the mean of the time series should be close to 0. The summary statistics includes the mean value and standard error of the time series, which showed how close the price difference to be 0.

1.4.2 Stationary check by autocorrelation function

From the description statistics of the price difference, we found the products have very different price reports from *Crows'* and *Random Lengths*. Then we focused on these products to study their changes over time using autocorrelation function.

The stationarity of time series tells whether the mean and autocorrelation functions change over time or have any trends. If price data are non-stationary, then conventional statistical procedures cannot be used to provide reliable hypothesis test on the true parameter values and this series' parameters such as the mean and variance will not change over time and do not follow any trends. Parameters such as the mean and variance, if they are present, also do not change over time and do not follow any trends. It is possible that individual variables are non-stationary, but a linear combination of them is stationary. The variables are then considered to be cointegrated which are still possible to test the relationship among the levels of economic variables (Engle and Granger, 1987).

If the prices from *Crows'* and *Random Lengths* both fairly represented the real price in the market, then we expect their difference time series to be stationary. If the price difference is

stationary, even the mean is not 0, however, at least we can tell that the price differences from *Crows'* and *Random Lengths* have constant mean and variance that does not change over time and have no trend.

The traditional way to detect the stationarity of time series is the unit root test. However, the unit root test for stationary loses power if structural breaks appeared in data. If the price data is non-stationary then the structural breaks or shocks will have a permanent effect, implying that data with structural breaks will cause errors on time series forecasting and other tests. Thus, Autocorrelation function (ACF) was applied to check the stationarity of the price data. The ACF measures correlation of a series with itself shifted by time delay; and if the ACF decay slowly and remains well above the significance range, this is indicative of a non-stationary series.

1.4.3 Binomial test

Random Lengths and *Crows'* both claim every week that they do hundreds of phone surveys to produce the average price and interviewed different sellers and buyers from random selection (Random Lengths publish, 2003). Thus we can assume the independence of each observation approximately. The autocorrelation of price depends on the accuracy of price difference between *Random Lengths* and *Crows'*. If the accurate value can be reduced into categorical variables, the effects of time dependence can also be reduced. So *Random Lengths* and *Crows'* should then have an equal chance to give a higher value.

Assuming the price from *Random Lengths* and *Crows'* are both from properly survey methods, *Random Lengths* and *Crows'* can have an equal chance to publish a higher price than each other every week in the past 20 year. The binomial test can tell if the price from *Random Lengths* has 50% chance to be higher than the price from *Crows'*.

In this case, a remarkably simple proof of the corresponding result can be shown in the theoretic probability.

Let a_i=the price of Crows', b_i=the price of Random Lengths, i=1,2, ...,816.

First, we removed the observation that $ai = bi$, and let n equals to the rest of the observations. The observations of $ai = bi$ are few, especially for the products price with high mean difference, then we consider the observation with $ai = bi$would not affect the result of binomial test.

State y_i is $\Omega \in \{1,0\}$.

$$y_i = \begin{cases} 1 & if\ ai > bi \\ 0 & otherwise \end{cases} \qquad (1)$$

Then the proportion of $a_i > b_i$ would be:

$$P(a_i > b_i) = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (2)$$

So, if the data is equal here, the expected value of

$$P(a_i > b_i) = 50\%, \text{ or } P(b_i > a_i) = 1 - P(a_i > b_i) = 50\%.$$

The binomial test was used to find out for which products, the $P(a_i > b_i)$ was significantly different from 50%.

The hypothesis of the test would be:

$$H_0: \ P(a_i > b_i) = 50\%$$

$$H_a: \ P(a_i > b_i) \leq 50\% \text{ or } P(a_i > b_i) \geq 50\%$$

1.5 Results and discussion

1.5.1 Summary statistics

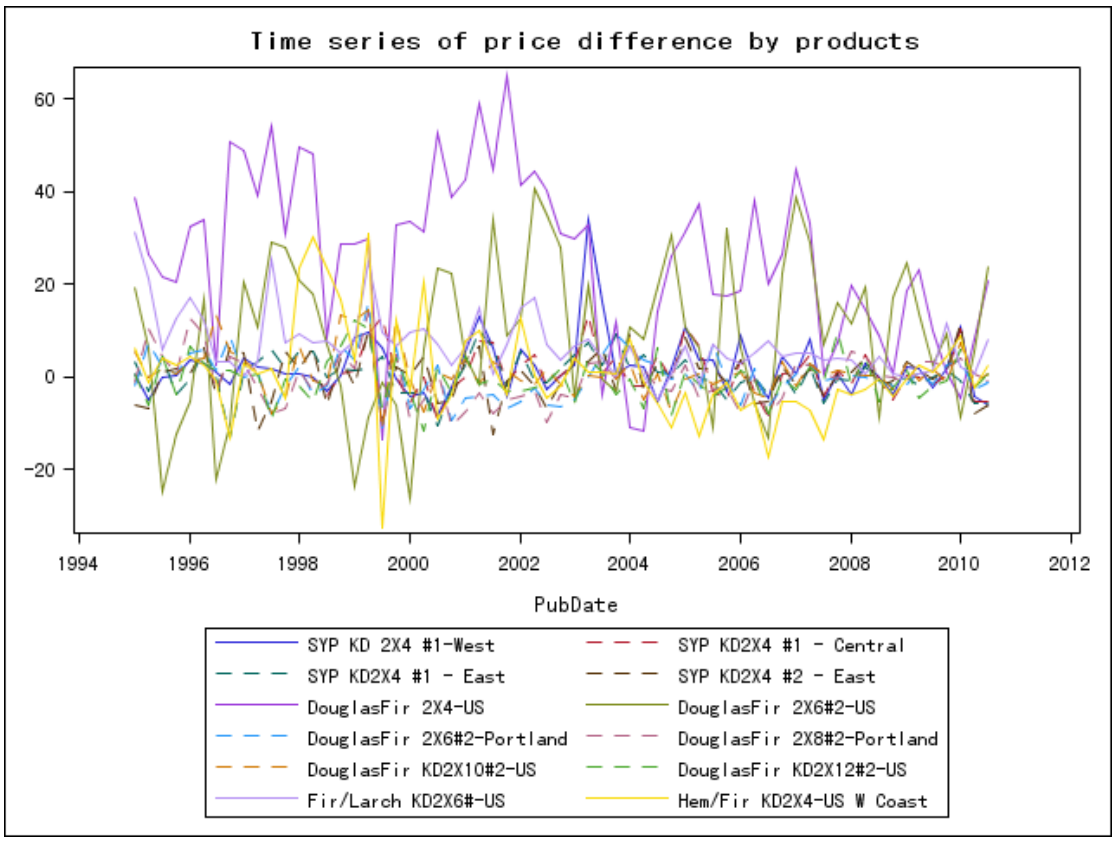


Figure 1.1 Time series of all 12 products price differences between *Crows'* and *Random Lengths*

For the 12 products' price differences between *Random Lengths* and *Crows'*, all differences have linear with mean 0 and a relatively small standard error over time is expected. In Figure 1.1, the picture reveals that for the 12 products' price difference changes during 1995 and 2011. This big picture gives us the understanding that not all time series lie in the range around 0 frequently. In Fig 1.1, the dashed lines show 7 products that have their price differences in a small range around 0. However, there are 5 time series with solid lines, indicating that they are far away from 0 mostly.

Similar results were also presented in description statistics from Table 1.2 (1) For Douglas fir products' price difference between *Random Lengths* and *Crows'*, two of them are distributed significantly different away from mean 0, and both of their standard error is larger than 0.02 which are larger than the other products. For Douglas fir Green 8' 2X4 Std&Btr-US, the price mean difference is the highest with 26 US dollars. Douglas Fir Green 8' 2X6 #2&Btr-US has 9 US dollars price mean difference which is the second highest. (2) For Southern Yellow Pine and Douglas fir products, the results show the mean difference are smaller than 1 US dollar, and the standard error are smaller than 0.018. Especially for SYP KD 8' 2X4 #2-East, this products' price mean difference between *Random Lengths* and *Crows'* is the smallest since the mean is 0.106 US dollars and standard error is 0.012. (3) For other species, Fir/Larch KD Random Tally 8'-20' 2X6 #3-US have average price difference with 6.783 US dollars and standard error 0.015, which is also higher than expecting.

In Fig 1.2, the box-plot reveals that products' mean of the price difference between *Random Lengths* and *Crows'*. The Douglas fir products show higher mean and larger variability than southern yellow pine products.

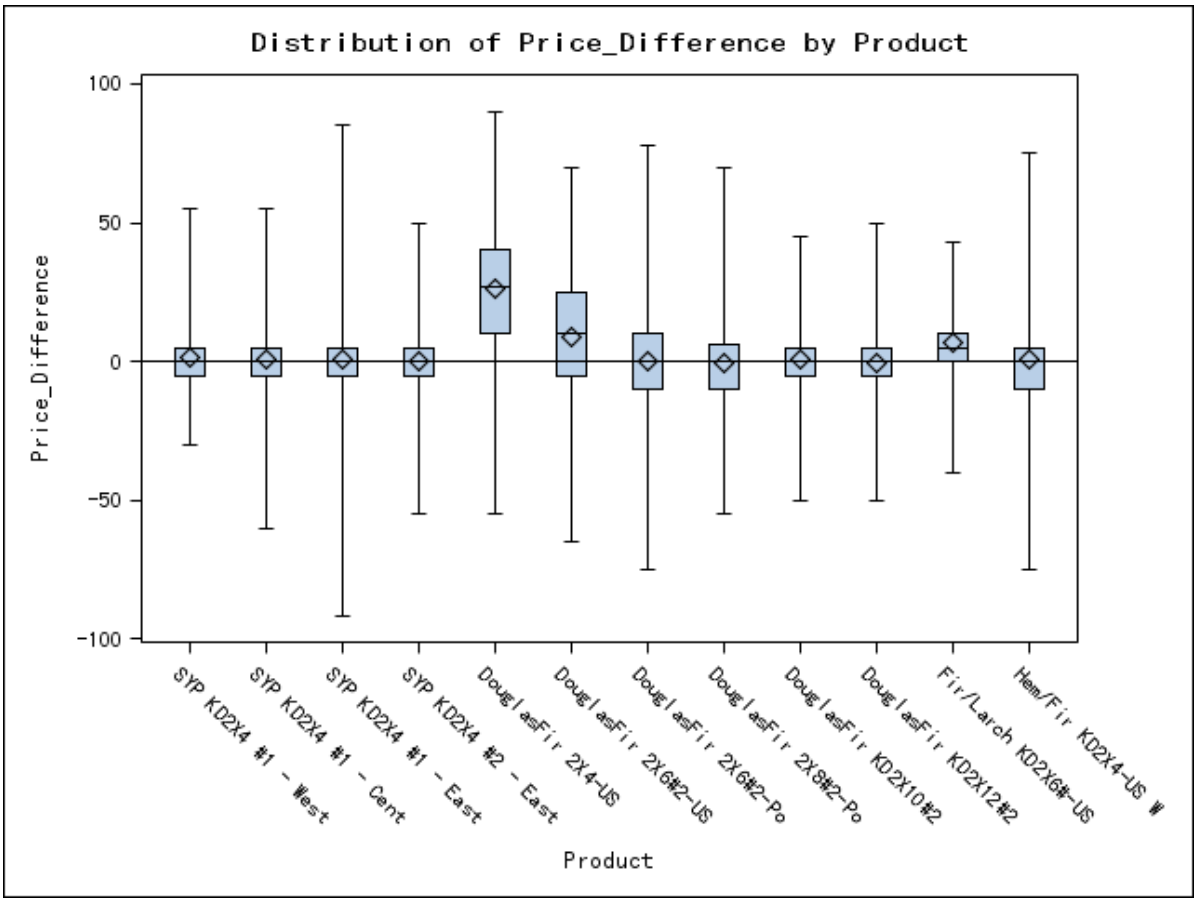


Figure 1.2 Box-plot of price difference for 12 products from *Crows'* and *Random Lengths*

Table 1.2. Mean and standard error of price difference between *Crows'* and *Random Lengths*

Products	Mean of Price difference	Standard Error
Southern Yellow Pine		
SYP KD 8' 2X4 #1-West	1.913	0.0185
SYP KD 8' 2X4 #1-Central	-0.9271	0.0189
SYP KD 8' 2X4 #1-East	0.448	0.0121
SYP KD 8' 2X4 #2-East	0.106	0.0128
Douglas Fir		
Douglas Fir Green 8' 2X4 Std&Btr-US	26.012	0.0275
Douglas Fir Green 8' 2X6 #2&Btr-US	9.052	0.0266
Douglas Fir Green Random Tally 8'-20' 2X6 #2&Btr-Portland	-0.146	0.0185
Douglas Fir Green Random Tally 8'-20' 2X8 #2&Btr-Portland	-0.552	0.0172
Douglas Fir KD Random Tally 8'-20' 2X10 #2&Btr-US	0.882	0.0136
Douglas Fir KD Random Tally 8'-20' 2X12 #2&Btr-US	-0.53	0.0134
Other		
Fir/Larch KD Random Tally 8'-20' 2X6 #3-US	6.783	0.0153
Hem/Fir KD 8' 2X4 Std&Btr-US West Coast	0.837	0.0205

1.5.2 Stationarity

In this study, the only concerned with the properties of autocorrelation for stationarity. The elements of price difference are the linear combinations, with complex coefficients, and their limits for mean square convergence. So if the ACF decay quickly, that means the time series is stationary. If the price difference is stationary, that means the price from *Random Lengths* and *Crows'* are stable. Even the mean difference is different, still the price from them are compatible over time. If the price difference between *Random Lengths* and *Crows'* is non-stationary, then the price difference has a trend over time. The original time series of *Crows'* price has a different trend from Random Lengths. The price difference between *Random Lengths* and *Crows'* is not stable with a constant mean.

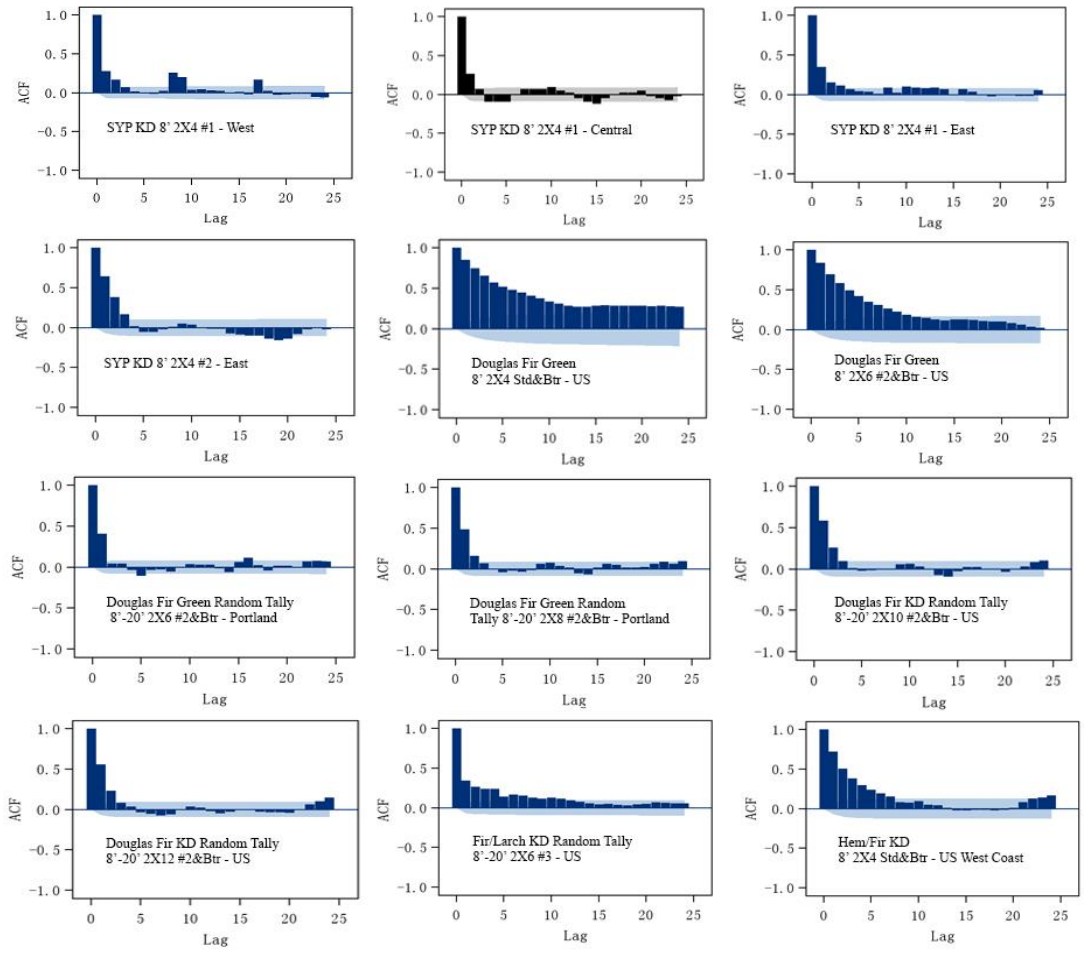


Figure 1.3 The ACF of price difference between *Crows'* and *Random Lengths*

Nonstationarity is reflected in a pattern of significant autocorrelations that do not decline quickly with increasing lag, not in the size of the autocorrelations. The ACF for Douglas Fir Green 8' 2X4 Std&Btr-US, Douglas Fir Green 8' 2X6 #2&Btr-US, Fir/Larch KD Random Tally 8'-20' 2X6 #3-US, and Hem/Fir KD 8' 2X4 Std&Btr-US West Coast were not decayed quickly with the increase of lag. The results showed the behaviors of the autocorrelation function for these products when the processes are non-stationary. The price difference between *Random Lengths* and *Crows'* for these four products are non- stationary. So for these four products, the price difference between *Random Lengths* and *Crows'* is changed over time, which conflict the expectation that *Random Lengths* and *Crows'* give same price over time.

1.5.3 Binomial test

Binomial test can be used to test if two sequences are match to each other. The binomial test is used for detect if *Crows'* price and *Random lengths* price have equal chance to be higher. Since *Crows'* price and *Random Lengths* price are both came from independent survey methods every week, we can assume their prices are not dependent on previous price report, and these prices should be randomly higher or lower to each other in a small range. Then we observe the chance P for *Crows'* price is higher than *Random Length*, and the binomial test can tell if P is equal to 0.5. If P equals to 0.5, then the *Crows'* price and *Random Lengths* price is distributed like the probability outcome of tossing a fair coin, either of them have equal chance to be higher.

Same method has been used in other field of study. Zo and Colwell (2008) studied the 16s RNA sequences. In this study a binomial test was performed if these two RNA sequences are duplicated. In Table 1.3 the binomial test reject null hypothesis for the 3 products which we focused on in previous methods. The statistics indicates the probability of *Crows'* price is higher than *Random Lengths* price for these 3 products are not equal to 50%. For these three products, Douglas Fir Green 8' 2X4 Std&Btr-US, Douglas Fir Green 8' 2X6 #2&Btr-US, and Fir/Larch KD Random Tally 8'-20' 2X6 #3-US, Crows' are more frequently to give higher price than Random Lengths. These three are also the products with highest price mean difference between *Random Lengths* and *Crows'*. The other interesting part is for KD Random Tally 8'-20' 2X12 #2&Btr-US, the p-value for binomial test is 0.9889. The large p-value represent that the alternative hypothesis should choose the other direction, which means for KD Random Tally 8'-20' 2X12 #2&Btr-US, the price from *Crows'* is more frequently to give lower price than *Random lengths*.

Table 1.3 Binomial test for price of *Crows'* is higher than *Random Lengths*

Species and Products	Number of total observations	Number of Observations have equal price	Number of Observations Crows' price is higher	P(Crows' price>Random Lengths)	Binomial test Z statistics	P-value
Southern Yellow Pine						
SYP KD 8' 2X4 #1-West	816	204	329	0.54	1.86	0.0344
SYP KD 8' 2X4 #1-Central	630	176	241	0.53	1.31	0.1025
SYP KD 8' 2X4 #1-East	816	253	275	0.49	-0.55	0.7224
SYP KD 8' 2X4 #2-East	816	152	315	0.47	-1.32	0.9128
Douglas Fir						
Green 8' 2X4 Std&Btr-US	816	34	708	0.91	22.67	<.0001
Green 8' 2X6 #2&Btr-US	816	61	527	0.7	10.88	<.0001
KD Green Random Tally 8'-20'2X6 #2&Btr-Portland	816	73	362	0.49	-0.7	0.7684
KD Green Random Tally 8'-20'2X8 #2&Btr - Portland	816	113	343	0.49	-0.64	0.7514
KD Random Tally 8'-20'2X10 #2&Btr-US	816	159	356	0.54	2.15	0.0175
KD Random Tally 8'-20'2X12 #2&Btr-US	816	173	293	0.46	-2.25	0.9889
Other						
Fir/Larch KD Random Tally8'-20' 2X6 #3-US	816	200	560	0.91	20.31	<.0001
Hem/Fir KD 8' 2X4 Std&Btr-US West Coast	816	177	317	0.5	-0.2	0.5938
Total	9606	1775	4626	0.59	16.06	<.0001

Combining all products together, then test if *Crows'* price has 50% chance to be higher than price from *Random Lengths* for all. We have a relatively large trail number of 7831. However, the binomial test still reveals the P is not equal to 50%. So for all the test result, we can say *Random Lengths* and *Crows'* are failed to have equal chance to give higher price. *Crows'* are more often to give higher price. So if we assume the price from *Random Lengths* and *Crows'* as two sides of a coin, then this coin is not fair.

1.6 Conclusions

When using third party external price recourse as the real marketing price, it is very necessary to find out how much difference between them. Since *Random Lengths* and *Crows'* refuse to expose any phone call survey information, we can compare their price to find how much difference between them to estimate the error rate.

During 1990 to 2010, for 9 of the products, the price differences between *Random Lengths* and *Crows'* for most products are less than $1 with standard error from 0.012 to 0.018. Southern Yellow Pine products have more similar price between *Random Lengths* and *Crows'* than Douglas Fir Green products. Two products of Douglas fir have very different price between *Random Lengths* and *Crows'*. The Douglas fir Green 8' 2X4 Std&Btr-US has the largest price difference between Random Lengths and Crows' with mean difference $26 with standard error 0.027. The Douglas Fir Green 8' 2X6 #2&Btr-US has the second largest price difference between *Random Lengths* and *Crows'* with mean difference $9 with standard error 0.026. Also these two product price difference is not stationary over time, means the price difference has trend. For these two products the price, Crows' are not only given higher value on average price, but also *Crows'* price is significantly more frequently to give higher price than *Random Lengths*. Since no information about the phone call survey from either of them, we cannot conclude which price is wrong, however, if the same editor made the price for both of these two products and none of others, then we can question there is something wrong with this price.

1.7 References

Engle, R.F., Granger, C.W.J., 1987. Cointegration and error correction: representation, estimation, and testing. Econometrica 55 251-276.

Gröger, J.P., Missong, M., Rountree ,R.A., 2011.Analyses of interventions and structural breaks in marine and fisheries time series: Detection of shifts using iterative methods. Ecological Indicators 11 1084-1092.

Harvey, D., Leybourne, S., Taylor, T., 2013. Testing for unit roots in the possible presence of multiple trend breaks using minimum Dickey–Fuller statistics. Journal of Econometrics 177 265-284.

Horikoshi, Y., Takemura, A., 2008. Implications of contrarian and one-sided strategies for the fair-coin game. Stochastic Processes and their Applications 118 2125-2142.

King, A., Ramlogan-Dobson, C., 2011. Nonlinear time-series convergence: The role of structural breaks. Economics Letters 110 238-240.

Kopalle, P.K., Lindsey-Mullikin, J., 2003. The impact of external reference price on consumer price expectations. Journal of Retailing 79 225-236.

Luca, P., 2014. Why are discounted prices presented with full prices? The role of external price information on consumers' likelihood to purchase. Food Quality and Preference 31 69-80.

Sen, A., 2008. Behaviour of Dickey–Fuller tests when there is a break under the unit root null hypothesis. Statistics & Probability Letters 78 622-628.

Salcedo, G.E., Porto,R.F., Morettin, P.A., 2012. Comparing non-stationary and irregularly spaced time series. Computational Statistics and Data Analysis 56 3921-3934.

Shook, S.R., 1999. Forecasting Adoption and Substitution of Successive Generations of Structural Wood Panel Products in the United States. Forest Science 45(2) 232-248.

Song, P.X., Freeland,R.K., Biswas, A., Zhang, S., 2013. Statistical analysis of discrete-valued time series using categorical ARMA models. Computational Statistics and Data Analysis 57 112-124.

Yin, R., Baek, J., 2004. The US–Canada softwood lumber trade dispute: what we know and what we need to know. Forest Policy and Economics 6 129-143.

Zo, Y., Colwell, R.R., 2008. A simple binomial test for estimating sequencing errors in public repository 16S rRNA sequences. Journal of Microbiological Methods 72, 166-179.

Chapter 2 Analyzing the effects of environmental and chemical factors for *Aphanizomenon flos-aquae* density in Willow Creek Reservoir, Oregon

2.1 Abstract

In Willow Creek Reservoir (WCR), Oregon, the *Aphanizomenon flos-aquae* density is significantly affected by nitrite, reactive phosphorus, Temperature, dissolved oxygen, and pH. For modeling the response of just one species, it is important to find the dominant factor which is quite difficult. If there is not reliable information to determine the dominant factor for dominant species, the selected data can used to build model. In WCR, for modeling *Aphanizomenon flos-aquae* density, the observation should choose only when its density is higher than 200 counts per liter. From the PCA analysis When eliminate the observation with *Aphanizomenon flos-aquae* density lower than 200 counts per liter, the model parameters estimation will of reactive phosphorus will have positive effects on *Aphanizomenon flos-aquae* density.

2.2 Introduction

Among blue-green algae, *Aphanizomenon flos-aquae* produce up to 6.6% dry weight of the toxin (Preu et al., 2008). Its toxin can reach high concentrations in water bodies if cyanobacteria dominant the phytoplankton community (Preu et al., 2008). *Aphanizomenon flos-aquae* can produce the toxins microcystins (MCs), anatoxins (ANTX), saxitoxins (STXs) and cylindrospermopsin (CYN) (Ledreux et al.2010). Monitoring of cyanobacteria and their potential toxicity can help to identify in lake areas in lake contaminated with these toxin (Ledreux et al., 2010). Cyanotoxins originating from reservoirs with toxic cyanobacterial blooms have caused human health concerns (Graham et al., 2012).

The Blue-green algae density in fresh water is dependent on temperature, and nutrient concentrations, particularly those of nitrogen (N) and phosphorus (P) (Paerl and Huisman, 2008). In fresh water ecosystems, the toxic cyanobacteria can cause health risks for animals and human beings (Hitzfeld et al., 2000).

Figure 2.1 Willow Creek Reservoir map. Source: Google Map; accessed 3 June 2014.

Nitrogen (N) and phosphorus (P) are usually considered as the most important nutrient elements for the growth of algae, as they are typically the limiting nutrients in freshwaters. They are required elements for the biological protein synthesis, synthesis of DNA, RNA molecules. So studies the N and P affections on algae density are important. For the algae density prediction modelling work, the significant effects of Nitrogen (N) and phosphorus (P) are expected. The total nitrogen to total phosphorus ratio (TN:TP) indicates cyanobacterial dominance in phytoplankton communities. TN:TP ratios >30 (by weight) were proposed as a threshold to reduce blue-green algae blooms (Harris, 2014, Pick and Lean, 1987). Adding

nitrogen to eutrophic systems is counterintuitive. Additionally, adding nitrogen to increase the TN:TP ratio to >50 in systems with high phosphorus concentrations (e.g., TP>100 μg/L) could result in extremely high concentrations of nitrite (Harris, 2014).

One research group in Dianchi, China showed that the blue-green algae bloom conditions depend on the Temperature and water nutrient concentrations. The algae bloom was caused by two species of cyanobacteria;Microcystis which prefers warm Temperature, while *Aphanizomenon flos-aquae* which prefers cooler Temperature. And also the levels (e.g. total nitrogen and total phosphorus) can decide which one of the species dominant the bloom (Liu, 2005).

Instead of predicting whether a blue-green algae bloom will happen in the lake, the purpose of this study is to predict the *Aphanizomenon flos-aquae* density in the willow creek based on the chemical conditions. Due to the competition between different algae species, the rich of N and P concentration does not guarantee the high *Aphanizomenon flos-aquae* density, although the total density of all algae species are high. Before the factor controls the dominant species is determined, the low *Aphanizomenon flos-aquae* density observations need to delete to reduce the dominant species effects. In order to predict the *Aphanizomenon flos-aquae* density, there should be a threshold for the *Aphanizomenon flos-aquae* density to reduce the competition effects on the model. From former study, *Aphanizomenon flos-aquae* is affected by temperature and the nutrient levels. However, different kinds of nutrient include nitrate, ammonia, total kjeldahl nitrogen, total phosphorus, soluble reactive phosphorus, and soluble reactive silicon would be test affect the density of *Aphanizomenon flos-aquae* is tested in the modelling process.

2.3 Data

2.3.1 Study site

Willow Creek Reservoir located in the eastern Oregon (Fig. 2.1). It is surface area is 0.52 km2 with volume 237.3 × 104 m3. The total Phosphorus concentration is 30 μg/L, summer chlorophyll *a* concentration is 10-15 μg/L) (Harris, et al. 2014). Willow Creek is the main (~90% of total inflow not including precipitation) perennial inflow from the south, while the smaller and seasonally intermittent Balm Fork Creek enters from the southwest (Fig. 2.1; USACE 2007). Since the completion of the dam in 1983, WCR has had annual cyanobacterial blooms during summer (>100,000 cells/mL; USACE 2007).

2.3.2 Sampling method

Data for this study originated from the regular monitoring of Willow Creek Reservoir undertaken by personnel of the US Army Corps of Engineers or their contractors starting in 1985. Collections originated primarily from 0 and 12 m of depth (epilimnetic waters) collected at a site just south of the dam typically at biweekly intervals between April and October and monthly at other times. Most years only the growing season was sampled, while occasionally sampling was continuous for the entire year. Depending on constituent of interest, analyses included profiles of temperature, dissolved oxygen, pH, conductivity and light, while samples for the analysis of zooplankton, algae biomass and species were collected from 0.5 m. Chemical analyses included, total and dissolved P, total Kejldahl nitrogen, ammonia, nitrate and nitrite, total and dissolved iron, and manganese. Analyses of algal species were done by Jim Sweet for the entire period of the study, so there is no risk of misidentification of species or counts.

For these analyses, I considered data for *Aphanizomneon flos-aquae*, the dominant species of cyanobacteria that occurred in WCR from 1990 to 2007, provided by the US Army corps of engineers from their database.

Data included 180 observations with *Aphanizomenon flos-aquae* density. Of these 103 observations include complete data and 77 observations have missing value (Table 2.1.).

Table 2.1 Data summary: The N Miss shows the number of missing values in each variable; environmental variables have a few missing values, while the chemical variables have more missing values

Variable	N Miss	N	Minimum	Maximum	Mean	Std Dev
Algae density	0	180	-2.958	9.676	5.318	2.088
Temperature	7	173	1.64	24.200	16.483	5.235
DO	6	174	0	14.620	8.394	3.194
pH	20	160	7	9.500	8.567	0.593
Cond	21	159	172	548	231.629	37.635
Total_Phos	74	106	0.035	0.461	0.116	0.085
Phosphorus	63	117	0.0012	0.294	0.055	0.061
TKN	81	99	0.0026	1.680	0.641	0.312
Ammonia	63	117	0.002	1.200	0.164	0.187
Nitrate	58	122	0	0.322	0.041	0.062
Silica	56	124	6. 0	50.750	23.505	8.555

2.4 Methods

2.4.1 Linear regression model with completed data

The linear regression model only with main effect is applied with complete data. This linear regression model includes parameter estimate, R-square, and Akaike Information Criterion (AIC) for stepwise model selection. The stepwise model selection is common procedure for model selection (Bengtsson, 2006). At the beginning, the scientific grounds were decided, multiple regression models to will be compared, that is, narrow down the number of variables to be considered for best fit.

The other benefits of the stepwise selection is reducing the variables numbers at the beginning, will help to dealing with missing data. The EM algorithm method is usually applied to parameter estimation with incomplete data (Park et al., 2007). However in this study some variables include 50% of missing values. If there are too many missing values,

the parameter estimation would depend on impute values as replacement of missing value more than real value. Reducing variables will help to avoid this risk.

Secondly, all possible models in the combination of variables from the data will be used to estimate the parameters for linear regression model with main effects. Maximum Likelihood estimation of parameters in all the models would be listed (Nie, 2007). Then for each model, calculate its AIC. Pick the model with the smallest AIC. That is the model in the suite with the best overall statistical properties and parameter balance (Bengtsson, 2006).

2.4.2 Principle component analysis

The principal components analysis (PCA) is a matrix based method for multiple dimensional analyses. The PCA can reduce the dimensions and give a clear view of the data distribution. Applying principal component analysis to the algae data patterns by to a matrix of seven columns (7 variables selected by AIC) and 103 rows (the data observation with complete data) (Godoy, 2014).

PCA is used to determine spatial and Temperature changes of physical and chemical conditions, also the associate with the total density of *Aphanizomenon flos-aquae*. In the PCA analysis, it included temperature, dissolved oxygen, conductivity, pH, nitrate ($N\text{-}NO_3^-$), ammonia ($N\text{-}NH_4^+$), total phosphorus, and reactive phosphorus (orthophosphate).

2.4.3 EM algorithm

In statistical parameter estimation problems involving missing values in data, the Expectation Maximization (EM) algorithm is a prime tool (Park, 2008). The EM algorithm is a bootstrap based method to compute the maximum likelihood estimates for parameters of data with missing values.

The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of data missingness. In MLE estimation, the idea is to

estimate the model parameters with the observed data are the most likely. Each iteration of the EM algorithm consists of two processes: The E-step, and the M-step. In the expectation (E-step), the missing data are estimated given the observed data and current estimate of the model parameters. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimates of the missing data from the E-step are used in lieu of the actual missing data. Convergence is assured because the algorithm is guaranteed to increase the likelihood at each of the iteration.

Suppose we wish to fit the linear regression parameter estimation with a probability density function (PDF) $f(\cdot|\theta)$. From an independent and identically distributed observed data Y = (Yi), we can find the likelihood function given by

$$L(\theta|Y)=\prod_{i=1}^{M} f(Yi|\theta)$$

The log likelihood is applied to find the maximized estimate of $\hat{\theta}$

$$l(\theta|y) = \log L(\theta|y) =\sum_{i=1}^{M} log\ f\ (yi|\theta),$$

Since the log likelihood is maximized, we have

$$\hat{\theta}= \text{argmax}\theta\in 2l(\theta|y).$$

In the next step, in addition to the observed data Y, some extra missing value set in data Z(*Zi*, $1\leq i\leq l$.)In that case, since the unobserved data is unknown, we need to maximize the conditional expected value of the complete log likelihood, given the value of the observed data Y = y and some current value θ(x) for the parameters. The E-step in the EM algorithm then requires the calculation of

$$Q(\theta, \theta_{(x)})=E_z(l_c(\theta|X)|Y,\theta_{(x)})$$

In the M-step $Q(\theta, \theta_{(x)})$ is maximized, that is, θ(x+1) is found so that

$$Q(\theta_{(x+1)}, \theta_{(x)}) \geq Q(\theta, \theta_{(x)})$$

2.5 Results and discussion

2.5.1 Linear regression model without missing value

There were 11 variables for water quality factors and *Aphanizomenon flos-aquae* density with 180 samples of each (the sample was collected from 1990 to2007). However, because the data was combined from different projects, and not all samples were collected on each sampling occasion, only 93 of the observations are without missing values. The lease square fit estimation requires the observations to be complete. So this linear regression model is based on 93 out of 203 observations

The linear regression model with diagnostic plot showed the response variable the total density of *Aphanizomenon flos-aquae* needs log transformation to fit.

The model would be

$$Log(Y_i)=B_0+B_1X_{1i}+B_2X_{2i}+\ldots\ldots+B_jX_{ji}+E_i$$

Also the stepwise variable selection based on Akaike Information Criterion (AIC) is processed to select variables.

$$AIC=2k-2\ln(L)$$

Where k is the number of variables selected in the model and L is the maximized likelihood function of the model.

Table 2.2 Stepwise variable selection by AIC

Variable Number	R^2	AIC	Variables in Model
4	0.3576	101.7139	Temperature pH Cond ReactivePhosphorus
4	0.3569	101.809	DO pH Cond ReactivePhosphorus
4	0.3564	101.8753	TemperatureDissolved Oxygen PH ReactivePhosphorus
4	0.3562	101.9111	pH Cond ReactivePhosphorus Ammonia
5	0.3916	98.9216	TemperatureDissolved Oxygen pH ReactivePhosphorus Nitrate**
5	0.3853	99.8299	Temperature pH ReactivePhosphorus Nitrate Silica
5	0.3815	100.3811	Temperature pH Cond ReactivePhosphorus Nitrate
5	0.3781	100.8546	Temperature pH ReactivePhosphorus TKN Nitrate
5	0.3779	100.8883	pH Cond ReactivePhosphorus Nitrate Silica
5	0.3724	101.6617	pH Cond ReactivePhosphorus TKN Nitrate
5	0.3715	101.7934	Temperature PH Cond ReactivePhosphorus Silica
6	0.3987	99.8876	Temperature DissolvedOxygen pH Cond ReactivePhosphorus Nitrate
6	0.3974	100.0826	Temperature DissolvedOxygen pH ReactivePhosphorus Nitrate Silica
6	0.3965	100.2109	Temperature pH Cond ReactivePhosphorus Nitrate Silica

The AIC can give a view of the model quality to the state dataset. The function is not evaluating the AIC for all possible models but uses a search method that compares models sequentially. The following variables: Temperature, DO, pH, Reactive Phosphorus, and Nitrate were selected in the linear regression model with minimum AIC (Table 2.2).

Table 2.3 the parameter estimation for data without missing value R^2=0.3916

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-14.006	2.9469	-4.75	<.0001
Temperature	-0.098	0.04333	-2.27	0.0254
DO	-0.117	0.06848	-1.71	0.0906
pH	2.575	0.42459	6.07	<.0001
Reactive Phosphorus	-8.831	3.0623	-2.88	0.0048
Nitrate	6.429	2.96613	2.17	0.0326

The selected model from Table 2.3 shows the temperature, DO, pH, reactive phosphorus, and Nitrate have significant effects in the model to predict the density of *Aphanizomenon flos-aquae*. Sheng and his team found Temperature, Nitrate, and pH had significant effect on blue-green algae bloom break out, Reactive Phosphorus has stead positive effects; the

influences of Total Phosphorus and DO can be ignored. (Sheng et al. 2012). The model for predicting *Aphanizomenon flos-aquae* total density (Yi=Number of counts per liter) is:

$$\text{Log(Yi)} = -14.006 - 0.098 * \text{Temperature (℃)} - 0.117 * \text{DO (PPM)} + 2.575 * \text{pH} - 8.831 * \text{Reactive Phosphorus } \left(\frac{\text{mg}}{\text{L}}\right) + 6.429 * \text{ Nitrate } \left(\frac{\text{mg}}{\text{L}}\right) + \varepsilon \qquad \text{Model (1)}$$

2.5.2 EM algorithm

The model (1) is based on the observations without missing values. In all 180 observations which capture the *Aphanizomenon flos-aquae* density information, only 103 of them have complete data without missing value. That means model (1) only use 50% of observations to estimate the model parameters. So The EM algorithm is processed to find if the

Table 2.4 The missing data pattern: The data displays a data set of three variables with a non-monotone missing pattern

Group	Yi	Temp	DO	pH	Cond	Phosphorus	Nitrate	Frequency	Percent
1	X	X	X	X	X	X	X	103	51.24
2	X	X	X	X	X	.	.	1	0.5
3	X	X	X	X	.	X	X	54	26.87
4	X	X	X	.	.	X	X	2	1
5	X	X	X	.	.	.	X	6	2.99
6	X	X	X	.	.	.	.	4	1.99
7	X	.	X	.	X	.	.	3	1.49
8	X	.	.	.	.	X	X	1	0.5

EM algorithm requires the data missing at random (MAR). Table 2.4 lists distinct missing data patterns with corresponding frequencies and percentages. The value of "X" means that the variable has a valid value in the corresponding group; and a "." means that the variable is missing. From Table 2.4, the missing pattern show the data missingness is MAR which allows EM algorithm to estimate the parameters.

Table 2.5 The EM algorithm for parameter estimation without missing value R^2=0.3911, EM Algorithm R^2=0.3415

	EM algorithm		**Estimate without missing value**	
Variable	Parameter Estimate	Pr > |t|	Parameter Estimate	Pr > |t|
Intercept	-12.97913	<.0001	-14.00639	<.0001
Temperature	-0.07522	<.0001	-0.09833	0.0254
DO	-0.04765	<.0001	-0.11702	0.0906
pH	2.35087	<.0001	2.57527	<.0001
Reactive Phosphorus	-8.99136	<.0001	-8.83127	0.0048
Nitrate	5.38887	<.0001	6.42968	0.0326

From the EM algorithm, we had model (2) parameter estimation

$$\text{Log(Yi)} = -12.97913 - 0.07522 * \text{Temperature (℃)} - 0.04765 * \text{DO (PPM)} + 2.35087 * \text{pH} - 8.99136 * \text{Reactive Phosphorus } \left(\frac{\text{mg}}{\text{L}}\right) + 5.38887 * \text{ Nitrate } \left(\frac{\text{mg}}{\text{L}}\right) + \varepsilon i \qquad \text{Model (2)}$$

Table 2.5 is the linear regression model parameter estimation with main effect only. The parameters have no big difference, except the Dissolved Oxygen is a more significant after the EM algorithm. The other potential problem is the reactive phosphorus has a negative effection on *Aphanizomenon flos-aquae* density, which is against the common research results.

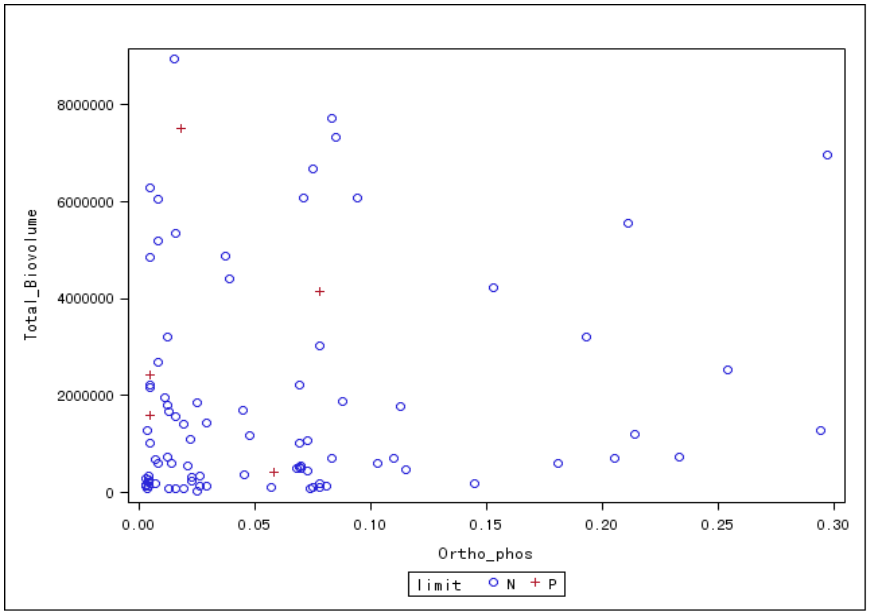
2.5.3 Dominant factor effects on model parameter estimation

Fig 2.2 (a) shows the distribution of total biovolume of all algae species versus reactive phosphorus; (b) shows total density of *Aphanizomenon flos-aquae* versus the reactive phosphorus. From the two distributions we can see a lot of the points has high biovolume content but lower density of the *Aphanizomenon flos-aquae*. In this case, the *Aphanizomenon flos-aquae* is not the dominant species. The large amount of bio volume belongs to other species including *Anabaena flos-aquae, Rhodomonas minuta, Cryptomonas erosa,* and so on. The nutrient would be consumed by the dominant species. The dominance of algae species depends on the type of algae and the certain environmental factors

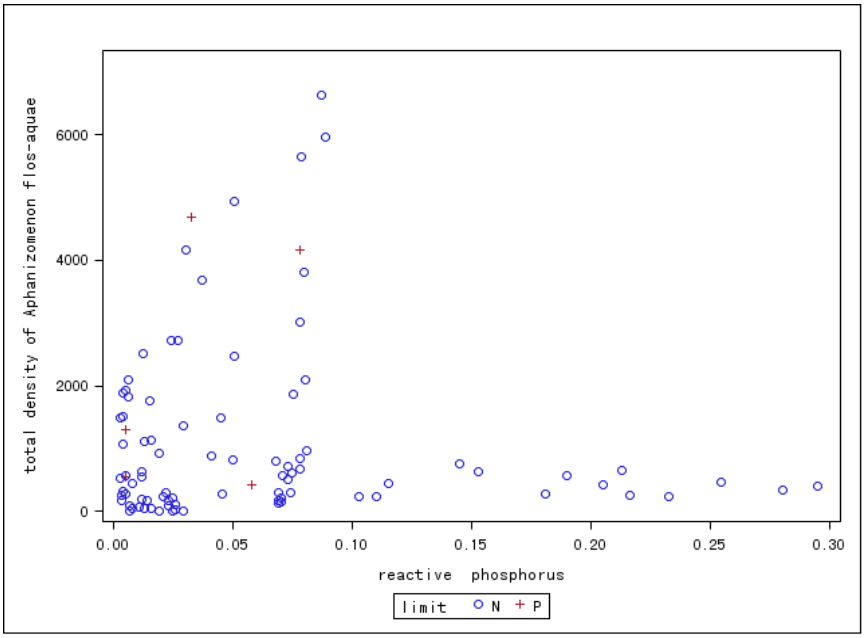
correlated to this species. For green algae, temperature and irradiance are the most important growth (Nelson et al. 2008). For example different Ulva species had different adaptabilities to temperature and irradiance which give different levels effects on their density. The seasonal variation of dominant species was due to their different ecological adaptabilities. However, the dominant algae species vary frequently (Han, 2013, Rucker, 2007). The reactive phosphorus may have positive effects on algae density, however the *Aphanizomenon flos-aquae* may not captured as dominant species at the high reactive phosphorus situation. This case will cause the error for predict *Aphanizomenon flos-aquae* density when most nutrient consumed by other species of algae.

The dominant species of algae is varying during short term study. The dominant species can be different from location, season, water type such as reservoir and water shed, and the distance from the dam (Ietswaart et al., 1999). The blue-green algae species of *Aphanizomenon flos-aquae* and *Anabaena* density can be reduced below the dam more quickly than the other species because they are specialized to live in stagnant water and the population density is negatively affected by turbulent conditions. However, in summer, some slowly-growing bacteria are still observed in the reservoirs with high density (Grabowska, 2012).

This study is trying to find a way to predict the *Aphanizomenon flos-aquae* density, and then to introduce a possible method to reduce the density of *Aphanizomenon flos-aquae* toxin in WCR or provide a management tool to decide when to close the reservoir to the public. When the *Aphanizomenon flos-aquae* density is low, or it is not the dominant species, toxin production in the water should be low, meaning there is no reason to close the lake. From Fig 2.2 (b), we can see that if we raise the baseline for data selection when the density of *Aphanizomenon flos-aquae* has to be higher than 200 per liter, there is a positive linear relationship between *Aphanizomenon flos-aquae* density and reactive phosphorus concentration.



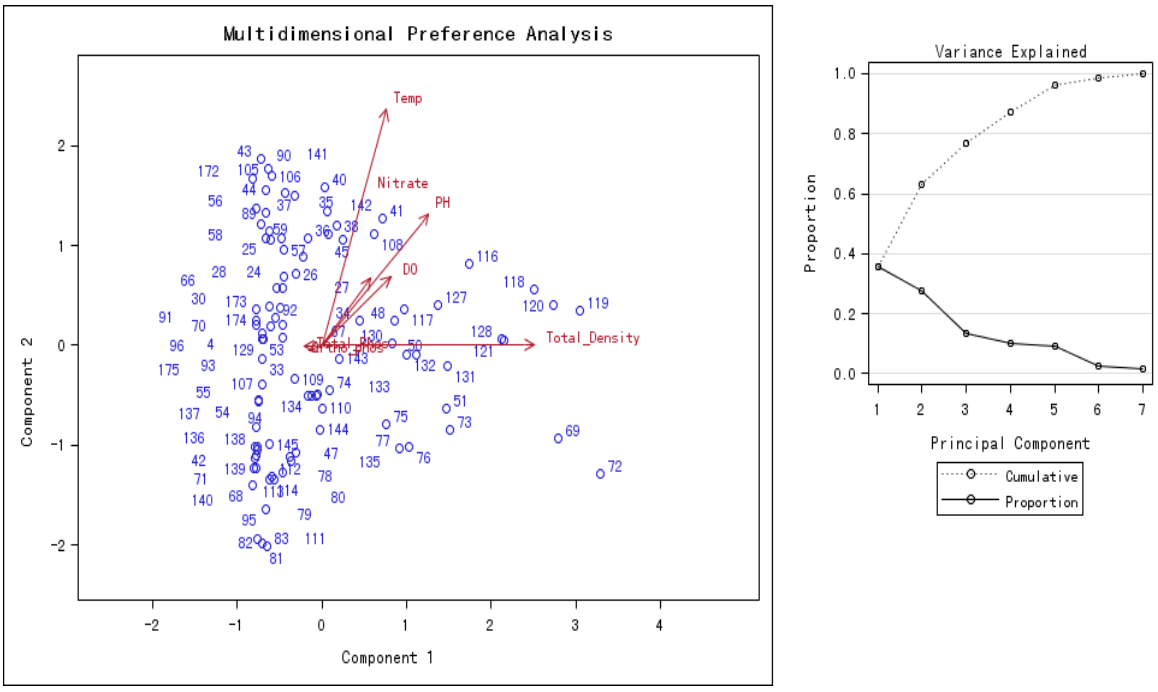
(a)



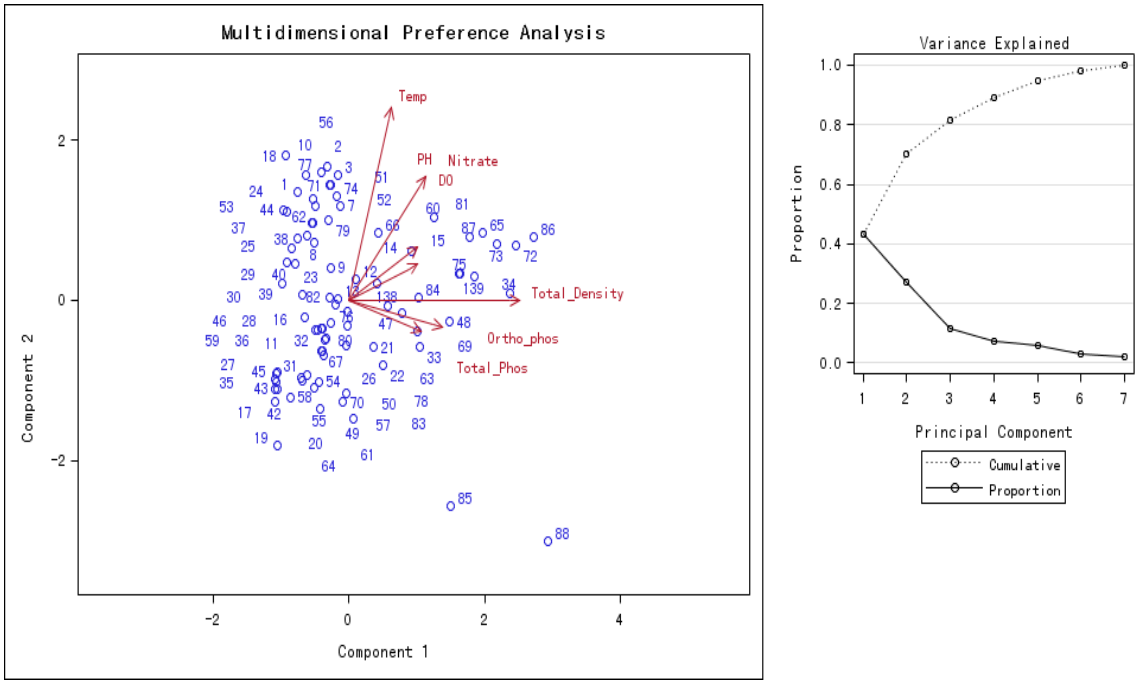
(b)

Figure 2.2 (a) Total biovolume versus reactive phosphorus. (b) total density of Aphanizomenon flos-aquae versus the reactive phosphorus

The PCA analysis takes Principle component 1 and Principle component 2 to build biplot for multivariate data visualization. Fig 2.3 (a) is the biplot for the whole data without missing value. The total density of *Aphanizomenon flos-aquae* from (a) and (b) both load on PC1 only. And the PC1 and PC2 both take about 55% from total. The biggest difference between (a) and (b) is the direction of total phosphorus and reactive phosphorus. If all data are included, the total phosphorus and reactive phosphorus are negatively related to the *Aphanizomenon flos-aquae* density. However if cell counts lower than 200 are removed the total phosphorus and reactive phosphorus will have a large positive correlation with *Aphanizomenon flos-aquae* density. The PCA analysis can give a visually idea to see the importance of data selection for modelling *Aphanizomenon flos-aquae* density. Building a model to predict one single species needs to concern the dominant factor which is different when predicting the whole algae density or the biovolume density. For single specie density estimation, the nutrients are a part of factors to predicting the algae density. The adaption for the location and the competition between algae are also important for density prediction. So if the *Aphanizomenon flos-aquae* is not the dominant species with really low density, the data would reduce the accuracy of the prediction model. The *Aphanizomenon flos-aquae* density observations which are lower than 200 hundred should be treated as outliers, even though they appeared in the dataset frequently. The purpose of this study focused on toxin of *Aphanizomenon flos-aquae* which may cause health risk, however, the low *Aphanizomenon flos-aquae* density will not cause this problem. Secondly, currently, no strong evidence indicate what factors can control the dominant species of algae in freshwater. This study only focused on the observations with high *Aphanizomenon flos-aquae* density.



(a)



(b)

Figure 2.3 PCA biplot (a) for the dataset with all algae density observations (b) for dataset with observations which the algae total density is higher than 200 per liter

Principle Component Analysis is found by calculating the eigenvectors and eigenvalues of the data covariance matrix. Also the eigenvectors and eigenvalues can be used to calculate the Variance Inflation Factor (VIF). VIF detects the severity of multicollinearity in least square regression modeling.

$$\text{VIF} = \sum_{i=0}^{k} \frac{A_{ij}^2}{L_i}$$

Where the A_{ij} were the eigenvector coefficients and the L_i were the eigenvalues from the principal component analysis of the independent variables in a regression model. Here is an illustration of that expression. Table 2.6. shows PCA results and VIF. There is no multicollinearity since the VIF are small.

Table 2.6 The eigenvectors, eigenvalues, and VIF for PCA

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	VIF
Total_Density	0.458387	0.355453	-0.154577	0.183693	-0.763899	0.14952	1.945346
Temp	0.383548	-0.458712	0.438581	0.443242	0.129795	0.486632	2.057404
DO	0.427102	-0.195598	-0.544468	-0.528517	0.224909	0.391107	1.859294
PH	0.500001	-0.372318	-0.133369	0.134703	0.037872	-0.757636	3.246768
Ortho_phos	0.29041	0.642789	-0.127781	0.374455	0.588138	-0.005755	1.521505
Nitrate	0.354988	0.270517	0.673195	-0.576283	0.041078	-0.117574	1.296461
Eigenvalue	2.732657	1.3067882	0.7655951	0.6316484	0.3715513	0.1917603	

2.5.4 EM algorithm without low *Aphanizomenon flos-aquae* density data

The new model estimation is based on the data without low *Aphanizomenon flos-aquae* density data observations. From Table 2.6, the estimate without missing value shows that the DO, pH, Reactive Phosphorus are not significant. The EM algorithm shows only DO is not a significant factor for *Aphanizomenon flos-aquae* density.

Table 2.7 The EM algorithm for parameter estimation without missing value R^2=0.2874, EM Algorithm R^2=0.2454

	EM algorithm		**Estimate without missing value**	
Variable	Parameter Estimate	Pr > |t|	Parameter Estimate	Pr > |t|
Intercept	1.23506	<.0001	2.27389	0.3116
Temperature	-0.04889	<.0001	-0.02649	0.3425
DO	-0.00533	0.1562	0.02512	0.6042
pH	0.7154	<.0001	0.5359	0.0864
Reactive Phosphorus	0.43897	0.0003	0.11595	0.9386
Nitrate	2.71256	<.0001	2.64022	0.006

The model (3) shows the Reactive Phosphorus has positive effect on *Aphanizomenon flos-aquae* density.

$$\text{Log(Yi)} = 1.23506 - 0.04889 * \text{Temperature (℃)} - 0.00533 * \text{DO (PPM)} + 0.71540 * \text{pH} + 0.43897 * \text{Reactive Phosphorus } \left(\frac{\text{mg}}{\text{L}}\right) + 2.71256 * \text{ Nitrate } \left(\frac{\text{mg}}{\text{L}}\right) + \varepsilon i \qquad \text{Model (3)}$$

2.6 Conclusions

Harmful algae blooms are becoming a serious environmental problem worldwide that potentially threatens the well-being of humans, pets and livestock via the production of potent toxins. By the increasing usage of fertilizer, algae blooms are increasing in frequency, intensity, and duration in all forms of aquatic environments (Routh et al., 2009). Most studies shows that if one is able to control the relevant environmental factors and nutrient, algal blooms can be reduced. However, the detailed mechanisms of algae bio adaption is unknown, due to the difficulty in the interactions of various environmental factors, nutrition, and even geology information that are known to change the dominant algae species (Moore et al., 2008).

Most studies for algae blooms modelling are focused on the total biomass, or predictingthe occurrence of a bloom outbreak. This study focused on *Aphanizomenon flos-aquae* density

only. These two type of study both need to explore the available nutrients of the water body, and other environmental variables.

The predicted results show for nutrition factors, nitrate has the highest positive affect on *Aphanizomenon flos-aquae* density; reactive phosphorus has lower positive effect. The environmental changes such as water temperature pH and dissolved oxygen are also significant. The future work for predicting *Aphanizomenon flos-aquae* density in WCR should be collecting samples cooperatively to avoid missing values to conclusively tease out the factors that can be used to predict the onset of blooms and potential toxin production.

2.7 References

Bengtsson. T., Cavanaugh. J. E., 2006. An improved Akaike information criterion forstate-space model selection. Computational Statistics & Data Analysis 50 2635-2654.

Conley, D., Paerl, H., Howarth, R., Boesch, D., 2009. Controlling eutrophication: nitrogen and phosphorus. Science (Washington) 323 (5917) 1014-1015.

Godoy, J.L., Vega, J.R., Marchetti .J.L., 2014. Relationships between PCA and PLS-regression. Chemometrics and Intelligent Laboratory Systems 130 182-191.

Grabowska, M., 2012. The role of a eutrophic lowland reservoir in shaping the composition of river phytoplankton. Ecohydrology Hydrobiology 12(3) 231-242.

Graham, J.L., Zielger, A.C., Loving, B.L., Loftin, K.A., 2012. Fate and transport of cyanobacteria and associated toxins and taste-and-odor compounds from upstream reservoir releases in the Kansas River, Kansas, September and October 2011. US Geological Survey Scientific Investigations Report 2012-5129.

Han, W., Chen, L., Zhang, J., Tian, X., Hua, L., He, Q., Huo, Y., Yu, K., Shi, D., Ma, J., He, P., 2013. Seasonal variation of dominant free-floating and attached Ulva species in Rudong coastal area, China. Harmful Algae 28 46-54.

Harris, T.D., Wilhelm, F.M., Graham, J.L., Loftin, K.A., 2014. Experimental manipulation of TN:TP ratios suppress cyanobacterial biovolume and microcystin concentration in large-scale in situ mesocosms. Lake and Reservoir Management 30:1 72-83.

Harris, T.D., Wilhelm, F.M., Graham, J.L., Loftin, K.A., 2014. Experimental additions of aluminum sulfate and ammonium nitrate to in situ mesocosms to reduce cyanobacterial biovolume and microcystin concentration. Lake and Reservoir Management 30:1 84-93.

Hautphenne, S., Fackrell, M., 2014. An EM algorithm for the model fitting of Markovian binary trees. Computational Statistics and Data Analysis 70 19-34.

Hodgkiss, I., 1998. Are changes in N:P ratios in coastal waters the key to increased red tide blooms? Conference on Science and Management of Coastal, pp. 141-147.

Hitzfeld, B.C., Hoger, S.J., Dietrich, D.R., 2000. Cyanobacterial toxins: removal during drinking water treatment, and human risk assessment. Environ. Health Perspect. 108 (S1) 113-122.

Ietswaart, T.h., Breebaart, L., van Zanten, B., Bijkerk, R., 1999. Plankton dynamics in the river Rhine during downstream transport as influenced by biotic interactions and hydrological conditions. Hydrobiol. 410 1-10.

Ledreux, A., 2010. Evidence for saxitoxins production by the cyanobacterium Aphanizomenon gracile in a French recreational water body. Harmful Algae 10 88-97.

Moore, S.K., Mantua, N.J., Hickey, B.M., Trainer, V.L., 2009. Recent trends in para-lytic shellfish toxins in Puget Sound relationships to climate, and capacity for prediction of toxic events. Harmful Algae 8 463-477.

Nelson, T.A., Haberlin, K., Nelson, A.V., Ribarich, H., Hotchkiss, R., Van Alstyne, K.L., Buckingham, L., Simunds, D.J., Fredrickson, K., 2008. Ecological and physiological controls of species composition in green macroalgal blooms. Ecology 89 1287-1298.

Nie, L., 2007. Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: Theory and applications mixed-effects models: Theory and applications. Journal of Statistical Planning and Inference 137 1787-1804.

Paerl, H.W., Huisman, J., 2008. Blooms like it hot. Science 320 57-58.

Park, J., Qian, G. Q., Jun. Y., 2008. Monte Carlo EM algorithm in logistic linear models involving non-ignorable missing data. Applied Mathematics and Computation 197 440-450.

Pick, F.R., Lean, D.R.S., 1987. The role of macronutrients (C, N, P) in controlling cyanobacterial dominance in temperate lakes. New Zeal J Mar Fresh. 21:425-434.

Rucker, J., Stuken, K., Nixdorf, A., Fastner, C., Chorus, I., Wiedner, C., 2007. Concentrations of particulate and dissolved cylindrospermopsin in 21 Aphanizomenon-dominantd temperate lakes. Toxicon. 50 800-809.

Routh, J., Choudhary, P., Meyers, P.A., 2009. A sediment record of recent nutrient loading and trophic state change in Lake Norrviken, Sweden. Journal of Paleolimnology 42 241-325.

[USACE] US Army Corps of Engineers. 2007. Long-term withdrawal of irrigation water, Willow Creek Lake, Morrow County, Oregon: draft environmental assessment. Portland (OR): USACE, Portland District.

Chapter 3 Conclusions and future works

Usually the T test for means is applied when comparing two variables. The independence of the observations is the basic requirement. So if the data are repeated measured over time, and the value is dependent on the previous values, the data formed into time series. In this case, T test or permutation based non parametric test are not satisfied. Also for time series analysis is designed for prediction and forecasting, also is not a straight forward method to compare two time series to find out which one is significantly higher. So it is meaningful to exploring how to find a properly method to compare time series.

For certain algae species density study, the more importance is to find the factors affect the dominant species rather than the factors affect the algae density. The water ecology system is complicate. The dominant species change frequently, and the factors are quite different from site to site. From limited references which study the dominant status of algae show that many factors might be significant, and for different lakes the factors are different. Focus on just one species is much harder than study the whole biomass. When the target specie is not dominant, even the nutrition is rich in water, the target specie density could be low. That is because most nutrition was consumed by the other species. In order to avoid this problem, the data should be selected first, to eliminate these low density observation with rich phosphors. The elimination starts from density 50 per liter, to 500 per liter. After eliminating the observations with lower than 200 per liter, the reactive phosphors will give positive affection on the algae density. So 200 per liter of the algae density can be chosen as the base line. The reason not to choose higher base line is we need to keep as many observation in the data, and 200 per liter is a safe level for public health.

From the linear model parameter estimation calibrations, EM algorithm is more powerful since it is able to estimate with missing values by iterations. In ecology sampling, data with large amount of missing value is common, because early ages techniques is not support some of the test, and the long run project is limited by funding of every year, and the target algae density cannot be always captured. Though EM algorithm is helpful to deal with

missing value, we still need to control the amount of missing values. For the variables contents large amount of missing values, such as over 40% or more of missing values, the variables are considered as lack of information. These information lacked variables should be removed from model even the statistical methods allowed to keep them. In order to get more valuable data with fewer missing values, the sampling schedule should be well organized. And if sampled on weekly base the time series analysis can be a useful method to apply.

The microbe community analysis could be another option to study the interaction between different algae. The data record the algae count on every sample, 15 types of algae are frequently viewed. *Aphanizomenon flos-aquae* is the most frequently viewed. Cluster analysis can find the relationship between *Aphanizomenon flos-aquae* and other blue-green algae or non-blue-green algae.