2016

Detecting the Spatial Patterns of Blue-green Algae in Harsha Lake using Landsat 8 Imagery

Jing Huang
Louisiana State University and Agricultural and Mechanical College, hjing0217@gmail.com

DETECTING THE SPATIAL PATTERNS OF BLUE-GREEN ALGAE IN HARSHA LAKE USING LANDSAT 8 IMAGERY

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Department of Geography and Anthropology

by
Jing Huang
B.S., Qufu Normal University, 2009
M.S., East China Normal University, 2012
August 2016

ACKNOWLEDGMENTS

First I would like to express my thanks to my advisor Dr. Lei Wang, who is very kind, patient, and knowledgeable, especially when I could not focus on my research and changed the topic of my thesis three times. He is not only my advisor in graduate study, but also my spiritual mentor who led me out of darkness and taught me how to get rid of negative emotion when bad things happened.

I would also like to thank my committee member Dr. Fahui Wang, who is always warm-hearted and quick to offer me and my family help, and let me know the importance of responsibility – I should be responsible for my study and work, be responsible for what I have said, and more importantly, be responsible for my family and my life.

Meanwhile, I am very thankful to my other committee member, Dr. Yijun Xu, who also helped me greatly in my thesis research. One of his courses, "watershed hydrology," offered me an opportunity to develop a detailed understanding of the concepts related to hydrology and freshwater ecology, which is the foundation of this thesis.

Special thanks to Luke Driskell for helping me proofread the draft very carefully, especially during the busy time when his little girl was just born. I also want to thank Dr. Hongxing Liu for providing me with the *in situ* data, and Mr. Joe Thompson, Mr. Frederick Fellner, and Mr. Michael Long for providing me graduate assistantships to support my master's research. Many thanks to my friends here at LSU, especially Yanqing Xu and Cong Fu, and for all the good memories with them.

I'd like to express my grateful thanks to my family for their infinite love, especially my father, who supported me to pursue my graduate study at LSU even though he knew he was not in good health, and my mother, who is helping me take care of my lovely son Benjamin so I can focus on my thesis. Finally, I would like to thank my husband Yujie for his companionship and encouragement these years; he has been the driving force keeping me going forward.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The incidence of harmful algal blooms (HABs) caused by blue-green algae has been increasing in coastal and freshwater ecosystems of the United States in recent years, and has had great influence on ecosystem, economic, and public health. This thesis aims at testing the feasibility of using machine learning methods in comparison to traditional regression models to detect and map the blue-green algae distribution in low-medium biomass waters (Chl-*a* < approx. 20 μg/L) from a Landsat 8 image with the support of some *in situ* Chl-*a* measurements in Harsha Lake, Ohio. Two algorithms were compared: one is the conventional empirical method – Stepwise Multiple Linear Regression – to see if there is a strong linear relationship between measured Chl-*a* concentrations and the Landsat 8 spectral data in the study area, and the other is one of the most popular machine learning methods–Random Forests. Major findings include: (1) both a conventional linear regression model and a Random Forests model worked well in mapping the extent and biomass of blue-green algae in Harsha Lake on September 21, 2015, but the Random Forests model outperformed the linear regression model; (2) the prediction surface from the Random Forests method illustrated that 89.30% of Harsha Lake's area had Chl-*a* values less than 10 μg/L on the sampling date, while only 10.70% of the entire study area had Chl-*a* concentrations between 10 μg/L and 20 μg/L. Higher Chl-*a* values (especially for Chl-*a* larger than 10 μg/L) were mostly distributed in the mouths of rivers or streams, which might be caused by the influx of nutrients from agricultural or urban land use by rivers and streams. The results show the utility of the Random Forests approach based on Landsat 8 imagery in detecting and quantitatively mapping low biomass HABs, which is considered to be a challenging task.

CHAPTER 1 INTRODUCTION

Harmful algal blooms (HABs) refer to the actual or potential harmful effects caused by the excessive growth of algae in water bodies (Shutler et al., 2012; Matthews et al., 2012). As for inland waters including lakes, ponds, and reservoirs, HABs are mostly made up of cyanobacteria (also called blue-green algae) and can cause harmful effects to: (1) freshwater ecosystems, such as pollution of beaches, taste and odor problems in drinking water, and depletion of oxygen levels causing fish kills (Braig IV et al., 2010); (2) the health of humans as well as other animals who use them for drinking or recreation (Matthews et al., 2010). That is because many blue-green algae species can produce toxins that affect the nervous system, liver, and skin (USACE, 2016), and hence cause illness, irritation, even death to humans, pets, and other animals (Braig IV et al., 2010). According to the World Health Organization (WHO) guidelines, chlorophyll *a* (Chl-*a*) concentrations between 10-50 µg/L represents a moderate human health risk from recreational contact caused by HABs (Braig IV et al., 2010), and the threshold may drop to between 10-25 µg/L for more toxic species (Matthews et al., 2012).

The incidence of HABs has been increasing in coastal and freshwater ecosystem of the United States in recent years (Lunetta et al., 2015). For example, Lake Erie in Ohio experienced a severe HABs event (three times greater than before) in 2011 due to unusually high runoff (Lunetta et al., 2015). HABs caused by blue-green algae have become a major water quality issue for inland waters in Ohio (Francy et al., 2015). HABs cost approximately $2.2 billion annually in the United States (Lunetta et al., 2015), and cost of treatment has been a burden on the already recessing economy in the past decade (Lunetta et al., 2015). Even though HABs have such a great influence on ecosystem, economic, and public health, HABs are generally assessed infrequently due to high cost and low efficiency of the ground-survey methods (Lunetta et al.,

2015). For example, the states only assessed 39% (14.8 million of 41.7 million acres) of inland waters in the United States according to the 2004 National Water Quality Inventory (Keith et al., 2012), and many of them only conducted event-based responses (Lunetta et al., 2015). That is because the typical water quality monitoring methods—visual assessment and point-scale water sampling—are time-consuming and costly (Randolph et al., 2008), and cannot accurately describe the spatial patchiness of HABs distribution in a water body (Hunter et al., 2008; Lunetta et al., 2015).

Therefore, satellite and airborne remote sensing can be regarded as the complementary approach to monitoring inland water quality, and are increasingly incorporated in the detection of freshwater HABs in recent years (Agha et al., 2012). Remote sensing images from hyperspectral or multispectral sensors can detect HABs below the water surface, or in turbid waters when the human eyes have difficulty interpreting the information (Kasich et al., 2015). Moreover, remote sensing images can be used to map the spatial patterns of HABs systemically over a very wide area and repeated in a short period (Jupp et al., 1994). For example, the State of Ohio first used satellite images to identify possible outbreaks of HABs in lakes, and then conducted ground sampling for phytoplankton and toxins in the suspicious areas (Kasich et al., 2015).

Research showed that hyperspectral data (such as AVIRIS, CASI, etc.) are very effective in recognizing and mapping freshwater HABs (Matthews et al., 2010). However, the use of hyperspectral data is limited with regard to the cost, availability, processing time and high dimensionality (Mutanga et al., 2011). A repeatable and cost-effective alternative is to use multispectral satellite imagery that are free to the public, such as MERIS (Wynne et al., 2008; Wynne et al., 2010; Matthews et al., 2012; Matthews and Odermatt, 2015; Lunetta et al., 2015), MODIS (Hu, 2009; Becker et al., 2009), and Landsat (Vincent et al., 2004; Tebbs et al., 2013).

MERIS and MODIS have better spectral resolution to identify the HABs, but they are not suitable for monitoring small inland water bodies because of their coarse spatial resolution (Hunter et al., 2008). To be more specific, the spatial resolution of MERIS is 300 m (Hunter et al., 2008) and that of MODIS medium-resolution data is 250 m and 500 m (Hu, 2009). Moreover, MERIS stopped working in April 2012 (Matthews and Bernard, 2015). Therefore, monitoring freshwater HABs often involves satellite images with higher spatial resolutions designed primarily for land applications, such as the Landsat TM and ETM+ images with 30 meters spatial resolution (Palmer et al., 2015). Many researchers have used Landsat images to study HABs (Vincent et al., 2004; Tebbs et al., 2013; Palmer et al., 2015), and illustrates that simple empirical algorithms (multiple regression models) using Landsat images and *in situ* measurements are effective approaches to mapping HABs biomass, particularly for small water bodies with severe HABs and for regions where data are limited to multispectral sensors (Tebbs et al., 2013). However, the remote sensing images and the *in situ* measurements of blue-green algae might not fit the simple linear regression models well when a low-medium biomass HABs event (Chl-*a* < approx. 20 μg/L) occurs (Matthews et al., 2012). At this point, machine learning methods can be employed as an alternative to building nonlinear prediction models to monitor and map the spatial distribution of HABs even at low biomass. And it is worthwhile to test whether the newly added bands on the Landsat 8 Operational Land Imager (OLI) sensor could provide better support for mapping HABs.

This thesis aims at testing the feasibility of using machine learning methods in comparison to traditional regression models by mapping the blue-green algae distribution in low-medium biomass waters (Chl-*a* < approx. 20 μg/L) (Matthews et al., 2012) from a Landsat 8 image with the support of some *in situ* Chl-*a* measurements in Harsha Lake, Ohio. Two

algorithms were compared; one is the conventional empirical method—Stepwise Multiple Linear Regression—to see if there is a strong linear relationship between measured Chl-*a* concentrations and the Landsat 8 spectral data in the study area. The other is one of the most popular machine learning methods –Random Forests.

There are five chapters in this thesis: Chapter 1 introduces the background and objectives of this study; Chapter 2 conducts a literature review of existing studies related to the detection and mapping of freshwater blue-green algae; Chapter 3 describes the study area, datasets and methods used in this thesis; Chapter 4 analyzes and discusses the prediction results of blue-green algae biomass by different methods (one is the best fitting Stepwise Multiple Linear Regression, and the other is the best fitting Random Forests Regression) along with the prediction surface generated by the best prediction model; Chapter 5 presents major findings.

CHAPTER 2 LITERATURE REVIEW

Remote sensing has been increasingly used for monitoring and mapping HABs in aquatic systems. The most commonly employed bioindicators to detect blue-green algal biomass include the concentration of chlorophyll a (Chl-*a*, a pigment of all photosynthetic phytoplankton) and phycocyanin (PC, the accessory pigment unique to blue-green algae) (Matthews et al., 2010). However, phycocyanin related spectral features are not detectable at small concentrations of blue-green algae (Kuster, 2009), and only clearly visible at high-biomass conditions (Chl-*a* concentrations greater than 20 μg/L) (Matthews and Odermatt, 2015).This may limit the accurate mapping of HABs in early warning systems (Kuster, 2009), and therefore the concentration of Chl-*a* is a better bioindicator of HABs at low biomass. There are several classical methods to map freshwater HABs in the literature.

2.1 Empirical approach

2.1.1 Band or Band ratio chlorophyll-retrieval algorithms

This kind of spectral algorithm is often developed by applying a statistical method (mainly Multiple Linear Regression) to show the relationship between experimental datasets (such as measured Chl-*a* concentrations and other water quality parameters) and the spectral values (or ratios) from remote sensing (Randolph et al., 2008). For example, Tebbs et al. (2013) developed a linear algorithm for mapping Chl-*a* in Lake Bogoria, Kenya, based on a time series of Landsat ETM+ images and monthly *in situ* measurement of Chl-*a* for the period from November 2003 to February 2005. They found that both the single near infrared band (835um) and the band ratio of near infrared (835um) with the red band (660um) had strong linear relationships with high biomass HABs (Chl-*a* concentrations up to 800 μg/L) (Tebbs et al., 2013).

Generally, in clear (oligotrophic) water conditions, this kind of spectral algorithm works well and yields accurate estimates of Chl-*a* concentrations using spectral characteristics in blue to green spectrums (Moses et al., 2009; Kuster, 2009).

As for the productive and turbid coastal and inland waters, where the dissolved organic matter and non-algal particles can cause absorptions in blue spectral region (Darecki and Stramski, 2004; Moses et al., 2009), spectral algorithms based on red and near infrared bands are preferable for estimating Chl-*a* concentrations and mapping HABs (Gower et al., 1999; Moses et al., 2009). What's more, a ratio of the red edge band to the red band can always enhance the detection results (Jupp et al., 1994). Many authors (Dekker, 1993; Jupp et al., 1994) confirm that Chl-*a* concentration has a strong relationship with band ratio of near infrared with red. That is because the Chl-*a* absorption occurs in the red band near 660-680 nm of the spectrum while the spectral reflectance peak due to phytoplankton scattering occurs in the red edge spectrum located at 685 - 715 nm (Dekker, 1993). However, water surface reflectance near the 700 nm wavelength is commonly measured by hand-held, shipboard remote sensing devices, or airborne hyperspectral sensors, but relatively rare in satellites (Kuster, 2009). MERIS, which stopped functioning in April 2012, was the only satellite sensor to estimate Chl-*a* concentration at the red edge spectrum (Kuster, 2009).

In waters with surface scums and floating algae, the detection of HABs can be treated as a kind of vegetation cover classification (Kutser, 2004). In such situations, the terrestrial standard products (e.g., Normalized Difference Vegetation Index (NDVI)) are considered a solution to estimating Chl-*a* concentrations (Kuster, 2009).

2.1.2 Band or Band ratio phycocyanin-retrieval algorithms

Many researchers use band or band ratio phycocyanin-retrieval methods for HAB mapping (Schalles and Yacobi, 2000; Vincent et al., 2004; Mishra and Mishra, 2014). This kind of algorithm is mainly based on two spectral features of pigment phycocyanin. One is the absorption occuring at approximately 620 to 630 nm, and another is the reflectance peak occuring near 650 nm. For example, the single reflectance band ratio algorithm used by Schalles and Yacobi (2000) employed a single band ratio algorithm to model the empirical relationship between the observed PC concentration and the band ratio of 650 um with 625 um in Carter Lake, Nebraska, USA (Ruiz-Verdú et al., 2008).

The above empirical models are easy and robust enough to be applied to a specific water body where *in situ* data can be regularly collected (Matthews et al., 2010). These methods work well with severe HABs that show strong spectral reflectance characteristics. However, as Matthews et al. (2010) mentioned, these empirical models are mostly used on airborne or hand-held sensors and cannot separate the signals of different constituents in water.

2.2 Semi-analytical algorithms

Semi-analytical algorithms have been proposed to detect the relationship of Chl-*a* or PC concentration with the retrieved inherent optical properties (IOPs) (such as the absorption coefficients of phytoplankton and other materials in waters, the total and particulate backscattering coefficients, and so on) (Matthews et al., 2010). The classical semi-analytical algorithms include the semi-analytical baseline subtraction algorithm (Dekker, 1993), the nested semi-analytical band ratio algorithm based on MERIS images (Simis et al., 2005), the three band

semi-analytical algorithm for MERIS-like sensors (Gitelson et al., 2003; Moses et al., 2009; Hunter et al., 2010) as well as the optimal PC_3 algorithm by Mishra and Mishra (2014), and the quasi-analytical algorithms (QAA) (Mishra et al., 2013; Mishra and Mishra, 2014; Ruiz-Verdú et al., 2008).

Semi-analytical algorithms are considered to be suitable for estimating PC or Chl-*a* concentration in turbid productive waters because they can solve several parameters simultaneously and hence separate the signals of different constituents in water (Matthews et al., 2010). However, they are sensitive to errors due to the many spectral bands and parameters involved (Matthews et al., 2010). Moreover, they are mostly designed for the remote sensing data from MERIS, MODIS, or hyperspectral sensors which have narrow bands in red and near infrared regions of the spectrum.

2.3 Analytical methods

Analytical methods use reflectance spectra instead of band ratios and statistics to map HABs (Kuster, 2009). For example, using derivative analysis to detect individual pigment signatures in absorbance or radiance spectra (Warner and Fan, 2013; Hunter et al., 2008; Kuster, 2009). Another kind of method, spectral shape methods using a computation process equivalent to the second derivative, mainly include fluorescent line height (FLH) (Gower et al., 1999); maximum chlorophyll index (MCI) (Gower et al., 2008), cyanobacteria index (CI) (Wynne et al., 2008; Wynne et al., 2010), the maximum peak-height algorithm (MPH) (Matthews et al., 2012), and the scattering line height (SLH) algorithm and aphanizomenon-microcystis index (AMI) (Kudela et al., 2015).

However, just like semi-analytical methods, most analytical methods work well with hyperspectral data and some ocean color sensors like MERIS, MODIS, or SeaWiFS (Kuster, 2009), and do not work well with multispectral data like Landsat images which have higher spatial resolution to map HABs in small water bodies.

2.4 Machine learning methods

In recent years, machine learning techniques (e.g., Decision Tree, Neural Networks, Random Forests) have been employed as more effective alternatives to conventional parametric algorithms for analyzing complex, high dimensional and nonlinear dataset (Olden et al., 2008; Rodriguez-Galiano et al., 2012), especially for ecological data which rarely require simple statistical analysis (Crisci et al., 2012). These algorithms are more efficient and accurate because they are less strict on data distribution assumptions (e.g., normal distribution), and can derive models from large, noisy datasets (Atkins et al., 2007; Rodriguez-Galiano et al., 2012). For example, a few studies have used neural networks (Schiller and Doerffer, 1999; Huang and Lou, 2003; Pozdnyakov et al., 2005) to retrieve Chl-*a* concentration and map phytoplankton blooms. Chen and Mynett (2004) used decision trees and nonlinear piecewise regression models to detect Phaeocystis globose bloom in Dutch coastal waters (Chen and Mynett, 2004).

The random Forests algorithm, which is reported to have a better performance than Support Vector Machine and most other machine learning methods (Crisci et al., 2012), has been widely used as a classification and/or a regression algorithm in a variety of fields (Mutanga et al., 2011). For example, the Random Forests algorithm has been widely applied in land-cover classification using multispectral and hyperspectral satellite images (Rodriguez-Galiano et al., 2012; Grinand et al., 2013; Ghosh et al., 2014). Recently, more studies have focused on the

regression application of Random Forests, such as the spatial distribution and prediction of aquaculture species (Vincenzi et al., 2011), the digital mapping of soil organic matter stocks (Wiesmeier et al., 2011), quantification of live aboveground forest biomass dynamics (Powell et al., 2010), and biomass estimation for wetland vegetation (Mutanga et al., 2011). However, the Random Forests approach has not been applied to HABs related studies yet.

As mentioned above, empirical, semi-analytical, and analytical methods have been widely applied in the quantitative detection and mapping of HABs, while only a few studies employed more complicated machine learning algorithms. Moreover, there is still relatively rare case studies to solve regression problems using machine learning algorithms and multispectral satellite data (Armston et al., 2009).

CHAPTER 3 MATERIALS AND METHODS

3.1 Study area

William H. Harsha Lake (Harsha Lake, or East Fork Lake) in Clemont County, located 25 miles east of Cincinnati in southwestern Ohio, is a reservoir project built in 1978 (Funk et al., 2003) and operated by the U.S. Army Corps of Engineers (USACE) for flood reduction, water supply, recreation, and a wildlife habitat (Beaulieu et al., 2014*a*). The reservoir is an open lake and covers an area of 8 km^2 and drains from a watershed of 890 km^2 (64% is land use in agriculture and 26% is forest cover), serving as the surface water source for the Bob McEwen Water Treatment Plant (10 MGD) (Green et al., 2010). Most areas of the lake have a water depth more than 8 m (Beaulieu et al., 2014*b*), and the maximum depth of the lake is 34 m (http://www.lakebrowser.com/ohio/william_h_harsha_lake.asp). Harsha Lake has generated about $32.7 million in visitor expenditures and prevented about $77 million in flood damages since its impoundment (Chang et al., 2014).

Over the past 10 years, HABs caused by blue-green algae have been increasing in Ohio inland waters (Francy et al., 2015), and Harsha Lake experienced HABs almost every year (Beaulieu et al., 2014*a*). The seasonal trend of blue-green algal biovolume and Microcystin (a kind of toxin produced by cyanobacteria) concentration in Harsha Lake has been revealed by USGS scientific investigation reports in 2013 and 2014 (Francy et al., 2015) —they increase from May to June, keep constant in July, and decrease slowly from August to October (Francy et al., 2015). Moreover, the phytoplankton community in Harsha Lake is complex and generally dominated by the blue-green algae (Francy et al., 2015).

Generally, the turbidity of Harsha Lake is quite low—Chang et al. (2014) reported that the average turbidity of Harsha Lake is 9.1 ntu (Chang et al., 2014). Similar evidence came from a report by USGS stating that (Francy et al., 2015) the turbidity values of the lake water were between 3.9 - 16.4 ntu in 2013 and 2014.

Harsha Lake is a typical small water body vulnerable to HABs. On the other hand, the low turbidity of the lake water makes Harsha Lake an ideal case study site for testing the algorithms for HABs mapping.

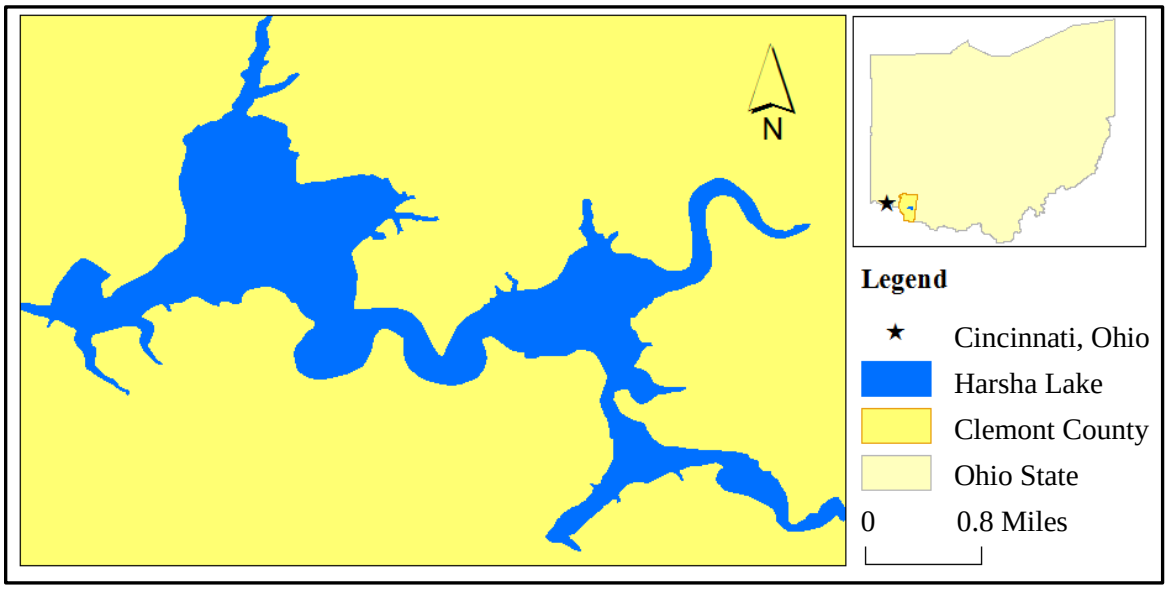

Figure 1. Harsha Lake, Ohio

3.2 Data

Two data sources were used in this thesis, a Landsat 8 OLI image acquired on September 21, 2015, and same-day *in situ* measurement of Chl-*a* in Harsha Lake.

3.2.1 In situ measurements of Chl-*a*

In situ measurements of Chl-*a* (µg/L) were conducted on September 21, 2015, deliberately synchronized with Landsat 8 satellite overpass. A total of 56 samples were collected across the lake (Figure 2) for analysis of Chl-*a* concentration. A range of Chl-*a* concentration was found between 3.6 - 15.2 µg/L, which was used to train and validate the models.

As for classifying the measured Chl-*a* concentrations, this thesis used the ranges of 0 - 6 µg/L, 6 - 8 µg/L, 8 - 10 µg/L and > 10 µg/L. In general, sample locations are evenly distributed in the area, except for southeastern Harsha Lake (only two sample sites) due to accessibility issue.

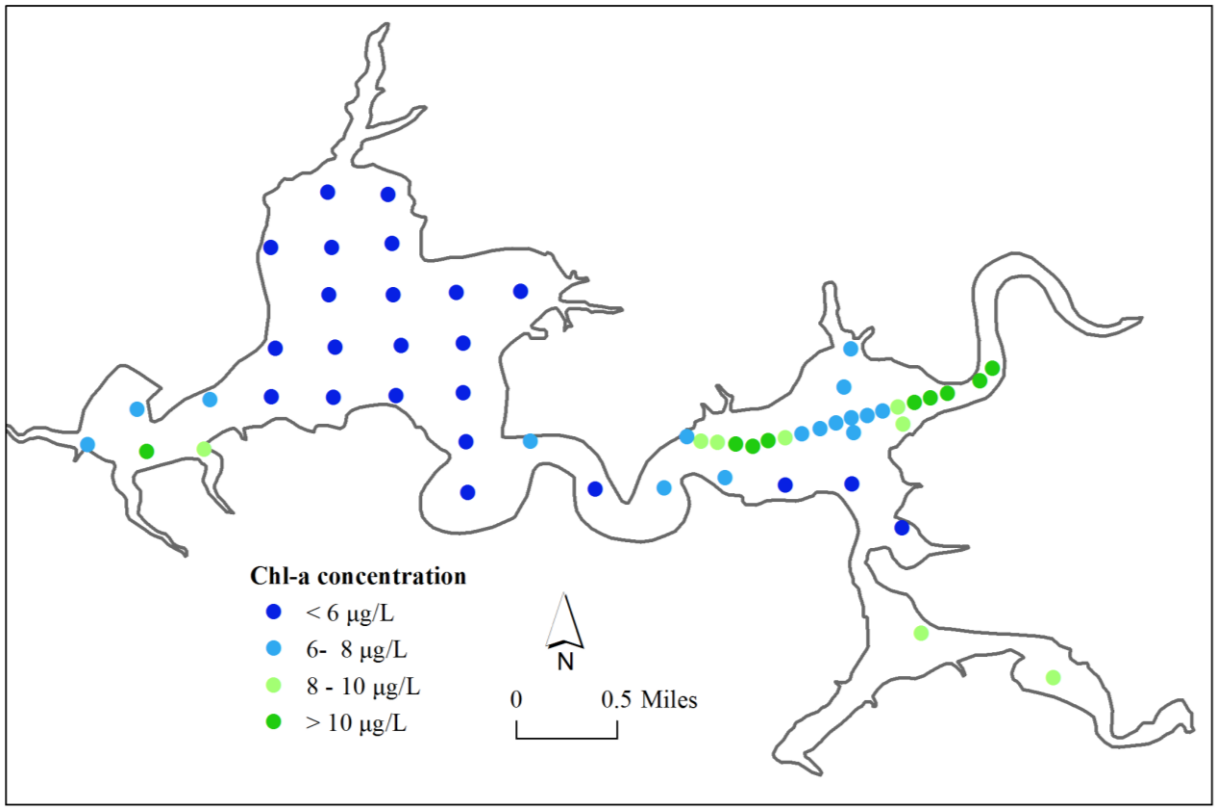

Figure 2. Sampling locations and measured Chl-*a* concentrations in Harsha Lake on 09/21/2015

The concentration of Chl-*a* (a pigment of all photosynthetic phytoplankton) is a common bioindicator to detect blue-green algae (Matthews et al., 2010). Jupp et al. (1994) indicated that algae would be present but with no suggestion of a bloom when Chl-*a* concentrations were less than 10 µg/L. In other words, there was algae present but with no severe HABs outbreak in Harsha Lake on September 21, 2015. The histogram in Figure 3 shows the descriptive statistics of the data.

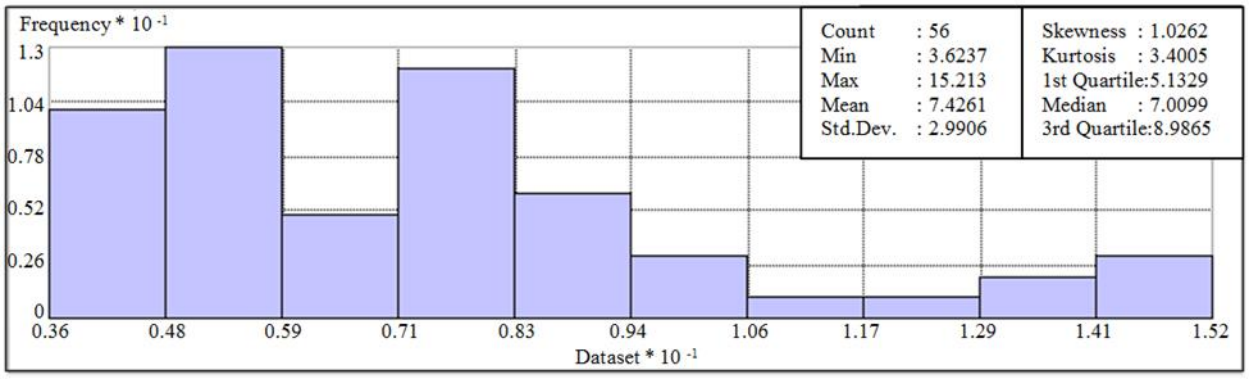

Figure 3. Chl-*a* concentrations histogram

3.2.2 Landsat 8 image

Landsat 8 OLI and Thermal Infrared Sensor (TIRS) images consist of nine spectral bands (bands 1 to 9) and two thermal bands (band 10 and 11). Bands 1 to 7 with a spatial resolution of 30 meters represent violet (0.43 - 0.45 µm), blue (0.45 - 0.51 µm), green (0.53 - 0.59 µm), red (0.64 - 0.67 µm), near infrared (0.85 - 0.88 µm), SWIR 1 (1.57 - 1.65 µm) and SWIR 2 (2.11 - 2.29 µm) regions of the spectrum, respectively. Band 8 (0.50 - 0.68 µm) which is panchromatic band, Band 9 (1.36 - 1.38 µm) which is designed to detect cirrus cloud, and thermal bands 10 and 11 which provide surface temperatures with a resolution of 100 meters were excluded in this study (USGS, 2016).

The Landsat 8 image was downloaded from USGS archives (http://earthexplorer.usgs.gov/). The image was corrected from atmospheric effects using the FLAASH module in ENVI. Spectral reflectance images of the bands were output from the FLAASH algorithm. The reflectance values were multiplied by 10,000 and converted to integers for further analyses.

Many studies (Vincent et al., 2004; Tebbs et al., 2013; Palmer et al., 2015) have illustrated that Landsat images can be used to map high biomass HABs for small water bodies based on red and near infrared spectral regions. Moreover, in clear water like Harsha Lake, Landsat 8 images should also be applicable to detect low-medium biomass of blue-green algae based on blue to green spectral regions (Moses et al., 2009; Kuster, 2009). Figure 4 shows the spectral reflectance values in each sampling site derived from the Landsat 8 image. Figure 4 indicates that the spectral reflectance characteristics in the 56 sample sites are similar but also could be used to distinguish between different Chl-*a* values, especially in the blue to red spectral region, which is important to detect blue-green algae in low-biomass waters.

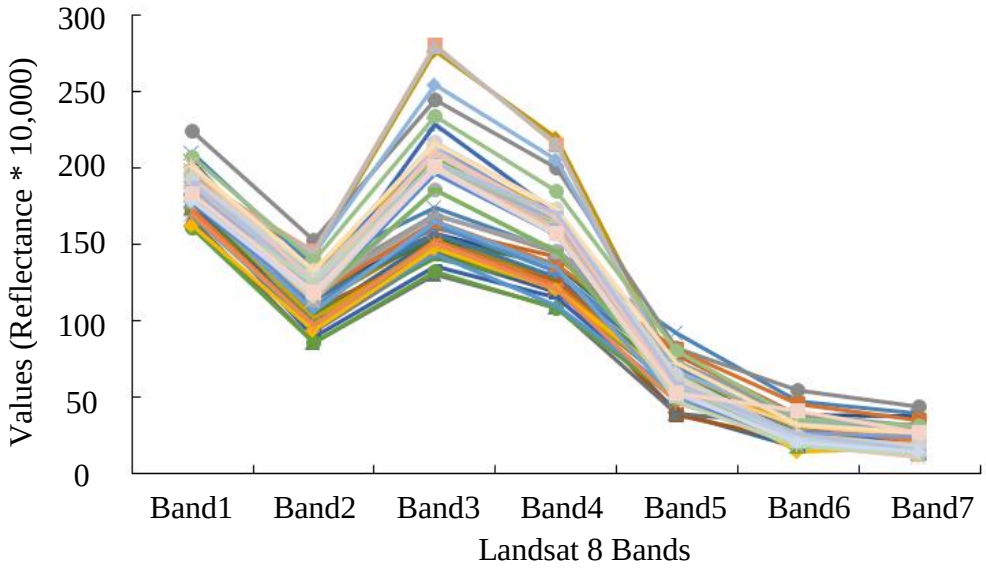

Figure 4. The Landsat 8 reflectance values of the 56 sampling sites in Harsha Lake on 09/21/2015

3.3 Methods

Based on the literature, it can be concluded that given the characteristics of the water condition of the study area Harsha Lake, even with low/medium Chl-*a* concentrations, it is possible to measure the biomass of blue-green algae from blue and green bands of the Landsat 8 image. This case study aims to find the empirical relationship between the spectral reflectance of Landsat bands and the *in situ* measured Chl-*a* concentrations, and to map the spatial distribution of blue-green algae in the study area. Two methods were employed; one is conventional Stepwise Multiple Linear Regression (SMLR) that is widely used in HABs studies, and the other is one of the popular machine learning method—Random Forests (RF) algorithm. Both methods were carried out in R software. To the best of my knowledge, this is the first time the feasibility of the RF algorithm has been tested with the use of Landsat 8 imagery to map HABs.

3.3.1 Stepwise Multiple Linear Regression (SMLR)

This thesis used Landsat 8 band values as the predictor variables, and the measured Chl-*a* concentrations as the dependent variable to build the Multiple Linear Regression model. The Multiple Linear Model can be written as Equation (1) (Armston et al., 2009):

$$Y = \sum_{i=1}^{p} A_i X_i + A_0 + \varepsilon \qquad (1)$$

where dependent variable Y is the measured Chl-*a* concentrations, predictor variable X_i is the band values of the Landsat 8 image, p is the total number of predictor variable X_i included in the model, parameter A_i is the regression coefficient for X_i, A_0 is the constant term, and ε is the unexplained variance by the model.

A series of candidate Multiple Linear Models can be generated for all possible combinations of predictor variables. This thesis used the "MASS" and "DAAG" packages in R software (R Core Team, 2013) to conduct SMLR. The criterion used to estimate model quality is AIC, and a smaller AIC value generally means a better model. The best SMLR model can be selected based on the comparison of AIC values obtained from each Multiple Linear Regression model.

3.3.2 Random Forests (RF)

Machine learning methods have been widely used to identify and map the spatial patterns of ecological data based on Remote Sensing images and measured geographical data (Chen and Mynett, 2004; Na et al., 2010). Currently, the RF approach is considered to be one of the most effective machine learning methods with regard to prediction accuracy (Na et al., 2010; Crisci et al., 2012). RF has been widely used in classification and regression problems (Mutanga et al., 2011).

RF is an ensemble machine learning algorithm developed by Breiman (2001) to minimize the over-fitting problem existing in the classification and regression trees (CART) method (Mutanga et al., 2011). In the RF regression algorithm, a forest is a large number of regression trees determined by randomization (Armston et al., 2009), and each tree is trained by selecting a random set of predictor variables from a bootstrap sample of the training data (Mutanga et al., 2011). The final estimate of the dependent variable is calculated by the ensemble mean of the decision trees in the forests (Armston et al., 2009). It is capable of dealing with complex, nonlinear and noisy datasets (Mutanga et al., 2011). Moreover, it offers an analysis of predictor variables' importance (Crisci et al., 2012). However, its "black box" nature makes it difficult to

interpret the relationships between the dependent and predictor variables (Wiesmeier et al., 2011). Furthermore, because of the randomized nature of the RF algorithm, it does not require users to provide extra samples for validation. There will be automated cross-validation results as a by-product of the algorithm.

In this thesis, the RF regression algorithm was used as a nonlinear variable selection and regression method for predicting the Chl-*a* concentrations of blue-green algae in Harsha Lake on September 21, 2015. The implementation of the RF model was through the "randomForest" package (Liaw and Wiener, 2002) in R software (R Core Team, 2013; Wiesmeier et al., 2011). No variable transformation and reduction technique were employed because RF is considered robust to collinearity among predictor variables (Powell et al., 2010).

3.3.3 Model validation and assessment

Models can be validated with three techniques: an independent test dataset, cross-validation, or (RF models only) Out-of-bag (OOB) estimation on the training data (Freeman et al., 2016). In order to make the above two methods comparable, this thesis performed independent test dataset validation. To be more specific, a proportion of the data (80% of the data) was set aside randomly as the training dataset, and 20% of the data as the independent test dataset. The reason for partitioning the data like this is that it can get the lowest prediction error (RMSE) on the test dataset for both SMLR and RF models. This thesis set the seed of R's random number generator as 1115 to make the model results reproducible.

With respect to the statistics to validate the regression models, Root Mean Squared Error (RMSE), Mean Error (ME), Mean Squared Error (MSE), Variance, and Coefficient of Model Determination (R^2) are the most commonly used statistics in previous studies (Armston et al.,

2009; Wiesmeier et al., 2011; Powell et al., 2010). This thesis applied RMSE, ME and R^2 to assess the prediction performance of the SMLR and RF models in predicting the Chl-*a* concentrations of blue-green algae.

RMSE can measure the accuracy of regression predictions, and is the most important statistic for model selection and validation in the literature. Generally, the best prediction model has the lowest value of RMSE on the test dataset. R^2 is expressed as the ratio of the explained variance to the total variance, which is also a commonly used index to assess regression models' performance, especially for the linear regression models. ME is to measure if the prediction error measures are biased by the overfitting problem (better fit than the real case). Even though RF models seldom over fit in practice, the overfitting problem may still exist if there are too many trees in the forest (Lin and Jeon, 2006). The formulae to calculate RMSE, ME and R^2 can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} E_i^2}{N}} \tag{2}$$

$$ME = \frac{\sum_{i=1}^{N} E_i}{N} \tag{3}$$

$$R^2 = \frac{\sum_{i=1}^{N} (y_i - Y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{4}$$

where E_i is the difference between the measured and predicted values of Chl-*a* concentration, N is the number of samples in the dataset, y_i is the measured values of Chl-*a* concentration, $\bar{y}$ is the mean of all measured values in the dataset, and Y_i is the predicted Chl-*a* concentration by regression models.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Best fitting SMLR model

The SMLR analysis can generate an equation to describe the linear relationship between one or more Landsat 8 bands (variable names are "Band1" to "Band7") and the dependent variable—measured Chl-*a* values. In the tests, SMLR models were built using the measured Chl-*a* concentrations of the 56 samples (variable name is "Chl-*a*") as dependent variable; then, the SMLR models were built again using the natural log-transformed Chl-*a* concentrations of the 56 samples (variable name is "LN(Chl-*a*)") as dependent variable. That is because the measured Chl-*a* concentration data are not normally distributed (see Figure 3 in section 3.2), and this thesis would like to test if the natural logarithm transformation can improve model performance. As mentioned in section 3.3, this thesis used the "MASS" and "DAAG" packages within R (R Core Team, 2013) to conduct SMLR and get the final SMLR model, and the criterion to select the final model was based on the AIC values.

4.1.1 The final SMLR model using "Chl-*a*" as dependent variable.

The final SMLR model consists of three predictor variables including Band2, Band3, and Band5, and R^2 of the model is 0.48. Table 1 lists the model coefficients, and we can know whether any of the predictors have significance in the model by analyzing the P value. The coefficients table below implicates that Band3 and Band2 are significant while Band5 is not statistically significant using the common confidence interval at 95%. Band 5 is only significant at confidence interval of 90%.

Table 1. The coefficients of the final SMLR model using "Chl-*a*" as the dependent variable

Model	Estimate	Std. Error	T	P value
Constant	6.10393	2.66532	2.290	0.02723 *
Band2	-0.22728	0.06700	-3.392	0.00155 **
Band3	0.12232	0.02336	5.235	5.25e-06 ***
Band5	0.08603	0.05095	1.688	0.09892

'***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1.

The final model using "Chl-*a*" as the dependent variable can be written as Equation (5):

$$Chl-a = 6.10393 - 0.22728 \times Band2 + 0.12232 \times Band3 + 0.08603 \times Band5 \quad (5)$$

4.1.2 The final SMLR model using "LN(Chl-*a*)" as the dependent variable.

The final SMLR model consists of two predictor variables including Band2 and Band3, and R^2 of the model is 0.50. Table 2 lists the model coefficients and the implication is that both Band2 and Band3 are significant at a significance level of 0.001. The final model using "LN(Chl-*a*)" as a dependent variable can be written as Equation (6).

$$LN(Chl-a) = 1.726157 - 0.017790 \times Band2 + 0.012288 \times Band3 \quad (6)$$

Table 2. The coefficients of the final SMLR model using "LN(Chl-*a*)" as the dependent variable

Model	Estimate	Std. Error	T	P
Constant	1.726157	0.326460	5.287	4.17e-06 ***
Band2	-0.017790	0.004978	-3.574	0.000901***
Band3	0.012288	0.002029	6.055	3.31e-07 ***

'***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1, ' ' 1.

4.1.3 Best fitting SMLR model

The above results suggested that the log transformation of the Chl-*a* observation could not significantly increase R^2 or lower the prediction error. In order to make the regression results more comparable to the RF model, the best SMLR model directly used Chl-*a* concentration ("Chl-*a*") as the dependent variable.

4.2 Best fitting RF model

In order to get the best fit of the RF model, the first step was to select a subset of predictor variables that can best predict the dependent variable. Then, based on the variables selected, built the best fitting RF model by using optimal parameters "mtry" and "ntree".

4.2.1 Subset size of the variables and variable importance measures

The RF model provided an evaluation of the subsets of variables from randomization. The principle of variable selection was to "select the fewest number of predictors that offer the

best predictive power of the dependent variable and help in the interpretation of the final model" (Mutanga et al., 2011). Based on previous studies (Díaz-Uriarte et al., 2006; Genuer et al., 2010), the variable selection procedure can be described as follows:

a. Partition the dependent variable (Chl-*a* concentrations of 56 samples) into two datasets randomly—80% of the data as the training dataset and 20% of the data as the test dataset.
b. Perform variable importance measures with the "randomForest" package in R using default parameters (mtry = 1/3 of the total number of input parameters, ntree = 500), and sort the variables in decreasing order of importance.

The variable importance measures can show the strength of each variable's relationship to the dependent variable (Wiesmeier et al., 2011). There are two types of variable importance measures provided by the "randomForest" package. Type one is to measure prediction accuracy (i.e., how worse the model performs) if a variable is permuted, and type two is to measure node purity (i.e., how pure a node is) at the end of a tree in the forest (Breiman, 2001). Type one can be performed by calculating the percent increase in Mean Standard Error (%IncMSE) as each variable is permuted while all others are unchanged, and a higher value represents greater variable importance (Genuer et al., 2010). %IncMSE is the most widely used importance measures of variables (Mutanga et al., 2011), and this thesis also uses it to measure variable importance. Figure 5 illustrates the %IncMSE calculated to measure the importance of all seven bands in predicting Chl-*a* concentrations.

c. Eliminate the smallest importance variable at each iteration and then build a new forest with the remaining variables.
d. Repeat steps b and c to fit RF models iteratively, and the prediction error (RMSE) on the test dataset at each iteration can be found in Figure 6.

e. Select the set of variables with lowest RMSE on the test dataset.

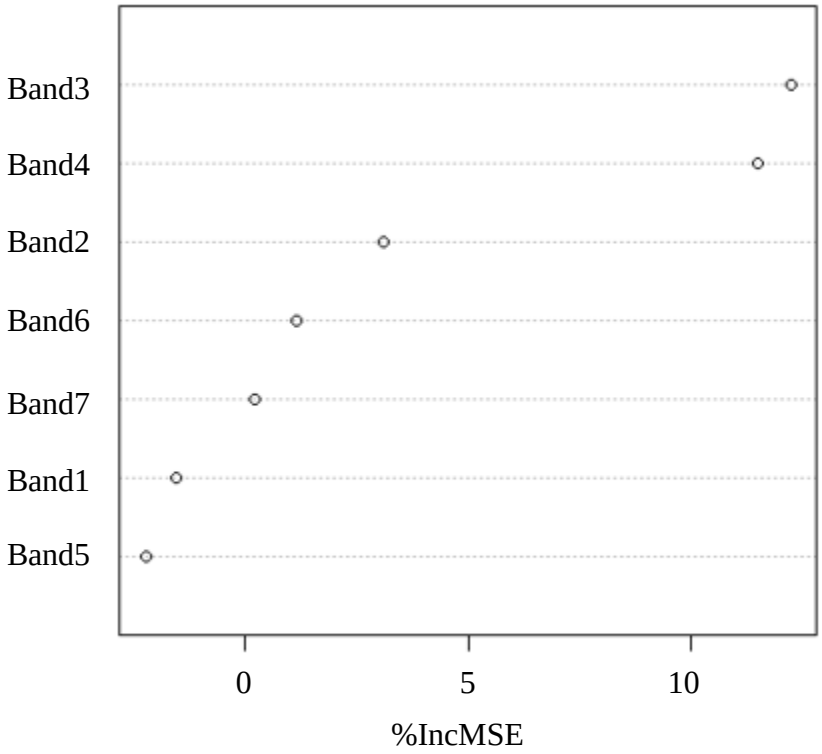

Figure 5. Variable importance measures (%IncMSE) of all variables in predicting Chl-*a* concentrations using the RF model

Figure 5 illustrates the importance ranking of all variables in decreasing order is Band3, Band4, Band2, Band6, Band7, Band1, and Band5. More precisely, this thesis defines the set of variables used in the RF model at each iteration as Set A, B, C, D, E, F, and G. The iterative process can be described as follows. Set A contains 7 variables—Band3, Band4, Band2, Band6, Band7, Band1, and Band5; Set B discards Band5 and contains 6 variables—Band3, Band4, Band2, Band6, Band7, and Band1; Set C discards Band1 and contains 5 variables—Band3, Band4, Band2, Band6, and Band7; Set D discards Band7 and contains 4 variables—Band3, Band4, Band2, and Band6; Set E discards Band6 and contains 3 variables—Band4, Band3, and

Band2; Set F discards Band2 and contains 2 variables—Band3 and Band4; Set E discards Band4 and contains only variable Band3. The RF model yields a RMSE on the test dataset at each iteration, and the results are summarized in Figure 6.

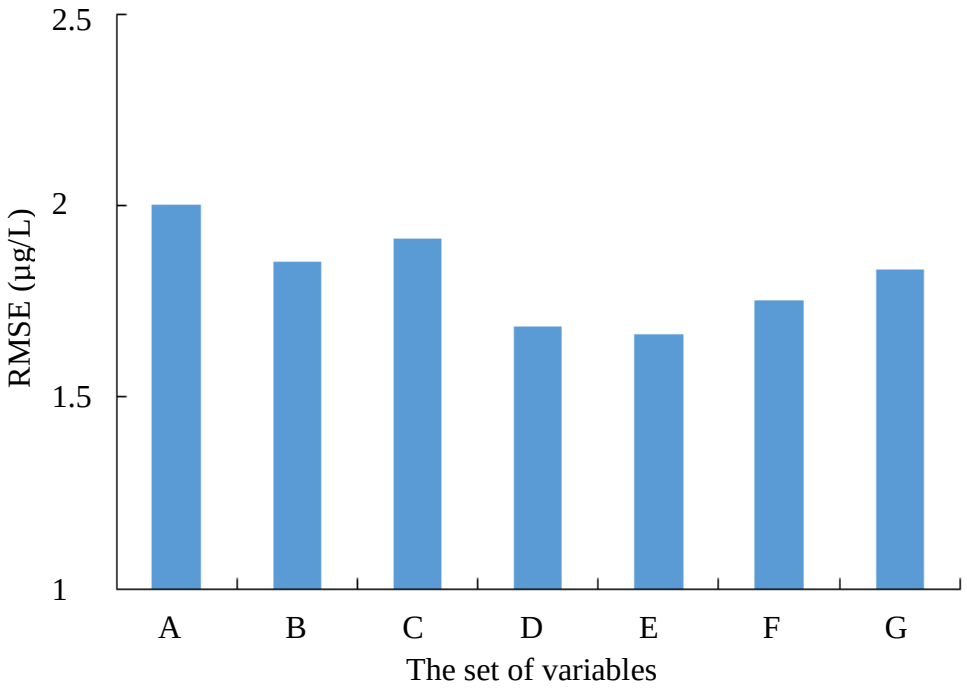

Figure 6. The RMSE on the test dataset yielded from the RF models at each iteration

In general, the RMSE decreases while the less important variables (Band5, Band7, Band6) are discarded progressively, and the RMSE increases while the more important variables (Band2, Band4) are discarded at the last two iterations. An exception is the increase in RMSE when the less important Band1 is discarded. The entire set of seven variables (Set A) yields the highest RMSE (2 µg/L) on the test dataset, while the set of three variables Band4, Band3 and Band2 (Set E) yields the lowest RMSE (1.66 µg/L) on the test dataset.

The result indicated that Band 4, Band3 and Band 2 were the most important variables in explaining the Chl-*a* concentrations of Harsha Lake on September 21, 2015. The variable selection procedure can identify three bands as the fewest number of variables that could offer the best predictive performance of the RF model. Therefore, Set E was selected to build RF regression model in predicting the Chl-*a* concentrations in Harsha Lake.

4.2.2 Optimal parameters for RF model

According to previous studies (Genuer et al., 2010; Mutanga et al., 2011; Vincenzi et al., 2011), mtry (the number of input variables randomly chosen for splitting) and ntree (the number of trees in the forest) are two parameters that need to be optimized in the RF algorithm. This section aims to find the optimal values of mtry and ntree by analyzing the RF model sensitivity to mtry and ntree.

Figure 7 below illustrates the sensitivity of the RF model (built by variable Set E including Band4, Band3, and Band2) to parameters mtry and ntree; in other words, it shows how RMSE on the test dataset varies with parameters mtry and ntree. This thesis tested seven values of ntree (30, 50, 100, 200, 500 the default, 1000, and 2000) and three values of mtry (from 1 to 3 using a single interval).

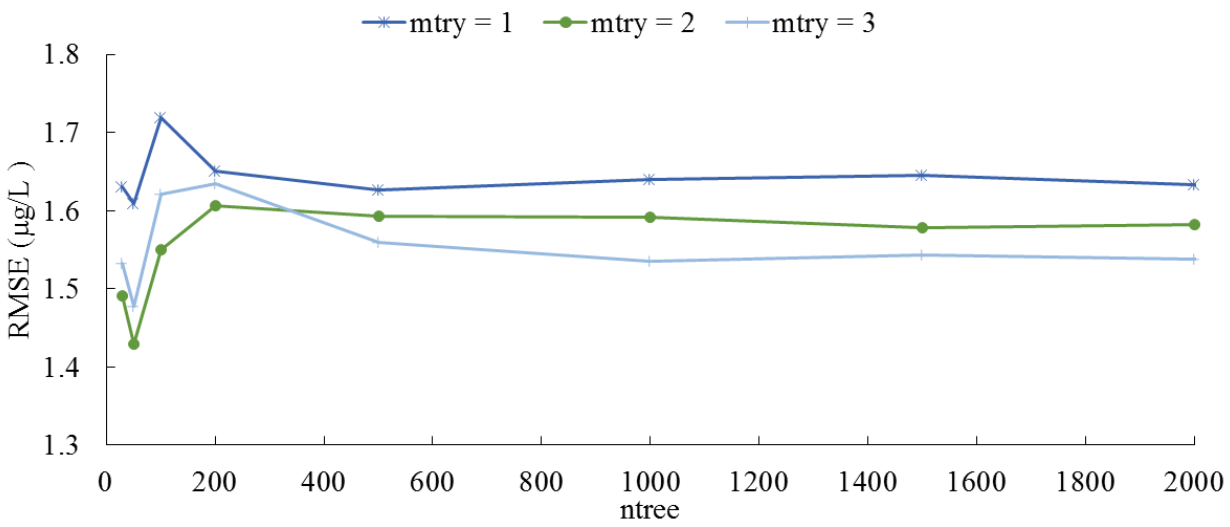

Figure 7. Sensitivity of the RF model (built by variable Set E including Band4, 3, 2) to parameters (mtry and ntree) based on RMSE.

The results implicate that mtry and ntree affect the prediction errors. To be more specific, the default mtry = 1 gives the worst prediction; mtry=3 produces the best prediction when ntree > 500, while mtry = 2 produces the best prediction when ntree < 500. Moreover, ntree = 50 is the curve trough and yields the lowest RMSE compared with other ntree values.

Overall, Parameter ntree = 50 and mtry = 2 yielded the lowest RMSE, and the RF model developed using 3 input variables (Band4, Band3, Band2), 2 mtry, and 50 ntree were selected as the best fitting model to predict the dependent variable the Chl-*a* concentrations of blue-green algae in Harsha Lake.

4.3 Comparison of the best fitting models

As mentioned in section 3.3, this thesis applied RMSE, ME and R^2 to assess the prediction performance of the best fitting models obtained in section 4.1 and 4.2. In order to

make the best fitting models comparable to previous studies, RMSE and ME were normalized using the range (the maximum value minus the minimum value) of the dataset (Vincent et al., 2004).

Table 3. The comparison of the best fitting SMLR and RF models

Model	SMLR Model	RF Model
RMSE (µg/L)	2.43	1.43
Normalized RMSE	20.93%	12.33%
ME (µg/L)	0.87	0.02
Normalized ME	7.52%	00.14%
R^2	0.48	0.82

Table 3 indicates that the best fitting RF model yields higher R^2 (0.82) and lower RMSE of prediction (1.43 µg/L, about 12.33% of the total range of the observed Chl-*a* concentrations) on the test dataset compared to the best fitting SMLR model which yields R^2 of 0.48 and RMSE of 2.43 µg/L (about 20.93% of the total range of the observed Chl-*a* concentrations). The uncertainty level is consistent with the previous study by Vincent et al. (2004) using Landsat TM imagery to map blue-green algae blooms in Lake Erie. In Vincent et al. (2004), they reported a RMSE of about 26% of the total range of Phycocyanin pigment. However, the machine learning algorithm greatly outperforms the SMLR model as well as previous research in similar approaches.

4.4 Prediction surfaces of Chl-*a* concentration by the best fitting RF model

The analysis results show that the RF model provides a better prediction of Chl-*a* concentration on the test dataset (R^2=0.82, RMSE is 1.43 µg/L), which suggests that the RF model is preferred for mapping low biomass HABs, especially with Chl-*a* concentrations less than 20 µg/L in aquatic systems. In this section, the best fitting RF model was applied to the entire study area of Harsha Lake to obtain the map of Chl-*a* concentrations. The prediction surface (Figure 8) was generated with a spatial resolution of 30 m ×30 m pixels from the Landsat 8 image. The spatial patterns can be summarized as follows.

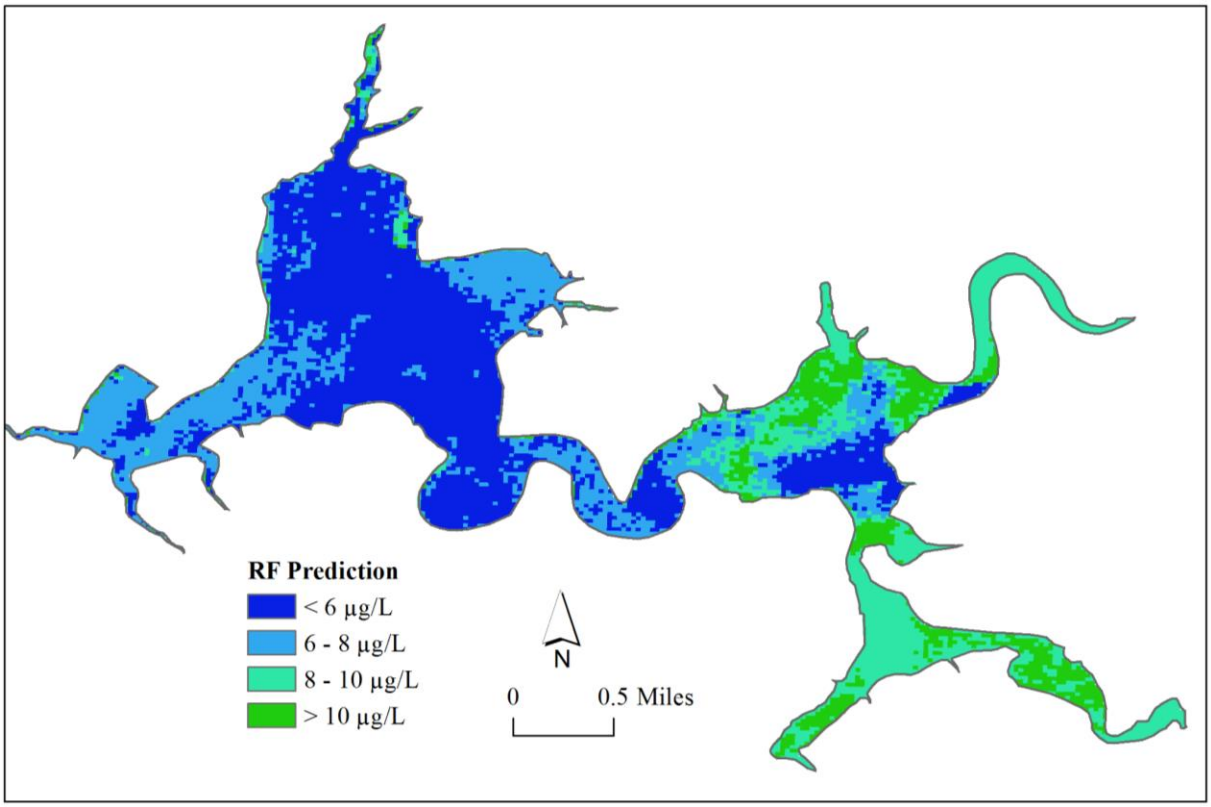

Figure 8. Prediction surface of Chl-*a* concentration generated by the best fitting RF model

The areas at different Chl-*a* concentration levels were calculated, and 89.30% of Harsha Lake area had Chl-*a* values less than 10 µg/L on the sampling date, while in only 10.70% of the entire study area were the Chl-*a* concentrations between 10 µg/L and 20 µg/L. The lowest Chl-*a* (Chl-*a* < 6 µg/L) accounted for 43.8% of the lake area, and was mainly distributed in the center of the lake area, while higher Chl-*a* values were closer to the shore. This is consistent with the previous conclusions about HABs' spatial distribution: "higher cyanotoxin concentrations are expected near shore" (Kasich et al., 2015). Moreover, the Chl-*a* concentrations in western Harsha Lake were generally less than 8 µg/L, and the Chl-*a* concentrations larger than 8 µg/L were mostly distributed in northeastern and southeastern Harsha Lake.

In order to analyze the probable causes of the Chl-*a* spatial patterns, the National Hydrography Dataset (NHD) containing the flow direction of the surface water system was downloaded the from USGS website (http://nhd.usgs.gov/data.html). NHD Flowline data of Ohio inland waters was used to trace the downstream and upstream of Harsha Lake (As shown in Figure 9). It should be noted that higher Chl-*a* values (especially for Chl-*a* larger than 10 µg/L) were mostly distributed close to the mouths of rivers or streams, such as the mouth of Cabin Run in northeastern Harsha Lake, the mouth of East Fork Little Miami River which crosses Harsha Lake east to west, and the mouth of Cloverlick Creek in southeastern Harsha Lake. Based on these patterns, it is suspected that the occurrence of blue-green algae in Harsha Lake might be attributed to the nutrient influx generated from the non-point sources of pollution in agricultural and urban lands through upstream rivers, although validating this hypothesis is beyond the scope of this thesis.

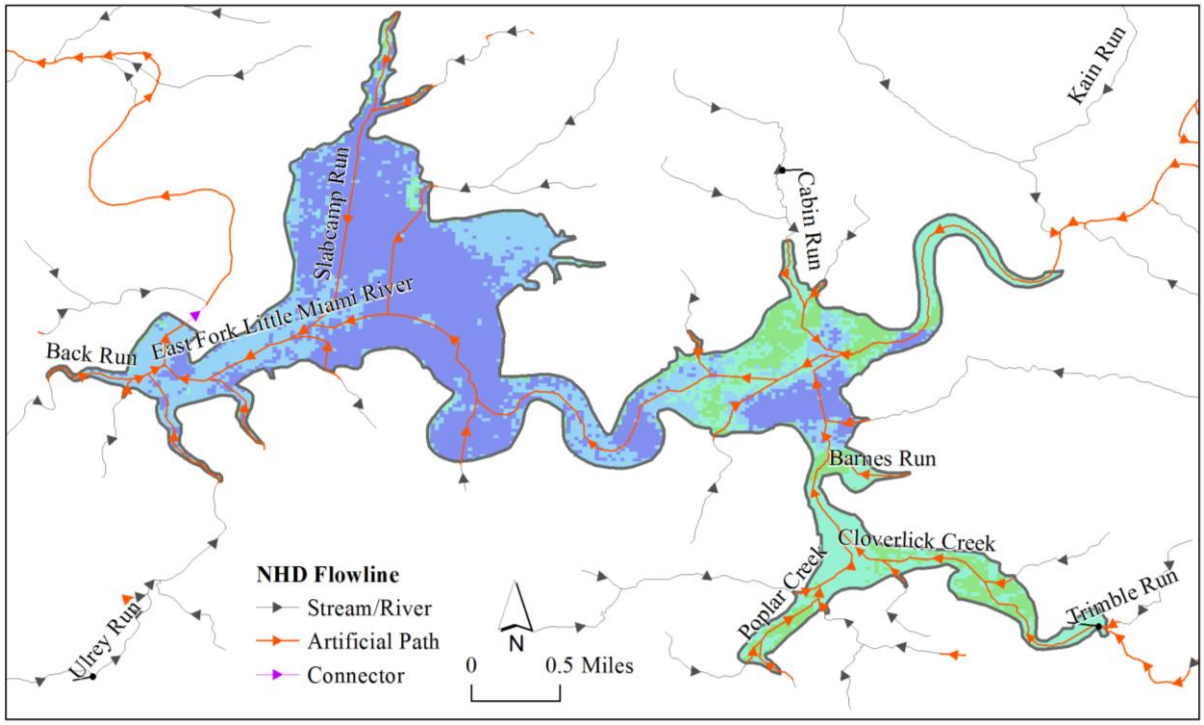

Figure 9. Flow direction of the surface water network from the National Hydrography Dataset (NHD)

CHAPTER 5 CONCLUSIONS

In comparing conventional linear regression and Random Forests machine learning algorithm for mapping blue-green algae in Harsha Lake using Landsat 8 imagery, I have come to the following conclusions.

(1) The study area, Harsha Lake, is a low-medium biomass water body, showing some spectral reflectance characteristics in blue and green spectral regions. One research hypothesis of this thesis is that Landsat 8 data is applicable to the detection and mapping of blue-green algae in Harsha Lake by analyzing the blue band (Band2) and green band (Band3). This thesis confirms this hypothesis through statistical analysis: the predictor variables selected by best fitting models to predict Chl-*a* concentrations are Band2, Band3, and Band5 (representing blue, green and near infrared spectral regions respectively) for the SMLR model, and Band4, Band3, and Band2 (representing red, green and blue spectral regions respectively) for the RF model. The models were all significant.

(2) The best fitting SMLR model includes 3 predictor variables—Band2, Band3, and Band5 at a significance level of 0.1. The R^2 is 0.48 and the produced RMSE on the test dataset is 2.43 μg/L—about 20.93% of the total range of the measured Chl-*a* concentrations.

(3) The best fitting RF Regression model also includes 3 predictor variables—Band4, Band3, and Band2, with parameter ntree = 50 and mtry = 2. The R^2 is 0.82 and the produced RMSE on the test dataset is 1.43 μg/L—about 12.33% of the total range of the measured Chl-*a* concentrations.

(4) The results show that compared to the conventional linear regression model, the performance of the RF model is better at predicting Chl-*a* concentrations of blue-green

algae, and the prediction accuracy of both models are sufficient to map the extent and biomass of the blue-green algae in Harsha Lake on September 21, 2015.

(5) The prediction surface by the besting fitting RF model illustrates that approximately 90% of Harsha Lake's area had Chl-*a* values less than 10 µg/L on the sampling date. Higher Chl-*a* values (especially for Chl-*a* larger than 10 µg/L) were mostly distributed close to the mouths of rivers or streams in northeastern and southeastern Harsha Lake, which might be caused by the influx of nutrients from agricultural or urban land use by rivers and streams.

(6) The results show the utility of the RF approach and Landsat 8 imagery in detecting and quantitatively mapping low-medium biomass HABs, which is considered to be a challenging task (Matthews and Odermatt, 2015).

REFERENCES

Agha, Ramsy, Samuel Cires, Lars Wörmer, José Antonio Domínguez, and Antonio Quesada. "Multi-scale strategies for the monitoring of freshwater cyanobacteria: Reducing the sources of uncertainty." *water research* 46, no. 9 (2012): 3043-3053.

Armston, John D., Robert J. Denham, Tim J. Danaher, Peter F. Scarth, and Trevor N. Moffiet. "Prediction and validation of foliage projective cover from Landsat-5 TM and Landsat-7 ETM+ imagery." Journal of Applied Remote Sensing 3, no. 1 (2009): 033540-033540.

Atkins, Jonathan P., Daryl Burdon, and James H. Allen. "An application of contingent valuation and decision tree analysis to water quality improvements." Marine Pollution Bulletin 55, no. 10 (2007): 591-602.

Beaulieu, Jake J., Rebecca L. Smolenski, Christopher T. Nietch, Amy Townsend-Small, Michael S. Elovitz, and Joseph P. Schubauer-Berigan. "Denitrification alternates between a source and sink of nitrous oxide in the hypolimnion of a thermally stratified reservoir." Limnology and Oceanography59, no. 2 (2014) a: 495-506.

Beaulieu, Jake J., Rebecca L. Smolenski, Christopher T. Nietch, Amy Townsend-Small, and Michael S. Elovitz. "High methane emissions from a midlatitude reservoir draining an agricultural watershed." Environmental science & technology 48, no. 19 (2014) b: 11100-11108.

Becker, Richard H., Mohamed I. Sultan, Gregory L. Boyer, Michael R. Twiss, and Elizabeth Konopko. "Mapping cyanobacterial blooms in the Great Lakes using MODIS." Journal of Great Lakes Research 35, no. 3 (2009): 447-453.

Braig IV, Eugene C., Joseph Conroy, Frank Lichtkoppler, William E. Lynch Jr., and Linda Merchant-Masonbrink. “Ohio Sea Grant Fact Sheets Harmful Algal Blooms in Ohio Waters.” (2010).

Breiman, Leo. "Random forests." Machine learning 45, no. 1 (2001): 5-32.

Chang, Ni-Bin, Benjamin W. Vannah, Y. Jeffrey Yang, and Michael Elovitz. "Integrated data fusion and mining techniques for monitoring total organic carbon concentrations in a lake." International Journal of Remote Sensing35, no. 3 (2014): 1064-1093.

Chen, Qiuwen, and Arthur E. Mynett. "Predicting Phaeocystis globosa bloom in Dutch coastal waters by decision trees and nonlinear piecewise regression." Ecological Modelling 176, no. 3 (2004): 277-290.

Crisci, C., B. Ghattas, and G. Perera. "A review of supervised machine learning algorithms and their applications to ecological data." Ecological Modelling 240 (2012): 113-122.

Darecki, Miroslaw, and Dariusz Stramski. "An evaluation of MODIS and SeaWiFS bio-optical algorithms in the Baltic Sea." Remote Sensing of Environment 89, no. 3 (2004): 326-350.

Dekker, Arnold Graham. "Detection of optical water quality parameters for eutrophic waters by high resolution remote sensing." (1993).

Díaz-Uriarte, Ramón, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." BMC bioinformatics 7, no. 1 (2006): 1.

Francy, Donna S., Jennifer L. Graham, Erin A. Stelzer, Christopher D. Ecker, Amie M. G. Brady, Pam Struffolino, and Keith A. Loftin. Water quality, cyanobacteria, and environmental factors and their relations to microcystin concentrations for use in predictive models at Ohio Lake Erie and inland lake recreational sites, 2013-14. No. 2015-5120. US Geological Survey, 2015.

Freeman, Elizabeth A., Tracey S. Frescino, and Gretchen G. Moisen. "ModelMap: an R Package for Model Creation and Map Production." (2014).

Funk, Jason M., David C. Reutter, and Gary L. Rowe. Pesticides and Pesticide Degradates in the East Fork Little Miami River and William H. Harsha Lake, Southwestern Ohio, 1999-2000. US Department of the Interior, US Geological Survey, 2003.

Genuer, Robin, Jean-Michel Poggi, and Christine Tuleau-Malot. "Variable selection using random forests." Pattern Recognition Letters 31, no. 14 (2010): 2225-2236.

Ghosh, Aniruddha, Richa Sharma, and P. K. Joshi. "Random forest classification of urban landscape using Landsat archive and ancillary data: Combining seasonal maps with decision level fusion." Applied Geography 48 (2014): 31-41.

Gitelson, Anatoly A., Yuri Gritz, and Mark N. Merzlyak. "Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves." Journal of plant physiology 160, no. 3 (2003): 271-282.

Gower, J. F. R., R. Doerffer, and G. A. Borstad. "Interpretation of the 685nm peak in water-leaving radiance spectra in terms of fluorescence, absorption and scattering, and its observation by MERIS." International Journal of Remote Sensing 20, no. 9 (1999): 1771-1786.

Gower, J., S. King, and P. Goncalves. "Global monitoring of plankton blooms using MERIS MCI." International Journal of Remote Sensing 29, no. 21 (2008): 6209-6216.

Green, Robert A., Eric Heiser, David Shank, and Vinod Korategere. "Meshing Treatment Objectives, Water Quality Goals and Regulatory Requirements Into A Plant Expansion Project." (2010).

Grinand, Clovis, Féty Rakotomalala, Valéry Gond, Romuald Vaudry, Martial Bernoux, and Ghislain Vieilledent. "Estimating deforestation in tropical humid and dry forests in Madagascar from 2000 to 2010 using multi-date Landsat satellite images and the random forests classifier." Remote Sensing of Environment 139 (2013): 68-80.

Hu, Chuanmin. "A novel ocean color index to detect floating algae in the global oceans." Remote Sensing of Environment 113, no. 10 (2009): 2118-2129.

Huang, W. G., and X. L. Lou. "AVHRR detection of red tides with neural networks." International Journal of Remote Sensing 24, no. 10 (2003): 1991-1996.

Hunter, Peter, Andrew Tyler, Nigel Willby, and David Gilvear. "The spatial dynamics of vertical migration by Microcystis aeruginosa in a eutrophic shallow lake: A case study using high spatial resolution time-series airborne remote sensing." (2008).

Hunter, Peter D., Andrew N. Tyler, Laurence Carvalho, Geoffrey A. Codd, and Stephen C. Maberly. "Hyperspectral remote sensing of cyanobacterial pigments as indicators for cell populations and toxins in eutrophic lakes."Remote Sensing of Environment 114, no. 11 (2010): 2705-2718.

Jupp, David LB, John TO Kirk, and Graham P. Harris. "Detection, identification and mapping of cyanobacteria—using remote sensing to measure the optical quality of turbid inland waters." Marine and Freshwater Research 45, no. 5 (1994): 801-828.

Kasich, J., M. Taylor, C. Butler, J. Zehringer, and R. Hodges. "State of Ohio Harmful Algal Bloom Response Strategy for Recreational Waters." Department of Health, Environmental Protection Agency and Department of Natural Resources (2015).

Keith, Darryl J., Bryan Milstead, Henry Walker, Hilary Snook, James Szykman, Michael Wusk, Les Kagey, Charles Howell, Cecil Mellanson, and Christopher Drueke. "Trophic status, ecological condition, and cyanobacteria risk of New England lakes and ponds based on aircraft remote sensing."Journal of Applied Remote Sensing 6, no. 1 (2012): 063577-1.

Kudela, Raphael M., Sherry L. Palacios, David C. Austerberry, Emma K. Accorsi, Liane S. Guild, and Juan Torres-Perez. "Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters."Remote Sensing of Environment 167 (2015): 196-205.

Kutser, Tiit. "Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters." International Journal of Remote Sensing 30, no. 17 (2009): 4401-4425.

Kutser, Tiit. "Quantitative detection of chlorophyll in cyanobacterial blooms by satellite remote sensing." Limnology and Oceanography 49, no. 6 (2004): 2179-2189.

Liaw, Andy, and Matthew Wiener. "Classification and regression by randomForest." R news 2, no. 3 (2002): 18-22.

Lunetta, Ross S., Blake A. Schaeffer, Richard P. Stumpf, Darryl Keith, Scott A. Jacobs, and Mark S. Murphy. "Evaluation of cyanobacteria cell count detection derived from MERIS imagery across the eastern USA." Remote Sensing of Environment 157 (2015): 24-34.

Matthews, Mark William, and Daniel Odermatt. "Improved algorithm for routine monitoring of cyanobacteria and eutrophication in inland and near-coastal waters." Remote Sensing of Environment 156 (2015): 374-382.

Matthews, Mark W., and Stewart Bernard. "Eutrophication and cyanobacteria in South Africa's standing water bodies: A view from space." South African Journal of Science 111, no. 5-6 (2015): 1-8.

Matthews, Mark W., Stewart Bernard, and Kevin Winter. "Remote sensing of cyanobacteria-dominant algal blooms and water quality parameters in Zeekoevlei, a small hypertrophic lake, using MERIS." Remote Sensing of Environment 114, no. 9 (2010): 2070-2087.

Matthews, Mark William, Stewart Bernard, and Lisl Robertson. "An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters." Remote Sensing of Environment 124 (2012): 637-652.

Mishra, S., and D. R. Mishra. "A novel remote sensing algorithm to quantify phycocyanin in cyanobacterial algal blooms." Environmental Research Letters 9, no. 11 (2014): 114003.

Mishra, Sachidananda, Deepak R. Mishra, Zhongping Lee, and Craig S. Tucker. "Quantifying cyanobacterial phycocyanin concentration in turbid productive waters: A quasi-analytical approach." Remote Sensing of Environment 133 (2013): 141-151.

Moses, Wesley J., Anatoly A. Gitelson, Sergey Berdnikov, and V. Povazhnyy. "Estimation of chlorophyll-a concentration in case II waters using MODIS and MERIS data—successes and challenges." Environmental Research Letters 4, no. 4 (2009): 045005.

Mutanga, Onisimo, Elhadi Adam, and Moses Azong Cho. "High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm." International Journal of Applied Earth Observation and Geoinformation 18 (2012): 399-406.

Na, Xiaodong, Shuqing Zhang, Xiaofeng Li, Huan Yu, and Chunyue Liu. "Improved land cover mapping using random forests combined with landsat thematic mapper imagery and ancillary geographic data." Photogrammetric Engineering & Remote Sensing 76, no. 7 (2010): 833-840.

Olden, Julian D., Joshua J. Lawler, and N. LeRoy Poff. "Machine learning methods without tears: a primer for ecologists." The Quarterly review of biology 83, no. 2 (2008): 171-193.

Powell, Scott L., Warren B. Cohen, Sean P. Healey, Robert E. Kennedy, Gretchen G. Moisen, Kenneth B. Pierce, and Janet L. Ohmann. "Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches." Remote Sensing of Environment 114, no. 5 (2010): 1053-1068.

R Core Team. "R: A language and environment for statistical computing. R Foundation for Statistical Computing." Vienna, Austria, www. R-project, org(2013).

Randolph, Kaylan, Jeff Wilson, Lenore Tedesco, Lin Li, D. Lani Pascual, and Emmanuel Soyeux. "Hyperspectral remote sensing of cyanobacteria in turbid productive water using optically active pigments, chlorophyll a and phycocyanin." Remote Sensing of Environment 112, no. 11 (2008): 4009-4019.

Rodriguez-Galiano, Victor Francisco, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. "An assessment of the effectiveness of a random forest classifier for land-cover classification."ISPRS Journal of Photogrammetry and Remote Sensing 67 (2012): 93-104.

Ruiz-Verdú, Antonio, Stefan GH Simis, Caridad de Hoyos, Herman J. Gons, and Ramón Peña-Martínez. "An evaluation of algorithms for the remote sensing of cyanobacterial biomass." Remote Sensing of Environment 112, no. 11 (2008): 3996-4008.

Schalles, John F., and Yosef Z. Yacobi. "Remote detection and seasonal patterns of phycocyanin, carotenoid and chlorophyll pigments in eutrophic waters." Ergebnisse Der Limnologie 55 (2000): 153-168.

Schiller, Helmut, and Roland Doerffer. "Neural network for emulation of an inverse model operational derivation of Case II water properties from MERIS data." International journal of remote sensing 20, no. 9 (1999): 1735-1746.

Shutler, Jamie D., Keith Davidson, Peter I. Miller, Sarah C. Swan, Michael G. Grant, and Eileen Bresnan. "An adaptive approach to detect high-biomass algal blooms from EO chlorophyll-a data in support of harmful algal bloom monitoring." Remote Sensing Letters 3, no. 2 (2012): 101-110.

Simis, Stefan GH, Steef WM Peters, and Herman J. Gons. "Remote sensing of the cyanobacterial pigment phycocyanin in turbid inland water." Limnology and Oceanography 50, no. 1 (2005): 237-245.

Tebbs, E. J., J. J. Remedios, and D. M. Harper. "Remote sensing of chlorophyll-a as a measure of cyanobacterial biomass in Lake Bogoria, a hypertrophic, saline–alkaline, flamingo lake, using Landsat ETM+." Remote Sensing of Environment 135 (2013): 92-106.

US Army Corps of Engineers (USACE) Louisville District. Accessed March 28, 2016. http://www.lrl.usace.army.mil/Missions/CivilWorks/WaterInformation/HABs.aspx.

USGS. Accessed March 29, 2016. http://landsat.usgs.gov/band_designations_landsat_satellites.php.

Vincent, Robert K., Xiaoming Qin, R. Michael L. McKay, Jeffrey Miner, Kevin Czajkowski, Jeffrey Savino, and Thomas Bridgeman. "Phycocyanin detection from LANDSAT TM data for mapping cyanobacterial blooms in Lake Erie."Remote Sensing of Environment 89, no. 3 (2004): 381-392.

Vincenzi, Simone, Matteo Zucchetta, Piero Franzoi, Michele Pellizzato, Fabio Pranovi, Giulio A. De Leo, and Patrizia Torricelli. "Application of a Random Forest algorithm to predict spatial distribution of the potential yield of Ruditapes philippinarum in the Venice lagoon, Italy." Ecological Modelling222, no. 8 (2011): 1471-1478.

Warner, Robert A., and Chunlei Fan. "Optical spectra of phytoplankton cultures for remote sensing applications: Focus on harmful algal blooms."International Journal of Environmental Science and Development 4, no. 2 (2013): 94.

Wiesmeier, Martin, Frauke Barthold, Benjamin Blank, and Ingrid Kögel-Knabner. "Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem." Plant and soil 340, no. 1-2 (2011): 7-24.

Wynne, T. T., R. P. Stumpf, M. C. Tomlinson, R. A. Warner, P. A. Tester, J. Dyble, and G. L. Fahnenstiel. "Relating spectral shape to cyanobacterial blooms in the Laurentian Great Lakes." International Journal of Remote Sensing 29, no. 12 (2008): 3665-3672.

Wynne, Timothy T., Richard P. Stumpf, Michelle C. Tomlinson, and Julianne Dyble. "Characterizing a cyanobacterial bloom in western Lake Erie using satellite imagery and meteorological data." Limnology and Oceanography 55, no. 5 (2010): 2025-2036.

VITA

Jing Huang received her B.S. in Geography Education from Qufu Normal University in 2009, and M.S. in Physical Geography from East China Normal University in 2012. Currently, she is a Master student majoring in Geography in the Department of Geography and Anthropology at Louisiana State University, and plans to graduate in August 2016.