4-1-2011

Development of Methods for Spatial Modeling of Sediments Quality in Lake Okeechobee, Florida

Yao Yan
Florida International University, yyan001@fiu.edu

DOI: 10.25148/etd.FI11050316
Follow this and additional works at: https://digitalcommons.fiu.edu/etd

Recommended Citation
https://digitalcommons.fiu.edu/etd/385

This work is brought to you for free and open access by the University Graduate School at FIU Digital Commons. It has been accepted for inclusion in FIU Electronic Theses and Dissertations by an authorized administrator of FIU Digital Commons. For more information, please contact dcc@fiu.edu.
DEVELOPMENT OF METHODS FOR SPATIAL MODELING OF SEDIMENTS
QUALITY IN LAKE OKEECHOBEE, FLORIDA

A dissertation submitted in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY
in
CIVIL ENGINEERING
by
Yao Y. Yan
2011
To: Dean Amir Mirmiran  
College of Engineering and Computing

This dissertation, written by Yao Y. Yan, and entitled Development of Methods for Spatial Modeling of Sediments Quality in Lake Okeechobee, Florida, having been approved in respect to style and intellectual content, is referred to you for judgment.

We have read this dissertation and recommend that it be approved.

_______________________________________
Assefa M. Melesse

_______________________________________
Reinaldo Garcia

_______________________________________
Walter Z. Tang

Fernando R. Miralles-Wilhelm, Major Professor

Date of Defense: April 1, 2011

The dissertation of Yao Y. Yan is approved.

_______________________________________
Dean Amir Mirmiran  
College of Engineering and Computing

_______________________________________
Interim Dean Kevin O’Shea  
University Graduate School

Florida International University, 2011
DEDICATION

I dedicate this dissertation to my family, especially to my beloved wife Jianqin and loving sons Michael, William and Alex for their unconditional love and making this effort meaningful. I also dedicate this work to my parents, Shaowen Yan and Quexiang Chen for their encouragement, love and understanding.
ACKNOWLEDGMENTS

Foremost, I would like to express my deepest gratitude to my major professor, Dr. Fernando R. Miralles-Wilhelm for providing his wisdom and guidance during my study at Florida International University (FIU). His encouragement and support have made my research much smoother and successful.

I would also like to thank Dr. Walter Z. Tang, Dr. Reinaldo Garcia and Dr. Assefa M. Melesse for serving in my research committee. Their insights and suggestions have been greatly helpful in my research. Special thanks also go to Dr. Fang Zhao, who was the first professor introduced me with all the information on the engineering department, and later served in my committee until she passed away in 2010. I appreciate all her advices and encouragement.

Special thanks go to Dr. Richard Tom James, my colleague at South Florida Water Management District (SFWMD). Dr. James reviewed most of my draft chapters, and greatly improved this research and the writing of the dissertation. Thanks also go to Dr. Hongying Zhao, who was very helpful with her discussion on regression and model validation. Special thanks go to SFWMD for providing financial support through the Education Reimbursement Program during my study.

Finally, and always, my special thanks to my family. It would not be possible for me to complete this dissertation without the loving support of my family. Special thanks also go to my parents. Their intelligence, hardworking and love always encourage me to chase my dreams.
The major objectives of this dissertation were to develop optimal spatial techniques to model the spatial-temporal changes of the lake sediments and their nutrients from 1988 to 2006, and evaluate the impacts of the hurricanes occurred during 1998-2006.

Mud zone reduced about 10.5% from 1988 to 1998, and increased about 6.2% from 1998 to 2006. Mud areas, volumes and weight were calculated using validated Kriging models. From 1988 to 1998, mud thicknesses increased up to 26 cm in the central lake area. The mud area and volume decreased about 13.78% and 10.26%, respectively. From 1998 to 2006, mud depths declined by up to 41 cm in the central lake area, mud volume reduced about 27%. Mud weight increased up to 29.32% from 1988 to 1998, but reduced over 20% from 1998 to 2006. The reduction of mud sediments is likely due to re-suspension and redistribution by waves and currents produced by large storm events, particularly Hurricanes Frances and Jeanne in 2004 and Wilma in 2005.

Regression, kriging, geographically weighted regression (GWR) and regression-kriging models have been calibrated and validated for the spatial analysis of the sediments TP and TN of the lake. GWR models provide the most accurate predictions for TP and TN
based on model performance and error analysis. TP values declined from an average of 651 to 593 mg/kg from 1998 to 2006, especially in the lake’s western and southern regions. From 1988 to 1998, TP declined in the northern and southern areas, and increased in the central-western part of the lake. The TP weights increased about 37.99%-43.68% from 1988 to 1998 and decreased about 29.72%-34.42% from 1998 to 2006.

From 1988 to 1998, TN decreased in most areas, especially in the northern and southern lake regions; western littoral zone had the biggest increase, up to 40,000 mg/kg. From 1998 to 2006, TN declined from an average of 9,363 to 8,926 mg/kg, especially in the central and southern regions. The biggest increases occurred in the northern lake and southern edge areas. TN weights increased about 15%-16.2% from 1988 to 1998, and decreased about 7%-11% from 1998 to 2006.
TABLE OF CONTENTS

CHAPTER ........................................................................................................................................ PAGE

CHAPTER 1 INTRODUCTION ........................................................................................................ xi
  1.1 Introduction .......................................................................................................................... 1
  1.2 Problem Statement .......................................................................................................... 3
  1.3 Relevance of this Investigation ...................................................................................... 8
  1.4 Research Objectives ........................................................................................................ 9
  1.5 Outline of the Dissertation ............................................................................................ 10

CHAPTER 2 PRINCIPLES AND METHODOLOGIES .............................................................. 12
  2.1 Universal Model of Spatial Variation ........................................................................... 12
  2.2 Kriging Analysis ............................................................................................................ 13
  2.3 Multi-linear Regression Analysis .................................................................................. 16
  2.4 Spatial Regression ......................................................................................................... 19
  2.5 Regression-Kriging ........................................................................................................ 26
  2.6 Spatial Model Selection, Assumptions and Uncertainty Analysis ............................ 30

CHAPTER 3 DATA COLLECTION AND DESCRIPTION ...................................................... 35
  3.1 Spatial and Temporal Changes of Sediment Zones .................................................... 35
  3.2 Integrated Digital Elevation Model ............................................................................. 37

CHAPTER 4 SPATIAL-TEMPORAL CHANGES OF MUD THICKNESS .............................. 40
  4.1 Data Description and Characteristics ........................................................................ 40
  4.2 Ordinary Kriging Model Calibration ........................................................................... 40
  4.3 Ordinary Kriging Model Validation ............................................................................ 44
  4.4 Spatial Changes of Mud Thickness Over-time ........................................................... 49
  4.5 Mud Surface Area and Volume Calculations ............................................................ 51
  4.6 Mud Weight Calculation .............................................................................................. 55
  4.7 Summary of Results ..................................................................................................... 61

CHAPTER 5 SPATIAL MODELING OF TOTAL PHOSPHORUS ........................................... 62
  5.1 Data Description and Characteristics ........................................................................ 62
  5.2 Regression Model Analysis ........................................................................................... 73
  5.3 GWR Models Analysis .................................................................................................. 88
  5.4 Kriging and Co-kriging Model Analysis ........................................................................ 105
  5.5 Regression-Kriging (R-K) Model Analysis ................................................................... 115
  5.6 Optimal Models and Weight Calculations .................................................................... 122
  5.7 Summary of Results ..................................................................................................... 136

CHAPTER 6 SPATIAL MODELING OF TOTAL NITROGEN .............................................. 138
  6.1 Data Description and Characteristics .......................................................................... 138
  6.2 Regression Model Analysis ........................................................................................... 145
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 3.1 Sediment zones and their area changes over-time</td>
<td>35</td>
</tr>
<tr>
<td>Table 4.1 Summaries of mud thickness and bulk density</td>
<td>40</td>
</tr>
<tr>
<td>Table 4.2 Calibration and validation results of mud thickness data</td>
<td>44</td>
</tr>
<tr>
<td>Table 4.3 Mud surface area and volume</td>
<td>51</td>
</tr>
<tr>
<td>Table 4.4 Changes of mud surface area and volume</td>
<td>51</td>
</tr>
<tr>
<td>Table 4.5 Mud weights and their changes over-time</td>
<td>61</td>
</tr>
<tr>
<td>Table 5.1 Data summaries of key elements of sediment data</td>
<td>67</td>
</tr>
<tr>
<td>Table 5.2 OLS model results and diagnosis of 2006 TP calibration data</td>
<td>74</td>
</tr>
<tr>
<td>Table 5.3 GWR model diagnosis with multiple key variables</td>
<td>75</td>
</tr>
<tr>
<td>Table 5.4 Statistically significant calibrated regression models</td>
<td>77</td>
</tr>
<tr>
<td>Table 5.5 Diagnosis of the GWR calibration model (TP vs. Th &amp; Elev)</td>
<td>89</td>
</tr>
<tr>
<td>Table 5.6 Diagnosis of the GWR calibration model (TP vs. Fe)</td>
<td>93</td>
</tr>
<tr>
<td>Table 5.7 Calibration and validation results of the OK and CK models</td>
<td>115</td>
</tr>
<tr>
<td>Table 5.8 TP model uncertainty and comparison</td>
<td>122</td>
</tr>
<tr>
<td>Table 5.9 TP weights (kg) and their changes over-time</td>
<td>136</td>
</tr>
<tr>
<td>Table 6.1 Statistically significant regression models for TN data</td>
<td>146</td>
</tr>
<tr>
<td>Table 6.2 Calibration result of the OLS model of 2006 TN</td>
<td>146</td>
</tr>
<tr>
<td>Table 6.3 Diagnosis of the GWR calibration model of 2006 TN</td>
<td>146</td>
</tr>
<tr>
<td>Table 6.4 Diagnosis of the GWR model calibration (TN vs. TC)</td>
<td>152</td>
</tr>
<tr>
<td>Table 6.5 TN model uncertainty and comparison</td>
<td>166</td>
</tr>
<tr>
<td>Table 6.6 TN Weights (kg) and their changes for OK and GWR models</td>
<td>176</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1 Major tributaries and sediment sample sites, Lake Okeechobee</td>
<td>2</td>
</tr>
<tr>
<td>Figure 1.2 Lake TSS mass (mt) by Water Year</td>
<td>5</td>
</tr>
<tr>
<td>Figure 1.3 Inflow (m$^3$) by Water Year</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1.4 Net TSS load (mt) by Water Year</td>
<td>6</td>
</tr>
<tr>
<td>Figure 1.5 Net TP load (mt) by Water Year</td>
<td>7</td>
</tr>
<tr>
<td>Figure 1.6 Net TN load (mt) by Water Year</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2.1 Variogram and its parameters</td>
<td>14</td>
</tr>
<tr>
<td>Figure 2.2 GWR weighting schemes and bandwidths</td>
<td>23</td>
</tr>
<tr>
<td>Figure 2.3 A schematic example of regression-kriging</td>
<td>28</td>
</tr>
<tr>
<td>Figure 2.4 Model selection processes for spatial prediction</td>
<td>32</td>
</tr>
<tr>
<td>Figure 3.1 Sediment zones distribution from 1988 to 2006, Lake Okeechobee</td>
<td>36</td>
</tr>
<tr>
<td>Figure 3.2 LiDAR DEM and bathymetry transects (a) and bathymetry DEM (b)</td>
<td>38</td>
</tr>
<tr>
<td>Figure 3.3 Digital elevation model and eco-zones (a) and 3D view (b)</td>
<td>39</td>
</tr>
<tr>
<td>Figure 4.1 Trend analyses of mud thickness data</td>
<td>41</td>
</tr>
<tr>
<td>Figure 4.2 Calibration and validation sample distributions of 2006 mud thickness</td>
<td>42</td>
</tr>
<tr>
<td>Figure 4.3 Fitted variogram models using mud thickness calibration data</td>
<td>43</td>
</tr>
<tr>
<td>Figure 4.4 Validation error distributions of the mud thickness data</td>
<td>47</td>
</tr>
<tr>
<td>Figure 4.5 Scatter plots of the mud thickness validation data</td>
<td>48</td>
</tr>
<tr>
<td>Figure 4.6 Fitted variogram models using the complete mud thickness data</td>
<td>49</td>
</tr>
<tr>
<td>Figure 4.7 Mud thickness maps for 1988, 1998 and 2006 data</td>
<td>50</td>
</tr>
</tbody>
</table>
Figure 4.8 ModelBuilder model for mud thickness data processing and volume calculation

Figure 4.9 Spatial changes of mud thickness over-time

Figure 4.10 Trend analyses of mud bulk density

Figure 4.11 Fitted variogram models for mud density

Figure 4.12 Spatial variations of mud density

Figure 4.13 ModelBuilder model for mud weight calculation

Figure 5.1 TP distributions (a, c, and e) and mean changes (b, d, and f) in sediment zones

Figure 5.2 TP trend analyses of the sediment data

Figure 5.3 Scatter plot matrix of sediment data

Figure 5.4 Total phosphorus maps for 1988 (a), 1998 (b) and 2006 (c) data

Figure 5.5 Spatial changes of TP during 1988-1998 (a) and 1998-2006 (b)

Figure 5.7 Validation error distributions of the OLS model (TP vs. Th & Elev)

Figure 5.8 Scatter plots of the OLS model (TP vs. Th & Elev) validation

Figure 5.9 Validation error distributions of the OLS model (TP vs. Fe)

Figure 5.10 Scatter plots of the OLS model (TP vs. Fe) validation

Figure 5.11 Coefficient maps of the GWR model (TP vs. Th & Elev) calibration

Figure 5.12 Coefficient maps of the GWR model (TP vs. Fe) calibration

Figure 5.13 Validation error distributions of the GWR model (TP vs. Th & Elev)

Figure 5.14 Scatter plots of the GWR model (TP vs. Th & Elev) validation

Figure 5.15 Validation error distributions of the GWR model (TP vs. Fe)

Figure 5.16 Scatter plots of the GWR model (TP vs. Fe) validation

xii
Figure 6.12 TC trend analyses (a, c and e) and estimation (b, d and f) using OK........... 169

Figure 6.13 TN concentrations (a, c and e) and weights (b, d and f) using GWR model (TN vs. TC).................................................................................................................................................................................. 172

Figure 6.14 TN concentrations (a, c and e) and weights (b, d and f) using OK............. 175
CHAPTER 1 INTRODUCTION

1.1 Introduction

Lake Okeechobee is a large shallow lake with an area of approximately 1730 km² and an average depth of 2.7 m. It is the second-largest freshwater lake wholly within the continental United States (after Lake Michigan) and the largest in the southern United States. The name Okeechobee comes from the Hitchiti words oki (water) and chubi (big). Lake Okeechobee formed out of the ocean about 6,000 years ago when the waters receded (Brooks, 1974). The floor of the lake is a limestone basin, and its water is turbid from mud sediments that cover a large portion of the lake. The lake is enclosed by a 20-foot (6 m) high dike built by the U.S. Army Corps of Engineers (USACE) after a hurricane in 1928 breached the old dike, flooding surrounding communities and claiming thousands of lives (Will, 1990). There are 32 inflows, including Taylor Creek, Fisheating Creek and the Kissimmee River, and a number of outflows including the St. Lucie River and West Palm Beach Canal to the east, the Caloosahatchee River to the west and the Miami, North New River and Hillsboro Canals to the south (Figure 1.1).

Since 1870, twenty five hurricanes of Category I or greater have passed within 50 miles of the center of Lake Okeechobee (James and Pollman, 2011). In 1926, the Great Miami Hurricane hit the Lake Okeechobee area, killing approximately 300 people. Two years later in 1928, the Okeechobee Hurricane crossed over the lake, killing over 2,500 people.
As a response to these disasters, the Florida State Legislature created the "Okeechobee Flood Control District", and it was authorized to cooperate with the U.S. Army Corps of Engineers in actions to prevent similar disasters. U.S. President Herbert Hoover visited the area personally, and afterward the Corps designed a new plan incorporating the construction of channels, gates and levees. The dike was named the "Herbert Hoover Dike" in honor of the president.

Figure 1.1 Major tributaries and sediment sample sites, Lake Okeechobee
Three recent hurricanes, Hurricanes Frances and Jeanne in 2004, and Hurricane Wilma in 2005 passed over Lake Okeechobee. Water levels increased to more than 18 feet (National Geodetic Vertical Datum, NGVD) after Hurricane Jeanne and over 16 feet after Hurricane Wilma. A drought began in 2006 and lake levels declined from 13 feet in 2006 to an all time low of 8.82 feet in July 2007. The low water level exposed a number of littoral and nearshore areas that have been covered with muck sediments. In an effort to restore the lake's natural sandy base and improve water clarity and enhance wildlife habitat, the South Florida Water Management District (SFWMD), in cooperation with the state of Florida and the Florida Fish and Wildlife Conservation Commission (FWC) and Glades County, completed a multi-million dollar effort to remove muck materials and restore six locations in the lake. Over two million cubic yards of muck (1.6 million cubic meters) have been removed from an estimated 2,000 ac (809 ha) of exposed shoreline, potentially restoring submerged aquatic vegetation (SAV) habitat. This project also removed an estimated 237 mt of phosphorus at an approximated cost of $11 million or $46.48 per kg (James & Zhang, 2008).

1.2 Problem Statement

The lake faces three major environmental challenges: (1) excessive phosphorus loads; (2) unnaturally high and low water levels; and (3) rapid spread of exotic and nuisance plants in the littoral zone (James et al., 2008). Multiple state and federal agencies have been working cooperatively to address these interconnected issues in order to rehabilitate the lake and enhance the ecosystem services that it provides (South Florida Water Management District et al., 2007).
Excessive phosphorus loads to the lake originate from agricultural and urban activities in the watershed (Flaig & Havens, 1995). Total phosphorus (TP) loads were 572 metric tons per year (mt/yr) averaged over WY2004–WY2009, which is more than four times higher than the Total Maximum Daily Load (TMDL) of 140 mt/yr (five-year average) considered necessary to achieve the in-lake TP target of 40 parts per billion (ppb) (McCormick et al., 2010). Despite a long history of regulatory and voluntary incentive-based programs to control phosphorus inputs into Lake Okeechobee, no substantial reduction in loading occurred during the 1990s. Consequently, the lake continues to become more eutrophic with blooms of noxious blue-green algae (cyanobacteria), loss of benthic invertebrate diversity, and spread of cattail (Typha sp.) in shoreline areas (South Florida Water Management District et al., 2007). As a result, the Florida legislature passed the Lake Okeechobee Protection Act (LOPA) [Section 373.4595, Florida Statutes, (F.S.)] in 2000, mandating that the TMDL be met by 2015 and that multiple state agencies work together to implement an aggressive program to address the issues of excessive TP loading and exotic species expansion.

In January 2004, the Lake Okeechobee Protection Plan (LOPP) was approved by the Florida legislature. The LOPP contains a phased, watershed-based, comprehensive approach to reduce TP loading to the lake. The cooperating agencies have implemented a large number of phosphorus reduction projects since 2000. In addition, a comprehensive monitoring program for water quality in the lake and watershed and ecological indicators in the lake has been implemented. The District and USACE are also in the process of implementing the Lake Okeechobee Watershed Project (LOWP), which is intended to
store approximately 273,000 ac-ft (336,740,541.65 m³) of water outside of the lake and it
will also reduce TP loads by approximately 60 mt/yr.

A $^{210}$Pb study on sediment cores from the center of Lake Okeechobee after the 2004 and
2005 hurricanes showed that the sediment layers were sequentially disturbed (Chang et
al. 2008). Hurricane Frances and Jeanne mixed sediment layers as deep as 10-12
centimeters. Hurricane Wilma mixed the sediment layers from 10-12 centimeters up to 25
centimeters. Almost one-third of the sediment bed was re-suspended during these
hurricanes, resulting in increased total suspended solids (TSS) (Figure 1.2). The TSS
spiked in 2005, which is eight time high of the values during 1985-1990. Inflows of the
Lake also increased during the 2004-2005 hurricanes (Figure 1.3).

![Figure 1.2 Lake TSS mass (mt) by Water Year](image-url)
Figure 1.3 Inflow (m$^3$) by Water Year

Figure 1.4 Net TSS load (mt) by Water Year
Mass balance of sediments in Lake Okeechobee has been monitored by SFWMD since 1970’s (SFWMD, 2011). The net TSS load lost a significant amount after 2004-2005 hurricanes (Figure 1.4) (Water-Year data of 2005 and 2006 include hurricanes of 2004 and 2005 respectively). Net TP and TN loads data indicate positive gains (Net TP loads > 0) but relatively low compared to the net loads before and after the 2004-2005 hurricanes (Figures 1.5-1.6).

Figure 1.5 Net TP load (mt) by Water Year
1.3 Relevance of this Investigation

Over 44% (as of 2006) of Lake Okeechobee is underlain with P enriched sediments (Yan and James, 2007). The lake bottom sediment has been the research subject since late 1960’s. Fifteen sediment samples were analyzed during 1969-1972 by the U.S. Geological Survey cooperated with the Central and Southern Florida Flood Control District (the early name of the SFWMD) (Joyner, 1974). In 1988, a comprehensive survey reported that the upper 10 cm of mud sediments within the lake contained an estimated 28,600 metric tons of P (Reddy et al., 1995). An additional survey in 1998 made similar estimates of sediment P within Lake Okeechobee (Fisher et al., 2001). Because of this large sediment pool of P, sediment-water interactions, including diffusive
flux and sediment re-suspension, can affect phosphorus dynamics within the Lake’s water column (Fisher et al., 2005; James et al., 2005; James et al., 2008). These sediment-water interactions can delay changes of in-lake phosphorus concentration to reduced external loads (James & Pollman, 2011). The historic and current distribution of these sediments, their compositions and changes over time, sediment re-suspension and interaction with water quality, are of direct concern for lake management and restoration efforts to improve the lake’s environmental quality.

Sediment re-suspension, driven by wind-induced water movement (especially during 2004-2005 hurricane seasons), may play a significant role in nutrient cycling in this large shallow lake, as particulate bound forms of P can be released into the overlying water column (Søndergaard et al., 2003). Hurricanes have large effects on Lake Okeechobee’s ecosystem through sediment re-suspension that resulted in dramatic changes in water quality (Havens et al., 2001; James et al., 2008), and the re-suspension of sediments, which redistributed them throughout the lake, increased the active layer thickness, and reduced the cohesiveness of the surface layers of sediments (James et al., 2008; Jin et al., 2011). All of these resulted in poorer water quality and increased turbidity throughout the lake.

**1.4 Research Objectives**

The aim of this dissertation is to develop optimal spatial models to estimate the mud weights and nutrient weights (TP and TN) in Lake Okeechobee over-time (1988-2006). In particular, their spatial changes were examined to identify the potential impacts of
extreme environmental forcing during 1998 and 2006 hurricane seasons. To improve our understanding of the sediment-water interactions within Lake Okeechobee, several unique spatial models were calibrated and validated to describe the sediment characteristics, then the validated models were used to calculate the mud, nutrients and their changes from 1988 to 2006. These models add new approaches to analyze other lakes or water bodies that are faced with similar challenges such as Lake George, Uganda and Lake Tai-Hou, China (Havens et al., 2007). Extreme weather conditions, such as hurricanes, which have affected the lake from 2004 to 2006 are the focus of this study. Three sediment surveys using similar techniques and from the same sampling locations were completed in 1988, 1998 (Reddy et al., 1995; Fisher et al., 2001) and 2006 (BEM & University of Florida, 2007) before and after major hurricane events. The objectives of this dissertation are to use the information from these surveys to answer the following questions:

1) Has the spatial distribution of sediment types changed over time, and what was the response to an extreme event?

2) Have mud thicknesses in the lake changed over time, and what was the response to an extreme event?

3) What is the amount of nutrients (TP, and TN) in the mud sediments and has it changed temporally and spatially?

1.5 Outline of the Dissertation

Chapter 1 gives a general introduction of the problems, research history and the significance of this study. Chapter 2 describes several spatial prediction models and their
assumptions. Traditional multi-linear regression (MLR), and classic kriging techniques (Ordinary Kriging and Co-kriging) are discussed first. Then the newly developed geographically weighted regression (GWR) and Regression-Kriging (R-K) are discussed in detail. Approaches of model selection, assumption, and uncertainty analysis are also discussed in the chapter. Chapter 3 introduces the sediment and elevation data collection and processing. All these data sets are used for nutrient weights calculations. Chapter 4 discusses the details of mud thickness modeling, their spatial and temporal changes. Mud weights and their changes are calculated using Ordinary Kriging and ModelBuilder models. Chapter 5 and 6 introduce the model calibration, validation and error analysis, and model comparison for TP and TN, respectively. Kriging, Ordinary Least Squares Regression, Geographically Weighted Regression, and Regression-Kriging were examined for TP analysis in Chapter 5. Kriging, OLS, and GWR models were examined for TN analysis in Chapter 6. Optimal models were selected to calculate the TP and TN weights and their changes.

Finally, overall conclusions and a general discussion for further studies are presented in Chapter 7.
CHAPTER 2 PRINCIPLES AND METHODOLOGIES

2.1 Universal Model of Spatial Variation

Geostatistics specializes in the analysis and interpretation of geographically referenced data (Goovarts, 1997; Webster and Oliver, 2001). It has been widely used in geosciences, soil sciences, water resources, and environmental sciences (Zhou et al., 2007). Geostatistical analysis uses computer programs to produce maps and predicted values at locations of interest from a combination of field and ancillary information.

Spatial variability of environmental observations is commonly a result of complex processes. Regionalized variable theory assumes that the spatial variation of any observation can be expressed as the sum of three major components (Matheron, 1969; Hengl, 2009): a structure component with a constant mean or trend; a random but spatially correlated component; and a spatially uncorrelated random error:

\[ Z(s) = Z^*(s) + \epsilon'(s) + \epsilon'' \]  

where \( Z^*(s) \) is the deterministic function describing the “structural” component of \( Z \) at \( s \), \( \epsilon'(s) \) is the stochastic, locally varying but spatially dependent residuals from \( Z^*(s) \); and \( \epsilon'' \) is a residual, spatially independent Gaussian noise with zero mean and variance, usually the result of the measurement error. This equation is often referred as the universal model of spatial variation. Many approaches have been developed to model this spatial variation and several selected methodologies will be discussed in the following sections.
2.2 Kriging Analysis

Kriging has been used as a synonym for geostatistical interpolation for many years (Burrough and McDonnell, 1998). The original idea came from the mining engineer D. G. Kridge and the statistician H. S. Sichel. Kridge published this technique in 1951, but the mathematician G. Matheron derived the formulas and established the linear geostatistics (Cressie, 1990; Webster and Oliver, 2001). The major contribution from Matheron (1962) and Gandin (1963) was the development of the semi-variance:

\[ \lambda(h) = \frac{1}{2} \sum_{i=0}^{n} [(z(s_i) - z(s_i + h))^2] \]  

(2)

where \( z(s_i) \) is the value of the target variable at sampled location and \( z(s_i + h) \) is the value of the variable at distance \( s_i + h \).

If there are \( n \) point observations, there will be \( n*(n-1)/2 \) pairs for which a semi-variance can be calculated. All semi-variance pairs can be plotted against a standard distance or lag to create a standard experimental variogram (Figure 2.1). The sill is where the fitted curve levels off at large lag \( h \). It implies that at these values of the lag there is no spatial dependence between the data points since all estimates of variance of differences will be invariant. Range (\( h \)) is the distance value at the sill. This is a critical measurement of the variogram because it describes how inter-site differences are spatially dependent. Within the range, closer sites are more similar to each other. The range also defines the size the search window should be for weighted moving average interpolation. The nugget (\( Co \)) is the positive \( \lambda(h) \) value when \( h \to 0 \). It is the estimate of \( \epsilon^" \), the residual, spatially unrelated noise. This is also the variance of measurements. The variogram provides a
quantitative description of the regionalized variation. Next the variogram is fitted to a standard variogram model such as linear, spherical, exponential, circular, Gaussian, Bessel, power etc. (Isaaks and Srivastava, 1989; Goovaerts, 1997). The variogram provides useful information for interpolation, optimizing sampling and determining spatial patterns. It also offers a measure of associated uncertainty, i.e. the estimated variance of the prediction error for a given model.

Figure 2.1 Variogram and its parameters

In addition to modeling of nugget and sill variation, variogram models can be extended to larger number of parameters to account for anisotropy or smoothness. The semivariance surface may be plotted as a 2D map or 3D surface representation. Given a strong directional bias the major axis provides the range in the primary direction, and the minor axis provides the range in the orthogonal direction. A single anisotropic model may be fitted to such datasets, or a series of separate models fitted to the data grouped into distinct directional bins. The 2D geometric anisotropy model in GSTAT (Pebesma and Wesseling, 1998), for example, replaces the range parameter with three parameters: range
in the major direction (direction of the strongest correlation), angle of the principal direction and the anisotropy ratio. Another three-parameter model is the Matérn variogram model, which adds additional parameter to describe smoothness (Stein, 1999; Minasny and McBratney, 2005). This model estimate both short and long distance variation. In reality, variogram models with more parameters are more difficult to fit automatically (Minasny and McBratney, 2005). The Stanford Geostatistical Modeling Software (SGEMS) provides a 3-D variogram fitting visually (Remy et al., 2009).

Kriging methods provide great flexibility for interpolations and different sub-methods were developed such as simple kriging, indicator kriging, universal kriging, etc. All these methods produce smoothly varying surfaces accompanied by an estimated variance surface. Simple Kriging assumes that the data have a known mean value throughout the study area and exhibits both first and second order stationarity. These assumptions are overly restrictive for most problems and hence this method is rarely used. De-trending and z-score (normal) transformation may help to remove some of these problems. Ordinary Kriging (OK) and its variants have more relaxed assumptions than Simple Kriging. Ordinary Kriging assumes second-order stationarity with an unknown mean.

In Ordinary Kriging (OK) the expected value of the random function is locally re-estimated from local data, while the covariance model is kept stationary:

\[ Z(s) = \mu + \varepsilon'(s) \]  

(3)

where \( \mu \) is the constant stationary function (global mean) and \( \varepsilon'(s) \) is the spatially correlated stochastic part of variation. The prediction at location \( S_0 \) is a weighted average:
where $w_i$ is the kriging weight at location $s_i$, $Z(s_i)$ is the observations at locations $s_i$. In a way, kriging can be seen as a sophistication of the inverse distance interpolation, and the weights are based on the spatial autocorrelation structure. Co-kriging allows samples of an ancillary variable (also called the co-variable), besides the target value of interest, to be used when predicting the target value at un-sampled locations. With Co-kriging the estimated value at an un-sampled location is a linear weighted sum of all of the variables being examined (i.e. two or more). The co-variables may be measured at the same points as the target (co-located samples), at other points, or both. Co-kriging is the extension of the kriging paradigm to estimate one attribute using a data set that contains observations related to other attributes (Goovaerts, 1997). For more detailed discussions, refer to the “classic” geostatistics textbooks by Isaaks and Srivastava (1989).

2.3 Multi-linear Regression Analysis

Regression analyses reveal the relationship between a dependent variable and a set of one or more independent variables. Regression analyses can be used to model, examine, and explore spatial relationships, and can help explain the factors behind observed spatial patterns. Ordinary Least Squares (OLS) is the best known of all regression techniques, and creates a single regression equation to estimate a dependent variable. The regression equation uses a mathematical formula and independent variables to predict a dependent variable:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \text{for } i=1 \ldots n$$  

(5)
where, $y_i$ is the dependent variable measured for observation $i$. The known $y$ values are often referred to as observed values. $x_i$ is the independent/explanatory variable for $i$. $\varepsilon_i$ is the error term at $i$. $\beta_0$ and $\beta_1$ are parameters and are estimated such that the value of $\sum_{k=0}^{n} (y_i - \hat{y})^2$ is minimized over the $n$ observations in the data set. The $\hat{y}$ is the predicted or fitted value for the $i$th observation. The term $(y_i - \hat{y})$ is called the residuals for the $i$th observation, and the residuals shall be both independent and form a normal distribution with a mean of zero.

A more general multiple linear regression model can be written as:

$$y_i = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik}(u_i v_i) + \varepsilon_i$$

where, $y_i =$ the predicted value at $i$th observation, $\beta_0 =$ model intercept, $\beta_k =$ coefficient of the $k$th explanatory variable, $n =$ the number of available ancillary variables, $x_k =$ $k^{th}$ explanatory variable, $k = 1, ..., n$, $\varepsilon_i =$ error associated with the $i$th observation; and $(u_i, v_i) =$ the coordinates of the $i$th observation in space. The regression coefficients can be estimated via regular ordinary least squared or, optimally, the generalized least squared estimators (Hengl et al., 2007).

Regression analysis is used to understand the variation in a dependent variable based on the variation of independent variable(s). While a strong relationship between independent and dependent variables suggests a relationship, it does not prove that one causes the other.
2.3.1 Regression model calibration and diagnosis

ArcGIS 9.3.1 implemented the regression function using OLS method, and a number of statistical indicators can be used to measure the model performance:

- The $R^2$ and adjusted $R^2$ values are used to assess model performance. The adjusted $R^2$ reflects model complexity (the number of variables) and its value is always a bit lower than the multiple $R^2$ value, and is a more accurate measure of model performance.

- The coefficient for each explanatory variable reflects both the strength and type of relationship the explanatory variable has to the dependent variable.

- The $T$ test is used to assess whether or not an explanatory variable is statistically significant.

- The variance inflation factor (VIF) measures redundancy among explanatory variables. Explanatory variables associated with VIF values larger than about 7.5 should be removed as recommended by ESRI (2009).

- The Koenker (BP) Statistic (Koenker's studentized Bruesch-Pagan statistic) is a test to determine if the explanatory variables in the model have a consistent relationship to the dependent variable both in geographic space and in data space. When the model is consistent in geographic space, the spatial processes represented by the explanatory variables behave the same everywhere in the study area (the processes are stationary). When the model is consistent in data space, the variation in the relationship between predicted values and each explanatory variable does not change with changes in explanatory variable magnitudes (there is no heteroscedasticity in the model).
• The Joint F-Statistic and Joint Wald Statistic are measures of overall model statistical significance. The Joint F-Statistic is trustworthy only when the Koenker (BP) statistic is not statistically significant. If the Koenker (BP) statistic is significant, the Joint Wald Statistic shall be considered to determine overall model significance.

• The Jarque-Bera statistic indicates whether or not the residuals are normally distributed and is a measure of model bias. Spatial Autocorrelation (Moran’s I) tool implemented in ArcGIS shall be examined on the regression residuals to ensure they are spatially random. Statistically significant clustering of high and/or low residuals (model under and over predictions) indicates a key variable is missing from the model (misspecification).

• The AIC (Akaike Information Criterion) is a relative estimator of the expected Kullback discrepancy between the true model and a fitted model, and is used to compare different models. Taking into account model complexity, the model with the lower AIC value provides a better fit to the observed data. If the AIC values for two models differ by more than 3, the model with the lower AIC is held to be better.

2.4 Spatial Regression

Spatial data exhibit two properties: spatial dependence and spatial heterogeneity and that make it difficult to meet the assumptions and requirements of traditional (non-spatial) statistical methods. Fotheringham (1997) summarized three reasons why a global model may not be appropriate for spatial data modeling: 1) random sampling variations result in spatial variations in observed relationships; 2) certain relationships intrinsically vary across space; and 3) a global model used to measure relationships is a gross
misspecification of reality. For example, a geographical property like temperature, pollution, and elevation varies from place to place and a local spatial model can be better than a global model to estimate the values of this property at unsampled locations based on the values at locally sampled locations and known values of ancillary variables.

There are different strategies to deal with spatial autocorrelation in regression model residuals, such as resampling until the input variables no longer exhibit statistically significant spatial autocorrelation, or isolating the spatial and non-spatial components of each input variable using a spatial filtering regression method. For the regional variation (non-stationarity) in OLS regression, one approach is to redefine/reduce the size of the study area so that the processes within it are all stationary. A better approach is to use methods that incorporate regional variation into the regression model such as Geographically Weighted Regression (GWR).

2.4.1 Geographically Weighted Regression (GWR)

GWR is a spatial regression method developed by Foterhingham and Brundson (2000, 2002), building on works of Hastie and Tibshirani (1990) and Loader (1999). It can calibrate a multiple regression model that allows different relationships between the dependent and independent variables to exist at different locations in space (non-stationarity). GWR is a local spatial regression approach based on the “First Law of Geography”: everything is related to everything else, but closer things are more related to each other (Tobler, 1970). As a local analysis technique, GWR tends to serve as a “microscope” that amplifies details of the data that are otherwise hidden. GWR provides
a local model of the predicted variable or process by fitting a regression equation to every feature in the data set. When used properly, these methods are powerful and reliable statistics for examining/estimating linear relationships. With these characteristics, GWR is recommended as an additional spatial interpolation technique that may increase interpolation accuracy.

GWR assumes that relationships between the target variable and ancillary variables are not constant over space, and the error terms are independent and identically distributed (i.i.d.) across space. \( \beta \)'s therefore vary from location to location, the interpolation can then be expressed as:

\[
y_i = \beta_0 (u_i, v_i) + \sum_{k=0}^{n} \beta_k (u_i, v_i) x_{ik} (u_i, v_i) + \epsilon_i
\]  

(7)

where \((u_i, v_i)\) is the coordinates of the ith observation and \( \beta_k (u_i, v_i) \) is a realization of the function \( \beta_k (u, v) \) at point i.

It is assumed that each prediction is generated through a Gaussian or Gaussian-like spatial process in which the strength of the correlation among observations declines as distance increases. With such assumption, it is possible to mimic the spatial process at any particular location in space, and generate a set of local predictions that are based on weighting of available observations according to the distance and the prediction location.

The coefficient raster surfaces created by GWR can be examined to evaluate regional variation in the model explanatory variables. The spatially consistent (stationary) relationships between the dependent variable and each explanatory variable and how they
change across the study area can be evaluated as well as the coefficient distribution which shows where and how much variation is present.

GWR can be used for prediction when it is applied to sampled data. First, a feature class needs to be specified with all of the explanatory variables for locations where the dependent variable is unknown. GWR then calibrates the regression equation using known dependent variable values from the input feature class, and creates a new output feature class with dependent variable estimates.

### 2.4.2 GWR bandwidth selection and weighting functions

The parameters estimated for a GWR model are dependent on the spatial weighting function and the bandwidth selected. The weighting function determines the importance of a local observation based on its proximity to the sample point (Figure 2.2). The bandwidth determines which nearby observations are considered when calibrating coefficients for a sample point. Bandwidths may be constant (fixed kernel) or variable (adaptive kernel) that vary by the location of a data point. An optimal bandwidth (or nearest neighbors) can be selected by satisfying either the least cross-validation (CV) score, or least Akaike Information Criterion (AIC).

In GWR, a corrected AIC (AICc) was used (Hurvich et al, 1998; Charlton, 2009). It was first suggested for normal linear regression by Sugiura (1978). Hurvich and Tsai (1989) demonstrated the small-sample superiority of AICc over AIC, and justified the use of AICc in the frameworks of nonlinear regression models and autoregressive models. The AICc can be calculated using the following equation:
\[
AIC_c = 2n \log_e(\hat{\sigma}) + n \log_e(2\pi) + n\left\{\frac{n+tr(S)}{n-2-tr(S)}\right\} 
\]

where \( n \) is the number of observations in the dataset, \( \hat{\sigma} \) is the estimate of the standard deviation of the residuals, \( S \) is hat matrix, and \( tr(S) \) is the trace of the hat matrix. Cross-validation minimization can be used to search for an optimal bandwidth:

\[
CV = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2
\]

The weighted function for parameter estimation was developed by Fotheringham, Brunsdon and Charlton (2002):
\[
\hat{\beta} (u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y
\]

where the bold symbol represents a matrix, \( \hat{\beta} \) is an estimate of \( \beta \), and \( W(u_i, v_i) \) is a \( n \) by \( n \) matrix whose off-diagonal elements are zero and whose diagonal elements denote the geographic weighting of the each of the \( n \) observed data for regression point \( i \).

The weights are chosen such that those observations near the point in space where the parameter estimates are derived have more influence on the result than observations further away. Two schemes have been used for the weight calculation: bi-square and Gaussian (Figure 2.2). In the case of the Gaussian scheme, the weight for the \( i^{th} \) observation \( w_{ij} \) is:

\[
w_{ij} = \exp \left[ -\frac{d_{ij}^2}{h^2} \right]
\]

In the bi-square function

\[
w_{ij} = \left[ 1 - \left( \frac{d_{ij}}{h} \right)^2 \right] \text{ if } d_{ij} < h
\]

\[
= 0 \text{ otherwise}
\]

where \( d \) is the Euclidean distance between the location of observation \( i \) and location \( j \), and \( h \) is the bandwidth.

In terms of influencing the fit of the model, the choice of a bandwidth is more important than the shape of the kernel. If the sample points are regularly spaced in the study area, then a kernel with a fixed bandwidth is a suitable choice for modeling. If the sample points are clustered in the study area, it is generally desirable to allow the kernel to
accommodate this irregularity by increasing its size when the sample points are sparser and decreasing its size when the sample points are denser. Both the Gaussian and the Gaussian-like functions (bi-square) were implemented in GWR3.0, ArcGIS 9.3.1 and R.

2.4.3 GWR model calibration and diagnosis

A common approach to regression analysis is to identify the best OLS model first, before moving to GWR regression. After one or more candidate regression models are identified using the ArcGIS OLS statistical regression tool, the GWR model is applied for the selected regression models. The ArcGIS GWR statistical tool creates coefficient raster surfaces for the model intercept and each explanatory variable.

The measure of goodness of fit used extensively in GWR is the corrected Akaike Information Criterion (AICc). The AICc compares models of the same target variable and it contains a penalty for the complexity (degrees of freedom) of the model. The AICc provides a measure of the information distance between the model which has actually been fitted and the unknown ‘true’ model. This distance is not an absolute measure but a relative measure known as the Kullback-Leibler information distance. AICc can be used to compare the global OLS model with a local GWR model. The AICc is also used in the software to determine the ‘optimal’ value for the bandwidth; the bandwidth with the lowest AICc is used in the estimation of the model parameters. Other model performance indicators are the sum of the squared residuals (SSR), sigma (the estimated standard deviation for the residuals), and $R^2$. 

25
In this study, the best GWR model will be selected based on AICc, SSR (or sigma) and adjusted $R^2$. If at least two of the SSR, AIC and adjusted $R^2$ values for Model A are better than those for Model B, while the remaining one (if any) is not too much worse than that for Model B, then Model A can be thought as better. The model output feature class table includes fields for observed and predicted y values, condition number (cond), Local $R^2$, residuals, and explanatory variable coefficients and standard errors and can be examined for local errors and local collinearity.

### 2.5 Regression-Kriging

#### 2.5.1 Introduction

Regression-Kriging (R-K) is one of the several hybrid spatial prediction models comprised of a combination of both correlation with ancillary predictors and spatial autocorrelation simultaneously. R-K was first named by Odeh et al. (1994 and 1995) and three sub-models were proposed.

In R-K models, the regression residuals represent uncertainty and are incorporated into kriging systems. In R-K model A, a regression model is first created between a target variable and the ancillary data, then the regression residuals are further interpolated using kriging. In model B, the regression part is similar to the first step of model A based on the multi-linear regression model fitted to points where both the target variable and the co-variables (such as the landform variables) are determined. The residuals of the target variable at the same locations are calculated, followed by ordinary kriging of both the residuals and the regressed values to a finer grid. This is based on the assumption that the
regression residuals retain some spatial structure inherent to the dependent variable. Both the regressed values and kriged residuals are added together at un-sampled locations as the predicted values. Model C is similar to Model B, and the difference is that the residuals are kriged into the same locations at nodes of the finer DEM (i.e., locations where the target variable is not determined). The residuals are then added to the values of the target variables obtained by regression models at the same locations to give the new prediction. Model C is usually called the Regression-Kriging model (R-K).

R-K is considered the Best Linear Unbiased Prediction (BLUP) model for spatial data (Gotway and Stroup, 1997; Stein, 1999; Christensen, 2001). Universal Kriging (UK) and Kriging with external drift (KED) are two variants of R-K, and are essentially the same or very similar techniques as R-K with computational differences. In the case of R-K, the deterministic (regression) and stochastic (kriging) predictions are done separately; in the case of kriging with external drift, both components are fitted simultaneously; the term universal kriging is often reserved for the case when the deterministic part is modeled as a function of coordinates. Goovaerts (1997) referred to R-K as Simple Kriging with varying local means. In many cases, kriging combined with regression yields more detailed results and higher accuracy of prediction than previous geostatistical techniques. Hudson and Wackernagel (1994) showed that kriging with use of elevation data improves mapping of temperature. Knotters et al. (1995) compared ordinary kriging with co-kriging and regression-kriging for soil mapping purposes, favoring the latter. Bourennane et al. (1996, 2000) showed that prediction of horizon thickness is more accurate with the use of a slope map as external drift. In several other studies (Odeh et al., 1994, 1995;
Goovaerts, 1999; Bishop and McBratney, 2001), combination of kriging and correlation with auxiliary data outperformed ordinary kriging, co-kriging and plain regression. Odel et al. (1994 and 1995) demonstrated that the RMSE for all soil variables are much smaller than those from multi-linear regression, ordinary kriging and co-kriging. Hengl et al. (2004 and 2009) described a framework for soil variable prediction based on regression-kriging (Figure 2.3) and compared it with ordinary kriging and plain regression using profile observations from the national survey in Croatia. Three target variables (organic matter, pH in topsoil and topsoil thickness) were predicted from six relief parameters and nine soil mapping units. His results show that the R-K method produce better predictions than OK and pure regression.

![Figure 2.3 A schematic example of regression-kriging](From Hengl, 2009)

### 2.5.2 Model equation and calibration

Both deterministic and stochastic components of spatial variation can be modeled separately, and can then be combined:

\[
\hat{Z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^{n} \lambda_i \cdot e(s_i)
\]

(13)
where $\hat{m}(s_0)$ is the fitted deterministic part, $\hat{e}(s_0)$ is the interpolated residual, $\hat{\beta}_k$ is the estimated deterministic model coefficients ($\hat{\beta}_0$ is the estimated intercept), $\lambda_i$ is the set of kriging weights determined by the spatial dependence structure of the residual and $e(s_i)$ is the residual at location $s_i$.

The regression coefficients $\hat{\beta}_k$ can be estimated from the samples by some fitting method, e.g. ordinary least squares (OLS) or Generalized Least Squares (Cressie, 1993). In this case, the weights are determined objectively to account for the spatial auto-correlation between the residuals.

Once the deterministic part of variation has been estimated, the residual can be interpolated with kriging and added to the estimated trend. Optimally, estimation of the residuals and their variogram model shall be performed in an iterative process (Hengl, 2009). However, several studies showed that the covariance function derived from the OLS residuals (i.e. a single iteration) is often satisfactory, and that is not much different from the function derived after several iterations (Kitanidis, 1994; Minasny and McBratney, 2007).

In this study, a computation procedure was implemented for this model using ArcGIS ModelBuilder, which includes the following steps: 1) estimate the deterministic part of variation and regression coefficients with ordinary least squares techniques (OLS) implemented in ArcGIS Statistical Tools; 2) compute the residuals from 1) using ordinary kriging; and 3) sum the trend (regression) and drift (residuals).
R-K can also be run in SAGA (Olaya, 2004) and gstat (implemented in R and Idrisi). SAGA has a user-friendly environment to enter prediction parameters, however, it does not offer possibilities for more extensive statistical analysis. R seems to be the most suitable computing environment for R-K as it allows largest family of statistical methods and supports data processing automation.

2.6 Spatial Model Selection, Assumptions and Uncertainty Analysis

2.6.1 Spatial model selection and comparison

Spatial prediction models (algorithms) can be categorized as either deterministic or geostatistic. For deterministic spatial models, empirical parameters estimate the dependent variable, and no estimate of error is available. Thiessen polygons, inverse distance interpolation, and spline methods belong to this group. For statistical spatial models, dependent variables are estimated from an assumed probability distribution, and are accompanied with estimate of the prediction error (uncertainty). Kriging, regression, Bayesian-based models, and mixed models (such as regression-kriging and Geographically Weighted Regression) belong to this category.

Development of generic and robust spatial interpolation techniques has been of interest for many years (Mitas and Mitasova, 1999). Kriging and its variants have been widely recognized as primary spatial interpolation techniques from the 1950s. In the 1990s, with the emergence of GIS and remote sensing technologies, exhaustively mapped secondary variables (DEM terrain attribute maps) and images (such as landuse) were used to
directly map soil variables by use of linear regression models (Gessler et al., 1995; Moore et al., 1993). This approach was termed environmental correlation (McKenzie and Ryan, 1999). In the last decade, a number of ‘hybrid’ interpolation techniques, which combine kriging and use of ancillary information, have been developed and tested. Kriging combined with regression (McBratney et al., 2000) is one of the most popular methods with few estimated model parameters (Knotters et al., 1995; Hengl et al., 2004).

Pebesma (2004 and 2009) selected models for spatial prediction based on three criteria: 1) if a physical model is defined; 2) if sampled variables correlate with environmental variables, and 3) if the residuals show spatial autocorrelation. Regression-kriging, ordinary kriging, multi-linear regression and GWR models can be chosen based on the set of conditions available for the study. Using these criteria, the type of prediction model to use can be determined (Figure 2.4). If a deterministic model is not defined, and correlations exist between sampled variables and environmental factors, then a statistically significantly regression model can be defined. The regression residuals are examined for spatial autocorrelation. If no spatial autocorrelation is identified with the residuals, the regression coefficients can be estimated using OLS. Otherwise, mixed models like regression-kriging and GWR can be applied for estimation. If the data shows no correlation with environmental factors, the variogram of the target variable can be analyzed and ordinary kriging and its variants can be used for estimation.
2.6.2 Spatial model assumptions

There are a number of assumptions associated with spatial statistical prediction models. Ignoring these assumptions leads to poor predictions (Hengl, 2009). Variables of interest must be stationary and that is the statistical properties of the variables (e.g. mean, standard deviation, covariance are the same over space and time for the area of interest). There are two orders of stationarity: first-order stationarity where the mean does not change and second-order where the covariance does not change. If a variable is non-stationary, the statistical properties change across space. There are several approaches that can be used. One is to create zones of homogeneity, and another is to allow parameters to vary constantly: creating measures of statistical relationships that are continuously varying across space, such as GWR (Fotheringham, 2002).

The mean and covariance stationarity and a normal distribution of values are required for ordinary kriging. For multi-linear regression, the error variation must be minimized,
constant and normally distributed around the model estimate (Goovaerts, 1997)
Transformations of the variables may be used to meet these conditions. For regression-
kriging, the target variable does not have to be stationary but its residuals must be
stationary.

2.6.3 Uncertainty analysis

A more systematic strategy to select the right spatial prediction technique is to use
objective criteria of mapping success. Model errors, robustness, reliability and the
computational burden are evaluated. Huang and Chen (2007) suggested a framework to
select the best predictor using an automated method to account for anisotropy, specified
observation errors, and extreme values.

Model variance is the statistical estimate of the model uncertainty. But the ‘true’
prediction power can only be assessed by using an independent (control) data set. The
prediction error is therefore often referred to as the precision of prediction. The true
quality of a map can be assessed by comparing estimated values with actual observations
at validation points. Commonly, two measures are most relevant:

- Mean prediction error (ME):

\[
ME = \frac{1}{l} \sum_{i=1}^{l} [\hat{z}(s_i) - \mathbb{Z}^*(s_i)]; \quad E\{ME\} = 0
\]

- Root mean square prediction error (RMSE):

\[
RMSE = \sqrt{\frac{1}{l} \sum_{i=1}^{l} [\hat{z}(s_i) - \mathbb{Z}^*(s_i)]^2}; \quad E\{RMSE\} = \sigma(h = 0)
\]

where \(l\) is the number of validation points.
In order to compare accuracy of prediction between variables of different types, the RMSE can also be normalized by the total variation, called the normalized root mean squared deviation or error (NRMSD or NRMSE):

\[
NRMSE = \frac{RMSE}{x_{\text{max}} - x_{\text{min}}}
\]  

(16)

The equation above determines the amount of global variation explained by the model. As a rule of thumb, a value of NRMSE that is close to 40% means a fairly satisfactory accuracy of prediction (R-square=85%). If NRMSE >71%, then the model accounts for less than 50% of variability at the validation points.

Another useful thing to do is to examine the histogram of errors at validation points and compare the errors estimated by the model and the true mapping error at validation points. This can help us to detect ‘unusual’ locations where the errors are much higher than at other locations. Scatter plots of predicted versus the measured values at control points can also be used to derive a correlation coefficient to test the model’s ability.
CHAPTER 3 DATA COLLECTION AND DESCRIPTION

3.1 Spatial and Temporal Changes of Sediment Zones

Sediment cores were taken at 170 locations throughout the lake in 1988, 1998 and 2006 (Figure 1.1, Reddy et al. 1995; Fisher et al. 2001; BEM & University of Florida 2007), respectively. Four sediment zones were identified in the lake: Mud, Peat, Sand and Rock. Sediment zone boundaries were delineated based on sampling site types, site locations, elevation data and high resolution aerial-photography (Figure 3.1).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MUD</td>
<td>171954.4268</td>
<td>38.52</td>
<td>193569.62</td>
<td>43.37</td>
<td>205497.6756</td>
<td>46.04</td>
</tr>
<tr>
<td>PEAT</td>
<td>122771.6477</td>
<td>27.51</td>
<td>97934.15</td>
<td>21.94</td>
<td>98132.7267</td>
<td>21.99</td>
</tr>
<tr>
<td>SAND</td>
<td>108679.4781</td>
<td>24.35</td>
<td>101351.41</td>
<td>22.71</td>
<td>97569.7962</td>
<td>21.86</td>
</tr>
<tr>
<td>Total</td>
<td>446354.3219</td>
<td>100</td>
<td>446354.3219</td>
<td>100</td>
<td>446354.3219</td>
<td>100</td>
</tr>
</tbody>
</table>

From 1988 to 1998, there were no major change for rock, and sand zones, but peat (littoral) zone reduced from 27.5% to 22%. From 1988 to 2006, mud zone increased from 38.5% to 46%. This is the zone changed mostly (Table 3.1 and Figure 3.1).
Figure 3.1 Sediment zones distribution from 1988 to 2006, Lake Okeechobee
3.2 Integrated Digital Elevation Model

An accurate bathymetric data of the Lake is very important to calculate lake volume, to define potential areas where aquatic plants may grow, and to model hydrodynamics, water quality and ecology. The district executed a contract to re-survey the lake from July 8 to September 29, 2008, which produced approximately 1,291 miles of surveyed bathymetric lines equivalent to 681,796 measured data points within the lake (Figure 3.2a). In 2007, the Florida Division of Emergency Management executed a contract to conduct for a LiDAR (Light Detection And Ranging) survey for the project area around Lake Okeechobee, Florida. Over 2,289 square miles of data were collected (Figure 3.2a). In this thesis, these two data sets were processed separately and then integrated into an accurate digital elevation model for the whole Lake Okeechobee using advanced GIS techniques. 2-ft contours (Figure 3.3a) and 3-D image (Figure 3.3b) were created using the digital elevation model. This new bathymetric map provides more current and accurate information compared to previous maps from 2001 and 1988.
Figure 3.2 LiDAR DEM and bathymetry transects (a) and bathymetry DEM (b)
Figure 3.3 Digital elevation model and eco-zones (a) and 3D view (b)
CHAPTER 4 SPATIAL-TEMPORAL CHANGES OF MUD THICKNESS

4.1 Data Description and Characteristics

Mud thickness data were measured at the 171 sample locations throughout the lake in 1988, 1998 and 2006. Mud thickness varied both on a temporal and spatial scale. The maximum mud thickness for 1988, 1998 and 2006 were 66, 74 and 51 cm respectively; and the mean thickness declined from 12.47 cm in 1988 to 8.27 cm in 2006 (Table 4.1). Spatially there was a first order trend of mud thickness in the E/W direction and a strong second order trend in N/S direction (Figure 4.1).

Table 4.1 Summaries of mud thickness and bulk density

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Median</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Mean</td>
<td>12.47</td>
<td>11.17</td>
<td>8.27</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>18.88</td>
<td>13.75</td>
<td>14.75</td>
</tr>
<tr>
<td>Max.</td>
<td>66.00</td>
<td>74.00</td>
<td>51.00</td>
</tr>
<tr>
<td>Skewness</td>
<td>1.41</td>
<td>1.66</td>
<td>1.43</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>3.68</td>
<td>4.51</td>
<td>4.05</td>
</tr>
</tbody>
</table>

4.2 Ordinary Kriging Model Calibration

Training and testing data sets were created randomly using GIS Geostatistical Analyst (GA) extension and the training data sets were used to calibrate the models and the testing data sets were used to validate the modes for accuracy.
Figure 4.1 Trend analyses of mud thickness data
(X, Y and Z represent E, N and vertical direction, respectively; red dots are samples; green and blue lines are trends)
Figure 4.2 Calibration and validation sample distributions of 2006 mud thickness

A spherical semivariogram model was fitted for each mud thickness data. The following parameters were selected based on data exploration and interactive visual check using ArcGIS Geostatistical Analyst: lag size 8000 ft, lag number 9, range 50000/45000 ft, and the second-trend removal.
1988: $249.93 \times \text{Spherical}(50000,45000,36.3) + 25.512 \times \text{Nugget}$

1998: $243.81 \times \text{Spherical}(50000,45000,62.3) + 36.485 \times \text{Nugget}$

2006: $83.419 \times \text{Spherical}(50000,45000,68.2) + 34.63 \times \text{Nugget}$

Figure 4.3 Fitted variogram models using mud thickness calibration data
4.3 Ordinary Kriging Model Validation

The validation errors were displayed in histogram charts (a, c and e of Figure 4.4) and Quantile-Comparison plots (b, d and f of Figure 4.4) using R. All the errors were normally distributed for the three mud thickness data sets. In the normal quantile plots, which plot the empirical quantiles of the errors (Y-axis) against theoretical quantiles (X-axis) of a comparison normal distribution, the 1988 mud thickness has errors limited in the 95% confidence envelopes; some of the validation sites have errors outside of the envelope for 1998 and 2006 data sets, which show under-estimates (positive errors) and over-estimates (negative errors). The scatter plots of the measured and predicted values of the validation sites (Figure 4.5) show very good fitting with determination of coefficients \(R^2\) of 0.8572, 0.829 and 0.8476 for 1988, 1998 and 2006 mud thickness, respectively.

Table 4.2 lists the calibration and validation errors for the three mud data sets. All the calibration errors are a little lower than the validation errors, partly due to smaller number of validation sites.

<table>
<thead>
<tr>
<th>Errors</th>
<th>2006</th>
<th>1998</th>
<th>1988</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>0.0195</td>
<td>0.0601</td>
<td>-0.0862</td>
</tr>
<tr>
<td>Root-Mean-Square</td>
<td>8.665</td>
<td>9.502</td>
<td>9.791</td>
</tr>
<tr>
<td>Mean Standardized</td>
<td>0.0019</td>
<td>0.0014</td>
<td>-0.005</td>
</tr>
<tr>
<td>Root-Mean-Square Standardized</td>
<td>1.051</td>
<td>0.8653</td>
<td>0.919</td>
</tr>
<tr>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>-0.1243</td>
<td>0.1014</td>
<td>-0.0638</td>
</tr>
<tr>
<td>Root-Mean-Square</td>
<td>7.821</td>
<td>10.24</td>
<td>9.387</td>
</tr>
<tr>
<td>Mean Standardized</td>
<td>-0.0272</td>
<td>0.0119</td>
<td>-0.0088</td>
</tr>
<tr>
<td>Root-Mean-Square Standardized</td>
<td>0.947</td>
<td>0.927</td>
<td>0.9224</td>
</tr>
</tbody>
</table>
Figure 4.4 Validation error distributions of the mud thickness data
Figure 4.5 Scatter plots of the mud thickness validation data
4.4 Spatial Changes of Mud Thickness Over-time

Validated kriging models were used to estimate the mud thickness distribution for 1988, 1998 and 2006, respectively. The fitted variograms were shown in Figure 4.6, and the mud thickness distributions were mapped in Figure 4.7.

1988: (248.72*Spherical(50000,45000,27.8)+29.521*Nugget)

1998: (282.03*Spherical(50000,45000,53.5)+21.604*Nugget)

2006: (93.167*Spherical(50000,45000,89.6)+32.647*Nugget)

Figure 4.6 Fitted variogram models using the complete mud thickness data
Figure 4.7 Mud thickness maps for 1988, 1998 and 2006 data
4.5 Mud Surface Area and Volume Calculations

ArcGIS 3D Analyst and Spatial Analyst provide Area/Volume and Cut/Fill tools to calculate 2D area, surface area, and volume and their changes. The 2D area of a rectangular patch of surface model is its length times its width. The surface area is measured along the slope of the surface. Unless the surface is flat, the surface area will always be greater than the 2D area. The difference between the values for the 2D area and surface area indicates the roughness or slope of the surface - the larger the difference between the values, the rougher or steeper the surface. The volume is the space between the surface and a reference plane set at a particular height. An ArcGIS ModelBuilder model was developed to integrate several tools to streamline the calculations of the mud area, mud volumes and their changes over-time (Figure 4.8 and Figure 4.9). For the mud sediments, the area differences are less than 3 square meters for all three data sets (Table 4.3), which suggest that the mud surfaces are nearly flat. Mud area decreased 13.78% during 1988-1998 and increased 0.74% during 1998-2006. The mud volumes reduced 10.26% and 26.62% during 1988-1998 and 1998-2006, respectively (Table 4.4).

Table 4.3 Mud surface area and volume

<table>
<thead>
<tr>
<th>Year</th>
<th>2D M²</th>
<th>3D M²</th>
<th>Acres</th>
<th>M³</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>1.65E+09</td>
<td>1.65E+09</td>
<td>408,232.18</td>
<td>2.19E+08</td>
</tr>
<tr>
<td>1998</td>
<td>1.42E+09</td>
<td>1.42E+09</td>
<td>351,978.94</td>
<td>1.97E+08</td>
</tr>
<tr>
<td>2006</td>
<td>1.43E+09</td>
<td>1.43E+09</td>
<td>354,577.85</td>
<td>1.44E+08</td>
</tr>
</tbody>
</table>

Table 4.4 Changes of mud surface area and volume

<table>
<thead>
<tr>
<th>Period</th>
<th>Area change (acres)</th>
<th>Area change (%)</th>
<th>Volume change (M³)</th>
<th>Volume change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1998-2006</td>
<td>2,600</td>
<td>0.74</td>
<td>-5.24E+07</td>
<td>-26.62</td>
</tr>
</tbody>
</table>
The spatial variations of mud thickness are displayed in Figure 4.9. From 1988 to 1998, mud sediments were continually moving toward the center of the Lake. Mud thicknesses increased up to 26 cm in the central lake area (Figure 4.7 and 4.9). Mud thickness reduction mainly occurred near the shore during this period (Figure 4.9a). From 1998 to 2006, the area of mud sediments declined slightly (0.74%) (Figure 4.9b, and Table 4.3-4.4). Mud thickness now only reached 51 cm (Figure 4.7). Mud depths declined by up to 41 cm in the central lake area and increased by up to 20 cm in surrounding areas, with small amounts of mud being deposited throughout the rest of the lake.

Changes in mud thickness between years are also informative regarding the potential effects of hurricane induced mixing. Between 1988 and 1998 (Figure 4.9a), the change vector is spatially heterogeneous, with local areas in primarily deeper water showing marked increases juxtaposed with areas in shallower regions showing equally marked decreases in mud depths. Notably, the areas in the center of the lake where mud is currently deepest accumulated large quantities, while outlying areas where mud is largely absent today (western lake) lost appreciable mud depth. Near the inflow of the Kissimmee River (northern lake), mud depths also declined. Overall, the weak spatial structure of the change pattern for 1988-to-1998 is suggestive of the major current gyres that exist in the lake (Jin and Ji, 2004). In contrast, mud depth changes between 1998 and 2006 occur over much larger scales (Figure 4.9b). The middle of the lake appears to have lost appreciable mud, while areas around the central zone appear to have accumulated mud. This is suggestive of re-depositional processes, perhaps in response to high wind mixing events.
Figure 4.8 ModelBuilder model for mud thickness data processing and volume calculation
Figure 4.9 Spatial changes of mud thickness over-time
4.6 Mud Weight Calculation

Mud weight was calculated by multiplying mud volume times mud density for each location. Because mud density changes over distance in the lake, it’s necessary to map the density changes first before calculating the total mud weights.

4.6.1 Bulk density mapping

2006 data has a total of 142 samples with valid density values and the max value of 5.43 g/cm$^3$ is an outlier (Site N3). The 1998 data has 149 valid density values and the 1988 data set has 134 valid values. The mean density values increased from 0.376 to 0.529 to 0.58 g/cm$^3$, from 1988 to 1998 to 2006, respectively. All the density values show a skewed distribution (Table. 4.1). Both the 1988 and 2006 density data have first order trend changes in both E/W and N/S directions (Figure 4.10) and the 1998 data has 2$^{nd}$ order trend changes in both E/W and N/S directions (Figure 4.10). All three data sets show weak spatial auto-correlation, and a spherical variogram was fitted for each data set (Figure 4.11).

Spatial changes of the mud density were mapped using Ordinary Kriging (Figure 4.12). The central mud area and southern peat areas have the lowest density (red colors in Figure 4.12) for all three data sets. The near-shore zones have the highest density values (blue colors in Figure 4.12).
Figure 4.10 Trend analyses of mud bulk density
1988: 0.1471*Spherical(45000)+0.080691*Nugget

1998: 0.077964*Spherical(45000)+0.16497*Nugget

2006: 0.089607*Spherical(45000)+0.15971*Nugget

Figure 4.11 Fitted variogram models for mud density
Figure 4.12 Spatial variations of mud density
4.6.2 Mud weight calculations

The total mud weight is calculated by summing all mud cell weights. The cell mud weight is equal to the cell volume multiplied by the cell density. The unit of the cell-edge is in feet, and the mud thickness in centimeter, the density in g/cm³. The mud cell weight (in kilogram) equation is:

\[ \text{Mud cell weight} = 232257.60 \times [\text{Mud Thickness}] \times [\text{Mud Density}] \]

where 232257.60 is a constant for unit conversion to produce mud weight in kilogram.

The calculation process is implemented using ArcGIS ModelBuilder and Map Algebra (Figure 4.13). The final calculation results for each data set are listed in Table 4.4. The mud weight is about 57.2 millions of tons for 1988 data, and 73.9 millions of tons for 1998, and 58.6 millions of tons for 2006 data. Over 29% of mud increased from 1988 to 1998, over 20% decreased from 1998 to 2006. However, the mud weight just increased 2.5% from 1988 to 2006. Where is the mud gone during 1998 and 2006? Numerous studies indicate that the mud sediments were re-suspended and re-disturbed during the four major hurricanes between 2004 and 2005. The major part of the lost mud could have been moved out of the lake with outflows after the hurricanes indicated by the spikes of Lake TSS mass after the 2004-2005 hurricanes (Figure 1.2). Small amounts of mud could have been re-distributed from the central area to the surrounding shallow area with a 0.75% of mud area increase from 1998 to 2006 (Figure 4.9b and Table 4.3).
Figure 4.13 ModelBuilder model for mud weight calculation
Table 4.5 Mud weights and their changes over-time

<table>
<thead>
<tr>
<th>Data</th>
<th>Cell #</th>
<th>Weight (Kg)</th>
<th>Weight (Ton)</th>
<th>Change %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(88-98) (98-06) (88-06)</td>
</tr>
<tr>
<td>1988</td>
<td>68815</td>
<td>5.19E+10</td>
<td>5.72E+07</td>
<td>Increase + 29.32</td>
</tr>
<tr>
<td>1998</td>
<td>60545</td>
<td>6.71E+10</td>
<td>7.40E+07</td>
<td>Decrease -20.73</td>
</tr>
<tr>
<td>2006</td>
<td>62465</td>
<td>5.32E+10</td>
<td>5.86E+07</td>
<td>Increase + 2.52%</td>
</tr>
</tbody>
</table>

4.7 Summary of Results

- From 1988 to 1998, mud sediments were continually focused toward the center of the Lake, and mud thicknesses increased up to 26 cm in the central lake area and reduced near the shore zone. The mud area and volume reduced up to 13.78% and 10.26%, respectively (Table 4.2 and Table 4.3).
- From 1998 to 2006, mud depths declined by up to 41 cm in the central lake area and increased by up to 20 cm in the surrounding areas, with small amounts of mud being deposited throughout the rest of the lake. The area of mud sediments increased slightly but the mud volume reduced about 27%. The reduction of mud sediments is likely due to re-suspension and redistributed by wind-induced waves and currents produced by Hurricanes Frances and Jeanne in 2004 and Wilma in 2005.
- Mud weight increased over 29% from 1988 to 1998, and reduced over 20% from 1998 to 2006. Overall, the mud weight increased about 2.5% from 1988 to 2006.
- The major part of the sediments accumulated during 1988 to 1998 was washed out of the lake during the four major hurricanes of 2004-2006 based on TSS mass change.
CHAPTER 5 SPATIAL MODELING OF TOTAL PHOSPHORUS

5.1 Data Description and Characteristics

Histograms of TP in sediments demonstrated a bi-normal distribution for each year sampled (a, c and e of Figure 5.1). TP varies in different zones, and with highest values in mud zones (> 850 mg/kg), lower values in peat zones (250-450 mg/kg) and lowest in sand and rocks zones (0-300 mg/kg) (b, d and f of Figure 5.1). The 2006 TP values range from 16.64-999.46 mg/kg, with mean and media values of 566 and 397 mg/kg, respectively. The 1998 TP values change from near zero to over 1,793 mg/kg, with mean and media values of 650 and 475 mg/kg, respectively. The 1988 TP ranges from 38.6 to 1,708 mg/kg, with 1st quantile of 220 mg/kg and 3rd quantile of 1,076 mg/kg. Its mean and median values are 670.5 and 664 mg/kg, respectively (Table 5.1). The TP concentrations of the three data sets show a 2nd order trend change along the N/S direction and relatively weak trend (1st order trend) in E/W direction (Figure 5.2).

The correlation plots display the relationship among different elements (Figure 5.3). TP is strongly and positively correlated with Fe. TP is also weakly and positively correlated with mud thickness, Ca, and TC and weakly and negatively correlated with bottom elevation.
Plot of Means

(a) Histogram showing frequency distribution for Sed88$TP.

(b) Scatter plot illustrating mean Sed88$TP across different SedZone categories: MUD, PEAT, ROCK, SAND.
Figure 5.1 TP distributions (a, c, and e) and mean changes (b, d, and f) in sediment zones
Figure 5.2 TP trend analyses of the sediment data
(Refer to Figure 4.1 for the symbols)
Table 5.1 Data summaries of key elements of sediment data

<table>
<thead>
<tr>
<th></th>
<th>TC</th>
<th>CA</th>
<th>MG</th>
<th>FE</th>
<th>TN</th>
<th>TP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min.</td>
<td>0.60</td>
<td>41</td>
<td>35</td>
<td>24</td>
<td>166</td>
<td>38.60</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>17.75</td>
<td>7297</td>
<td>691</td>
<td>643</td>
<td>1541</td>
<td>220.70</td>
</tr>
<tr>
<td>Median</td>
<td>107.80</td>
<td>54377</td>
<td>4243</td>
<td>1700</td>
<td>8621</td>
<td>664.00</td>
</tr>
<tr>
<td>Mean</td>
<td>118.93</td>
<td>53799</td>
<td>7184</td>
<td>2695</td>
<td>9094</td>
<td>670.50</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>151.80</td>
<td>85154</td>
<td>11790</td>
<td>4787</td>
<td>11319</td>
<td>1076.00</td>
</tr>
<tr>
<td>Max.</td>
<td>481.90</td>
<td>176020</td>
<td>40019</td>
<td>10524</td>
<td>36107</td>
<td>1708.10</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min.</td>
<td>1.40</td>
<td>220</td>
<td>40</td>
<td>112</td>
<td>0</td>
<td>2.20</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>25.93</td>
<td>10396</td>
<td>775</td>
<td>1449</td>
<td>400</td>
<td>165.70</td>
</tr>
<tr>
<td>Median</td>
<td>143.60</td>
<td>51325</td>
<td>3534</td>
<td>3506</td>
<td>8000</td>
<td>474.70</td>
</tr>
<tr>
<td>Mean</td>
<td>145.06</td>
<td>60705</td>
<td>11402</td>
<td>7060</td>
<td>9363</td>
<td>650.20</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>182.45</td>
<td>89439</td>
<td>20767</td>
<td>14352</td>
<td>12500</td>
<td>1149.80</td>
</tr>
<tr>
<td>Max.</td>
<td>487.10</td>
<td>328000</td>
<td>95470</td>
<td>21123</td>
<td>47800</td>
<td>1793.30</td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min.</td>
<td>1.15</td>
<td>114</td>
<td>36</td>
<td>39</td>
<td>70</td>
<td>16.64</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>25.69</td>
<td>6813</td>
<td>936</td>
<td>716</td>
<td>610</td>
<td>141.39</td>
</tr>
<tr>
<td>Median</td>
<td>140.77</td>
<td>64349</td>
<td>6796</td>
<td>2013</td>
<td>8170</td>
<td>397.42</td>
</tr>
<tr>
<td>Mean</td>
<td>139.09</td>
<td>64217</td>
<td>13037</td>
<td>3429</td>
<td>8926</td>
<td>566.22</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>175.81</td>
<td>94162</td>
<td>21069</td>
<td>6947</td>
<td>12500</td>
<td>999.46</td>
</tr>
<tr>
<td>Max.</td>
<td>490.55</td>
<td>347790</td>
<td>91451</td>
<td>8827</td>
<td>41030</td>
<td>1285.31</td>
</tr>
</tbody>
</table>
1998
Figure 5.3 Scatter plot matrix of sediment data
Figure 5.4 Total phosphorus maps for 1988 (a), 1998 (b) and 2006 (c) data
From 1988 to 1998, total phosphorus (TP) values slightly decreased about 20 mg/kg averagely, declined from 670.5 mg/kg to 650.20 mg/kg (Table 5.1). Different changes occurred in different part of the lake during this period: TP declined in the northern and southern areas, and increased in the central-western part of the lake (Figure 5.5a). From 1998 to 2006, TP values showed an average decrease of 84 mg/kg, the decrease mainly occurred in the lake’s western and southern regions (Figure 5.4 and Figure 5.5b). Some area experienced increase, such as the northern and western edges, with maximum increase up to 1,000–3,000 mg/kg.

5.2 Regression Model Analysis

5.2.1 Key variable selection and regression model calibration
Training and validation data were created randomly using Geostatistical Analyst extension of ArcGIS 9.3 with 75-25% split. The “best” subset of predictors were determined using ArcGIS OLS tool and Stepwise regression tools of R, by removing redundant (collinear) variables. Time and/or money can be saved by not measuring redundant predictors. Three statistically significant models with different independent variables were identified for TP: Fe only, Th & Elev together, and Th, TC, Ca, Fe together, respectively (Table 5.2 and 5.4).

Table 5.2 OLS model results and diagnosis of 2006 TP calibration data

<table>
<thead>
<tr>
<th>OLS Model Results</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Intercept</td>
<td>55.839436</td>
</tr>
<tr>
<td>TH</td>
<td>3.334015</td>
</tr>
<tr>
<td>TC</td>
<td>0.276333</td>
</tr>
<tr>
<td>CA</td>
<td>0.000554</td>
</tr>
<tr>
<td>FE</td>
<td>0.120419</td>
</tr>
</tbody>
</table>

OLS Model Diagnosis

<table>
<thead>
<tr>
<th>OLS Model Diagnosis</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations: 114</td>
<td>Number of Variables: 5</td>
</tr>
<tr>
<td>Degrees of Freedom: 109</td>
<td>Global Moran's I Sum:</td>
</tr>
<tr>
<td>Multiple R-Squared [2]: 0.90</td>
<td>Adjusted R-Squared [2]: 0.901777</td>
</tr>
<tr>
<td>Joint F-Statistic [3]: 260.36</td>
<td>Moran's Index: 0.020488</td>
</tr>
<tr>
<td>Joint Wald Statistic [4]: 2147.88</td>
<td>Expected Index: -0.00885</td>
</tr>
<tr>
<td>Koenker (BP) Statistic [5]: 15.58</td>
<td>Variance: 0.005105</td>
</tr>
<tr>
<td>Jarque-Bera Statistic [6]: 28.28</td>
<td>Z Score: 0.410596</td>
</tr>
</tbody>
</table>

For 2006 TP, Total C, Total Fe, Total Ca and mud thickness are significant independent variables that explain over 90% of the variation in TP. From the coefficients and t-statics, Total Fe explains the greatest amount of the variation while Total Ca explains the least (Table 5.2).

GWR is a well-suited tool to explore the changes of spatial relationship and can help to identify which variable (s) is the most significant one to the target (TP for this case, table
5.3). Fixed models are better than adaptive models based on model fit and performance. Coefficients maps of the model show the relationship and their changes with TP data (Figure 5.5).

### Table 5.3 GWR model diagnosis with multiple key variables

<table>
<thead>
<tr>
<th>GWR Models</th>
<th>Neighbours</th>
<th>ResidualSquares</th>
<th>Effective #</th>
<th>Sigma</th>
<th>AICc</th>
<th>$R^2$</th>
<th>$R^2_{\text{Adj}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adaptive</td>
<td>66</td>
<td>1108979.84</td>
<td>18.47</td>
<td>107.75</td>
<td>1407.11</td>
<td>0.95</td>
<td>0.94</td>
</tr>
<tr>
<td>Fixed</td>
<td>38198 ft</td>
<td>985618.99</td>
<td>25.06</td>
<td>105.27</td>
<td>1405.83</td>
<td>0.95</td>
<td>0.94</td>
</tr>
</tbody>
</table>

![Coefficients maps of the model show the relationship and their changes with TP data.](image-url)
Figure 5.6 Coefficient maps of the key variables of the 2006 TP GWR calibrated model

Total Fe is positively related to TP over the whole lake area. The correlation is strongest in the SW area and weakest in the northern region (Figure 5.6b). Total Carbon is positively related to TP in the north-western zone and negatively in the south-eastern zone. (Figure 5.6c). Mud thickness is positively related to TP in the southern peat zone and western edge of the littoral zone and negatively related to TP in most other areas of lake (Figure 5.6d). Total Ca is positively related in the north-eastern pelagic zone and negatively in most other area of the lake (Figure 5.6e).

Total Fe can explain over 88% of the model variation, and it is one of the “cheapest” models. The residuals are not normally distributed as shown by the Jarque-Bera statistic. The BP statistic is not significant; indicating the correlation between TP and Total Fe is stationary in the study area (Table 5.4). Also, mud thickness and lake bottom elevation data are available and the elevation data has much higher sampling density. Numerous
studies demonstrated that the “cheap” DEM data can improve correlated target variable estimates such as TP and TN. Regression analysis show that both thickness and elevation are statistically independent. The residuals are not normally distributed as shown by the Jarque-Bera statistic.

Table 5.4 Statistically significant calibrated regression models

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fe</td>
<td>Th &amp; Elev</td>
<td>Fe</td>
</tr>
<tr>
<td>AIC</td>
<td>1580.781885</td>
<td>1657.348992</td>
<td>1563.969487</td>
</tr>
<tr>
<td>R²</td>
<td>0.719529</td>
<td>0.460541</td>
<td>0.813803</td>
</tr>
<tr>
<td>AdjR²</td>
<td>0.717025</td>
<td>0.450821</td>
<td>0.812141</td>
</tr>
<tr>
<td>F-Stat</td>
<td>287.328962</td>
<td>47.380926</td>
<td>489.514772</td>
</tr>
<tr>
<td>F-Prob</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>K(BP)</td>
<td>2.012166</td>
<td>2.998024</td>
<td>1.558775</td>
</tr>
<tr>
<td>K(BP)-Prob</td>
<td>0.156042</td>
<td>0.223351</td>
<td>0.211845</td>
</tr>
<tr>
<td>JB</td>
<td>22.710739</td>
<td>10.300433</td>
<td>447.173681</td>
</tr>
<tr>
<td>JB-Prob</td>
<td>0.000012</td>
<td>0.005798</td>
<td>0.000000</td>
</tr>
<tr>
<td>Sigma</td>
<td>246.071287</td>
<td>342.802448</td>
<td>228.579184</td>
</tr>
<tr>
<td>Neighbors</td>
<td>15795.23739</td>
<td>15795.23739</td>
<td>13665.41159</td>
</tr>
<tr>
<td>Stationary</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Residuals</td>
<td>Clustered</td>
<td>Clustered</td>
<td>Random</td>
</tr>
</tbody>
</table>

The BP statistic is not significant, indicating the correlations between TP and thickness and elevation are stationary in the study area (Table 5.4). For 1998 and 1988 TP data sets, similar models were identified. Two models for each data set were evaluated using validation data sets (Table 5.4)
5.2.2 Regression model validation

5.2.2.1 OLS model (TP vs. Mud thickness and Elevation)

For the model of TP vs. mud thickness and elevation, the OLS model equation calculates TP concentration from a raster calculator or map algebra functions. The 500 ft grid estimates at the validation points are compared to observed data (Figure 5.7-Figure 5.8). This process is implemented in ArcGIS using Model Builder.

In the histograms (a, c and e of Figure 5.7), the errors showed a skewed distribution for 2006 data, near normal distributions for 1998 and 1988 data. Two validation sites have large positive errors which are under-estimated in 2006 data (f of Figure 5.7). For 1988 and 1998 models, the errors are distributed in narrow zones. (b and d of Figure 5.7)

After removing the outliers which have large positive or negative values (as shown in Figure 5.7), the scatter plots of the observed values and estimated values of the validation samples showed good fitting with $R^2$ of 0.62-0.66 (Figure 5.8).
(a) 

(b)
Figure 5.7 Validation error distributions of the OLS model (TP vs. Th & Elev)
Figure 5.8 Scatter plots of the OLS model (TP vs. Th & Elev) validation

(a) $y = 0.64x + 253.37$
$R^2 = 0.6268$

(b) $y = 0.6482x + 267.04$
$R^2 = 0.6556$

(c) $y = 0.5568x + 265.91$
$R^2 = 0.6224$
5.2.2.2 OLS model (TP vs. Total Fe)

For the 2006 TP vs. Fe OLS model, the regression equation is, \( TP = 99.7153 + 0.1284*Fe \), and TP concentrations are calculated using the raster calculator. The model errors distributed normally and vary in the narrow range (Fig 5.9); (e and f of Figure 5.9); the errors for 1998 and 1988 showed non-normal distributions and there were several outliers in the normal quantile charts, which are under-estimated (a, b, c and d of Figure 5.9). The 2006 model matches the validation data very well with an \( R^2 \) of 0.8548 (Figure 5.10). However, for 1988 and 1998 the comparison is not as good to the validation data, in part due to the outliers in the validation data set (Figure 5.9). After removing the outliers, \( R^2 \) improved to 0.87 and 0.85 for 1988 and 1998 data, respectively (a and b of Figure 5.10)
(c) Frequency

(1998)

(d) TP98Val_OLS_\$Error vs norm quantiles
Figure 5.9 Validation error distributions of the OLS model (TP vs. Fe)
Figure 5.10 Scatter plots of the OLS model (TP vs. Fe) validation
5.3 GWR Models Analysis

5.3.1 GWR model calibration

5.3.1.1 GWR Model (TP vs. Mud thickness and Elevation)

For all data sets the Fix CV model was provided the best overall results (Table 5.5). For the 2006 data set, the Fixed CV model has the lowest errors (SSR, Sigma), better model performance (lower AICc value) and the model explained much of the variability (higher R², Table 5.5. This sub model was validated. Its residuals are dispersed based on the Moran Index value. For the 1998 TP data, all the GWR sub models performed the same with small differences and weakly dispersed residuals. Fixed CV model has higher R² and lower RSE, and will be used for model validation. For the 1988 TP data, the Fixed CV performed better with the lowest SSR, and its residuals are distributed randomly.

The model coefficient maps show some very strong spatial gradients of the relationships to TP concentrations (Figure 5.11). In the 2006 data, mud thickness has positive correlation with TP, and the correlation decreases from the eastern lake zones to the western littoral zones (c of Figure 5.11); Site elevation has strong negative correlation in the central lake zone and positive correlation in most of the lake. The 1988 and 1998 data show similar correlations (b and c of Figure 5.11).
Table 5.5 Diagnosis of the GWR calibration model (TP vs. Th & Elev)

<table>
<thead>
<tr>
<th>Year</th>
<th>GWR Models</th>
<th>Adaptive AICc</th>
<th>Adaptive CV</th>
<th>Fixed AICc</th>
<th>Fixed CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neighbours</td>
<td>49</td>
<td>54</td>
<td>30916.05</td>
<td>28774.00</td>
</tr>
<tr>
<td>2006</td>
<td>ResidualSquares</td>
<td>4738304.05</td>
<td>4895445.12</td>
<td>4211677.18</td>
<td>4010589.60</td>
</tr>
<tr>
<td></td>
<td>EffectiveNumber</td>
<td>17.7256</td>
<td>16.0334</td>
<td>25.5135</td>
<td>28.0932</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>221.8482</td>
<td>223.5409</td>
<td>218.1670</td>
<td>216.0680</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>1569.8017</td>
<td>1570.1055</td>
<td>1571.1211</td>
<td>1571.4100</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.7821</td>
<td>0.7748</td>
<td>0.8063</td>
<td>0.8155</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.7442</td>
<td>0.7403</td>
<td>0.7526</td>
<td>0.7574</td>
</tr>
<tr>
<td></td>
<td>Neighbours</td>
<td>66</td>
<td>66</td>
<td>44943.0081</td>
<td>42883.6271</td>
</tr>
<tr>
<td>1998</td>
<td>ResidualSquares</td>
<td>12781567.51</td>
<td>12781567.51</td>
<td>12902300.62</td>
<td>12735948.86</td>
</tr>
<tr>
<td></td>
<td>EffectiveNumber</td>
<td>11.7449</td>
<td>11.7449</td>
<td>13.0234</td>
<td>13.8792</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>353.5490</td>
<td>353.5490</td>
<td>357.4565</td>
<td>356.6593</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>1671.5057</td>
<td>1671.5057</td>
<td>1673.8917</td>
<td>1673.9266</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.5933</td>
<td>0.5933</td>
<td>0.5895</td>
<td>0.5948</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.5506</td>
<td>0.5506</td>
<td>0.5406</td>
<td>0.5426</td>
</tr>
<tr>
<td></td>
<td>Neighbours</td>
<td>32</td>
<td>27</td>
<td>30777.37</td>
<td>21021.36</td>
</tr>
<tr>
<td>1988</td>
<td>ResidualSquares</td>
<td>5694144.0</td>
<td>5233680.9</td>
<td>6224683.5</td>
<td>4734358.9</td>
</tr>
<tr>
<td></td>
<td>EffectiveNumber</td>
<td>26.45482689</td>
<td>31.0222</td>
<td>24.4740</td>
<td>41.0187</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>255.0340212</td>
<td>251.1438</td>
<td>263.6842</td>
<td>254.6976</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>1610.105276</td>
<td>1611.9063</td>
<td>1613.4679</td>
<td>1625.7858</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.76450822</td>
<td>0.7836</td>
<td>0.7426</td>
<td>0.8042</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.696036113</td>
<td>0.7052</td>
<td>0.6751</td>
<td>0.6968</td>
</tr>
</tbody>
</table>
1988 data
1998 data
2006 data

Figure 5.11 Coefficient maps of the GWR model (TP vs. Th & Elev) calibration
5.3.1.1 GWR model (TP vs. Total Fe)

For 1988 and 2006 data, all GWR models are found to perform better than the corresponding OLS model based on AICc and $R^2$. Among the GWR sub-models, CV models (Adaptive CV and Fixed CV) are better than AICc models. The two CV models are very close in both AICc and $R^2$, but the Fixed CV has much lower SSR. So Fixed CV is the better model for TP estimation (Table 5.6). For 1998 data, the Fixed AICc model performed the best. Total Fe has positive correlation with TP in all three data sets and the correlations increase from east to west (Figure 5.12).

Table 5.6 Diagnosis of the GWR calibration model (TP vs. Fe)

<table>
<thead>
<tr>
<th>Year</th>
<th>Neighbours</th>
<th>ResidualSquares</th>
<th>Sigma</th>
<th>AICc</th>
<th>$R^2$</th>
<th>$R^2_{Adjusted}$</th>
<th>Adaptive CV</th>
<th>Adaptive AIC</th>
<th>Fixed CV</th>
<th>Fixed AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td>38</td>
<td>25048.4583</td>
<td>28060.8468</td>
</tr>
<tr>
<td></td>
<td>Neighbours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>32</td>
<td>38</td>
<td>1474545.9874</td>
<td>1370525.42</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1559401.91</td>
<td>1525901.91</td>
<td>127.1493</td>
<td>125.8744</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1442.3589</td>
<td>1442.0453</td>
<td>1445.8776</td>
<td>1445.1828</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9322</td>
<td>0.9283</td>
<td>0.9370</td>
<td>0.9320</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9178</td>
<td>0.9160</td>
<td>0.9176</td>
<td>0.9155</td>
</tr>
<tr>
<td></td>
<td>$R^2_{Adjusted}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9178</td>
<td>0.9160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>114</td>
<td>46</td>
<td>2583312.66</td>
<td>38985.12</td>
</tr>
<tr>
<td></td>
<td>Neighbours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>114</td>
<td>46</td>
<td>5721282.0780</td>
<td>4772089.89</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4772089.89</td>
<td>4772089.89</td>
<td>228.3922</td>
<td>228.5793</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>228.3922</td>
<td>228.5793</td>
<td>1566.6582</td>
<td>1566.1879</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8180</td>
<td>0.8482</td>
<td>0.8138</td>
<td>0.8487</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8124</td>
<td>0.8270</td>
<td>0.8121</td>
<td>0.8296</td>
</tr>
<tr>
<td></td>
<td>$R^2_{Adjusted}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8124</td>
<td>0.8270</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>38</td>
<td>21433.62</td>
<td>24336.93</td>
</tr>
<tr>
<td></td>
<td>Neighbours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13</td>
<td>38</td>
<td>2256156.71</td>
<td>4669867.28</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2256156.71</td>
<td>4669867.28</td>
<td>189.9278</td>
<td>205.7668</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>220.3011</td>
<td>205.7668</td>
<td>1567.6434</td>
<td>1564.5454</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1582.8175</td>
<td>1567.6434</td>
<td>1564.5454</td>
<td>1563.3411</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.9067</td>
<td>0.8069</td>
<td>0.8593</td>
<td>0.8424</td>
</tr>
<tr>
<td></td>
<td>$R^2_{Adjusted}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8314</td>
<td>0.7732</td>
<td>0.8021</td>
<td>0.7936</td>
</tr>
</tbody>
</table>
5.3.2 GWR model validation and error analysis

5.3.1.2 GWR model (TP vs. Mud thickness and Elevation)
GWR can predict TP values at un-sampled locations. The fixed CV GWR model produced the best calibration results. Using elevation and mud thickness as independent variables, TP values were estimated on a 500 ft by 500 ft grid. The predicted values at the validation sites were compared to observed data. The residual errors (differences between the observed and predicted values) for the 2006 validation data were normally distributed within a narrow region (e and f of Figure 5.13). Both residual errors of 1998 and 1988 were skewed with some large positive errors (under-estimates) in the 1988 data (a and b of Figure 5.13). The quantile-comparison plots demonstrated weak fits for the measured and estimated values (Figure 5.14).
(a) 

(b) 

1988
Figure 5.13 Validation error distributions of the GWR model (TP vs. Th & Elev)
Figure 5.14 Scatter plots of the GWR model (TP vs. Th & Elev) validation

(a) $y = 0.6344x + 248.35$
$R^2 = 0.6903$

(b) $y = 0.6489x + 239.85$
$R^2 = 0.6698$

(c) $y = 0.6137x + 254.95$
$R^2 = 0.6918$
5.4.2.2 GWR model (TP vs. Fe) validation

The coefficient values and intercepts change over the study area, and both values are needed to predict TP. The validation results were calculated through a series of GIS processes. All the residual errors are skewed distributions with higher positive errors (under-estimates) for 2006 (e and f of Figure 5.15) and higher negative errors (over-estimated) for both 1988 (a and b of Figure 5.15) and 1998 data sets (c and d of Figure 5.15). Modeled vs. predicted comparisons of the 2006 validation data were very good, with $R^2$ of 0.88 (c of Figure 5.16). Once the outliers were removed from the 1988 and 1998 validation data sets, the comparisons of the modeled vs. predicted were good as well with $R^2$ of 0.87 and 0.86 respectively (b and c of Figure 5.16).
Figure 5.15 Validation error distributions of the GWR model (TP vs. Fe)
Figure 5.16 Scatter plots of the GWR model (TP vs. Fe) validation

(a) \[ y = 0.8519x + 75.282 \]
    \[ R^2 = 0.8682 \]

(b) \[ y = 0.9266x + 34.596 \]
    \[ R^2 = 0.8512 \]

(c) \[ y = 0.9504x + 82.522 \]
    \[ R^2 = 0.887 \]
5.4 Kriging and Co-kriging Model Analysis

5.4.1 Ordinary kriging model calibration and validation

Calibration and validation are performed together using Ordinary Kriging of ArcGIS 9.3.1. The validation errors formed near normal distributions for all three data sets (a, c and e of Figure 5.16), and some sites were over-estimated with large negative errors for 2006 and 1988 validation data in the normal quantile charts (c and a of Figure 5.16, respectively). The predicted values matched the observed values quite well after removing the under-, and over-estimated sites (Figure 5.18).
Figure 5.17 Validation errors distribution of the Ordinary Kriging
Figure 5.18 Scatter plots of the Ordinary Kriging validation
5.4.2 Co-kriging calibration and validation

Followed procedures similar to Ordinary Kriging, the validation results can be determined for Co-kriging using Fe as co-variable. The error distributions and patterns are very similar to the Ordinary Kriging methods with $R^2$ of 0.64–0.68 after the removal of several outliers (Figures 5.19-5.20, Table 5.7).
Figure 5.19 Validation error distributions of the Co-kriging (TP vs. Fe)
Figure 5.20 Scatter plots of the Co-kriging (TP vs. Fe) validation
Table 5.7 Calibration and validation results of the OK and CK models

<table>
<thead>
<tr>
<th>Methods</th>
<th>Data Year</th>
<th>2006</th>
<th>1998</th>
<th>1988</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OK</td>
<td>CK:Fe</td>
<td>OK</td>
<td>CK:Fe</td>
</tr>
<tr>
<td>Calibration</td>
<td>Mean</td>
<td>-0.65</td>
<td>-1.53</td>
<td>6.07</td>
</tr>
<tr>
<td></td>
<td>Root-Mean-Square</td>
<td>282.70</td>
<td>238.20</td>
<td>403.90</td>
</tr>
<tr>
<td></td>
<td>Average Standard Error</td>
<td>274.80</td>
<td>267.20</td>
<td>401.80</td>
</tr>
<tr>
<td></td>
<td>Mean Standardized</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Root-Mean-Square Std</td>
<td>1.02</td>
<td>0.89</td>
<td>1.00</td>
</tr>
<tr>
<td>Validation</td>
<td>Mean</td>
<td>-27.38</td>
<td>-27.54</td>
<td>-89.46</td>
</tr>
<tr>
<td></td>
<td>Root-Mean-Square</td>
<td>331.00</td>
<td>325.40</td>
<td>415.10</td>
</tr>
<tr>
<td></td>
<td>Average Standard Error</td>
<td>274.40</td>
<td>272.70</td>
<td>400.10</td>
</tr>
<tr>
<td></td>
<td>Mean Standardized</td>
<td>-0.10</td>
<td>-0.10</td>
<td>-0.23</td>
</tr>
<tr>
<td></td>
<td>Root-Mean-Square Std</td>
<td>1.21</td>
<td>1.20</td>
<td>1.04</td>
</tr>
</tbody>
</table>

5.5 Regression-Kriging (R-K) Model Analysis

5.5.1 OLS model (TP vs. Th and Elev) calibration

Following the proposed procedure in section 2.5, chapter 2, the “trend” is first calculated using the OLS regression equation (6): “trend” = Intercept + Coefficient_1* Elev (Grid) + Coefficient_2*Th (Grid) using Raster Calculator or Map Algebra (Figure 5.20a). Then the residuals of the regression equation were kriged (Figure 5.10b), and finally sum the “trend” and the kriged residuals (Figure 5.21b), and the result is a raster data (500 ft cell). Figure 5.22 is the model result for 2006 TP data.

5.5.2 Model validation and error analysis

First export the prediction raster from the calibrated R-K model as points data set, and then perform spatial Join (Match option: closest) with the validation data set to retrieve the values for each validation point (Figure 5.21b).
Figure 5.21 ModelBuilder model of the R-K model calibration and validation
Figure 5.22 R-K model calibration result of the 2006 TP data

The errors show a skewed distribution for 1988 validation data (a of Figure 5.2) and normal distributions for 1998 and 2006 data (c and e of Figure 5.23). Most points are located in narrow zones in the norm quantiles charts (b, d and f of Figure 5.23). The scatter plots indicate good fits between the observed and predicted values at the validation data sites after removal of the outliers (Figure 5.24).
Figure 5.23 Validation error distributions of the R-K model
Figure 5.24 Scatter plots of the R-K model validation (TP vs. Th & Elev)
5.6 Optimal Models and Weight Calculations

5.6.1 Error analysis and model comparison

For all models evaluated, the GWR models (TP vs. Fe and TP vs. Th & Elev) have lower RMSE and lowest mean errors for TP 2006 and TP 1988 data sets (Table 5.8). For 1998 TP data, OK and CK have the lowest mean errors (ME) but higher RMSE errors than OLS and GWR models. The best models are highlighted in yellow.

Table 5.8 TP model uncertainty and comparison

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>-27.38</td>
<td>-27.54</td>
<td>31.19</td>
<td>30.19</td>
<td>40.21</td>
<td>20.82</td>
<td>65.15</td>
</tr>
<tr>
<td>RMSE</td>
<td>331</td>
<td>325.4</td>
<td>169.96</td>
<td>204.61</td>
<td>325.23</td>
<td>307.6</td>
<td>325.1</td>
</tr>
<tr>
<td>NRMSE</td>
<td>26.09</td>
<td>25.65</td>
<td>13.4</td>
<td>16.13</td>
<td>25.64</td>
<td>24.25</td>
<td>25.63</td>
</tr>
<tr>
<td>1998</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>-89.46</td>
<td>-90.64</td>
<td>136.93</td>
<td>132.81</td>
<td>106.64</td>
<td>100.58</td>
<td>126.48</td>
</tr>
<tr>
<td>RMSE</td>
<td>415.1</td>
<td>415.8</td>
<td>374.75</td>
<td>359.83</td>
<td>442.98</td>
<td>389.38</td>
<td>419.14</td>
</tr>
<tr>
<td>NRMSE</td>
<td>23.18</td>
<td>23.22</td>
<td>20.92</td>
<td>20.09</td>
<td>24.73</td>
<td>21.74</td>
<td>23.4</td>
</tr>
<tr>
<td>1988</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Error</td>
<td>-41.74</td>
<td>-45.27</td>
<td>110.94</td>
<td>109.59</td>
<td>64.15</td>
<td>24.39</td>
<td>51.64</td>
</tr>
<tr>
<td>RMSE</td>
<td>402.1</td>
<td>407.4</td>
<td>368.36</td>
<td>375.88</td>
<td>371.97</td>
<td>234.34</td>
<td>371.99</td>
</tr>
<tr>
<td>NRMSE</td>
<td>0.24</td>
<td>0.24</td>
<td>0.22</td>
<td>0.23</td>
<td>0.22</td>
<td>0.14</td>
<td>0.22</td>
</tr>
</tbody>
</table>


GWR models performed the best for most of the data since there are strong positive correlation between TP and total Fe, and moderate positive correlation between TP and mud thickness and site elevation. GWR models also use smaller search windows and only local samples used for estimation, which make it more effective than global OLS models. OK and CK models utilize only the weak spatial autocorrelation among TP data and lead to highest RMSE; their smaller mean errors are mainly due to smaller search widows used in variogram fitting. Regression-Kriging models utilize both weak spatial
autocorrelation among TP values and positive correlation between TP and mud thickness, and they performed better than OK and CK models.

5.6.2 TP Concentration and weight distribution for selected models

Based on model performance comparison and validation errors, the GWR (Fe) and GWR (Thickness and Elevation) models were selected for TP concentration prediction and weight calculation. The OK model also is used as a benchmark for comparison purposes. TP has weak spatial auto-correlation in the lake sediments and TP concentrations were first estimated using OK models (a, c and e of Figure 5.25). Then the TP weights (mass) were calculated using GIS methods and the mud weight data calculated in Chapter 4.

For the GWR (TP vs. Fe) model, total Fe was estimated using OK first. Total Fe has strong trends in N/S direction and weaker trends in E/W directions (a, c, and e of Figure 5.26). TP concentrations (a, c and e of Figure 5.27) and weights were calculated (b, d and f of Figure 5.27) using equation (7) (Chapter 2) implemented in ArcGIS.

For the GWR (TP vs. Th + Elev) model, the mud thickness raster created in Chapter 4 and the lake bottom elevation data created in Chapter 3 were used to calculate the TP concentrations (a, c, e of Figure 5.28) and weights (b, d and f of Figure 5.28).
Figure 5.25 TP concentrations (a, c and e) and weights (b, d and f) estimated using OK
Figure 5.26 Total Fe trend analyses (a, c and e) and estimation (b, d and f) using OK
Figure 5.27 TP concentrations (a, c and e) and weights (b, d and f) estimated using GWR model (TP vs. Fe)
Figure 5.28 TP concentrations (a, c and e) and weights (b, d and f) estimated using GWR model (TP vs. Th & Elev)
5.6.3 TP weights and their changes

TP weights were calculated for the three data sets using three selected models, OK, GWR (TP vs. Fe) and GWR (TP vs. Th & Elev; Table 5.9). The results showed the same trend for the TP weights depending on the model: increased from 1988 to 1998 by 38 to 44%; decreased from 1998 to 2006 by 30% to 34%. There are small declines in TP weights between 1988 and 2006: 1.1E+06 kg or 2% for the OK model, 1.0E+06 kg or 2% for the GWR (Fe) model, and 4.2E+06 kg or 10% for the GWR (Th + Elev) model.

<table>
<thead>
<tr>
<th>Year</th>
<th>OK Model Weight</th>
<th>OK Model Change</th>
<th>OK Model Change%</th>
<th>GWR (Fe) Model Weight</th>
<th>GWR (Fe) Model Change</th>
<th>GWR (Fe) Model Change%</th>
<th>GWR (Th + Elev) Model Weight</th>
<th>GWR (Th + Elev) Model Change</th>
<th>GWR (Th + Elev) Model Change%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988</td>
<td>4.25E+07</td>
<td></td>
<td></td>
<td>4.20E+07</td>
<td></td>
<td></td>
<td>4.43E+07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998</td>
<td>5.89E+07</td>
<td>1.64E+07</td>
<td>38.59%</td>
<td>6.04E+07</td>
<td>1.84E+07</td>
<td>43.68%</td>
<td>6.11E+07</td>
<td>-1.94E+07</td>
<td>-32.16%</td>
</tr>
<tr>
<td>2006</td>
<td>4.14E+07</td>
<td>-1.75E+07</td>
<td>-29.72%</td>
<td>4.10E+07</td>
<td>-1.94E+07</td>
<td>-32.16%</td>
<td>4.01E+07</td>
<td>-2.10E+07</td>
<td>-34.42%</td>
</tr>
</tbody>
</table>

5.7 Summary of Results

Three years of sediment TP concentration data from Lake Okeechobee were analyzed using a variety of spatial models in this Chapter. TP has the highest concentrations in the mud zone and demonstrated first order of trend change in E/W and second order in S/N directions. TP also has strong positive correlation with total Fe and moderate positive correlation with mud thickness. From 1988 to 1998, TP values declined from 670.5 mg/kg to 650.20 mg/kg, just about 20 mg/kg decrease in average (Table 5.1). Different changes occurred in different parts of the lake during this period: TP declined in the
northern and southern areas, and increased in the central-western part of the lake (Figure 5.5a). From 1998 to 2006, TP values showed an average decrease of 84 mg/kg, and the decrease mainly occurred in the lake’s western and southern regions (Figure 5.4 and Figure 5.5b). Some areas experienced increase, such as the northern and western edges, with maximum increase up to 1,000–3,000 mg/kg.

Based on TP’s spatial auto-correlations and correlations with other variables, OLS, GWR, OK and CK, and Regression-Kriging models were applied for the studies of the three data sets. These models were calibrated and validated using separate data sets. All models were compared based on model performance and error analysis. Based on model accuracy (RMSE), the best models were GWR (Total Fe) and GWR (mud thickness and site elevation) since these models use both spatial auto-correlation and positive correlation between TP and independent variables. OK/CK models utilized only the weak spatial autocorrelation in the data and produced higher RMSE. Regression-Kriging models performed better than OK/CK models since positive correlation with mud elevation was considered in the estimation (Table 5.8). GWR (Fe), GWR (Thickness and Elevation) and OK models were selected for final TP concentration calculations.

TP weights were calculated using the mud weight and TP concentration for the three data sets separately. All these processes were implemented in ArcGIS using ModelBuilder. The TP weight increased about 37.99%–43.68% from 1988 to 1998 and decreased about 29.72%–34.42% from 1998 to 2006.
CHAPTER 6 SPATIAL MODELING OF TOTAL NITROGEN

6.1 Data Description and Characteristics

TN of 1988 varies from 166-36,107 mg/kg, with median and mean of 8,671 and 9,094 mg/kg, respectively. TN of 1998 changes from near zero to 47,800 mg/kg with median and mean of 8,000 and 9,363 mg/kg, respectively. TN of 2006 ranges from 70 to 41,030 mg/kg, and the median and mean are 8,170 and 8,926 mg/kg, respectively (Table 5.1). Total nitrogen data show skewed distribution in all three data sets (Figure 6.1). Most sites have TN range of 100-15,000 (mg/kg), and some sites can reach up to 36,000 to 50,000 mg/kg. The sites with higher concentration were located in the peat zone, next is the sand zones (Figure 6.1, Table 5.1).

Total nitrogen has strong positive correlation with total carbon, positive correlation with Fe and site elevation, and negative correlation with Ca (Figure 5.3). A 2nd order of trend changes appeared in the N/S directions for 2006 and 1988 TN values (Figure 6.2). On maps created using Ordinary Kriging (Yan and James 2007), total nitrogen was more homogenous in 2006 than in 1998 and declined from an average of 9,550 to 9,056 mg/kg, especially in the central and southern regions (Figure 6.3 and Figure 6.4). The biggest increases occurred in northern lake and southern edge areas. From 1988 to 1998, total nitrogen decreased in most areas, especially in northern and southern lake regions; Western littoral zone had the biggest increase, up to 47,800 mg/kg. These changes could be attributed to remixing and redistribution of sediments throughout the lake.
1998

(c)

(d)
Figure 6.1 TN histograms (a, c and e) and boxplots by sediment zones (b, d and f)
Figure 6.2 Trend analyses of TN concentrations
Figure 6.3 TN maps for 1988 (a), 1998 (b) and 2006 (c) data using OK
6.2 Regression Model Analysis

6.2.1 Key variable selection and regression model calibration

2006 TN data were subset into training and validation groups by 75-25% split using ArcGIS GA tool. Two statistically significant regression models were identified (Table 6.1). With Fe, Ca, TC and elevation values as independent variables, this regression model can explain up to 96% of the variation for 2006 TN, 95% for 1998 TN, and 93% for 1988 TN, respectively. All these variables are statistically significant from t-statics (Table 6.2). GWR models can be used to explore the relationship between TN and correlated variables and their relationship changes in space. From the coefficient raster created by GWR method, the relationship between TN and these variables were changing
### Table 6.1 Statistically significant regression models for TN data

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>2090.19</td>
<td>2027.1057</td>
<td>2142.1652</td>
</tr>
<tr>
<td>R²</td>
<td>0.9315</td>
<td>0.9620</td>
<td>0.9037</td>
</tr>
<tr>
<td>AdjR²</td>
<td>0.9309</td>
<td>0.9609</td>
<td>0.9028</td>
</tr>
<tr>
<td>F-Stat</td>
<td>1522.27</td>
<td>927.0777</td>
<td>1050.7152</td>
</tr>
<tr>
<td>F-Prob</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>JB</td>
<td>135.94</td>
<td>903.2436</td>
<td>147.1549</td>
</tr>
<tr>
<td>JB-Prob</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Sigma²</td>
<td>5281819</td>
<td>2985548</td>
<td>8332364</td>
</tr>
<tr>
<td>neighbors</td>
<td>14728.81</td>
<td>14728</td>
<td>13665</td>
</tr>
<tr>
<td>Residuals</td>
<td>Clustered</td>
<td>Clustered</td>
<td>Clustered</td>
</tr>
<tr>
<td>Stationary</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

### Table 6.2 Calibration result of the OLS model of 2006 TN

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>StdError</th>
<th>t-Statistic</th>
<th>Probability</th>
<th>Robust SE</th>
<th>Robust t</th>
<th>Robust Pr</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-821.0381</td>
<td>418.0021</td>
<td>-1.9642</td>
<td>0.052029</td>
<td>376.956466</td>
<td>-2.178071</td>
<td>0.031529*</td>
<td>---</td>
</tr>
<tr>
<td>TC</td>
<td>63.3496</td>
<td>1.5414</td>
<td>41.0995</td>
<td>0.000000*</td>
<td>1.882007</td>
<td>33.660641</td>
<td>0.000000*</td>
<td>1.1705</td>
</tr>
<tr>
<td>Ca</td>
<td>-0.0174</td>
<td>0.0029</td>
<td>-6.0779</td>
<td>0.000000*</td>
<td>0.003682</td>
<td>-4.735599</td>
<td>0.000008*</td>
<td>1.2447</td>
</tr>
<tr>
<td>Fe</td>
<td>0.4814</td>
<td>0.0710</td>
<td>6.7771</td>
<td>0.000000*</td>
<td>0.092908</td>
<td>5.181811</td>
<td>0.000001*</td>
<td>1.3501</td>
</tr>
<tr>
<td>Elevation</td>
<td>180.7021</td>
<td>52.0059</td>
<td>3.4747</td>
<td>0.000741*</td>
<td>60.071585</td>
<td>3.008113</td>
<td>0.003262*</td>
<td>1.4131</td>
</tr>
</tbody>
</table>

### Table 6.3 Diagnosis of the GWR calibration model of 2006 TN

<table>
<thead>
<tr>
<th>Neighbours</th>
<th>Adaptive AIC</th>
<th>Adaptive CV</th>
<th>Fixed AIC</th>
<th>Fixed CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>79</td>
<td>37609.6335</td>
<td>53446.70122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>288372446</td>
<td>282223936</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1714.7092</td>
<td>1631.2917</td>
<td>1706.5739</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2054.8016</td>
<td>2051.4147</td>
<td>2052.9035</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Global Moran's I Sum: Search distance: 11984.78 ft

| MorD Index | 0.022317     | 0.022468    | -0.004982  |          |
| Expected Index | -0.008772 | -0.008772  | -0.008772 |          |
| Variance   | 0.007628    | 0.007583    | 0.007619   |          |
| Z Score    | 0.355968    | 0.357681    | -0.500913  | 0.043418 |
| p-value    | 0.721864    | 0.720582    | 0.616432   | 0.965369 |
across the lake (Table 6.3 and Figure 6.5). Using TC as independent variable alone, the regression model can explain over 93% of variations for 2006 TN, 90% for 1998 TN and 91% for 1988 TN, respectively (Table 6.1). TC will be used as the only independent variable for calibration and validation of regression, Co-kriging and GWR models.
6.2.2 Regression Model Validation

OLS model assume a uniform global variation in the study area. Once the calibration model is established, the estimated values and errors can be calculated for the validation sites. All the validation errors are near normally distributed (a, c and e of Figure 6.6). The scatter plot demonstrated excellent fitting between the measured and estimated values for the validation data sets (b, d and f of Figure 6.6) with $R^2$ of 0.88-0.94.
\[ y = 0.9173x + 1164.7 \]

\[ R^2 = 0.8755 \]
\[ y = 0.7743x + 1619.4 \]

\[ R^2 = 0.9194 \]
Figure 6.6 Error histograms (a, c and e) and scatter plots (b, d and f) of the OLS model (TN vs. TC) validation

$y = 1.0437x + 51.81$

$R^2 = 0.9396$
6.3 GWR Model Analysis

6.3.1 GWR model (TN vs. TC) calibration

TC was used as the independent variable for the GWR model calibration as the OLS model. For 2006 TN data, the two fixed models (AICc and CV) are very close and better than the two adaptive models with lower SSR, sigma, and higher $R^2$. Fixed CV model performs the best with lowest errors (SSR and sigma) and highest $R^2$ (Table 6.4).

Table 6.4 Diagnosis of the GWR model calibration (TN vs. TC)

<table>
<thead>
<tr>
<th>Data</th>
<th>GWR Models</th>
<th>AdaptiveAICc</th>
<th>Adaptive CV</th>
<th>Fixed AICc</th>
<th>Fixed CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>Neighbours</td>
<td>47</td>
<td>74</td>
<td>28256.67</td>
<td>27142.25</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td>469619336</td>
<td>522314480</td>
<td>395683424</td>
<td>385821059</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>2163.8366</td>
<td>2224.2531</td>
<td>2069.4622</td>
<td>2059.0357</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>2104.7141</td>
<td>2107.2116</td>
<td>2098.6763</td>
<td>2098.6776</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.9469</td>
<td>0.941</td>
<td>0.9553</td>
<td>0.9564</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.9397</td>
<td>0.9363</td>
<td>0.9448</td>
<td>0.9454</td>
</tr>
<tr>
<td>1998</td>
<td>Neighbours</td>
<td>38</td>
<td>28</td>
<td>32734.68</td>
<td>30195.89</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td>533073193</td>
<td>482474162</td>
<td>552433293</td>
<td>532323639</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>2353.9782</td>
<td>2311.4882</td>
<td>2399.3373</td>
<td>2384.2005</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>2108.1714</td>
<td>2109.3908</td>
<td>2110.8526</td>
<td>2111.0464</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.9569</td>
<td>0.9610</td>
<td>0.9553</td>
<td>0.9570</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.9494</td>
<td>0.9512</td>
<td>0.9474</td>
<td>0.9481</td>
</tr>
<tr>
<td>1988</td>
<td>Neighbours</td>
<td>32</td>
<td>16</td>
<td>15793.65803</td>
<td>15793.65803</td>
</tr>
<tr>
<td></td>
<td>ResidualSquares</td>
<td>4.33E+08</td>
<td>2.82E+08</td>
<td>1.66E+08</td>
<td>1.66E+08</td>
</tr>
<tr>
<td></td>
<td>Sigma</td>
<td>2155.793312</td>
<td>1967.990881</td>
<td>1643.679739</td>
<td>1643.679739</td>
</tr>
<tr>
<td></td>
<td>AICc</td>
<td>2090.670946</td>
<td>2094.643198</td>
<td>2065.260311</td>
<td>2065.260311</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.949719696</td>
<td>0.967245621</td>
<td>0.980695358</td>
<td>0.980695358</td>
</tr>
<tr>
<td></td>
<td>$R^2$Adjusted</td>
<td>0.938983301</td>
<td>0.949151209</td>
<td>0.964529372</td>
<td>0.964529372</td>
</tr>
</tbody>
</table>
For the 1998 TN data, the best model is the adaptive CV with lowest errors and highest $R^2$ (Table 6.4). For the 1988 TN data, the fixed CV sub-model performed the best with lowest errors and highest $R^2$ (Table 6.4). Figure 6.7 shows the model coefficients and their change across the lake area.

### 6.3.2 GWR model (TN vs. TC) validation

A series of processes were performed using ArcGIS tools, and the model errors were calculated and plotted in Figure 6.8 (a, c and e). The error distributions of the GWR model are close to normal for 1988 and 2006 TN data, and sort of skewed for the 1998 TN data. The fittings between the observed and estimated values for the validation data are excellent with $R^2$ of 0.8938-0.9763 (b, d and f of Figure 6.8), which are better than the corresponding OLS models (Figure 6.6).
$y = 0.9503x + 298.95$

$R^2 = 0.8938$
\[ y = 0.995x - 107.03 \]

\[ R^2 = 0.9763 \]
Figure 6.8 Error histograms (a, c and e) and scatter plots (b, d and f) of the GWR model (TN vs. TC) validation

\[ y = 1.0091x + 84.346 \]

\[ R^2 = 0.96 \]
6.4 Kriging and Co-kriging Model Analysis

6.4.1 Ordinary Kriging calibration and validation

Ordinary Kriging for TN data were performed using ArcGIS 9.3.1, and weak autocorrelations were identified in all three data sets. A variogram was fitted for each calibration data set (Figure 6.9). The model errors are near normally distributed (a, c and e of Figure 6.10), but the scatter plots show poor fitting between the measured and predicted values (b, d and f of Figure 6.10).

![Figure 6.9 Fitted variograms for TN calibration data](image)

1988 TN: $2.4197 \times 10^7 \times \text{Spherical}(50000) + 2.8097 \times 10^7 \times \text{Nugget}$

1998 TN: $1.8419 \times 10^7 \times \text{Spherical}(50000) + 7.2478 \times 10^7 \times \text{Nugget}$

2006 TN: $1.5358 \times 10^7 \times \text{Spherical}(50000) + 4.3412 \times 10^7 \times \text{Nugget}$
1988

(a) Frequency distribution of TNBBVal_OKS error.

(b) Scatter plot showing the relationship between Predicted and Measured values with the equation:
\[ y = 0.6401x + 4501.1 \]
and the coefficient of determination:
\[ R^2 = 0.6544 \]
$y = 0.3936x + 5300$

$R^2 = 0.6537$

1998
Figure 6.10 Error histograms (a, c and e) and scatter plots (b, d and f) for OK validation
6.4.2 Co-kriging model calibration and validation

TC was used as co-variable for Co-kring of TN analysis. As the same as Ordinary Kriging, a variogram was fitted for each calibration data and then the errors were calculated using the estimated values and the measured values for the validation data. The errors are normally distributed (a, c and e of Figure 6.11) for all three data sets. The measured and estimated values matched moderately well at the validation sites with $R^2$ varies from 0.67-0.70 (b, d and f of Figure 6.11) after the removal of some outliers.
The data can be modeled by the equation $y = 0.6733x + 4262.7$, with a coefficient of determination $R^2 = 0.6664$. This indicates a moderate fit between the predicted values and the actual measured data.
$y = 0.4725x + 4276.4$

$R^2 = 0.6796$
Figure 6.11 Error histograms (a, c and e) and scatter plots (b, d and f) of CK (TN vs. TC) validation

\[ y = 0.5845x + 3879.7 \\ R^2 = 0.7019 \]
6.5 Optimal Models and Weight Calculation

6.5.1 Error analysis and model comparison

Optimal models can be identified based on the independent validation results. For the 2006 TN data, the GWR model performed the best with much lower mean errors and RMSE comparing to OK and CK. OLS is the second best model (Table 6.5). For the 1998 TN and 1988 TN data, the regression models have the lowest mean errors and the GWR models have the lowest RMSE errors. Both models are much better than the Ordinary Kriging and Co-kriging models, which are based on the weak spatial autocorrelations of the data.

Table 6.5 TN model uncertainty and comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Error</th>
<th>OK</th>
<th>CK with TC</th>
<th>OLS with TC</th>
<th>GWR with TC</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>ME</td>
<td>1505</td>
<td>1555</td>
<td>-415</td>
<td>-160</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>6028</td>
<td>6055</td>
<td>2071</td>
<td>1568</td>
</tr>
<tr>
<td>1998</td>
<td>ME</td>
<td>1351</td>
<td>1279</td>
<td>82</td>
<td>149</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>8850</td>
<td>9029</td>
<td>2946</td>
<td>1547</td>
</tr>
<tr>
<td>1988</td>
<td>ME</td>
<td>1124</td>
<td>1141</td>
<td>-405</td>
<td>157</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>7410</td>
<td>7242</td>
<td>2874</td>
<td>2655</td>
</tr>
</tbody>
</table>

6.5.2 TN concentration and weight estimation

Total carbon is the only independent variable for the GWR model. It has a second order trend in N/S direction for all three data sets (a, c and e of Figure 6.12). It has higher values in the southern edges of the lake, lower values in the near-shore regions and western littoral zones, and the central mud zones have values between for all three data sets (b, d and f of Figure 6.12). Then the GWR model was applied to estimate the model parameters and TN concentrations over-time. Figure 6.13 showed the TN concentration changes and TN weight distribution in the lake for 1988, 1998 and 2006, respectively. Ordinary Kriging was also used to estimate the TP concentrations and TN weight and their changes over-time (Figure 6.14).
Figure 6.12 TC trend analyses (a, c and e) and estimation (b, d and f) using OK
Figure 6.13 TN concentrations (a, c and e) and weights (b, d and f) using GWR model (TN vs. TC)
Figure 6.14 TN concentrations (a, c and e) and weights (b, d and f) using OK
6.5.3 TN weights and their changes over-time

Total TN weights were calculated for the two selected models for the three data sets (Table 6.6). Both models, OK and GWR (TC), demonstrated that the TN weights increased 15.02-16.20% from 1988 to 1998, and decreased 7.53-10.94% from 1998 to 2006. The differences between the two models are small, and vary from 0.31-3.51%.

Table 6.6 TN Weights (kg) and their changes for OK and GWR models

<table>
<thead>
<tr>
<th>Year</th>
<th>OK Model</th>
<th>GWR (TC) Model</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Weight</td>
<td>Change</td>
<td>Change %</td>
</tr>
<tr>
<td>1988</td>
<td>4.12E+08</td>
<td></td>
<td>-2.46 %</td>
</tr>
<tr>
<td>1998</td>
<td>4.74E+08</td>
<td>6.19E+07</td>
<td>15.02 %</td>
</tr>
<tr>
<td>2006</td>
<td>4.38E+08</td>
<td>-3.57E+07</td>
<td>-7.53 %</td>
</tr>
</tbody>
</table>

6.6 Summary of Results

TN concentrations for the years of 1988, 1998 and 2006 were discussed in detail in this chapter. TN data have skewed distributions with highest values in peat zone located in the western littoral zone and southern lake edges. TN has weak spatial autocorrelation and strong positive correlation with total carbon for all three data sets.

From 1988 to 1998, TN decreased in most areas, especially in the northern and southern lake regions; western littoral zone had the biggest increase, up to 40,000 mg/kg. The large TN increase in the littoral zone may be due to accumulation of organic matter such as vegetation. From 1998 to 2006, TN declined from an average of 9,363 to 8,926 mg/kg, especially in the central and western littoral regions, which could be due to the disturbing
and redistribution of the top mud layer (in the central lake area) and peat (in the littoral zone) caused by hurricanes and other large storms occurred during 2004-2005. The biggest increases occurred in the northern lake and southern edge areas, which could be partially due to organic matter input from Kissimmee River (in the northern lake) and organic matter accumulation in the outflows in the southern lake areas.

Based on its spatial auto-correlations and strong correlations with total carbon, OLS, GWR, and OK and CK models were examined for the studies of the three data sets. All these models were calibrated first, then validated using separate data sets. All models were compared based on model performance and uncertainty analysis. The GWR model (TN vs. TC) performed the best, next is the OLS model (TN vs. TC) based on model accuracy (RMSE). OK and CK models have the lower model bias (mean errors) and higher RMSE (Table 6.5). GWR (TN vs. TC) and OK models were selected for final TN concentration calculations. TN weights were calculated using the mud weight and TN concentration for the three data sets separately. All these processes were implemented in ArcGIS using ModelBuilder. TN weights increased about 15.02-16.20% from 1988 to 1998, and decreased about 7.53-10.94% from 1998 to 2006 (Table 6.6).
CHAPTER 7 CONCLUSIONS AND RECOMMENDATIONS

The major objectives of this dissertation were to develop optimal spatial techniques to model the spatial-temporal variations of the lake sediments, to map and calculate the mud and nutrients weights and their changes from 1988 to 2006; especially to evaluate the potential impacts of the hurricanes during 1998-2006.

Major sediment zones and their changes were identified and mapped using the sediments survey data collected in 1988, 1998 and 2006, respectively. Mud thickness was mapped using calibrated Ordinary Kriging models and mud weights were calculated for each data using spatial models implemented in ArcGIS. The following are the major conclusions from this study:

- From 1988 to 1998, mud sediments were continually focused toward the center of the Lake, and mud thicknesses increased up to 26 cm in the central lake area and reduced near the shore zone. The mud area and volume increased up to 13.78% and 10.26%, respectively.

- From 1998 to 2006, the area of mud sediments increased slightly but the mud volume reduced about 27%. Mud depths declined by up to 41 cm in the central lake area and increased by up to 20 cm in the surrounding areas, with small amounts of mud being deposited throughout the rest of the lake.

- Mud weight increased up to 29.32% from 1988 to 1998, but reduced over 20% from 1998 to 2006. The reduction of mud sediments is likely due to re-suspension and redistribution by waves and currents produced by large storms during this period, particularly Hurricanes Frances and Jeanne in 2004 and Wilma in 2005.
Multi-linear regression, kriging, geographically weighted regression (GWR) and Regression-Kriging (R-K) models have been calibrated and validated for the spatial analysis of the sediment nutrients (TP and TN) of the lake. Sediment TP and TP have different spatial properties and are correlated with different environmental factors, the same model could perform different for TP and TN. Data exploration demonstrated that TP has weak spatial autocorrelation and strong positive correlation with total Fe, and moderate positive correlation with mud thickness. Model validation results suggested that the GWR (Fe) and GWR (Th & Elev) models provide the most accurate predictions. TN data also has weak spatial autocorrelation but strong positive correlation with total carbon, GWR model with total carbon (TC) as independent variable has the best performance.

The spatial and temporal changes of TP and TN concentrations were calculated using the selected models. From 1988 to 1998, total phosphorus (TP) values just decreased about 20 mg/kg from average, declined from 670.5 mg/kg to 650.20 mg/kg (Table 5.1). Different changes occurred in different part of the lake during this period: TP declined in the northern and southern areas, and increased in the central-western part of the lake (Figure 5.5a). From 1998 to 2006, TP values showed an average decrease of 84 mg/kg, the decease mainly occurred in the lake’s western and southern regions (Figure 5.4 and Figure 5.5b). Some area experienced increase, such as the northern and western edges, with maximum increase up to 1,000–3,000 mg/kg. TP weights were calculated using the mud weight and TP concentration for the three data sets separately and ModelBuilder
implemented in ArcGIS. The TP weight increased from 37.99% to 43.68% from 1988 to 1998 and decreased from 29.72% to 34.42% from 1998 to 2006.

From 1988 to 1998, TN decreased in most areas, especially in the northern and southern lake regions; western littoral zone had the biggest increase, up to 40,000 mg/kg. The large TN increase in the littoral zone may be due to accumulation of organic matter such as vegetations. From 1998 to 2006, TN declined from an average of 9,363 to 8,926 mg/kg, especially in the central and western littoral regions, which could be due to the disturbing and redistribution of the top mud layer (in the central lake area) and peat (in the littoral zone) caused by hurricanes occurred during 2004-2004. The biggest increases occurred in the northern lake and southern edge areas, which could be partially due to organic matter input from Kissimmee River (in the northern lake) and organic matter accumulation in the outflows in the southern lake areas. TN weights were calculated using a ModelBuilder model implemented in ArcGIS. TN weights increased about 15% - 16.2% from 1988 to 1998, and decreased about 7%-11% from 1998 to 2006.

Geographically weight regression is a newly developed technique and has great potential to be used widely in spatially related fields, including social science, ecological, environmental and soil sciences. This dissertation is one of the first applications for lake sediment research. The approaches and models developed in this study can be applied to other similar lakes and basins to track the spatial-temporal changes of the sediments and its nutrients. Since there are limited samples for the spatial analysis of TP and TN concentrations, more tests with more data points will improve this technique.
REFERENCES


VITA

YAO Y. YAN

EDUCATION:
• Doctoral Candidate in Civil and Environmental Engineering, FIU, Miami, FL, expected 2011
• M.S. in Geosciences (GIS), University of Houston (Main campus), Houston, TX
• M.S. in Geosciences, China University of Geosciences, Beijing
• B.E. in Geosciences, Hunan University of Science and Technology, Hunan

PROFESSIONAL EXPERIENCE:
• Geographer/Scientist, Sr. (2005-)
  Staff Geographer (2003-2005)
  Restoration Science Department, South Florida Water Management District
  3301 Gun Club RD, West Palm Beach, FL 33406
• Coordinator, GIS/Computer Applications (2002)
  GeoData Center, University of West Florida
  11000 University Parkway, Pensacola, FL 32514
• GIS/Remote Sensing Specialist (2000-2001)
  Johnson Control World Service, Inc., USGS National Wetlands Research Center,
  700 Cajundome Boulevard, Lafayette, LA 70506
• Sr. Research Fellow (Sr. Geologist) (1995-)
  Assistant Research Fellow (1990-1995)
  Tianjin Institute of Geology and Mineral Resources
  Chinese Academy of Geological Sciences, Tianjin, 300170, P. R. China

PUBLICATIONS/POSTERS (Selected):
• Yan, Y., Tom James, Fernando Miralles-Wilhelm and Fang Zhao. Spatial modeling of mud sediments and its changes over-time for the central mud zone (1988-2006), Lake Okeechobee, Florida. South Florida GIS Expo, 2010
• Yan, Y., et al., LiDAR and bathymetry data integration and 3-D visualization, Lake Okeechobee, Florida. South Florida GIS Expo, 2009
• Yan, Y. and John F. Casey, Geochemical Characteristics of Siqueiros Transform, East Pacific Rise. Geological Survey and Research (Special Issue), Vol. 29, No. 4, pages 279-293, 2006
• Ramsey E., III, Y. Yan, Palo Alto battlefield national historic site landscape classification and historical analysis. Proceedings 20th Biennial workshop on Aerial photography, Videography, and High Resolution Digital imagery for Resource Assessment. Weslaco, Texas. October 4-6, 2005


REWARDS:
- 1st Analytical Prize, South Florida GIS Expo, 2010. Yao Yan, Tom James, Fernando Miralles-Wilhelm and Fang Zhao. Spatial modeling of mud sediments and its changes over-time for the central mud zone (1988-2006), Lake Okeechobee, Florida
- 3rd Project Prize, South Florida GIS Expo, 2009. Yao Yan, et al., LiDAR and bathymetry data integration and 3-D visualization, Lake Okeechobee, Florida

MEMBERSHIP:
- American Society for Photogrammetry and Remote Sensing (ASPRS)
- American Association of Geographers