A GIS-BASED EXPERT SYSTEM TO IMPROVE THE ACCURACY OF WETLAND CLASSIFICATION

by

Jing Deng

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Geography

Charlotte

2020

Approved by:

____
Dr. Eric Delmelle

____
Dr. Gang Chen

____
Dr. Craig Allan

____
Dr. Craig Depken, II

ABSTRACT

JING DENG. A GIS-based expert system to improve the accuracy of wetland classification. (Under the direction of Dr. Eric Delmelle)

Wetlands play a critical role in our natural environment, such as improving water quality, controlling erosion and flooding, and protecting biodiversity. To better protect wetland systems, a comprehensive knowledge of their spatial distribution is important to minimize potentially devastating impacts and help with improving wetland function. For instance, accurate wetland delineation guides the mitigation plan in the transportation and construction work to protect wetlands as the US 1970 National Environmental Policy Act (NEPA) requires.

Determining the precise location and extent of wetlands across large-scale regions requires a substantial amount of fieldwork. Therefore, automatic image classification with the support of remote sensing data has become a trend in studying wetland distribution. Current wetland classification studies leverage statistical or machine learning methods to build spatial models based upon the training dataset. They apply these models to predict the occurrence of wetlands, which can later be evaluated through actual fieldwork. However, current studies often face challenges introduced by the data quality. For example, the process of collecting data may introduce inaccuracy and the samples may not reflect the characteristics of the objective region. These factors have the potential to bias the identification of boundaries among different wetland types.

Therefore, a flexible framework that can take into consideration the data quality and produce promising results for different scenarios is necessary.

This dissertation focuses on the development of such a wetland type classification framework that can predict the spatial distribution of different types of wetlands in North Carolina. The overall objective is to build a robust and reliable expert system that can accurately classify wetland types, using training datasets of various quality. To be more specific, this system should be able to tolerate unbalanced and less representative data samples in the training data.

To improve the quality of the classification model, I use various data sources to generate detailed topographic information, such as high-resolution Light Detection and Ranging (LiDAR) data, satellite data, and soil data. I also develop an integrated method to combine advantages of different models and compensate for unbalanced and limited data samples. Lastly, I construct a GIS-based orchestration system to facilitate the replication of the modelling process in a different region.

Leveraging this framework, I conduct different experiments to test the model performance responding to various sampling conditions. The results reveal that the machine learning based methods mainly rely on the quality of the data over the quantity. Under a representative distribution, a sampling data set using five percent of the population proves as accurate as a sampling data set using eighty percent. In the opposite scenario, the proposed integrated method can produce better prediction accuracy than any individual model.

DEDICATION

To my parents Mr. Chuanmei Deng and Mrs. Weiping Yin for their unconditional love and full support.

ACKNOWLEDGEMENTS

There are many people I would like to thank for their contributions and support. Specifically, I would like to thank three groups of people, without whom my research would not have been possible: my advisors and professors in the Geography Department at the University of North Carolina at Charlotte (UNCC), my funding agency and financial support, and my friends and family.

First, I would like to thank my previous advisor, Wenwu Tang, who provided me such a great opportunity to begin my Ph.D. study here at UNCC. Wenwu taught me a lot of fundamental skills for conducting research in the area of spatial modeling and high-performance computing. This knowledge has greatly benefited me throughout my time in graduate school and will continuously benefit my future professional career. I also greatly appreciate my advisor, Eric Delmelle, for his continuous support towards my research interest: wetland classification through machine learning. We have collaborated on several research papers and I have learned a great amount from him. He showed me the level of detailed work and continuous effort required for transforming research into publications. Another professor I would like to thank is Sheng-Guo Wang from the Engineering Department. I have worked with him on wetland projects as a research assistant, and this experience influenced my choice for this dissertation topic. I was also inspired by Dr. Wang’s passionate attitude towards research, and his creative ways of thinking. He always encourages me to think outside of the box and describe the idea using concise mathematical language. I also would like to thank the rest of my dissertation committee members besides my advisors for their great support and valuable suggestions. To summarize, I thank all the great professors in UNC Charlotte that provided me excellent support in my research and inspired me in my academic, professional, and personal life.

Next, I would like to thank the NCDOT Research and Development Unit, which supported this work through the NCDOT Research Grants RP2016-16 and RP2016-19. Thanks to UNC Charlotte and the Geography Department for providing financial support for my research. I would also like to thank Sandy and Scott at Axiom Environmental Inc., for their expertise in wetlands and support conducting ground truthing field work.

Lastly, I would like to express my deepest gratitude to my friends and family. I would like to thank Shanshan Jiang, Wenpeng, Ran, Alex, Adam, Mike, Meijuan, Huifang, and Jiyang. I remember those days of going to classes together, discussing research questions and helping each other out on academic challenges. I will greatly miss those fun days and those hours we spent together. I thank my dear parents, who give me endless support and warm love all the time, even though I have stayed so far away from home and could not visit them often. I deeply appreciate my husband Josiah, your love gives me courage to keep going. Without all your support, this dissertation will not be possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

GIS	Geographic Information System
RS	Remote Sensing
GLM	Generalized Linear Model
RF	Random Forest
GBM	Gradient Boosting Machine
ML	Machine Learning
ES	Expert System
NWI	National Wetland Inventory
WAM	Wetland Assessment Manual
LiDAR	Light Detection and Ranging
NC	North Carolina

CHAPTER 1: INTRODUCTION

1.1 Background

Wetlands are transitional areas between upland and aquatic systems; water has periodically or continuously saturated the soil and vegetation within and therefore, wetland vegetation has adapted to the saturated environment (EPA 2012). According to the Cowardin classification system of wetlands, the definition of wetland is (Cowardin et al. 1979):

> *"…lands transitional between terrestrial and aquatic systems where the water table is usually at or near the surface of the land is covered by shallow water. For purposes of this classification wetlands must have one or more of the following three attributes: (1) at least periodically, the land supports predominantly hydrophytes; (2) the substrate is predominantly undrained hydric soil; and (3) the substrate is nonsoil and is saturated within water or covered by shallow water at some time during the growing season of each year."*

In the United States Code (16 U.S.C., Section 380[a] [18]), the term "wetlands" refers to:

> *"…land that (A) has a predominance of hydric soils, (B) is inundated or saturated by surface or groundwater at a frequency and duration sufficient to support a prevalence of hydrophytic vegetation typically adapted for life in saturated soil conditions and (C) under normal circumstances supports a prevalence of such vegetation."*

According to the National Wetlands Working Group of Canada (1988), a wetland is an area that is:

> *"...saturated with water long enough to promote wetland or aquatic processes as indicated by poorly drained soils, hydrophytic vegetation, and various kinds of biological activity that are adapted to a wet environment."*

These definitions emphasize the physical components of wetlands, including hydrology, soil, and vegetation. Wetlands are protected by public laws and regulations due to their important ecological functions benefiting the natural environment, namely carbon sequestration (Gorham 1991), flood mitigation (Olhan et al. 2010), wildlife habitat (Kennedy and Mayer 2002), water quality improvement (Keddy 2010), and biodiversity maintenance (Li and Chen 2005). For example, the National Environmental Policy Act (NEPA 1970) of the United States regulates activities that lead to possible long-term and short-term impacts associated with the destruction of wetlands. Mitigation activities should compensate wetland loss or degradation due to human construction and development (Boyd 2002).

The process of wetland inventory essentially maps the distribution of wetlands and monitors the dynamics of wetland systems, which represent significant health indicators for the natural environment (Hess et al. 2003). Institutions can construct wetland inventories to monitor the wetland resources for local regions or larger areas. In the United States, government agencies at the federal level have produced wetland maps for many years. One such example is the U.S. Fish and Wildlife Service (USFWS), which has maintained the National Wetland Inventory (NWI) since the mid-1970s. The USFWS most recently updated its dataset in 2016, including geospatial datasets of wetlands and surface water. In these datasets, they classified wetlands under the national standard—the Cowardin classification system—which provides consistent ecological descriptors (Cowardin et al. 1979).

Trained analysts conducted the classification and mapping practices for the NWI by visually interpreting the presence or absence of wetlands based on high-altitude aerial photography overlaid with other data such as contour maps and land use maps. Although the NWI greatly benefits the general public, it proves inadequate for regional or site-specific applications for several reasons. First, the wetland classification system in the NWI operates at the national level and may overlook the characteristics of the local environment. Second, given the coarse resolution of the NWI, analysts may miss small wetlands. Third, the manual mapping process is time consuming and—many times unintentionally—analysts may bias or influence the results. For instance, they cannot distinguish the wetlands under dense tree canopies in remote sensing images and may make incorrect decisions based on prior knowledge. Due to these reasons, the NWI suffers from several issues including low accuracy in terms of wetland location and type classification. More importantly, this manual mapping is a one-time evaluation approach, which has no ability to provide additional information of new environmental changes. Therefore, the map is incongruous with ever-changing wetlands, resulting in data becoming outdated for many applications (Melloh et al. 1999).

To identify the wetland types that occur in North Carolina and assess their corresponding functions, an interagency team composed of federal and state staff was established in 2003 (N.C. Wetland Functional Assessment Team [WFAT]). Specifically, the WFAT set the objectives to develop a scientific and systematic process for classifying and evaluating NC wetlands. WFAT identified 16 general wetland types within a two-level structure wetland system (see Table 1). The trained field assessors then follow a written manual—the North Carolina Wetland Assessment Manual (NCWAM)—to conduct the classification process. The fifth version of the manual was recently updated in 2016.

Table 1. NCWAM wetland types (N.C. Wetland Functional Assessment Team 2010)

Level I Category	**Wetland Type**
Salt/Brackish Marsh	Salt/Brackish Marsh
Riparian	Estuarine Woody Wetland
	Tidal Freshwater Marsh
	Riverine Swamp Forest
	Bog
	Non-Tidal Freshwater Marsh
	Floodplain Pool
	Headwater Forest
	Bottomland Hardwood Forest
Non-riparian	Seep
	Hardwood Flat
	Non-Riverine Swamp Forest
	Pocosin
	Pine Savanna
	Pine Flat
	Basin Wetland

1.2 Problem Statement

Wetland delineation and type classification generally require a consequential amount of fieldwork. Experts need to visit specific sites, collect soil samples, check the hydrologic conditions, and observe the type of vegetation community present at the sites to determine the wetland type. While this process may be accurate, it reveals substantial challenges. First, some areas may be extremely difficult to access. Second, the process is time consuming; wetland delineation of a large-scale region is nearly impossible.

Given these constraints, Remote Sensing (RS) technologies present a natural solution for efficiently acquiring information on the earth surface at a larger scale. In addition, Geographic Information System (GIS) provides methods to further analyze this data for wetland delineation purposes. This approach can partially, or in some instances, completely replace fieldwork. Existing studies have documented success in wetland classification supported by RS and GIS. The commonly used workflow includes analyzing relationships between environmental characteristics and wetland occurrence based on existing field sampling data, and then extrapolating the relationships to other regions to predict wetland occurrence (Rebelo et al. 2009). However, this workflow is very sensitive to the quality of the sample data and it usually can only apply to a small region in order to maintain high classification accuracy. This limitation will hinder studies and decision making in grand-scale applications. For instance, people who are planning for the construction of new state highways may not be able to avoid wetlands without understanding their accurate distributions.

The objective of this research is therefore to improve current RS-based wetland classification workflow to build an efficient and reliable wetland classification framework. I apply GIS technology to extract critical information of topography, hydrology, soil, and vegetation, mainly from LiDAR data. I then leverage machine learning algorithms to mine and model the decision rules for identifying wetland types through a supervised training process. I also improve model accuracy by incorporating expert knowledge (e.g., soil type for supporting wetlands formation) in the models (N.C. Wetland Functional Assessment Team 2010).

1.3 Research Questions

This research attempts to address the following questions:

1. How might one efficiently extract environmental information from high resolution LiDAR data for detailed wetland types identification?
2. How might one leverage various data sources and classification techniques to improve wetland prediction accuracy?
3. How might one build a flexible classification system that can produce reliable classification results and can be easily applied to a different region?

By answering these questions, this research will facilitate the automatic wetland prediction process and improve the accuracy of wetland type classification. The following tasks are necessary to achieve the aim of the study:

1. Investigate environmental characteristics corresponding to different types of wetland in North Carolina and represent them as spatial data layers through GIS.
2. Examine and evaluate the strengths and merits of several commonly used machine learning algorithms in wetland classification.
3. Construct a GIS-based expert system to integrate different modeling approaches in order to better support wetland modeling.

1.4 Significance and Contributions

My dissertation mainly contributes to the field of Geography in two aspects. First, it contributes to spatial modeling methodology by integrating and merging two different types of modeling philosophies, deductive and inductive modeling. I conduct theoretical analysis by reviewing several of the most commonly-used models in wetland classification studies. I then summarize the benefits and limitations of those models. For example, statistical models are constrained by many assumptions, and machine-learning models neglect the reasoning process behind the data. I also conduct experiments to illustrate the different performance of these models under "near-real-world" situations. According to the experiment results, the new modeling method proposed in this dissertation can provide a better solution in the real-world application.

The second contribution is to the understanding of "model performance" in the spatial modeling field, which is often represented by the notion of "accuracy". Current studies evaluate model performance by dividing the entire dataset into multiple subsets, and cross-validating each dataset through the calculation of accuracy. I argue that the spatial distribution of the sampling data will greatly affect the outcome of model validation, which has not been widely considered in existing literatures. These studies tend to over-estimate model performance when using stratified sampling method to create training and testing datasets which probably end up representing similar distribution pattern. In this dissertation, I introduce methods to avoid the high similarity between test and training datasets thus to generate more objective evaluations for the model performance.

In terms of the specific wetland classification models, the contribution lies in several aspects:

1. Introduction of a classification system for detailed NC wetland types based on North Carolina Wetland Assessment Manual (NCWAM) definitions.
2. Integration of various data sources (under different resolutions and data types) to derive as much information as possible from different perspectives, such as the surface micro-topography from LiDAR, vegetation structure, soil, land use, and hydrology information.
3. Integration of different modeling methods to construct the core of the expert system—the knowledge base.
4. Improvement of the efficiency of the classification workflow to increase the flexibility for applying it to different datasets or different study areas.
5. Support of scenario analysis for the decision-making process to conduct wetland protection practice, such as wetland mitigation and restoration.

CHAPTER 2: LITERATURE REVIEW

The traditional and most accurate way for wetland delineation is to send pre-trained assessors to corresponding locations to conduct fieldwork. Such an approach is time-consuming and lacks consistency due to the difference in operational preferences of the assessors. For instance, the results may differ based on different standards and methods used by the assessors, or due to the season and time that the surveys take place, where climate and environment change considerably (Milton and Hélie 2003, Cihlar et al. 2000). Instead, related studies have considered remote sensing technology and computer-based classification methods to be a more practical and applicable method for wetland delineation, classification, and management for relatively large areas.

Remote sensing technologies provide a synoptic view of the earth. Image analysis can be conducted to extract features for identifying wetland types. This approach enables wetland assessors to work remotely, alleviating the constraints of location accessibility. There exist different methods to process collected datasets, such as manual interpretation, semi-automation or complete automation. For instance, the National Wetland Inventory (NWI) in the U.S. is a product of manual interpretation, which uses color-infrared aerial photography and black-and-white photography (Tiner 1990). Many wetland identification and mapping studies prefer the manual interpretation method since it directly incorporates expert knowledge and experience (Sohl et al. 2004). However, one can question the efficiency and uncertainty for this method. A myriad of studies has focused on the development of methods that facilitate the automatic detection of wetlands (Jean and Bouchard 1991, Jensen et al. 1995, Munyati 2000).

In this chapter, I conduct a literature review related to the development of this automated technique that supports wetland detection and classification modeling through GIS and remote sensing. The literature review focuses on three aspects: types of wetlands studied by current literature; spatial datasets which serve as the foundation of the modeling; spatial modelling methods that help with detecting wetland distributions.

2.1 Targeted Wetland Types

According to the National Wetland Inventory (NWI), the most common wetland classes include emergent, forested, and scrub/shrub wetlands (Corcoran et al. 2013). For more detailed wetland types, Ozesmi and Bauer (2002) summarized the tasks in the order of increasing difficulty level: water, marshes and swamps, deciduous forest wetlands, evergreen forested wetlands, and scrub-shrub wetlands. The wetness level in wetlands plays an important role since water reveals very distinct spectral reflectance (Butera 1983, Henderson and Lewis 2008). Wetland classification can be difficult if different types have similar wetness level, such as swamps and marshes. Therefore, other features such as the occurrence of a characteristic vegetation community becomes a key factor for distinguishing swamp and marsh.

Different vegetation species have different impacts on backscatter characteristics; further study of this relationship can help with the identification of different vegetation types. The complexity has been identified in terms of classifying forest wetlands and scrub-shrub wetlands, as the optical signatures for these wetlands reveal similar features (Reese et al. 2002, MacAlister and Mahaxay 2009).

2.1.1 Coastal and Riverine Wetlands

A large number of studies have focused on coastal and riverine wetlands, including tidal marshes, swamps, and mangroves, some of which also belong to the forest wetland category. For this type of wetland, the hydrogeomorphic condition is a key detector (Hupp and Osterkamp 1996)(Butera 1983) .

Studies applied spectral signatures to identify water regimes; however, it may fail to separate terrestrial water regimes within the wet regions due to the massive similarity of spectral features (Augusteijn and Warrender 1998). Brisco et al. (2011) applied C-band radar imagery to clearly separate open-water, deep-marsh, and shallow-marsh areas based on the scatter plots of different polarimetric decomposition parameters. Furthermore, the application of multi-temporal data can improve the classification accuracy, since the combination of data collected at different seasons (such as winter and spring) can help exclude the impacts of climate change and distinguish different plant species, such as emergent vegetation and floating vegetation (Ramsey III and Laine 1997).

There exists significant confusion when classifying swamps and marshes using radar data (Grenier et al. 2007, Gosselin et al. 2014). Instead of focusing on water regimes, one can also extract spectral characteristics aiming at distinguishing vegetation.

Several studies have analyzed the differences in leaf spectral reflectance to support wetland type classification, especially for marshes and swamps (Anderson and Perry 1996, Spanglet et al. 1998). There are also applications to extract the information of vegetation from multiple spectral bands to identify mires (Bronge and Näslund-Landenmark 2002, Bronge 1999). However, it is noted that there may exist a high correlation among different bands. To choose different band combinations may result in different classification performance.

2.1.2 Forested Wetlands

Forested wetland type poses a unique mapping problem due to the canopy coverage and similarity among trees (Augusteijn and Warrender 1998). Several studies attempted to integrate the identification of vegetation type with analyzing hydrology condition under the canopy to improve the classification accuracy for forest wetlands. Hess et al. (1990) reviewed studies applying radar for detecting flooding on forested floodplains from 1971 to 1990, highlighting the capability of using radar polarimetry to investigate the scattering matrix of different forest types, and summarized the useful bands for identifying them. Franklin (1991) documented the improvement of forest classification by analyzing topographic environments and hydrological characteristics. Researchers applied more complex techniques and more various data sources to tackle these challenges. For instance, Augusteijn and Warrender (1998) built neural network classifiers based on radar data together with multi-spectral datasets. They classified different levels of wetness in forested wetland due to the rich spectral information provided by 15 channels in the visible, near-infrared, mid-infrared, and thermal-infrared spectrums.

2.1.3. Scrub/Shrub Wetlands

The challenges for distinguishing scrub-shrub wetlands are similar to forested wetlands. It is even harder to identify this type of wetland for several reasons. First is the complexity of the ecosystem for scrub and shrub wetlands. They may coexist with other types of wetlands, such as marshland with shrub patches, or co-occur under forest tree canopies, which will give the same optical response as forest wetlands. Therefore, traditional optical images and spectral information are no longer sufficient to discern this type of wetland. Second, scrub/shrub wetlands vary dramatically, in terms of environmental conditions, such as vegetation species and hydrologic conditions. Two scrub/shrub wetland sites can appear very differently on the remote sensing image. Therefore, research on vegetation type identification can benefit the classification of certain wetland types. Zhang (2014) combined hyperspectral data with LiDAR data to increase the accuracy of mapping diverse vegetation in complex wetlands. Third, the size for this type of wetland is relatively small, thus the detection is greatly limited by the spatial resolution of the remote sensing data. To tackle this resolution limitation, several studies provided potential solutions by researching mixed pixels (Hurd et al. 2006). For instance, they applied approaches such as spectral unmixing to discern the fractional composition of wetlands in the classification process (Rogers and Kearney 2004, Wei et al. 2008).

To summarize the literature on wetland types, one will mostly use spectral information to detect and classify different types of wetlands. Studies have also applied multi-spectral data, hyperspectral data, and radar data to extract spectral signatures for different vegetation species. Furthermore, LiDAR data have provided the possibility to detect vegetation species by analyzing vertical structural characteristics. However, one should keep in mind the limitation introduced by the resolution of the data while using the spatial data to analyze the ground features. For example, Wang et al. (1998) observed that, in the classification of multiple marsh wetland types, the classification accuracy did not keep increasing past a certain level as the number of SAR channels increased.

2.2 Data Used in Wetland Classification

Spatial data and information collected by remote sensing techniques play a beneficial role in wetland classification, due to their broad spatial coverage of earth surface and the timely manner of results production (Ozesmi and Bauer 2002). Identifying wetland types based on remote sensing data boils down to categorizing image pixels based on their dissimilarity in features. For instance, on photography images, the same type of ground surface objects reveals similar optical features: vegetation is green, water is dark, etc. Manual identification relies on the visual interpretation of the images and this often represents a large investment of time, labor, and expense (Lunetta and Balogh 1999). The semi-automatic or automatic methods refer to processing images in a batch through software programs. This approach does not require the same level of expert knowledge, where researchers can explicitly program the standards and rules into the classifiers. However, the automatic procedure faces difficulties in taking regional environmental differences, such as topographic and hydrologic conditions, into consideration and adjusting the methods accordingly. The performance of both manual and automatic methods relies significantly on the quality of the remote sensing data.

There exist two major types of sensors based on the type of signal used by the remote sensing platform. Passive sensors respond to external stimuli (e.g., sunlight) while active sensors leverage internal stimuli, such as laser beams, to record information. The platforms carry sensors include aircraft and satellites, characterized by different flying altitudes and cycles. Satellite platforms cost less time and resources to gather data for large aerial extents than airborne platforms. However, they cannot provide information as detailed as aerial photography (Ozesmi and Bauer 2002).

2.2.1 Data from Passive Remote Sensing

The community of wetland researchers first applied optical remote sensing (e.g., aerial photography) for wetland classification. They used visual interpretation method on the aerial images to map wetland spatial distributions in the National Wetland Inventory project (Anderson and Wobber 1973). With the launch of the Landsat satellite series in the early 1970s, more studies increasingly adopted data from related products like the Multispectral Scanner System (MSS) to identify wetlands. The MSS sensor can detect light in both the visible and near-infrared (NIR) spectrum (Lyon et al. 2001). NIR has proven particularly useful in detecting wetland vegetation due to the strong relationship between NIR reflectance and the biomass of photosynthetic tissue (Jensen et al. 1984, Hardisky et al. 1986). However, the coarse resolution of MSS data (80 meters) limited the accuracy of wetland classification in large study regions.

The advancement of remote sensing techniques introduced data with finer resolutions. For instance, Landsat Thematic Mapper (TM) data can provide more accurate information due to finer spatial resolution (30 m) and additional information in the mid-infrared and thermal-infrared bands. The use of Landsat TM data has improved the accuracy of wetland classification (Reese et al. 2002, Sader et al. 1995). The *Systeme Pour l'Observation de la Terre* (SPOT), an earth resource satellite launched by the French government in 1986, uses a high resolution visible (HRV) system composed of green, red, and near-infrared spectral bands at a 20m spatial resolution. This can provide better information for detecting marsh wetlands than the TM images (Hardisky et al. 1986, Marceau et al. 1990).

Landsat MSS, TM, and SPOT represent the most commonly used satellite systems for wetland classification. Repeat coverage and updated information greatly facilitate detailed classification of wetlands and improved model accuracy (Wright and Gallant 2007). Other than these common platforms, wetland studies also applied some other passive sensors to collect imagery data. For instance, the NOAA satellite with Advanced Very High-Resolution Radiometer (AVHRR) instruments can observe night cloud patterns, sea surface temperatures, and terrestrial vegetation. One can use this type of data to identify vegetation, distinguish wetland types, and estimate the biomass of forested wetlands (Moreau et al. 2003, Llewellyn et al. 1996). However, the coarse spatial resolution of 1.09 km makes it unsuitable for identifying small wetlands patches.

The Indian Remote Sensing Satellite (IRS), started by India in the early 1980s, represents another source of multispectral satellite imagery applied to wetland identification. It targeted several natural resources, such as agriculture, water, forestry, and geology (Ozesmi and Bauer 2002). The IRS-1B Linear Imaging Self-scanning Sensor (LISS-II) has four bands similar to Landsat TM bands—blue, green, red, and the near-infrared band of the spectrum—at a spatial resolution of 36.5 meters. Kindscher et al. (1997) applied IRS-1B LISS-II data to analyze the spectral reflectance of meadows, thus to identify wetlands based on the percentage of wetland plant species. It emphasizes the capability of remote sensing data for identifying wetland vegetation communities. Chopra et al. (2001) visually interpreted land cover categories of wetland ecosystems using IRS data (including 1A LISS-I, 1A LISS-II, and 1B LISS-II). The authors obtained the environmental information from multiple perspectives through the multi-date and multi-season data, such as seasonal variation of water and vegetation. However, the spectral overlap between different land-use types brought challenges to achieve more accurate classification results.

Spanning over a few years, Adam et al. (2010) further illustrated the trend of using multispectral and hyperspectral data. The European Space Agency (ESA) launched the Sentinel-2 mission in the European Copernicus program as the follow-up for the Landsat. Sentinel-2 provides 13 spectral bands and various resolutions from 10 to 60 meters. It yields 5 days between revisit, while the Landsat-7 has 16 days and SPOT enables 26 days for revisit (Drusch et al. 2012). Studies have applied Sentinel-2 data to generate a vegetation index to serve as training data for object-oriented classification in wetland mapping (Kaplan and Avdan 2018). Hyperspectral data contain a large number of relatively narrow spectral bands, either visible or infrared. Hyperspectral remote sensing data hold particularly value for wetland vegetation mapping due to their rich radiometric content. However, high dimensionality also introduces massive amounts of redundant spectral information (Thenkabail et al. 2016, Adam et al. 2010, Zhang and Xie 2014). The authors cited above generally observed that the increased use of spectral channels offer the potential to improve classification accuracy. However, with the introduction of more channels, classification accuracy first increases to a peak and then starts to decrease—a pattern referred to as the *Hughes Phenomenon* (Hughes 1968). Studies have applied feature extraction, such as the Principal Component Analysis and Minimum Noise Fraction method, to maintain the most representative information (Zhang et al. 2007, Zhang 2014).

With the aforementioned satellite data, cycles may repeat on a relatively long scale, however, slowing the frequency of wetland classification map updates and affecting the monitoring of wetland dynamics. This constitutes an important limitation. Furthermore, the atmospheric constituents, such as clouds or canopy, can greatly affect the quality of the imagery products (Richards and Richards 1999).

2.2.2 Active Remote Sensing Data

The active remote sensing technique applies its own source to emit energy and record reflected energy instead of relying on natural emissions. Examples of active remote sensing include satellite radar imaging systems and Light Detection and Ranging

(LiDAR). Radar uses electromagnetic energy in the radio frequency range (microwave) while LiDAR uses much shorter wavelengths (visible and near-infrared). Compared to passive remote sensing techniques, active remote sensing has three major advantages. First, it can collect the data at almost any time of day since the weather affects them less. Second, it can easily detect ground surface information in areas covered by tree canopies. Radar transmits signals with a longer wavelength that will have greater penetration in forest canopies. LiDAR can "see around" trees, since the high density of laser beams can penetrate through the physical gaps of the canopies. Third, these data can provide supplementary information other than spectral information (Ozesmi and Bauer 2002). For instance, radar measures backscatter to detect water and LiDAR can depict the three-dimensional extent of the object.

(1) Radar

Radar applications emerged for studying wetlands in the late 1960s (Waite and MacDonald 1971). Several studies have summarized the knowledge of detecting wetlands using the features revealed in images in terms of dielectric and geometric attributes (Hess et al. 1990, Kasischke and Bourgeau-Chavez 1997, Schmullius and Evans 1997, Ramsey III 1998, Henderson and Lewis 2008). Synthetic aperture radar (SAR) applies microwaves to penetrate canopies and clouds to detect the dielectric properties of the surface. SAR can also detect hydrologic characteristics, such as surface inundation and soil moisture (Henderson and Lewis 2008, Lang et al. 2008). The most often-utilized satellites in wetland detection include the European Remote Sensing

Satellite (ERS-1, 26 m range resolution, launched in 1991), the Japanese Earth Resource Satellite (JERS-1, launched in 1992), and the Canadian RADARSAT Satellite (launched in 1995) (Wang et al. 1998). These radar systems have different parameters in terms of wavelength, polarization, spatial resolution and so on. SAR remote sensing serves as an important tool for monitoring surface water due to its sensitivity to water-related backscatter. Researchers find it especially useful to detect forest flooding where the area appears bright in the images due to the double-bounced reflections from tree trunks and water surfaces. However, other situations—such as slopes of hills, rough surfaces, plowed rows in agricultural lands, buildings, and streets—can also generate corner reflections and cause confusions (Hess et al. 1990).

Space-borne imaging radar, such as ERS-1, can provide greater ground coverage than air-borne radar. ERS-1 SAR data can detect soil moisture, water presence, and vegetation types (Kasischke and Bourgeau-Chavez 1997). Water under a canopy can increase the radar backscatter for woody wetlands while decreasing the backscatter for herbaceous wetlands. Compared with ERS-1, which uses the shorter C-band sensors, JERS-1 utilizes longer L-band sensors, resulting in better penetration of tree canopies. JERS-1 data makes it easier to distinguish flooded and non-flooded areas than ERS-1 does (Townsend and Walsh 1998). Similarly, RADASAT data have frequently assisted the study of wetland hydrology and forest types (Töyrä et al. 2001, Li et al. 2007). However, the RADARSAT-1 sensor cannot penetrate high-density forest due to the use of C-band sensors with short wavelengths (Townsend 2001).

To summarize, radar remote sensing enhances the efficiency of detecting water presence and certain types of vegetation; however, it requires background knowledge of radar scattering from different types of vegetated surface and interaction with soil (Ghedira et al. 2000). It requires massive efforts in studying the relationship between backscattered energy and the characteristics of observed objects. Gaining complete understanding of wetland complexity in terms of vegetation density and heterogeneity through radar data remains a challenge.

(2) LiDAR

Wetland studies have increasingly applied another active remote sensing technique, Light Detection and Ranging (LiDAR), to model wetland distributions (Goetz 2006). The LiDAR technique obtains data by emitting a large number of laser pulses per second, then records multiple returns per pulse after the light has bounced off the ground objects. Based on the time delay between emission and return, it calculates the distances between objects and measures the dimensions of the objects. The LiDAR platform carries Inertial Measurement Units (IMU) and Geographic Positioning System (GPS), facilitating the generation of geo-referenced points with X, Y, and Z information. Similar to other remote sensing methods, LiDAR provides information for ground surface, though it produces high-resolution data for a large-scale region in a relatively shorter time period at a lower cost. Using LiDAR data, we can further generate various products such as bare-earth DEMs, intensity images, canopy models and building models.

The LiDAR technique provided great support for the high-resolution topographic mapping applications in the late 1990s. The use of LiDAR to assist with wetland mapping has occurred more recently. Automated wetland mapping can greatly benefit from LiDAR-derived DEMs for topographical analysis (Wang et al. 2015, Shaeffer 2008). This fine-resolution topographical information can aid identification of the landscape position of ground objects, such as on a side slope or within a depression. Researchers derived terrain metrics to quantify the landscape pattern and evaluate the possibility of wetland occurrence. For example, they use a wetness index to represent the possibility of a region being wet according to surface hydrological mechanisms and localized depression processes (Tenenbaum et al. 2006). Additionally, the fine resolution of topographical information enables the detection of relatively small wetlands, such as vernal pools, which other types of imagery data might easily omit (Lang et al. 2009, Lichvar et al. 2006).

LiDAR data not only provide elevation and topographical information but can also reveal hydrological conditions under vegetative canopies. Since water strongly absorbs the near-infrared energy that most LiDAR sensors use, one can use the intensity image of bare-earth LiDAR returns to detect inundation. For example, Hofle et al. (2009) analyzed the reflection characteristics of water and surface roughness information using airborne laser scanning systems. They proposed an object-based classification workflow to distinguish water areas based on the high number of laser shot dropouts and the predominantly low backscatter energy.

Another advantage of LiDAR lies in its ability to provide information about the vertical dimension. LiDAR performs well in capturing structural characteristics of ground objects, such as height, biomass, and canopy shape, which can further facilitate the separation of vegetation types (Vierling et al. 2008). Allen et al. (2011) applied the first-return LiDAR points to construct a vegetation canopy model and integrate it with multi-date SAR images for discerning key vegetation classes.

2.2.3 Data Fusion

Remote sensing techniques have greatly facilitated automated wetland classification, especially for relatively large spatial expansions. However, the complexity of the wetland system itself remains a challenge for the pixel-based image classification. These difficulties involve variations of features within a single pixel, including texture characteristics, water saturation in soil, vegetation coverage, and habitat characteristics. Due to the similarity of these features represented in the images, the spectral confusion issue exists when distinguishing wetland types. Under this situation the spectral signatures for specific wetland types show great diversity and overlap with each other, making it a challenge to separate different wetland types on the image.

Both passive and active remote sensing techniques have their merits and shortcomings. To better take advantage of different types of data and overcome their limitations, researchers have used data fusion techniques to improve performance through the integration of data and information acquired from different sources (Klemas et al. 1974, Zhang and Xie 2014, Yang et al. 2009). In wetland classification applications, Augusteijn and Warrender (1998) compared the suitability of different remote sensing datasets in wetland classification, including visible/near-infrared (NIR), thermal-infrared (TIR), and radar. They found that applying these data in isolation yields similar performance for wetland classification, but the combination of these data can achieve better performance. Multiple other studies have drawn similar conclusions about the advantages of data fusion techniques for wetland classification (Li and Chen 2005, Töyrä and Pietroniro 2005, Lichvar et al. 2006, Vierling et al. 2008).

(1) Multi-sources

Previous studies have reported limitations of classification performance based on single-date radar images (Aschbacher et al. 1995, Wang et al. 1998). Aschbacher et al. (1995) compared the classification accuracies of wetland using different sensors and found significantly lower classification accuracy when using radar data alone compared to combining multiple single-dated radar images, or integrating radar data with optical remote sensing images (Augusteijn and Warrender 1998, Allen et al. 2011, Townsend and Walsh 1998, Henderson and Lewis 2008). Therefore, wetland classification applications commonly integrate radar data with other data.

Castañeda and Ducrot (2009) extracted soil surface characteristics related to wetness and roughness from SAR images, which improved Landsat classification by enhancing the contrast of radiometric features among different locations. They applied two fusion methods to their research: (1) concatenating radar channels to Landsat bands, and (2) integrating the classification results. Huang et al. (2010) applied radar data to gain complementary vegetation information. They fused optical and radar to estimate the percent of ground cover of different vegetation types in Yellowstone National Park. However, in their approach, the application of radar data had a negative impact on the classification performance due to the speckle and noise effect in the images (Li and Chen 2005, Chust et al. 2004).

However, both optical and radar data lack sufficient information in terms of terrain characteristics and vegetation structural features. LiDAR data can help with filling this gap by supplying detailed information on terrain and canopy structure (Lefsky et al. 2002, Lim et al. 2003). To classify different savannah tree species, Naidoo et al. (2012) constructed and trained Random Forest models using a hybrid dataset composed of hyperspectral data and LiDAR-derived structural information. Based on the modeling results, the most important variables included structural parameters such as tree height and spectral variables such as NDVI. Similarly, Hill and Thomson (2005) derived thematic classes through integrated LiDAR and spectral data. The classes contain both information of species composition and canopy structure which reflects the underlying ecological processes of woodlands. Geerling et al. (2007) fused spectral image and LiDAR-derived datasets on the pixel level to classify floodplain vegetation. The classification achieved the highest accuracy by using the fused data rather than using spectral or LiDAR data alone, especially for bush and forest lands. These applications of LiDAR represent a step forward in distinguishing structural subdivisions for different vegetation species—also useful in wetland mapping.

The data product generated through the process of analyzing remote sensing images represents another type of major data source for improving wetland classification. The remote sensing images with spectral signatures and texture characteristics can generate such information. For example, a tasseled-cap method can extract information based on three transformed components: greenness, brightness, and wetness. This approach has proven suitable for detecting wetlands, especially forest wetlands at the spring season when the water table is high (Hodgson et al. 1987, Crist and Cicone 1984). Other analyses can produce spectral signatures, such as measuring difference or ratios between different bands.

Other than remote-sensing data, other data sources—such as vegetation mapping, soil distribution, and hydrological conditions—can also significantly support wetland mapping and improve its accuracy by further identifying the heterogeneity among or within wetland types (Bronge and Näslund-Landenmark 2002). For example, environmental datasets can refine classification maps (Bolstad and Lillesand 1992, Sader et al. 1995). Feature dimensions such as vegetation, soil, landform, and bedrock geology can facilitate the construction of classification models (Wright and Gallant 2007).

Table 2 summarizes the type of information that different types of data can provide in supporting wetland classification according to three major aspects: soil, vegetation, and hydrology.

Table 2. Data perspectives in support of wetland classification

Information	Potential Data	Wetland features		
		Soil	Vegetation	Hydrology

Spectral Characteristics	photography images;
	multi-spectral data;
	hyper-spectral data
		level of moisture	identify vegetation type;
			vegetation indices
				detection of flooding and inundation area

Spatial Structure	LiDAR data (intensity, geometry);
	photography image
		-	canopy structure model;
			tree structure model;
			canopy surface model
				surface modeling and hydrological simulation;
				terrain derivatives;
				geometrical analysis;
				intensity image analysis to detect water surface

Ancillary Information	GIS layers;
	spatial database;
	non-spatial datasets
		soil type;
		soil components;
		hydric information
			vegetation type;
			land cover type
				climate data;
				precipitation;
				hydrology dataset

(2) Multi-dates

The rationality of integrating multi-date data has two main components: (1) to generate new wetland maps according to the application needs for mapping frequency, and (2) to compensate for data limitations. Data achieved at different time periods can reveal different information. For example, the data can divulge more under-canopy information during the leaf-off season.

Due to the changes of hydroperiod in space and time, climate and weather change will significantly affect the data acquisition for both remote sensing and field collection. The hydrology and vegetation conditions within wetlands will also vary accordingly (Johnston and Barson 1993). Therefore, one can rely on multi-temporal data to monitor the dynamics of wetland systems. For instance, Lang et al. (2008) applied edge-detection filters to detect North Carolina coastal wetland shoreline change based on multisensory SAR imagery over 12 years.

The integration of multi-date remote sensing data under singular technology can significantly improve classification accuracy. Many scholars have documented the integration of multi-date radar data has for assisting with wetland identification (Shang 1996, Wang et al. 1998, Milne et al. 2000, Sokol et al. 2000). These studies applied false color composite images to enhance spectral details generated through intensity, hue, and saturation (IHS) transformations (Kushwaha et al. 2000).

2.3 Classification Methods for Wetland Identification

Classification methods / modeling serve as the key component for wetland mapping. Using the remote sensing and spatial data as the training inputs, these models can automatically identify wetland distributions through machine-based image processing algorithms. Automatic image classification methods include three categories:

unsupervised, supervised, and hybrid method (Ozesmi and Bauer 2002). This section will summarize the methods that are commonly used in wetland mapping applications.

2.3.1 Unsupervised Methods

The unsupervised method directly groups objects into different classes, in which the objects will share certain common features with each other. This method requires little effort in data training, and the resulting classes are distinct from each other. However, the result of grouping greatly relies on the feature variables used for classification, whether these features represent the distinguishable characteristics between different classes. The commonly used unsupervised methods include the *k*-means approach and the Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA) (Pope et al. 1994). These algorithms adopt different mechanisms for measuring the similarity among different data samples.

To apply the *k*-means approach, one needs to specify the number of classes as a *priori*. This approach then iteratively adjusts the membership of classes by assigning data samples to their nearest group and recalculating the centroids of the groups based on their updated members. Similar to the *k*-means algorithm, ISODATA adopts an iterative method governed by predefined threshold values for parameters. Certain parameters mainly refer to statistics of the classification results, including the average inter-center distance, the maximum number of classes, the standard deviation within each cluster, and the maximum number of iterations (Ball and Hall 1965). However, without well-defined thresholds, the algorithm may not generate appropriate results in a timely manner.

Generally, the unsupervised methods require very little effort from the users, and the algorithms can be effective and efficient with appropriate parameter settings. Furthermore, they perform well for classifying objects that appeared in nature according to their spectral variability in the images, such as distinguishing among different vegetation covers (Kindscher et al. 1997, Allen et al. 2011). Therefore, researchers usually conduct unsupervised methods as a pre-classification step to increase the classification accuracy. For instance, Huang and Jensen (1997) first pre-classified SPOT multispectral data into 50 spectral classes using ISODATA before other classification methods. Principal Component Analysis (PCA) is also used as a pre-processing step to reduce redundancy. Gluck et al. (1996) extracted different principal components from Landsat TM data to highlight different aspects of information, including vegetation differences, wetness differences, and differences between wetlands and uplands. They further employed the ISODATA classification approach, based on the results from the principal components to distinguish more detailed wetland types. Similarly, Lang et al. (2008) isolated the dominant temporal information relative to hydrological period by conducting PCA on multi-temporal SAR data, suggesting significant correlations between the first principal component and the hydrologic features, such as soil moisture.

2.3.2 Supervised Methods

Supervised methods require the users to specify the class labels for the training data in the classification modeling. The models then use a programmatic way to generate classification results based on the input data. Increasing applications applied the supervised methods in wetland classification for the efficiency and accuracy of these methods.

(1) Conventional supervised methods

The conventional supervised classifiers used in wetland classifications mainly include regression models, discriminant function analysis, parallelepiped classifier, and Maximum Likelihood Classification (MLC) (MacAlister and Mahaxay 2009, Munyati 2000, Crawford et al. 1999).

Regression models can quantify the explanatory power of the independent variables using the variance of the dependent variable. In wetland classification, the dependent variable can present as a binary form, representing wetland occurrence (Toner and Keddy 1997). Similar method prior to the regression model was known as discriminant analysis (Harper and Ross 1983). Franklin et al. (1994) applied discriminant function analysis to distinguish kalmia (shrub) from other land cover types and wetland vegetation. They compared the classification results of discriminant analysis based on different data sources, reaching an overall accuracy of 96% using aerial remote sensing and 85% using satellite data. In a regression model, the coefficients of the variables indicate their contributions in explaining the dependent variable. For example, Shaeffer (2008) constructed regression models using a set of various DEM-derivatives to predict the occurrence of wetlands while identifying the significant impacting factors.

The parallelepiped classifier often fits the applications that use multispectral data; it classifies the data samples based on the definition of class boundaries. For each dimension, the boundary represents a standard deviation surrounding the mean value of the defined class. If the value of a pixel falls within the range for all the spectral bands of the class, it will be assigned to that class. The parallelepiped classifier computes fast, especially when combined with a table look-up scheme. However, it lacks the capability to distinguish among spectral signatures that are naturally similar (Hines et al. 1992).

The Maximum Likelihood Classifier (MLC) uses statistics (mean, variance, and covariance), and especially the likelihood function, to capture the features of the classification in the training samples as they happen, resulting in a probability that a pixel belongs to a particular class (Short 2010). Forgette and Shuey (1997) conducted MLC and minimum-distance-to-means methods based on 14 signature files and 6 spectral bands, processed from spring and summer SPOT images collected at different years. Huguenin et al. (1997) compared MLC to the unsupervised ISODATA method to classify cypress and tupelo wetlands in Georgia and South Carolina, where MLC yielded more accurate results. MacAlister and Mahaxay (2009) successfully applied MLC in a challenging classification task, to distinguish 31 wetland and 23 non-wetland categories. Brisco et al. (2011) applied MLC to evaluate polarization diversity and polarimetry of SAR data for wetland type classification, using three different image polarimetric decomposition parameter sets (Freeman-Durden, Pauli, and Cloude-Pottier). Using these decomposition approaches, MLC performed well for distinguishing water and vegetation boundaries; it produced desired accuracies for more detailed wetland vegetation species. This research demonstrated the potential capability of C-band SAR data for the classification of individual wetland vegetation communities.

Conventional methods oftentimes hold a number of underlying assumptions, which may not apply to the real-world scenarios. Furthermore, the mathematical theories behind these methods could remain challenging for modelers and researchers to understand, which affects the efficiency of model construction. Comparably, another major category of supervised classification is machine learning-based method, which provides a data-driven approach to build the relationship between features and categories. This type of method has gained great popularity due to its simplicity and efficiency while applied to the real-world applications.

(2) Machine learning-based supervised methods

As a subfield of artificial intelligence, machine learning has attracted wide attentions from various applications related to spatial analysis and modeling (Walker and Moore 1988, Aspinall 1992). This approach does not require assumptions on data distribution as certain statistical approaches. It also can perform well with complex data structure. Machine learning methods enable automatic knowledge extraction through an iterative learning process based on training data samples. The most commonly used methods in wetland classification include Artificial Neural Network (ANN), Support Vector Machine (SVM), and Decision Tree (DT) based methods (Breiman 2001, Pal and Mather 2003, Gislason et al. 2006, Rodríguez-Galiano et al. 2011, Loh 2011, Corcoran et al. 2013).

a. Artificial Neural Network (ANN)

In ANN, the neural network contains at least three layers: one input layer, one output layer, and at least one hidden layer. Each layer includes a number of nodes, which gather information inputs from nodes in the previous layer, calculate the output using an activation function and transmit the results to the nodes in the succeeding layer. This information communication pattern reveals an intricate network to enable information passing through any pair of nodes between two contiguous layers. In a study to delineate forested wetlands, Augusteijn and Warrender (1998) employed a feed-forward neural network on radar data and multi-spectral data to distinguish among different levels of wetness in a forested wetland in Maryland. Ghedira et al. (2000) constructed back-propagation neural networks and applied multi-temporal datasets of radar images to successfully distinguish more detailed forest wetland types in Quebec, Canada, such as forest, woody, and shrubby wetlands.

b. Support Vector Machine (SVM)

As a non-parametric machine learning method derived from statistical learning theory, SVM divides the input dataset into subspaces through the identification of optimal boundaries (hyperplane), which have the least error comparing to all the other possible boundaries for separating the classes in the SVM scheme (Vapnik 2013, Mountrakis et al. 2011). SVM works especially well at handling a large number of input variables and thus can achieve great performance in hyperspectral image classification (Melgani and Bruzzone 2004, Huang et al. 2002). However, SVM does not work efficiently in cases with a relatively small number of input features. For instance, Pakhale and Gupta (2010) compared ANN and SVM classifiers using Landsat-7 ETM data for identifying wetlands, and ANN provided higher accuracy than SVM.

c. Decision Tree (DT)

DT-based methods have become increasingly popular for wetland mapping and classification due to their capability to deal with complex and non-linear relationships (Running et al. 1995). They represent complex classification rules as a number of consecutive simple decision-making processes (Safavian and Landgrebe 1991). In DT methods, an upside-down tree represents the classification rules and knowledge used in the decision process. The tree construction process will select partial input features to recursively divide the training dataset into smaller but more homogeneous subsets. The tree nodes represent the rules and criteria for splitting the dataset. For wetland classification, the schemes are often built based on the composition of different feature abundances extracted from the data (water, vegetation, and soil).

For wetland applications, Wright and Gallant (2007) applied classification trees with multi-year satellite images to explore different combinations of predictors to model palustrine wetlands. Wei et al. (2008) applied a decision tree to analyze the best combination of wetland features (water, vegetation, and soil) for wetland classification. Hui et al. (2009) constructed a single decision tree to identify wetland from TM images using water features. All these applications observed significant accuracy improvement using DT methods compared to other supervised learning methods such as SVM.

During the construction of the DT, researchers can transfer the expert knowledge into the rules to affect the tree structure. Hurd et al. (2006) applied a decision tree to implement the rules to combine the classification results generated by pixel-based and object-based methods. However, they operated manual intervention to visually check the data layers and determine the appropriate threshold values in the decision rules.

d. Random Forest (RF)

To build a single classification tree may introduce an overfitting issue, where the classifier is tuned to a specific study area and only able to produce high accuracy for the trained study area. Therefore, studies questioned the applicability of the single tree model (Hui et al. 2009). To tackle this issue, they developed an assembling strategy to integrate the results from multiple classifiers. Random Forest (RF) method belongs to such strategy, it uses the majority vote among multiple decision trees in the forest to determine the final decision/classification. As such, the RF can generate robust estimations to handle the noise in the training dataset, as well as to reduce the risk of overfitting.

RF offers a more robust performance by introducing bootstrap aggregation and random selection of feature subset for its optimization choice at each node. RF can utilize high-dimensional datasets to produce better results than other traditional classifiers such as MLC and minimum distance; it thus fits the application of hyperspectral image classification (Zhang and Xie 2012, Crawford et al. 2003, Naidoo et al. 2012, Zhang and Xie 2014). Naidoo et al. (2012) used RF to classify wetland tree species based on the integration of both spectral and structural datasets, revealing excellent performance.

Similarly, Corcoran et al. (2013) built several RF models by combining data from different sources, evaluating the impacts of multisource and multi-temporal datasets on the performance of RF. They demonstrated the significance of the quality of input data for RF classification accuracy.

Supervised method proves to reveal great accuracy with the support of high quality and large data sample. However, the challenges remain when the data examples that have been assigned to the same category present various characteristics projected on the selected feature variables. Also, the categories predefined in the training process can limit the classification results.

2.3.3 GIS Rule-based Expert System

In prior to the wide application of unsupervised and supervised image classification, researchers have tested rule-based methods to improve the accuracy of wetland classification (Bolstad and Lillesand 1992). This method explicitly program the rules to guide the classification process; while the machine learning-based methods extract the rules from the training data. The rule-based methods often adopt GIS technique to translate the rule implementation process into spatial operations.

The generic form of a rule is an "if-then scenario," which means that certain "**condition**" can result in different decisions. In GIS rule-based expert system, spatial data layers represent the conditions and environmental information. Applying the expert system to the classification problem, models that built upon those explicitly defined rules play the role of a human expert to make decisions in terms of classification. For example, a rule model can answer a question such as: what the wetland type for a specific location is according to its surrounding environmental characteristics.

In related studies, researchers built the expert systems through the construction of knowledge base, which contains hypotheses or complicated rule sets derived from the training examples and human knowledge, and stored as a computer-usable format (Xu 2014). Skidmore (1989) introduced the expert system to classify forest types, where he applied rules to quantify the relationships between forest types and terrain characteristics of the study area, such as gradient, aspect, and topographic position. Challenges lie in the process of formulating the human knowledge into a set of reliable and explicit rules to guide machines to make decisions (Argialas and Harlow 1990, Kontoes et al. 1993). Multiple studies have taken advantage of GIS to facilitate the representation of such information, by transferring knowledge into spatial data layers and spatial operations (Xu 2014, Zhang et al. 2011, Hui et al. 2009, Li and Chen 2005).

Bolstad and Lillesand (1992) integrated spatial environmental data (roads, land cover, soil, and terrain) with Landsat TM data in a rule-based approach, and significantly improved the accuracy of wetland classification. Similarly, Sader et al. (1995) compared mapping accuracies of forest wetlands between image classification methods and a GIS rule-based model. In the GIS model, they conducted a pixel-level analysis to generate five data layers, including the unsupervised classification result, NWI, soil, DEM, and hydrography. The results show that the rule-based model has the highest overall accuracy.

However, it did not achieve significant improvement compared to the conventional methods.

Lunetta and Balogh (1999) implemented a simple rule-based model in wetland mapping. They first generated a land cover map and vegetation category map based on the unsupervised classification using multi-date Landsat TM data. They then integrated other raster data layers to construct the rule-based model, conducting an overlay process to mesh hydrologic soil conditions with vegetation and land cover types to further classify wetland types. The rule-based method significantly improved the classification accuracy from 69 percent to 88 percent.

Li and Chen (2005) documented the successful use of knowledge-based decision rules for detecting five types of wetlands in Canada. The conditional data layers included DEM, slope gradient, SAR images, Landsat ETM images, and NDVI. They applied object-based classification methods on the SAR images to generate different classes. They further applied the supervised classification method on optical data to generate land cover maps. Lastly, they used decision rules to integrate all the classification results and spatial data layers. Their research indicated that the application of rule-based method on top of supervised classification can improve the classification accuracy to a new level (higher than 90%).

Aforementioned studies exemplified the applications of the rule-based methods for boosting the classification accuracy. This method provides the capability to integrate information collected from various data sources and prediction methods; thus to improve the classification accuracy by involving more data layers (Sader et al. 1995, Bolstad and Lillesand 1992, Li and Chen 2005). However, the process of transferring the expert knowledge into classification rules can be sophisticated and difficult to validate due to the involvement of human intervention.

2.3.4 Hybrid Methods

Each aforementioned method plays a valuable role in identifying certain wetland types. However, one cannot identify a singular method to apply to all wetland types due to the great variability among different wetland types (Milton and Hélie 2003). To achieve better accuracy, hybrid methods have revealed increasing advantages over singular method in wetland classification by combining the strengths of both supervised and unsupervised methods, as well as better targeting different features for identifying different types of wetlands.

One type of hybrid method develops classifiers at multiple stages and iteratively refine previous classification results. For example, one may first apply unsupervised classifiers to filter the data or to conduct a simple classification task, and then use supervised classifiers to identify more detailed and specific wetland types. For instance, Hurd et al. (2006) first performed a “cluster-busting” process using the ISODATA algorithm on Landsat ETM data to identify 150 spectral clusters, labeled into three specific land use types and one “others” category. They further classified pixels within the “others” category into different clusters. They repeated the procedure for each category to form clusters and further applied object-based classifiers to identify vegetative composition based on the cluster results.

The hybrid method has another advantage in processing complex datasets which are derived from multiple data sources. For instance, multiple factors such as the presence of speckle and different spectral separability of classes in SAR data can affect the accuracy of classifiers. The hybrid approach can utilize different techniques to improve the quality of input data and thus to mitigate the negative effects. Crawford et al. (1999) integrated a set of techniques to extract more information out of SAR data in terms of polarimetric and topographic features to classify coastal wetlands in Texas. They first applied a supervised feature selection technique to filter irrelevant features and keep effective ones for distinguishing specific classes. They then integrated a neural network and Bayesian pairwise classifier to conduct the classification. The results showed that this hybrid framework can achieve good classification accuracy handling multisource and multi-resolution SAR datasets. The accuracy of different types of marshes were all above 90%.

In another study, Zhang (2014) applied hyperspectral and LiDAR data to map wetland vegetation in Florida. They first conducted two cascaded PCA to filter the noise in the data and extract useful information by creating coherent eigenimages. Based on these datasets, they integrated three other techniques, including RF, SVM, and NN to conduct the classification. The synergy of all the datasets and the hybrid methodology leads to good classification performance with a Kappa value of 0.82. This statistic coefficient measures agreement for categorical data with the consideration of the agreement by chance.

2.4 Limitations of Current Studies

Studies have applied abundant techniques in automatic wetland classification focusing on different research perspectives. Previous section summarized several major techniques and analyzed their merits and disadvantages in wetland classification. Unsupervised classification can suffer from low classification performance. Statistical methods require many underlying assumptions that may not apply well to the real-world applications. Rule-based methods emerged along with the development of GIS and spatial analysis techniques; however, its implementation and validation remain a challenge. Due to the simplicity of application, machine-learning based methods become very popular for mapping the distributions of different wetland types. Meanwhile, the hybrid methods perform well in many applications since they perform well in distinguishing different wetland types that are covered by similar vegetation communities.

Although the current studies have reported high classification accuracy for all these methods, there exists several limitations. First, the high accuracy revealed at the training process tends to give a false sense of a well-performed model. Which means, even though the models can reach a relatively high classification accuracy during the training phase, they are less likely to reproduce the same performance when applied to a different study area. Studies also consider this situation as the overfitting issue, under which a trade-off exists between the generality of the model and high accuracy for predicting specific region.

Second, the training data samples can greatly constraint ML-based models, since only the types that existed in the training samples will occur in the classification results. The type distribution in the classification result also tend to follow similar proportions as observed in the training data, which means wetlands with smaller samples could be easily omitted in the classification.

Third, current studies lack emphasis on the repeatability and automation of the overall procedure, including data processing, model construction, model prediction, and post-processing. Automation plays an significant role in situation such as the need to rerun the classification workflow multiple times to conduct wetland mapping for different environmental conditions.

This research aim at improving above listed limitations using the proposed methodology illustrated in the following chapter.

CHAPTER 3: METHODOLOGY

This section introduces the methodology developed for this research. Methods include: (1) developing a GIS-based expert system for the wetland type classifications; (2) constructing a fusion database through data processing techniques; (3) building a knowledge base through two types of modeling methods, open-loop and closed-loop methods; (4) automating the workflow and evaluating classification results. The methodology aims to tackle specific objectives: (1) the expert system should perform better in terms of classification accuracy than traditional methods; (2) the methods should suit different study regions, with the consideration of local environmental characteristics such as topography and hydrology; (3) the method can be executed as an automatic process at a relatively low computational cost.

3.1 Framework of the GIS-based Expert System

This study proposes a hybrid expert system that combines the strengths of two classification approaches, which are the machine learning-based classifier approach and GIS rule-based approach.

3.1.1 Two Modeling Perspectives

Applying classification rules to categorize data samples (pixels on imageries) into different groups represents the core process of machine-based classification methods.

There mainly exist two common inference strategies to achieve knowledge: induction and deduction. The process of inductive learning depends on generalization from existing data and examples for the purpose of training, and the results are represented as production rules (Bratko 2001, Ahmad 2001). The deductive process requires more human inputs, which mostly include manual generalization and explicit summarization, to extract the rules from domain knowledge and expert experience.

Corresponding to the two inference strategies, one can construct the classification models in different manners. Figure 1 shows the proposed workflows of the classification modeling that focus on different perspectives. The closed-loop modeling applies the inductive method to construct models based on existing data samples. The open-loop model relies on the knowledge deducted from expert experience. Respectively, the mechanisms behind the two modeling methods reveal differences. In the closed-loop modeling, certain evaluation processes provide feedbacks inheritably. For instance, the closed-loop model can apply accuracy metrics to guide the further improvement of the classification, such as using a regression line to fit the data through minimizing squared errors. This process relies on an algorithm instead of manual adjustments. In the open-loop modeling, expert knowledge drives the decision-making process, which requires the explicit rules defined in advance. GIS introduces spatial analysis and spatial operations to generate data layers for representing the environmental features and translate human decision rules. Therefore, the modeling process in the closed-loop method represents a "black box" process guided by machine learning algorithms. To the contrary, the open-loop modeling method explicitly defines the path to take.

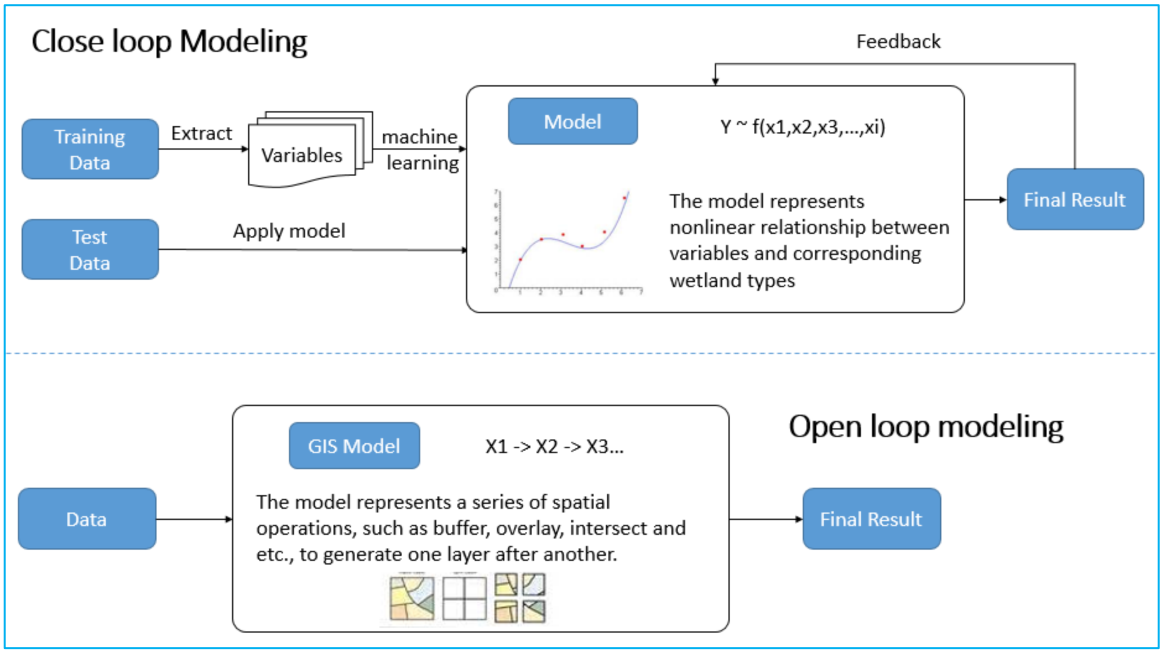

Close loop Modeling
Feedback
Training Data
Extract
Variables
machine learning
Model
Y ~ f(x1,x2,x3,...,xi)
The model represents nonlinear relationship between variables and corresponding wetland types
Final Result
Test Data
Apply model
Open loop modeling
Data
GIS Model
X1 -> X2 -> X3...
The model represents a series of spatial operations, such as buffer, overlay, intersect and etc., to generate one layer after another.
Final Result

Figure 1. Two modeling perspectives

These two modeling perspectives can be complementary to each other. The closed-loop model mainly relies on the inductive process to learn from input training data. It extracts rules in a form that may not be intuitively straightforward for users but can provide high prediction accuracy. The open-loop model on the other hand provides rules based on reasoning and can be explained.

3.1.2 Generating Decision Rules

The traditional rule-based spatial models usually reveal decision rules as a deterministic and exclusive manner, for example:

```
Define decision rule $R1$:
if ($A$ >= $a$ AND $B$ >= $b$):
        wetland type = $c$;
else:
        wetland type = $d$;
```

In this model, we consider both A and B as key variables for identifying wetland types. The condition for supporting decision making is a binary statement, representing whether it has met certain criteria or not. Therefore, this model requires the process of transferring continuous real-world data into binary states. It also rely on the appropriate setup for parameters *a* and *b* to effectively separate different classes. This study uses probability-based statements instead of deterministic ones to represent the contributions of different variables. The form of the logic is then shown as below:

Define decision rule *R*2:

```
if ($A >= a$):
    probability of wetland type $c = c + c1$;
if ($B >= b$):
    probability of wetland type $c = c + c2$;
```

Through the above process, when certain conditions are met, the confidence level of being a specific wetland type will increase. For example, in the above expression, c1 and c2 represent probability variation on top of the previous probability. As a result, numerical values will be generated based on the sum of all relevant data layers, to indicate the potential for a wetland type to occur at the corresponding location. In general, a higher score indicates a higher possibility for the wetland type to occur. Users can then decide the cut-off threshold for determining wetland type in the final step. It relaxes the risk of inappropriate parameter calibrations in the earlier stage.

3.1.3 System Framework

In this research, the expert system integrates both modelling perspectives. Figure 2 shows the overall framework of the expert system for supporting wetland type classification.

First, the system contains a core database to store data inputs from various sources, such as LiDAR data, satellite image (Sentinel-2 data), soil, and land cover data. I collected these spatial and nonspatial datasets for the study area, and further processed the source data to generate spatial data layers to represent environmental characteristics, mainly including the vegetation, hydrology, and topographical features. One data source can derive multiple data layers, for instance, we can extract geometrical features of LiDAR data to generate terrain derivatives or use intensity information to detect inundation or flooding under tree canopies. Other than environmental variables, the database also contains the wetland inventory information collected through fieldwork, which serves as ground truth information and will be used as a reference of wetland occurrences.

Another core element of this system is the knowledge base, also described as the model base in this research. It stores rules constructed from both deductive and inductive methods—for instance, the classification models (tree structure rules) constructed from the closed-loop model training process, as well as decision rules derived from the open-loop method based on the expert knowledge. I then derived an overall rule-based model to integrate rules from different sources, of which the quality depends on both the database and knowledge base.

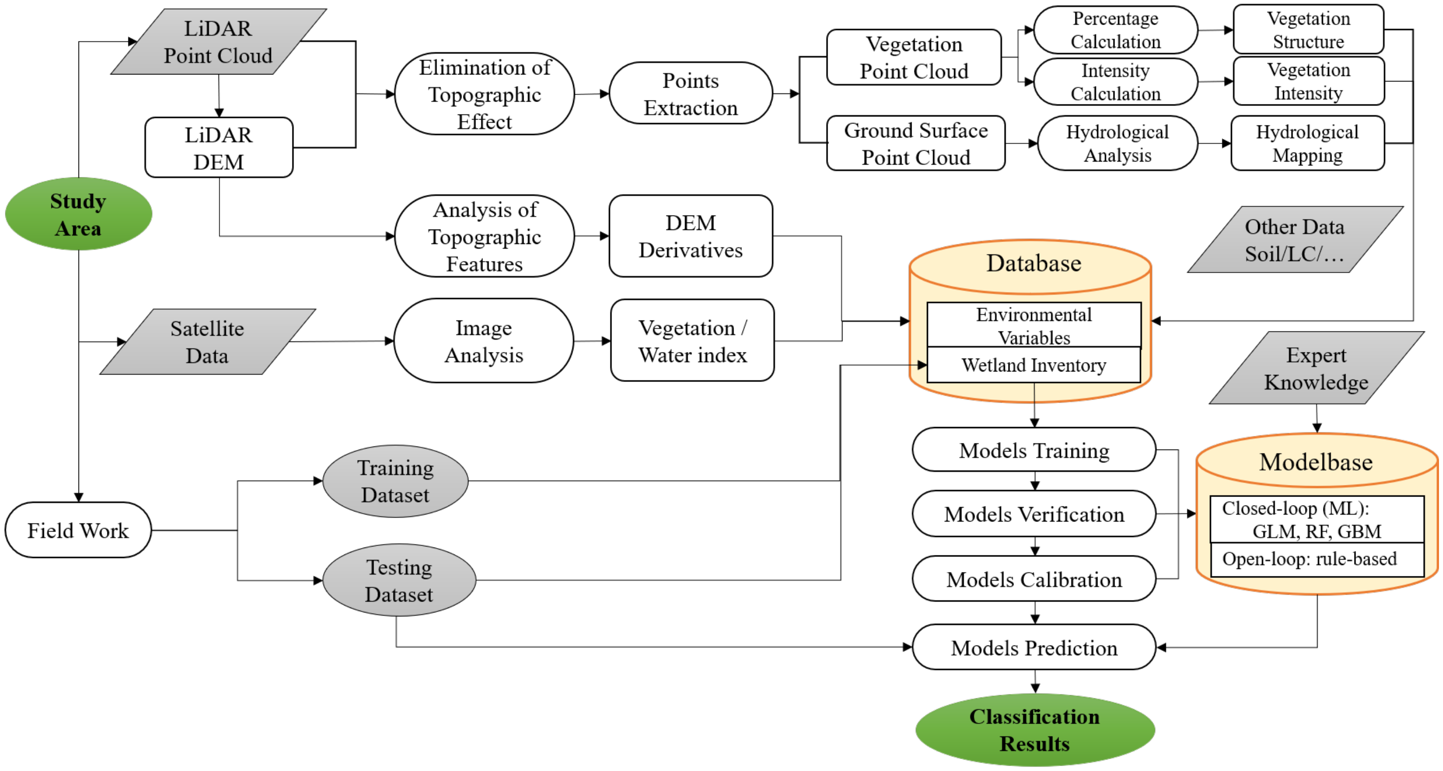


Figure 2. Expert system for wetland type classification (DTM: Digital Terrain Model; DEM: Digital Elevation Model)

3.2 Construction of Knowledge Base

3.2.1 Extract Knowledge through Closed-Loop Modeling

The major objective of the closed-loop modeling lies in extracting rules from algorithms to form a knowledge base for future wetland prediction. In this dissertation, I introduced automatic machine learning-based algorithms to this methodology framework as the knowledge extraction process. I selected three machine learning-based methods, including General Linear Model (GLM), Random Forest (RF), and Gradient Boosted Machine (GBM). The first two methods have occurred in several existing wetland classification studies and revealed good performance, while the GBM method represents a relatively new method. They all belong to the closed-loop modeling methods by extracting classification rules from the training datasets. GLM is the generalized form of the regression model, RF and GBM belong to the category of decision tree-based algorithms.

(1) Generalized Linear Model

A linear regression model represents the linear relationship between independent variables and a dependent variable. The traditional linear regression model holds several assumptions, such as the normal distribution assumption for the response variable and constancy of the variance. GLM is a generalization of the traditional linear model that unifies different types of regression models by relaxing data assumptions and allowing a non-normal form of the errors, it thus introduces more flexibility to regression model structure for a wider range of applications (Nykodym et al. 2016, McCullagh and Nelder

1989). GLM has three core components: probability distribution, linear prediction, and link function.

In this study, we consider *K* categories, including one non-wetland type and *K*-1 wetland types. Equation (1) shows the relationship between the response variable and independent variables for the *k*-th category. GLM assumes the dependent variable $y_{i,k}$ to follow a probability distribution that belongs to the exponential family, such as Gaussian, Poisson, and gamma distributions. Different distributions will introduce different transformations for probability calculation. One can assume the mean μ_k to be certain function of linear combination $\beta_k{}^T X_i$, and call the function for connecting the linear predictor with the mean as link function. In this study, I consider multinomial distribution for the dependent variable and applied softmax function as the link function. Formula (2) shows the probability calculation for each category (Nykodym et al. 2016). All *K* probabilities sum to one; the prediction type goes to the one with the highest probability.

$$y_{i,k} = \beta_{k0} + \beta_{k1}x_{1i} + \beta_{k2}x_{2i} + \cdots + \beta_{kp}x_{pi} + \epsilon_i$$

$$for\ i \in \{1, \cdots, n\}\ and\ k \in \{0,1, \cdots, K-1\} \tag{1}$$

$$P(Y_i = k) = \frac{e^{\beta_k{}^T x_i}}{\sum_{t=0}^{K-1} e^{\beta_t{}^T x_i}} \tag{2}$$

where: *i* is the index for an observation (data sample; it corresponds to a grid cell in this study);

k is the index for the category group which the dependent variable falls into;

$x_{ji}, j \in \{1, \cdots, p\}$, is the j-th explanatory variable value for the i-th observation;

$X_i = [1, x_{1i}, x_{2i}, \dots, x_{pi}]^T$, is a vector of independent variables for the i-th observation;

$\beta_k = [\beta_{k0}, \beta_{k1}, \beta_{k2}, \dots, \beta_{kp}]^T$, is a vector of parameters for category group k;

$\beta_k^T X_i = \beta_{k0} + \beta_{k1}x_{1i} + \beta_{k2}x_{2i} + \cdots + \beta_{kp}x_{pi}$, is known as the linear predictor;

$Y_{i,k} = \beta_k^T X_i$ is the dependent variable for cell *i* which falls into the group *k*;

$P(Y_i = k)$ is the probability of observation i to be predicted as *k*-th wetland type based on measured features of the observation.

(2) Random Forest

In the modeling process, RF method uses the bootstrap aggregation (bagging) technique to train all decision trees in the forest by randomly selected sub-datasets as training source data for different decision trees and to average all classification results from all decision trees together to vote for the final results. To build each tree, RF randomly selected a number of data samples from the overall training dataset. It chooses the variable with the highest information gain among the feature subset to split the node of the decision tree. It also evaluates all the trees using an internal accuracy estimation—out-of-bag (OOB) error while constructing them.

RF suits a wide range of datasets without the restriction of data distribution, and it handles well with data noise or overtraining issues. RF has another great merit as providing variable importance by quantifying changes in classification accuracy by taking out the variable from the OOB data sample. However, RF lacks transparency due to the large number of trees in the model.

(3) Gradient Boosting Model

The Gradient Boosting Model (GBM, also known as Stochastic Gradient Boosting) is an improved technique based on CART methods. It combines gradient-based optimization and boosting strategies together. Gradient-based optimization aims at minimizing a model loss function for the training data, while the boosting strategy gathers a number of consecutively fitted trees to create a robust classification system (Click et al. 2016). Compared to RF, GBM applies a boosting strategy instead of a bagging method. It builds each tree to enhance prior trees and reduce the net error (Freund and Schapire 1996).

For a *K*-class classification task, GBM can build *K* trees, each targeting at one class. The process of building each tree includes: (1) compute the residuals (the gradient values of user-specified loss function); (2) fit a regression tree to the gradients; (3) add current model to the fitted regression tree and improve it through a descending step, which gives the misclassified observations higher weights in the next iteration. The model is thus “boosted” over successive iterations.

GBM uses a gradient-descent-based formulation to maximize the correlation between newly trained models with the negative gradient of the loss function. Users can define the loss function based on specific tasks in a flexible way, in terms of adjusting the loss function to achieve the appropriate model design. However, one needs to use an appropriate stopping point to reduce overfit since GBM is sensitive to data noise. Studies has applied GBM in land cover classification and wetland change detection but not in wetland type identification yet (Baker et al. 2006, Baker et al. 2007).

3.2.2 Summarize Expert Knowledge through Open-Loop Modeling

In this study, I built an expert system applying open-loop modeling to introduce the expert knowledge for constructing the knowledge base. I used NCWAM as the source of truth for related expert knowledge of North Carolina wetlands classification. I translated the criteria in the classification rules into digital data through GIS spatial analysis. These data layers represent the influential environmental information for determining wetland types of a given location. I then calculated the probability for each wetland type to occur by translating the decision rules into spatial operations using previously generated data layers. In the end, I incorporated the prediction results from the closed-loop models into the final probability calculation.

(1) Probability calculation for wetland types

I first describe several basic definitions in this expert system, and then introduce the process of calculating probabilities. The expert system aims to answer a research question: “which wetland type is most likely to occur at a given wetland location?” Assuming there are *K*-1 wetland types in total in the classification task, and the spatial database contains a total number J of $x_j^o (j \in \{1, \cdots, J\})$ processed variable layers with the same spatial extent (the number of variables used here may differ from the closed-loop method), and georeferenced to the same coordinate system to support the rule-based model. Let x_{ji}^o denote the value of cell *i* for input variable layer j, assuming there are *I* cells in total in one data layer. Let H_k be the hypothesis that wetland type k occurs at cell i, where $k \in \{1, \cdots, K-1\}$ types. I do not consider non-wetland type here in the rule-based classification process (H_0). Formula (3) denotes the format of a simple rule; that under certain conditions, such as cell values from different layers fall within specific value ranges, then we can infer hypothesis H_k. R_j refers to the condition using variable data layer j, where $j \in \{1, \cdots, J\}$ variable data layers. In this example, I use logic "and" (&) to connect all the conditions; however, in a more complicated situation, logic "or" operators may also be involved. The research objective is to identify the set of conditions that support each wetland type and formulize their contributions in the calculation of wetland type probability.

$$x_{1i}^{o} \in R_1 \;\&\; x_{2i}^{o} \in R_2 \;\&\; \ldots \;\&\; x_{ji}^{o} \in R_J => H_k, \text{ for cell } i \text{ where } i \in \{1, \cdots, I\} \qquad (3)$$

Figure 3 illustrates the decision-making process of determining wetland types as an upside-down tree structure. This decision tree starts from a basic assumption that the location is considered as wetland based on observation or a prediction result from another model; it then diverges into a number of routes towards different wetland types. From each splitting node, it generates two alternative branches based on whether certain conditions have been met. A data layer with initial binary values of 0 and 1 represents whether the required condition is met or not met at the splitting node, and it is called a conditioning layer C_i. We define a path as the route starting from the top node to one bottom leaf node (belonging to one wetland type), which goes through a series of splitting nodes. We call an individual path evidence E_{kt}, the *t*-th evidence to support wetland type *k*. In total, assume that *m* multiple paths can lead to the same wetland type *k*.

In this study $m = 1$ for most wetland types, but few include multiple paths, e.g., $m = 2$. Figure 3 shows that two paths lead to wetland Type 1. The probability of evidence is shown in Formula (4)—it is the multiplication of all the probabilities of conditions C_i along this evidence path because it needs to fulfill all the criteria to reach the bottom leaf node. Formula (5) shows the probability calculation for each wetland type to occur, which summarizes the probabilities of all evidence, as an example to show wetland type k. $P(H_k|E_{kt})$ is the conditional probability for H_k to be true given evidence E_{kt} is true. This probability depends on the reliability of this rule. In this research, we don't consider uncertainty in the correctness of the expert knowledge, the value of $P(H_k|E_{kt})$ is 1. $P(C_i)$ denotes the probability of the splitting node going towards the direction that is in the path (E_{kt}) towards corresponding wetland type (k). The sum of the probabilities for all the *K*-1 wetland types will be 1, see Formula (6). In this example, the probabilities for the two wetland types shown in Figure 3 are $P(C_1)P(C_2) + (1 - P(C_1))$ and $P(C_1)(1 - P(C_2))$ respectively.

$$P(E_{kt}) = \prod_{i=1}^{l} P(C_i) \tag{4}$$

$$P(H_k) = \sum_{t=1}^{m} P(H_k|E_{kt})P(E_{kt}) \tag{5}$$

$$\sum_{k=1}^{K-1} P(H_k) = 1 \tag{6}$$

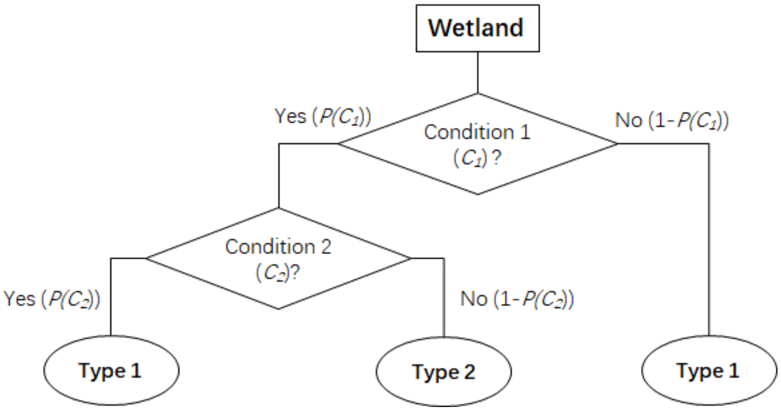


Figure 3. Decision tree example

Figure 4 shows the overall decision tree for North Carolina wetland types proposed by Axiom Environmental (Wang et al. 2014) based on NCWAM (N.C. Wetland Functional Assessment Team 2010). Here lists 15 wetland types (pine savannah and pine flat are combined as one type) and labels each wetland type with a code, shown in the parentheses beside the type name. NCWAM categorizes wetlands that are affected by tides as tidal wetlands, while the remainder falling into the non-tidal wetland group. It further classifies wetlands in the non-tidal group as riparian if they located within a geomorphic crenulation, floodplain, or adjacent to a 20-acre or larger lake. Non-tidal wetlands that are not associated with the above features are classified as non-riparian type. P_i denotes the probability at each condition node. The total of probabilities for paths that are derived from the same splitting node must sum up to 1.

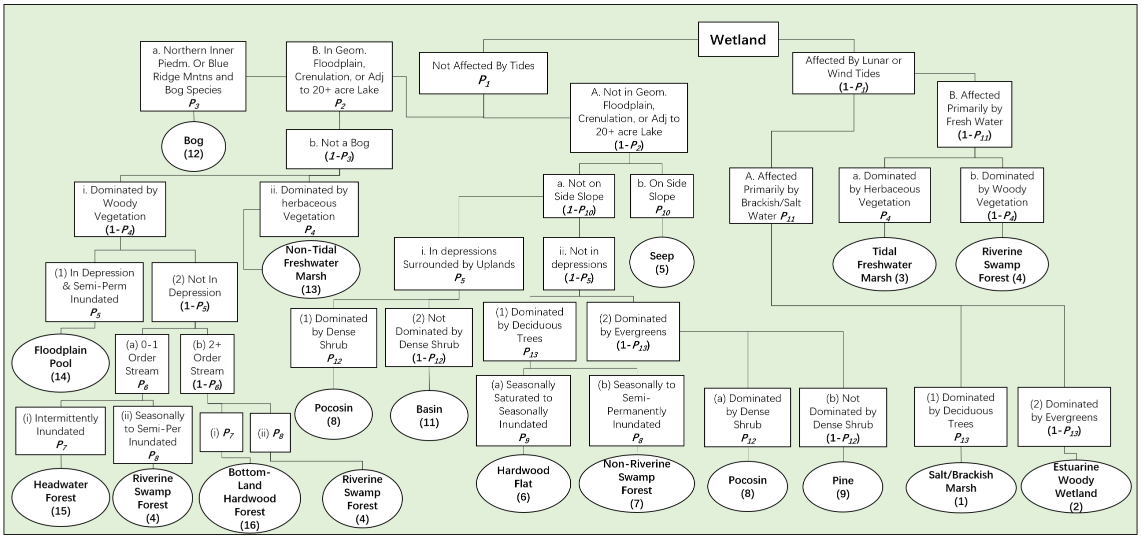

Wetland
Not Affected By Tides P_1
Affected By Lunar or Wind Tides $(1-P_1)$
a. Northern Inner Piedm. Or Blue Ridge Mntns and Bog Species P_3
B. In Geom. Floodplain, Crenulation, or Adj to 20+ acre Lake P_2
A. Not in Geom. Floodplain, Crenulation, or Adj to 20+ acre Lake $(1-P_2)$
B. Affected Primarily by Fresh Water $(1-P_{11})$
Bog (12)
b. Not a Bog $(1-P_3)$
i. Dominated by Woody Vegetation $(1-P_4)$
ii. Dominated by herbaceous Vegetation P_4
a. Not on Side Slope $(1-P_{10})$
b. On Side Slope P_{10}
A. Affected Primarily by Brackish/Salt Water P_{11}
a. Dominated by Herbaceous Vegetation P_4
b. Dominated by Woody Vegetation $(1-P_4)$
Non-Tidal Freshwater Marsh (13)
i. In depressions Surrounded by Uplands P_5
ii. Not in depressions $(1-P_5)$
Seep (5)
Tidal Freshwater Marsh (3)
Riverine Swamp Forest (4)
(1) In Depression & Semi-Perm Inundated P_5
(2) Not In Depression $(1-P_5)$
(1) Dominated by Dense Shrub P_{12}
(2) Not Dominated by Dense Shrub $(1-P_{12})$
(1) Dominated by Deciduous Trees P_{13}
(2) Dominated by Evergreens $(1-P_{13})$
Floodplain Pool (14)
(a) 0-1 Order Stream P_6
(b) 2+ Order Stream $(1-P_6)$
(a) Seasonally Saturated to Seasonally Inundated P_9
(b) Seasonally to Semi-Permanently Inundated P_8
(a) Dominated by Dense Shrub P_{12}
(b) Not Dominated by Dense Shrub $(1-P_{12})$
(1) Dominated by Deciduous Trees P_{13}
(2) Dominated by Evergreens $(1-P_{13})$
Pocosin (8)
Basin (11)
(i) Intermittently Inundated P_7
(ii) Seasonally to Semi-Per Inundated P_8
(i) P_7
(ii) P_8
Headwater Forest (15)
Riverine Swamp Forest (4)
Bottom-Land Hardwood Forest (16)
Riverine Swamp Forest (4)
Hardwood Flat (6)
Non-Riverine Swamp Forest (7)
Pocosin (8)
Pine (9)
Salt/Brackish Marsh (1)
Estuarine Woody Wetland (2)

Figure 4. Decision process for all wetland types based on expert knowledge

In the probability Equation, $P(C_i)$ denotes the chance of the decision tree going towards a particular path, given the data observed. It uses the environmental data to analyze the probability of each branch at the splitting node. The ideal situation is, when certain criteria are met, the value of the cell on this raster layer equals 1; otherwise 0. For instance, the first splitting node is “whether it is affected by tides.” My solution here is to collect Tidal Influence Zone (TIZ) data and Tidal Water Amplitude (TWA) data from the National Oceanic and Atmospheric Administration (NOAA), to analyze the possibility that of tides affecting a specific area. In the end, I divided the region to areas that are affected by tides and that are not affected.

Ideally, assuming that the original data is correct, and the analysis process has no error, we can draw clear boundaries on the map between tidal and non-tidal areas, with corresponding cell values of 1 and 0. However, In reality, abrupt boundaries do not exist as illustrated on the simplified maps. In this example, the areas on different sides of the boundary and near the boundary may not significantly differ. Furthermore, errors can exist in data collecting and processing steps. All these factors can result in ambiguity on the maps. The same situation applies to the formation of the decision tree. Misguided by the inaccurate splitting criteria, classification for a data sample could follow the wrong path and end up with a incorrect wetland type. To better deal with this situation, I set a tolerance at each splitting node. Instead of using 0 or 1 to denote the possibility of taking a path, I applied a probability value between 0 and 1 for each branch. I then calculated the values basing on the percentage of misclassified examples caused by the data map at each splitting node. I consider this process as the calibration for the open-loop method.

In this research, I used the standards shown in Table 3 to assign a pair of probability values to the two branches at a splitting node. The probability values for each path corresponds to $P(C_i)$ in Formula (4). I then calculated a final probability result between 0 to 1 for each wetland type by multiplying all the probability values of the paths taken. Higher probability value represents a higher possibility for the wetland type to occur given a series of evidence observed according to current environmental features. This process can also help with evaluating the quality of the data layers. If the error rate for the classification is too high, such as larger than 50%, the corresponding data layer is considered as unreliable, and the generation process of the data layer needs further examination.

Table 3. Value adjustment for each data layer

Percentage of misclassified pixels	**Values update**
> 50%	Recheck the generation of data layer
40% ~ 50%	(0.4, 0.6)
30% ~ 40%	(0.3, 0.7)
20% ~ 30%	(0.2, 0.8)
5% ~ 20%	(0.1, 0.9)
0% ~ 5%	(0, 1)

3.2.3 Rules Integration from Different Models

The models from model base will generate different classification results, which requires an integration mechanism to conclude the final result. Common strategies to integrate model results include majority vote, Bayesian average, the fuzzy integral method and so on (Zhang 2014, Moreno-Seco et al. 2006). By applying the average or majority vote method, one needs to take into consideration of results from all the models. However, if multiple models sharing similar essence and results, votes from these models may dominate all the voting. For instance, in our case, results among the three ML models could reveal higher similarity compared to the result of the rule-based model. Considering this case, a weighted summary tends to be fairer. Another potential issue for the current integration method relates to the reliability of different models. It is hard to determine which one has higher reliability and higher weight in the integration. In this research, I applied a strategy to conduct a precheck and eventually proceed with a weighted sum to determine the result.

In this expert system, results generated by this system include one rule-based (open-loop) model and three results from ML models (the closed-loop method). First, I combined the ML results based on a weighted sum process. Secondly, I applied the combined closed-loop result to impact the probability calculation process in the rule-based model. However, whether the second integration step happens depends on a criterion related to the consistency among the ML models. Formula (7) shows the calculation of integrated probability $PI(H_k)$. Modifying the weights will adjust the relative contribution between the open-loop method and the closed-loop method, as well as among different machine-learning models. However, in the integration process, the closed-loop models can only provide the contribution to the wetland types that have occurred in the training dataset.

$$PI(H_k) = \begin{cases} w_o P(H_k) + \ w_c \sum_{i=1}^{n} w_i P_{MLik} , & if \sum_{i=1}^{n} w_i P_{MLik} \geq t \\ P(H_k), \quad if \sum_{i=1}^{n} w_i P_{MLik} < t \end{cases} \tag{7}$$

$$P_{MLik} = \begin{cases} 1, & if\ predicted\ as\ type\ k\ by\ ML\ method\ i \\ 0, & if\ predicted\ as\ other\ type\ by\ ML\ method\ i \end{cases} \tag{8}$$

$$w_o + \ w_c = 1 \tag{9}$$

$$\sum_{i=1}^{n} w_i = 1 \tag{10}$$

where: $PI(H_k)$ is the integrated probability for the pixel to be assigned to wetland type *k*;

$P(H_k)$ is the probability for the pixel to be assigned to wetland type *k*, calculated based on the rule-based model, see Formula (5);

$k \in (1 \cdots K - 1)$;

t is the threshold for integrating the closed-loop models in the probability calculation;

P_{MLik} is the binary result map generated by the machine-learning (ML) method *i* for wetland type *k*;

w_o is the weight for the probability calculated from the open-loop method;

w_c is the weight for the integrated result of all the closed-loop models.

The first two steps derived the final probability for each wetland type. Finally, I conducted a competition process to compare all the probabilities among different wetland types and consider the type with the highest value as the classification result. If multiple wetland types reveal the same highest probability value, I will use the neighboring pixels within a three-by-three window to determine the most likely wetland type. Under this situation, I will calculate the sums of probability values of all the neighboring pixels for these wetland types and compare them. And choose the wetland type that has the highest probability as well as highest neighboring sum probability.

3.3 Construction of a Fusion Database

To support the construction of the model base in the expert system, an important step consists of preparing the input datasets for both the closed-loop and open-loop models. In this section, I summarize the key variables and data used by these two models. The corresponding module for database preparation is the data processing and variable generation module. This module contains a number of GIS functions and processes, which together play an important role in integrating data from different sources and constructing a fusion database.

3.3.1 Variable Generation for the Closed-Loop Model

Variables here refer to the representations of key features and critical spatial characteristics for the algorithms to determine wetland types. In the closed-loop model, variables serve as the basis for machine learning algorithms to extract knowledge and to build a relationship between features and classified types. Table 4 lists the variables generated for this study (Wang et al. 2015). Such variables are organized and stored in a GIS geodatabase as either raster layers or vector data files. Each variable represents a dimension in the feature space, and the measurement unit represents the aggregation level. Each variable is associated with multiple raster layers with different resolutions representing different levels of aggregation.

Table 4. Variables generated for the closed-loop modeling

Variable Name	Full Name	Data Source and Illustration
DEM Derivatives		
ELVPCD[1]	Elevation	Elevation of each cell: z(x, y)
SLP	Slope	In degree:$\mathrm{slp}(x, y) = 57.29578 \times$ $\mathrm{atan}\ (\sqrt{(dz/dx)^2 + (dz/dy)^2}\ )$
ASP	Aspect	ASP = 57.29578 * atan2 ([dz/dy], -[dz/dx])
HILL	Hillshade	Hillshade = 255.0 * ((cos(Zenith_rad) * cos(Slope_rad)) + (sin(Zenith_rad) * sin(Slope_rad) * cos(Azimuth_rad - Aspect_rad))) Zenith_rad: zenith angle in radians Slope_rad: slope radians calculated in 3 by 3 window Azimuth_rad: azimuth angle in radians Aspect_rad: aspect calculated in 3 by 3 window
RF[2]	Roughness Index (Jacek 1997)	The roughness in a continuous raster within a specified window. RI = std * std std: the standard deviation within a specified window
IMI[2]	Integrated Moisture Index (Iverson et al. 1997)	IMI = 0.15*curv + 0.35*flac + 0.5*hill curv:curvature flac: flow accumulation hill: hillshade
CTI[2]	Compound Topographic Index	CTI can indicate the water accumulation possibility at specific locations. Small value denotes upper catenary positions (Gessler et al. 2000). It is a function of both the slope and the upstream contributing area per unit width to the flow direction. CTI = ln (AS / (tan(beta)))

[1] These variables are further normalized to the same value range (0-100).

[2] These variables are generated using a tool box downloaded from: http://evansmurphy.wix.com/evansspatial#!arcgis-gradient-metrics-toolbox/crro

		where "AS" is the area calculated as (flow accumulation + 1) *(pixel area m^2) and beta is the slope in radians http://arcscripts.esri.com/details.asp?dbid=11863
SRR^2	Surface Relief Ratio	It describes rugosity in a continuous raster surface within a specified window (Pike and Wilson 1971). SRR=(z_{mean} - z_{min}) / (z_{max} - z_{min})
LS	Slope Length and Steepness Factor	Index LS can reflect the complexity of the surface (Olaya 2009). LS = Power("flowacc"* cellSize /22.1,0.4) * Power(Sin("slp"*0.01745))/0.09, 1.4) * 1.4 "flowacc" = Flow accumulation cellSize = Resolution of DEM in meters "slp" = slope in degrees
FLACD	Flow Accumulation	Accumulated weight of all cells flowing into each downslope cell in the output raster. The calculation is based on the DEM with "fill" preprocessing. FLAC = $\sum_{i=0}^{n} c_i w_i$ where c_i is a neighboring cell that has water flows into the cell of interest w_i is the weight of the neighbor cell
FLATDC	Flatness Index	Sum of the absolute value of the difference between a cell and its eight neighbors.
MDECD	Maximum Downslope Elevation Change	Maximum difference of z(x,y) between the cell and its neighbor cells. MDEC = Max(z_i - z) where z_i is the elevation of a neighbor cell
WEID	NCDOT Wetland Elevation Index	Series of increasingly larger neighborhoods used to determine the relative landscape position of each cell. WEI = $\frac{\sum_{i=1}^{n} dz_i}{n}$ where d_{zi} = z – Mean(z_i) d_{zi} is the difference of elevation between the cell and the mean elevation of its neighboring cells at certain window size; z_i is the elevation value at the cell location
WEIRE	Reclassification of NCDOT WEI	WEIR = $\begin{cases} 1, & if\ WEI > 0 \\ 0, & if\ WEI \leq 0 \end{cases}$
DROPD	Elevation Drop	The ratio of the maximum elevation change along flow direction DROP = (z_i - z) / (1.5 * cell size) where z_i is the elevation of the neighbor cell that lies along the flow direction
Vegetation		
LOWI	Intensity of Low	The classification of vegetation point is labeled as:

	Vegetation Returns	Low Vegetation (3); Medium Vegetation (4); High Vegetation (5) High intensity values represent photosynthetically active vegetation, while lower intensity values are likely to represent wet surface condition or less photosynthetically active vegetation.
MEDI	Intensity of Medium Vegetation Returns	
HIGHI	Intensity of High Vegetation Returns	
LOWP	Percentage of Low Vegetation Returns	Number of low vegetation returns divided by the total number of point returns (the total points counted here include ground, low vegetation, medium vegetation, high vegetation and water).
MEDP	Percentage of Medium Vegetation Returns	Number of medium vegetation returns divided by the total number of point returns.
HIGHP	Percentage of High Vegetation Returns	Number of high vegetation returns divided by the total number of point returns.

EVI	Enhanced Vegetation Index	2.5 * ((b8-b4) / (b8 + 6* b4 - 7.5* b2 + 1))
		Where: b2 is Band 2 – Blue;
		b4 is Band 4 – Red;
		b8 is Band 8 - NIR

NDVI	Normalized Difference Vegetation Index	(b8 - b4) / (b8 + b4) Where: b4 is Band 4 – Red; b8 is Band 8 - NIR
NDWI	Normalized Difference Water Index	(b3 - b8) / (b3 + b8) Where: b3 is Band 3 – Green; b8 is Band 8 - NIR

MSAVI	Modified Soil-Adjusted Vegetation Index	b8 + 0.5 - (0.5 * sqrt($(2 * b8 + 1)^2$ - 8 * (b8 - (2 * b4))))
		Where: b2 is Band 2 – Blue;
		b4 is Band 4 – Red;
		b8 is Band 8 - NIR

MSAVI2	Modified Soil-Adjusted Vegetation Index II	(2 * (b8 + 1) - sqrt($(2 * b8 + 1)^2$ - 8 * (b8 - b4))) / 2
		Where: b2 is Band 2 – Blue;
		b4 is Band 4 – Red;
		b8 is Band 8 - NIR

Other Data		
SOIL	Soil Type	Categorical data, soil type based on Natural Resources Conservation Services (NRCS) soils

		database files.
TIDAL	Riparian Information	Binary data, 0 denotes non-riparian and 1 denotes riparian area, data is provided by NCDOT
LC	Land Cover Type	Reclassified 2011 NLCD dataset. The reclassified types are: Water (1), Built-up (2), Barren (3), Deciduous (4), Evergreen (5), Mixed Forest (6), Shrubs (7), Herbaceous (8), Agriculture (9), Woody Wetlands (10), Herbaceous Wetlands (11)

The variables mainly focus on three aspects of environmental characteristics: vegetation, hydrology, and terrain, corresponding to the key indicators to distinguish different wetland types. For example, terrain derivatives play important roles in detecting the possibility of water inundation; they are typically estimated based on a digital elevation model (DEM) to simulate hydrologic processes at a watershed scale. In this study, I applied LiDAR data to generate the DEM layer.

DEM layer provides the foundation to generate terrain derivatives. Existing applications applied various terrain derivatives and may vary in specific algorithm. For example, there exist multiple versions of surface roughness to depict the variation of the terrain elevation at the local scale. The roughness of the surface has impacts on the capability of surface water storage; therefore, this metric can provide great help in providing information about hydrological condition and thus has occurred in wetland related studies. This researched adopted the calculation from: http://gis4geomorphology.com/roughness-topographic-position.

We can summarize vegetation characteristics by the composition of different vegetation types (e.g., herbaceous, evergreen or deciduous), vegetation density, and vertical structure (profile percentages). One can use LiDAR intensity (amplitude) to provide important information on vegetation conditions, such as differentiating vegetation species through the characteristics of surface reflections (Song et al. 2002). LiDAR return intensity can help with quantifying the amount of energy returned to the sensor relative to the emitted energy per laser pulse. In this study, I calculated the composition of different types of vegetation, as well as the vertical strata, the percentage of returns from different heights of plants.

I applied multispectral satellite data to generate several metrics, to represent vegetation and hydric conditions. I used four Sentinel-2 bands (Blue, Green, Red, and Near-infrared) to produce several indicators, including the Enhanced Vegetation Index (EVI), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Modified Soil-adjusted Vegetation Index (MSAVI), and MSAVI2.

Other than terrain and vegetation, I collected additional data such as soil category and land cover types as other variable layers. I conducted a set of GIS operations to process the datasets to ensure the consistency in the spatial extent, common resolution, and projection. For more complex spatial algorithms, I developed more advanced GIS tools to assemble the functionalities that can automate the variable generation process.

Table 4 shows the variables that will serve as the input data for the closed-loop modeling. To compare variables among different study regions, I normalized several variables to the range of 0 – 100 (Formula 11). In this way, the value represents the relative level compared to the local overall condition. For the vegetation metrics, the calculation is based on different bands obtained from Sentinel-2 data. Several examples of the variables are shown in Figure 5 to Figure 8.

$$V_i' = 100 \times \frac{(V_i - V_{min})}{(V_{max} - V_{min})} \tag{11}$$

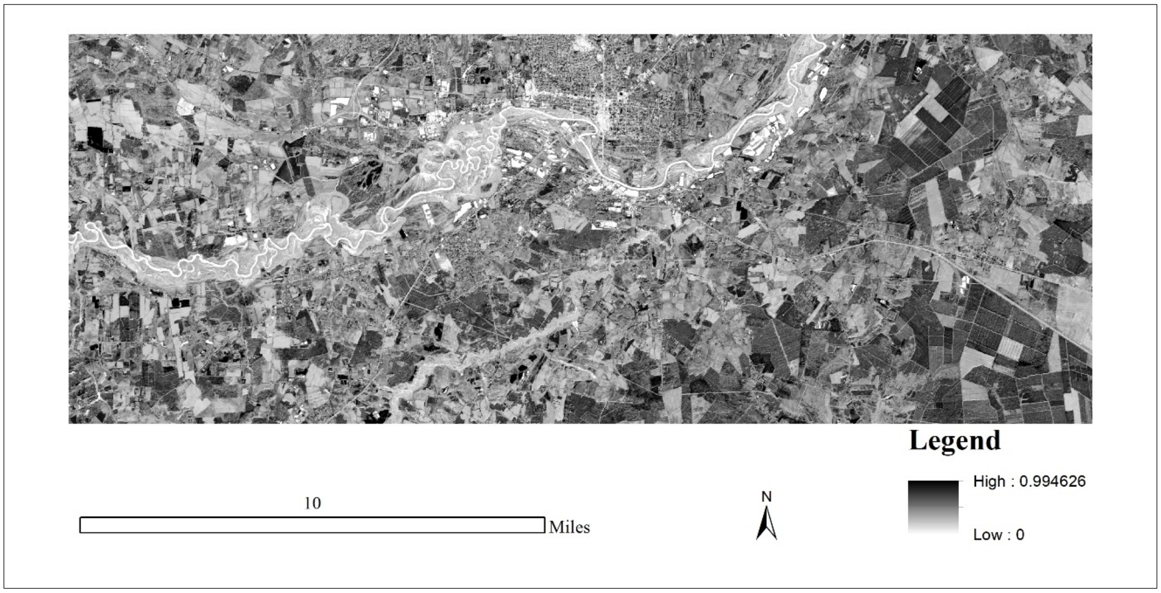


Figure 5. EVI based on Sentinel data for the study area

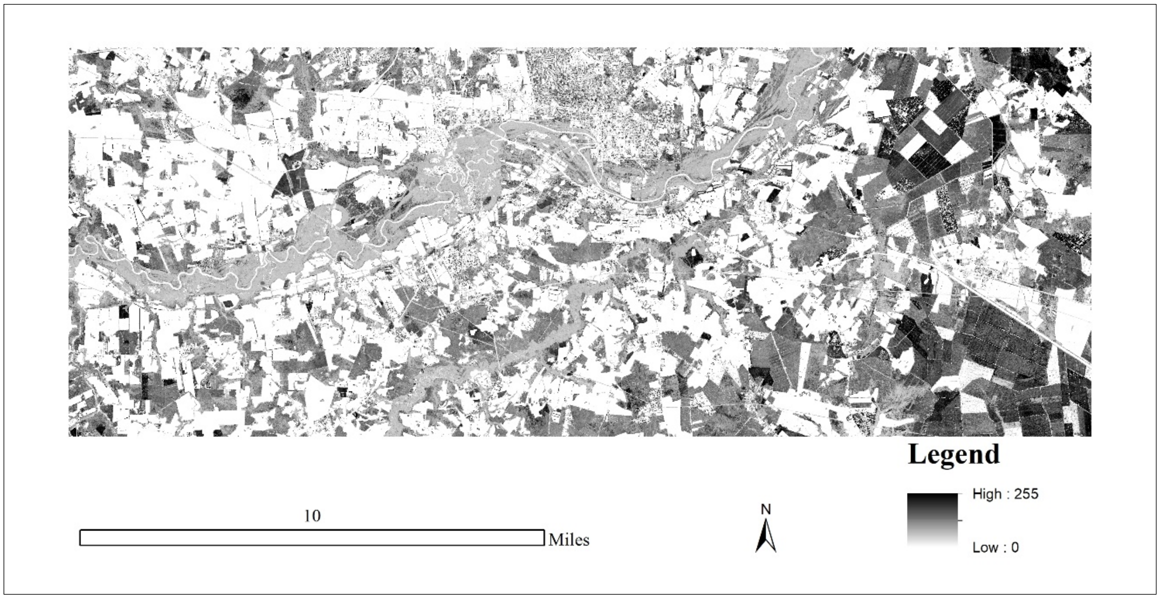


Figure 6. Intensity of the high vegetation LiDAR points

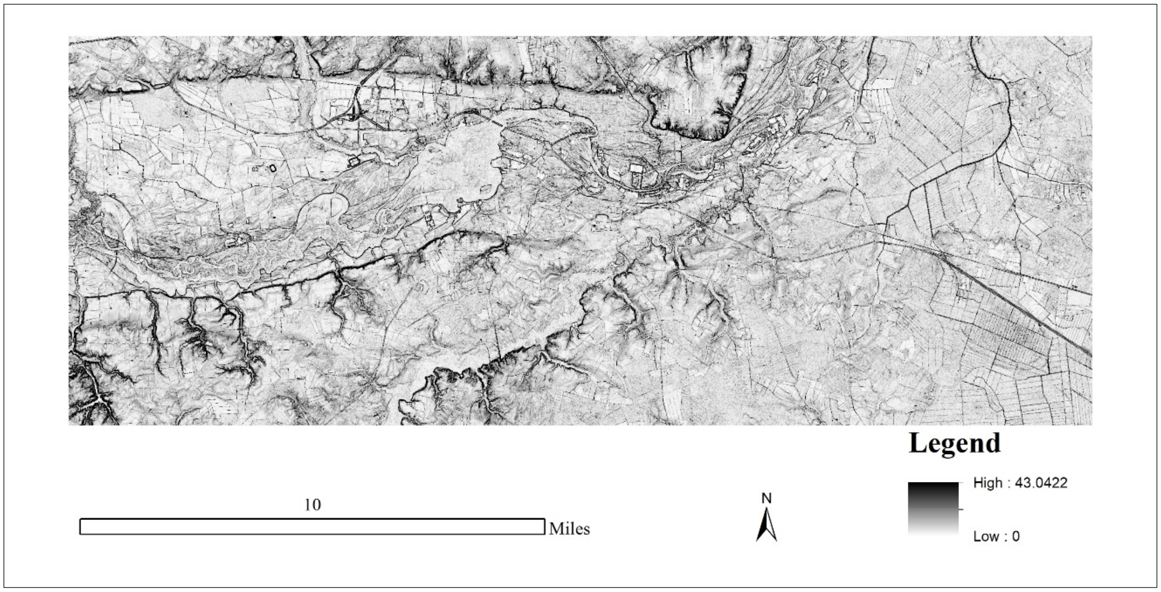


Figure 7. Slope layer based on DEM

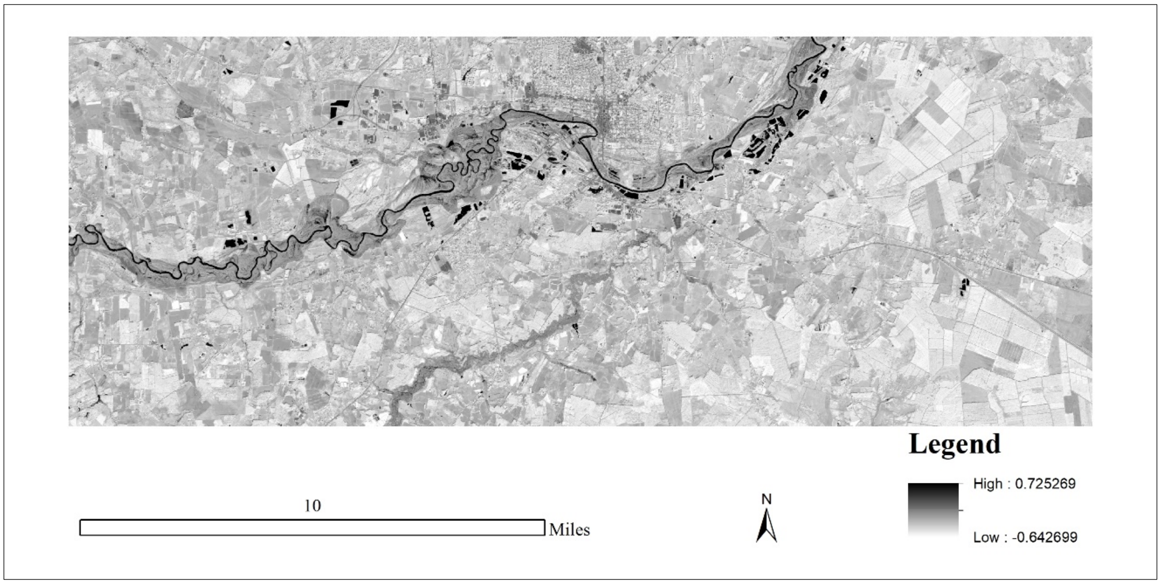


Figure 8. NDWI layer based on Sentinel data for the study area

3.3.2 Variable Generation for the Open-Loop Model

For the open-loop model, I generated spatial variable layers using several input datasets to represent the conditional layers in the decision tree. A number of environmental characteristics then collectively lead to the evidence to support specific wetland types.

Table 5 summarizes the basic characteristics of these wetland types from the main aspects of landscape position, water, soil, and vegetation. Based on the landscape positions where the wetlands occur, there exist three major categories, including salt/brackish marsh, riparian, and non-riparian type. Riparian wetlands refer to wetland types typically found in one of the following situations: (1) in a geomorphic floodplain; (2) at a natural topographic crenulation; (3) contiguous with open water covering 20 acres or larger; (4) subject to tidal flow regimes, excluding salt/brackish marsh. The four non-wetland open water types—natural waterbodies, artificial waterbodies, estuarine waters, and ocean—are not discussed here.

To measure and quantify the features through GIS tools, I gathered the datasets that can represent the characteristics. For each wetland type, I extracted the key factors for distinguishing this wetland type (as listed in Table 5). In the final step, I converted all the data to a binary raster layer, of which the value denotes whether the cell satisfies the condition of the key factor.

Table 5. Characteristics of different wetland types and related key factors

Wetland Type (code)	Characteristics	Key Factors and Criteria

Salt/Brackish Marsh (1)	**Landscape Position**: areas subject to regular or occasional flooding by tides, including wind tides;
	Ecoregion: in the tidewater region of the Middle Atlantic Coastal Plain ecoregion;
	Soil: organic and mineral soils;
	Vegetation: dominated by herbaceous vegetation (less than 50 percent coverage by woody species);
	Hydrology: with water salinities equal to or exceeding 0.5 parts per thousand during the period of average, annual low flow.
		soil: salt/brackish marsh soil;
		tidal affected area;
		water salinity: affected by salt/brackish water;
		vegetation (herb): percentage of low vegetation >= *t*[1];
		ecoregion: Middle Atlantic Coastal Plain

Estuarine Woody Wetland (2)	**Landscape Position**: wetlands occur on the margins of estuaries, they are typically fringing tidal marshes, and they are subject to occasional flooding by tides;
	Ecoregion: in the tidewater region of the Middle Atlantic Coastal Plain ecoregion;
	Soil: organic or mineral soils;
	Vegetation: dominated by woody vegetation including shrubs and trees.
		soil: estuarine woody wetland soil;
		tidal affected area;
		riparian area;
		water salinity: affected by salt/brackish water;
		vegetation (woody): percentage of middle and high vegetation >= *t*;
		ecoregion: Middle Atlantic Coastal Plain

Tidal Freshwater Marsh (3)	**Landscape Position**: wetlands occur on the margins of estuaries and in lower reaches of streams and rivers where they are saturated most of the time and are also subject to regular or occasional flooding by tides, including wind tides; **Ecoregion**: in the tidewater region of the Middle Atlantic Coastal Plain ecoregion; **Soil**: organic or mineral soils; **Vegetation**: dominated by herbaceous vegetation; **Hydrology**: with water salinities below 0.5 parts per thousand, but possibly	soil: tidal freshwater marsh soil; tidal affected area; water salinity: affected by fresh water; riparian area; vegetation (herb): percentage of low vegetation >= *t*; ecoregion: Middle Atlantic Coastal Plain

	exceeding this threshold as a result of storm events.	
Riverine Swamp Forest (4)	**Landscape Position**: wetlands occur throughout the state, but are most extensive and abundant in Coastal Plain ecoregions; **Soil**: organic or mineral soils; **Vegetation**: dominated by woody vegetation; **Hydrology**: seasonal to semi-permanent inundation.	soil: riverine swamp forest soil; tidal affected area; water salinity: affected by fresh water; vegetation (woody): percentage of middle and high vegetation >= *t*; riparian area; non-depression; hydrology: Axiom[2] Soil Tables Hydrology Code E5, E6

Seep (5)	**Landscape Position**: located throughout the state where groundwater is discharged to the surface on a slope not in a geomorphic floodplain or a natural topographic crenulation;
	Soil: organic or mineral soils;
	Vegetation: variable;
	Hydrology: semi-permanently to permanently saturated by ground water.
		soil: seep soil;
		non-riparian area;
		slope >=1%;
		non-depression

Hardwood Flat (6)	**Landscape Position**: mostly found on poorly drained, interstream flats;
	Soil: mineral soils;
	Ecoregion: mainly in Coastal Plain ecoregions;
	Vegetation: commonly dominated by hardwood tree species including various oaks;
	Hydrology: seasonally saturated or intermittently to seasonally inundated by a high water table or poor drainage, but have a shorter hydro period than Non-Riverine Swamp Forests.
		soil: hardwood flat soil;
		non-riparian region;
		slope <=1%;
		non-depression area;
		ecoregion: Coastal Plain;
		vegetation: deciduous in NLCD;
		hydrology: no evidence of flooding, the number of ground points >= *t*;
		Axiom Soil Tables Hydrology Code E1 – E5

Non-Riverine Swamp Forest (7)	**Landscape Position**: occurs primarily in the embayed region on poorly drained, interstream flats not contiguous with streams, rivers, or estuaries;
	Soil: mucky mineral or organic soils;
	Ecoregion: northeastern Middle Atlantic Coastal Plain ecoregion;
	Vegetation: forest vegetation;
		soil: non-riverine swamp forest soil;
		non-riparian area;
		slope <=1%;
		non-depression area;
		ecoregion: Middle Atlantic Coastal Plain region;
		vegetation: deciduous in

	Hydrology: seasonally to semi-permanently inundated with hydrology driven by groundwater discharge, overland runoff, and/or precipitation rather than overbank or tidal flooding.	NLCD; hydrology (evidence of flooding): the number of ground points; Axiom Soil Tables Hydrology Code E6

Pocosin (8)	**Landscape Position**: occurs on poorly drained, interstream flats and in basins of various sizes such as peat-filled Carolina bays;
	Soil: mineral or organic soils; special pocosin soil;
	Ecoregion: Coastal Plain ecoregion or southeastern plains;
	Vegetation: dominated by dense, waxy evergreen shrubs;
	Hydrology: seasonally saturated or inundated by a high or perched water table.
		soil: pocosin soil type only;
		non-riparian area;
		ecoregion: Middle Atlantic Coastal Plain or Southeastern Plain;
		depression area;
		vegetation: evergreen and shrub in NLCD;
		vegetation density: QL2 the percentage of middle and low vegetation falls within the highest quantile

Pine Savanna / Flat (9/10)	**Landscape Position**: occurs on poorly drained, interstream flats;
	Soil: mineral soils;
	Ecoregion: Coastal Plain ecoregion or Southeastern Plain;
	Vegetation: dominated by long-leaf and pond pine, with scattered, low shrubs and grassy ground cover;
	Hydrology: seasonally saturated by a high water table or poor drainage, but have a shorter hydroperiod than Non-Riverine Swamp Forest; little surface storage;
	Other: maintained by frequent, low-intensity fires.
		soil: pine soil;
		non-riparian area;
		ecoregion: Middle Atlantic Coastal Plain or Southeastern Plain;
		slope <=1%;
		non-depression area;
		vegetation: evergreen in NLCD;
		vegetation pattern: QL2 percentage of high vegetation falls within the highest quantile; middle vegetation intensity <= *t*

Basin Wetland (11)	**Landscape Position**: occurs throughout the state in depressions surrounded by uplands (usually on interstream flats or in localized depressions); may also occur on the fringe of small open waters (less than 20 acres in size);
	Soil: mineral soils;
	Vegetation: varies widely from forest to herbaceous or emergent;
	Hydrology: seasonally to semi-
		soil: basin soil;
		non-riparian area;
		slope <=1%;
		depression area;
		proximity: distance to small open waters <= *t*

	permanently inundated but may lose surface hydrology during later portions of the growing season; **Other**: seasonal waterlines are often apparent on the vegetation; this type is very heterogeneous.	

Bog (12)	**Landscape Position**: occurs in geomorphic floodplains or natural topographic crenulations and is typically located on flat or gently sloping ground;
	Soil: organic or mucky mineral soils;
	Ecoregion: typically in the Blue Ridge and Northern Inner Piedmont ecoregions;
	Vegetation: 1) dominated by dense herbaceous or mixed shrub/herbaceous vegetation; 2) tree cover over much of the wetland area and dense herb cover limited to small openings;
	Hydrology: at least semi-permanently saturated, but typically not inundated;
	Other: there are specific soils that support bog.
		soil: bog soil;
		riparian area;
		ecoregion: Blue Ridge and Northern Inner Piedmont;
		vegetation: percentage of low and medium vegetation points >= *t*

Non-Tidal Freshwater Marsh (13)	**Landscape Position**: throughout the state in geomorphic floodplains, in natural topographic crenulations, or contiguous with open waters 20 acres or larger;
	Soil: organic or mineral soils;
	Vegetation: predominantly herbaceous;
	Hydrology: semi-permanent inundation or saturation, but are typically not subject to regular or occasional flooding by tides.
		riparian area;
		non-tidal area;
		soil: organic or mineral soil;
		vegetation: percentage of low vegetation >= *t*

Floodplain Pool (14)	**Landscape Position**: throughout the state in geomorphic floodplains, may occur in abandoned stream or river channels (oxbows) or in localized depressions near the toe of slopes;
	Soil: mineral soils;
	Vegetation: trees are commonly found around the edge of the pool rather than growing within the pool; vegetation within the pool can be sparse or variable;
		soil: floodplain pool soil;
		riparian area;
		vegetation: percentage of middle and high vegetation >= *t*;
		depression: localized depression;
		hydrology: Axiom Soil Tables Hydrology Code E6

	Hydrology: semi-permanent inundation.	

Headwater Forest (15)	**Landscape Position**: throughout the state in geomorphic floodplains of first-order or smaller streams and in topographic crenulations without a stream;
	Soil: mineral soils;
	Vegetation: hardwood tree and shrub species are the predominant vegetation;
	Hydrology: relatively flat ground surface that provides little water storage; it frequently has surface flow, especially through ephemeral channels; intermittently inundated by surface water or seasonally saturated to semi-permanently saturated.
		soil: headwater forest soil;
		riparian area;
		vegetation: percentage of middle and high vegetation >= *t*;
		non-depression;
		stream: order <=1;
		surface flow: percentage of ground points >= *t*;
		hydrology: Axiom Soil Tables Hydrology Code E1 – E4

Bottomland Hardwood Forest (16)	**Landscape Position**: throughout the state in geomorphic floodplains of second-order or larger streams;
	Soil: mineral soils;
	Vegetation: dominated by a variety of hardwood tree species;
	Hydrology: generally intermittently to seasonally inundated; has ground surface relief that provides good water storage.
		soil: bottomland hardwood forest;
		riparian area;
		vegetation: percentage of middle and high vegetation >= *t*;
		non-depression;
		stream: order >= 2;
		surface flow: percentage of ground points >= *t*;
		hydrology: Axiom Soil Tables Hydrology Code E1 - E5;

Note: 1. *t* denotes a predefined threshold as a criterion for data classification; 2. Soil data is processed by the Axiom Company, and in the data table, hydrology code implements the inundation level.

3.4 Model Evaluation Method

There exist various selections of metrics to assess the performance of the wetland model by comparing the prediction results with ground truth classification. In most cases, we consider classification validated by fieldwork as the source of truth. For a more consistent reference, one can adopt the method to conduct fieldwork and divide the collected data into two separate datasets, one for training and the other for evaluating the results generated by the training model (Franklin and Moulton 1990, Lauver and Whistler 1993).

Researchers have developed various metrics for the evaluation process; most of the metrics are based on the confusion matrix, which is a table for tracking the prediction result of each class by quantifying the number of samples which are correctly classified and those that are incorrectly classified to other classes. This paper mainly used the evaluation metrics such as overall accuracy, precision, recall, and Kappa index.

Overall accuracy represents the percentage of all the correctly classified samples over the total samples, with the value range between 0 and 1 and the higher value denotes that more samples are predicted correctly. It is the most straightforward metric. However, it cannot exclude the cases that are randomly predicted correctly due to imbalanced samples among different classes. Based on the confusion matrix we can also calculate the recall and precision for each wetland type. Additionally, another four important concepts can help with better illustrating the definition of recall and precision. For each wetland type, “true positive” means the samples were predicted as the wetland type that they belong to; “true negative” represents the samples which do not belong to this wetland type that were correctly predicted as other wetland types; “false positive” denotes the samples that actually belong to other wetland types that were incorrectly predicted as this wetland type; and false negative corresponds to the samples that belong to this type were omitted and predicted as other types. Based on these concepts, precision denotes the rate of true positive samples over the sum of true positive and false positive samples. The precision is also called “user’s accuracy”, as the complementary of the commission error. Precision of a wetland type essentially tells how often the type presented on the map actually correspond to the true type on the ground. This metric refers to the reliability of the map (prediction data). Another metric recall, also known as the “producer’s accuracy”, represents the fraction of true positive samples among the sum of true positive and false negative samples. Recall represents as complementary to the omission error, it equals 100% minus omission error percentage. It describes the accuracy from the perspective of the mapmaker or data producer, and it tells us how often the wetland type that occurred on the ground can be predicted on the map as such.

Cohen’s Kappa index evaluates the performance of the prediction model compared to randomly assigning prediction results to samples. It adjusts the prediction accuracy by considering the probability of correct prediction by chance. According to the confusion matrix, Kappa considers not only the diagonal agreements used in accuracy but also the off-diagonal instances. Formula (12) shows the calculation process (Cohen 1960, Cohen 1968). The observed agreement was the total number of instances that appear in the diagonal of the confusion matrix. For each class, multiply the proportion of agreement with the marginal rates of the two classifiers for predicting this class.

$$Kappa = \frac{OA - AC}{T - AC} \tag{12}$$

where OA is observed agreement (i.e., the total number of instances where both classifiers agree), T means the total number of samples, and AC denotes agreement by chance (agreement with a random classifier).

The Kappa index ranges from -1 to 1, i.e., 0 means the classification is no better than a random classification. A negative value means the classification is worse than random prediction. A value close to 1 indicates that the prediction is significantly more accurate than random classification and the prediction matches the ground truth observations.

CHAPTER 4: CASE STUDY

In this chapter, I first introduce the study area and data used in this research. In the second section, I illustrate the implementation of the expert system, including the steps for both modeling methods. In the end, I also introduce the implementation of the calibration process which will be further applied in the experiments.

4.1 Information

4.1.1 Study Area

The study area represents a potential Kinston bypass roadway corridor in Lenoir and Jones counties in eastern North Carolina (Figure 9). Within the corridor, the NC Department of Transportation (NCDOT) conducted fieldwork in 2015 that resulted in the identification of wetland/upland boundaries and the identification of wetland types. In 2017, they revisited the nearby regions and conducted further wetland delineation for several sites. They digitalized all of the fieldwork results in ArcGIS and labeled the wetland polygons with the wetland type code according to the NCWAM wetland classification system, developed by the WFAT for identifying the specific wetland types that occur in North Carolina (N.C. Wetland Functional Assessment Team 2010).

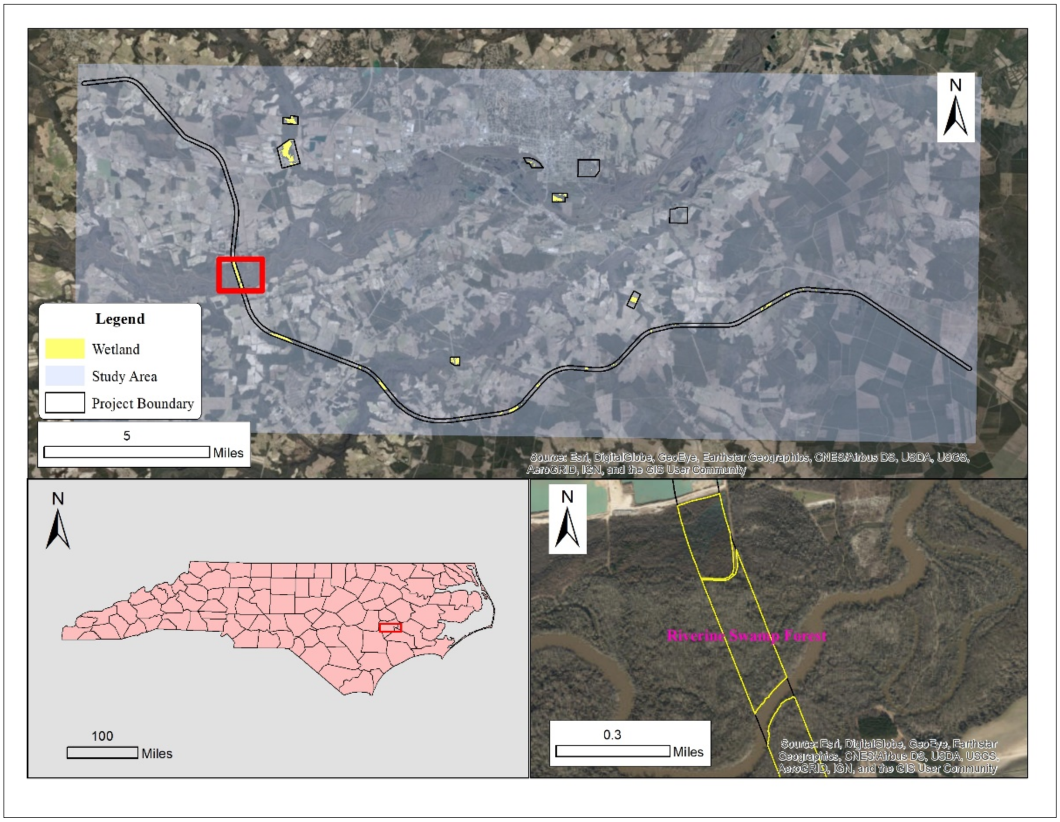

N
Legend
Wetland
Study Area
Project Boundary
5
Miles
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community
N
100
Miles
N
Riverine Swamp Forest
0.3
Miles
Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Figure 9. Study area of Kinston bypass roadway corridor

4.1.2 Data Sources

This dissertation applied several major datasets for the study area. I adopted Level 2 (QL2) LiDAR data which were collected during winter, and provided by NCDOT project as LAS files. The major types of information stored in this data file format include the x, y, z location of each point, class label of each point, four types of pulse returns (first, second, third, and last) and intensity of each return type (Heidemann 2012). The data has two pulse points per square meter, and 0.18-meter fundamental vertical accuracy. Based on the LiDAR data files, one can further produce other various data derivatives, such as DEM layers with different resolutions. Other than LiDAR data, I downloaded some free Sentinel-2 satellite data to take advantage of the multi-spectral bands information. I identified one Sentinel scene that contains the study area to build several vegetation-related metric layers, such as EVI, NDVI, and NDWI. I selected the scene with the capture date of February 1st, 2016 due to data quality (less cloud coverage) and proximity to field sample collection time (the capture date is near winter and between the years of the two collection dates). I also collected some other spatial data layers for auxiliary information, such as land cover information, soil data, and hydrological information (Table 6).

Table 6. Spatial data used in the research

Data	Main Source	Information
LiDAR	NCDOT	Source: NC QL2 project Resolution: 2 points per square meter Date: 2014 to 2017
Sentinel-2	European Space Agency (ESA)	Source: https://sentinel.esa.int/web/sentinel/sentinel-data-access Resolution: 10 meters (32.81 feet) Date: February 1st, 2016
Land Cover Data	National Land Cover Dataset	Source: https://www.mrlc.gov/ Resolution: 30 meters (98.43 feet) Date: 2011
Soil Data	Natural Resources Conservation Services (NRCS) Soils Database and published county soil surveys	Source: http://websoilsurvey.sc.egov.usda.gov Resolution: ranging from 1:12,000 to 1:63,360 Publication Date: 1994
Hydrology Data	National Hydrography Dataset	Source: https://nhd.usgs.gov/ Resolution: 1:24,000 or better Produced and updated: from 1950s to the present

4.1.3 Wetland Classification System

NCWFAT developed the wetland classification system applied in this dissertation, with the purpose of providing a unified list of wetland types for North Carolina to distinguish their inherent differences in ecological functions. There are 16 wetland types in total (see Table 5). Figure 10 summarizes the dynamic relationships among different wetland types based on the type definitions in NCWAM. The wetland can transit from one type to another due to the long-term change of environmental conditions. The arrows denote the possible transitions. This figure reveals the dynamic nature of wetland systems, as well as the challenge in type classification.

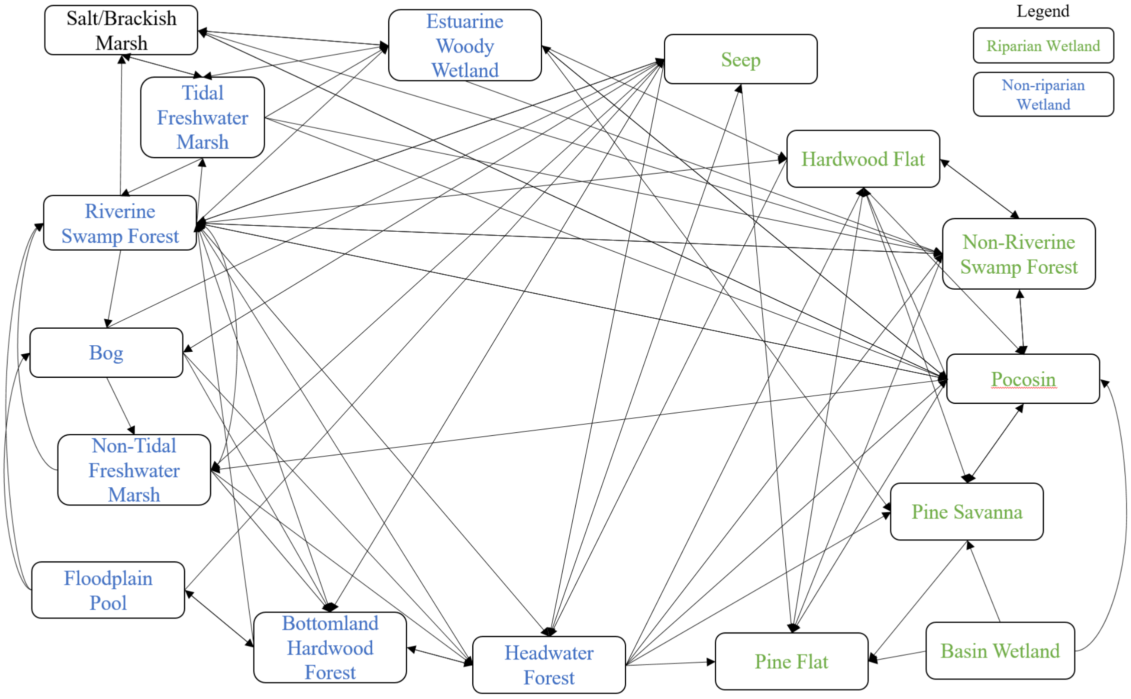


Figure 10. Transition types within the wetland system (N.C. Wetland Functional Assessment Team 2010)

4.1.4 Wetland Type Distribution

According to the field sample data, we detected 9 wetland types in the study area. Table 7 lists the number of wetland patches (minimum continuous area with same wetland type) and the total area for each wetland type. The dominant wetland types include riverine swamp forest and bottomland hardwood forest. They have the highest area percentage. Other than these two types, pocosin and pine have a relatively high percentage. Seep and hardwood flat are rarely shown in this case. As for the average size for each wetland patch, hardwood flat has the smallest size, followed by seep. Riverine swamp forest has the largest average size. Pocosin, non-tidal freshwater marsh, bottomland hardwood forest, and pine also tend to occur in a continuous large area.

Table 7. Wetland distribution in the study area

Wetland Code	**Wetland Type**	**Area (square meters)**	**Area Percentage (%)**	**Number of Polygons**
4	Riverine Swamp Forest	415,194.456	33.778	16
5	Seep	155.536	0.013	1
6	Hardwood Flat	47.884	0.004	2
8	Pocosin	108,268.920	8.808	7
10	Pine Flat	169,055.520	13.754	16
11	Basin Wetland	1,940.141	0.158	9
13	Non-Tidal Freshwater Marsh	25,908.250	2.108	2
15	Headwater Forest	38,728.181	3.151	23
16	Bottomland Hardwood Forest	469,879.197	38.227	40
Total		1,229,178.084	100	116

4.2 Implementation

I implemented the Expert System using ArcGIS-based scientific workflow. Corresponding to the construction of the database and knowledge base, I developed modules of generating data variables and constructing models. In this section, I mainly introduce the design and implementation of the major modules in the expert system. The ArcGIS platform provides us an automation solution through scientific workflow to orchestrate spatial operations. Figure 11 demonstrates the atom unit of a GIS workflow, which includes a core function, input and output data. This encapsulation method thus enables the repetition of function executions and the connections among multiple functions through a workflow. In this component, a GIS function can represent an existing tool in ArcGIS, or user-developed script. In this dissertation, I developed several new GIS tools based on Python to construct the new workflow. These tools inherited the user interface style of ArcGIS and can be added to a workflow through drag-and-drop operations through "ModelBuilder" platform.

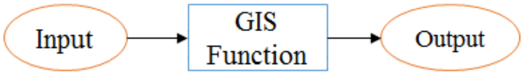


Figure 11. The unit component of GIS workflow

4.2.1 Variable Generation Module

The variable generation module contains three major tools. The first tool can generate DEM derivatives (referring to Table 4) based on the DEM data, satellite data, soil data, vegetation data, and so on, with the output results of GIS raster layers. The second tool can then generate sampling points for the entire study area, in order to extract values from the raster layers generated by the first tool. The third tool can then convert the attribute table of the point shapefile into a training data table in csv format, of which each column refers to value extracted from one data layer.

4.2.2 Models Construction Module

Based on the newly generated data variables, the next step is to construct models. For the closed-loop method, I developed tools and workflows to train machine learning models. For the open-loop method, I translated expert knowledge into spatial variables through similar workflows in ArcGIS.

(1) The closed-loop method

I implemented three machine learning models (GLM, RF, GBM) through an R package “H2O”. This package supports efficient parallel executions of machine learning algorithms. I developed python scripts to call functions from this R package for the automation of machine learning procedures, see Figure 12 as the workflow.

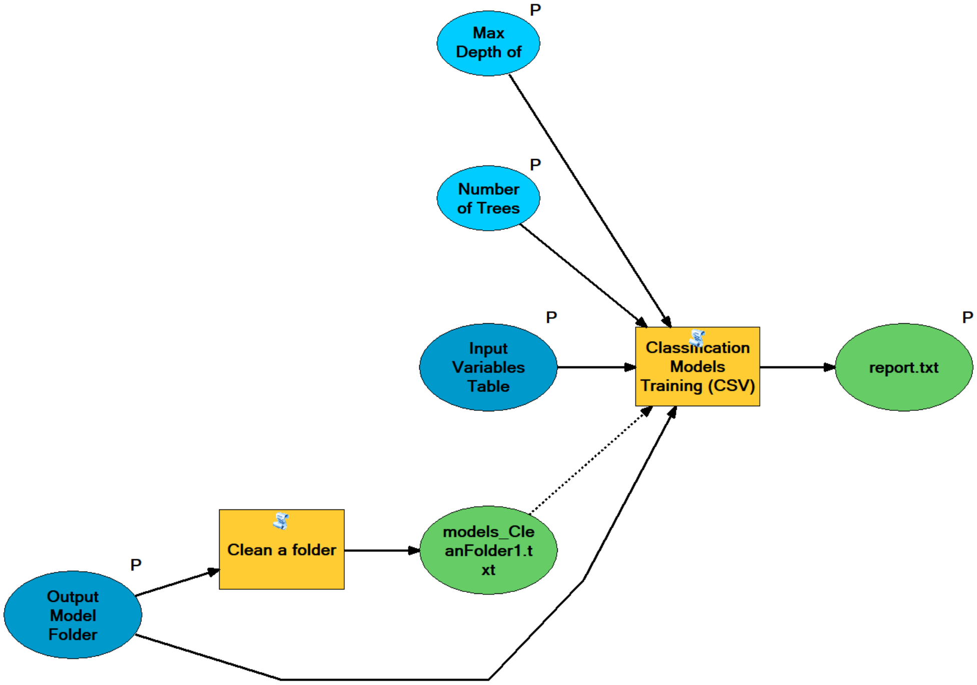
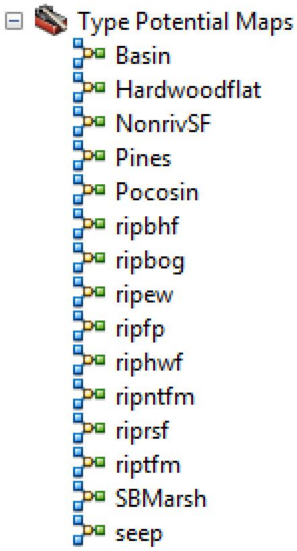


Figure 12. Module for training machine learning models

(2) The open-loop method

According to the probability calculation method illustrated in section 3.2.2, I developed a workflow to calculate the probability for each wetland type (see the toolbox in Figure 13). Each workflow contains the data variables that are important for this wetland type. For example, Figure 14 shows the variables used for seep wetland type, and the tool “raster calculator” specifies the formula of calculating the probability. For this case, it is calculated as: (1 - "%riparian%") * "%slpsteep%" + 0.01 * "%seepsoil2%". It means, when the wetland locates in the non-riparian region and on a side slope, it is then considered as seep. The calculation follows the definition of seep wetland (see the overall decision tree for all wetland types in Figure 4). It is noted that for each wetland type, I used a soil data layer to adjust the probability calculation of the corresponding wetland type.



Type Potential Maps
- Basin
- Hardwoodflat
- NonrivSF
- Pines
- Pocosin
- ripbhf
- ripbog
- ripew
- ripfp
- riphwf
- ripntfm
- riprsf
- riptfm
- SBMarsh
- seep

Figure 13. Toolbox for calculating wetland types

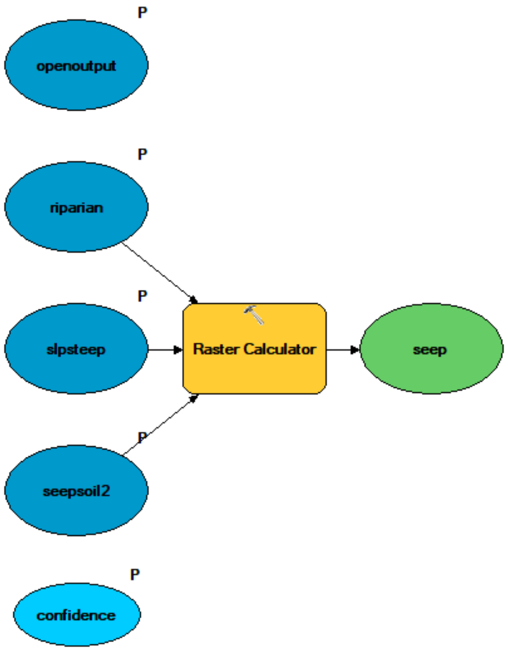


Figure 14. Probability calculation for seep wetland

4.2.3 Prediction Module

In this module, I used the models trained in previous modules to the entire study area to predict wetland type distribution. Considering that the entire study area can be large, I applied a decomposition process to divide the whole study area into sections of, at most, 100 rows in the raster layer. I ran the constructed models on each section individually to generate classification results. In the final step, I merge all the results together as the result of the entire study area (see Figure 15).

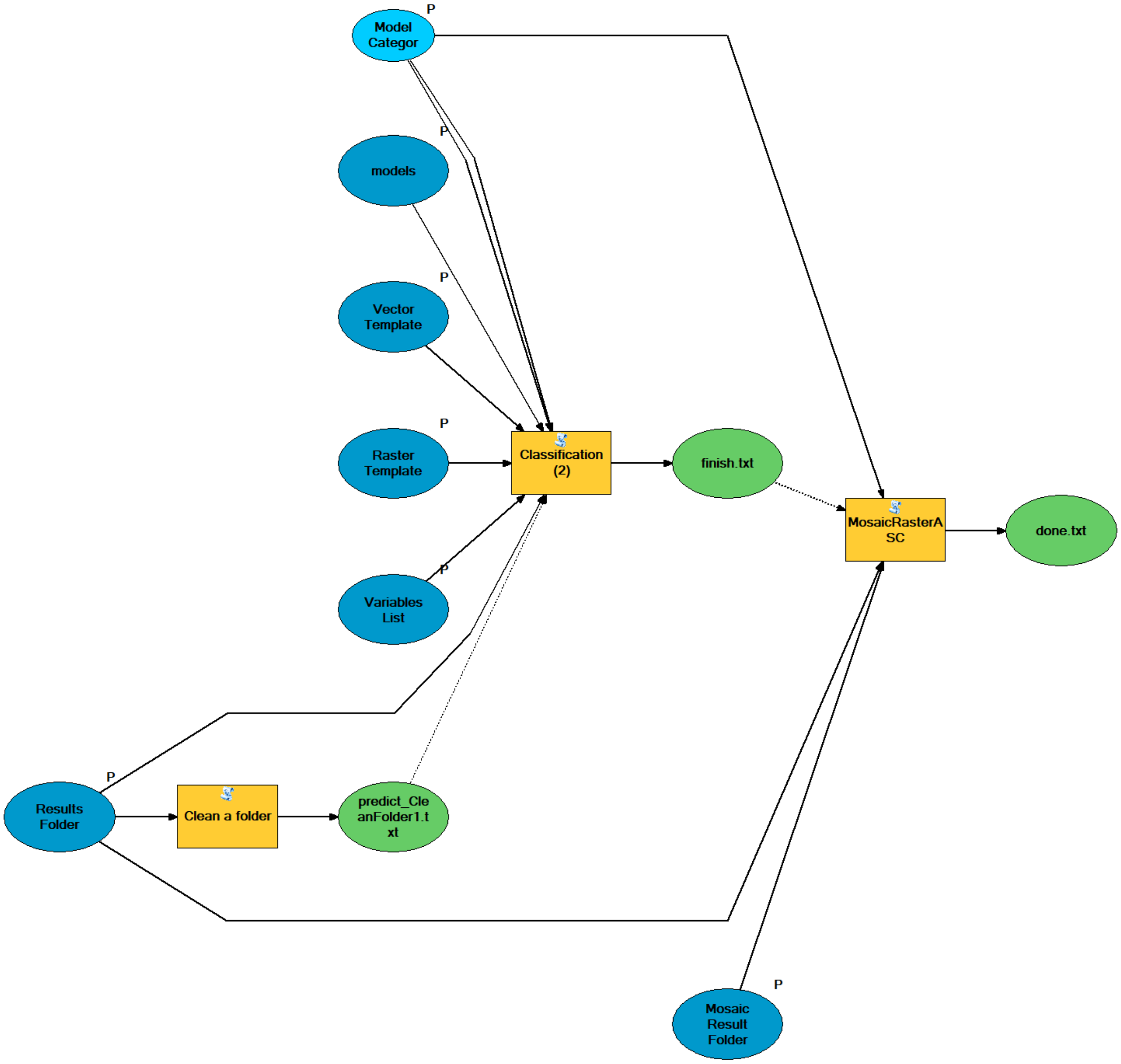


Figure 15. Workflow to generate prediction results

4.2.4 Model Integration

This module integrates the results generated by the closed-loop and open-loop methods. In this process, I introduced GLM, RF, and GBM prediction results to affect the wetland probability calculation from the open-loop method.

This workflow includes two key parameters, “Threshold for Integration” (WT) and “Weight for the Closed-Loop Model” (WC). They both have value with a range of 0 to 1. I use one wetland type “riverine swamp forest” as an example to show how the integration works. Formula (13) denotes the logic process of the integration. In this formula, one layer (“riprsf”) represents the integration of the three machine learning prediction results using corresponding weights—see Formula (7). Another layer (“rsf”) denotes the probability calculated through the open-loop method. I first compare whether the closed-loop integration (“rsf”) is larger than a threshold (“Threshold for Integration”). If yes, I consider the machine learning methods as highly consistent with each other, and can then be integrated into the final probability calculation a weight. Otherwise, if the ML models do not agree with each other, I only rely on the open-loop result.

Con("%rsf%" >=float(%WT%),"%riprsf%" * (1 - float(%WC%)) + float(%WC%) * "%rsf%","%riprsf%")   (13)

After calculating the final probability value of each wetland type for each location, one can then execute a spatial competition process among wetland types to determine the most probable type for each location. This pixel-based assignment process does not consider the situation of neighboring cells leading to a “salt and pepper” effect. One can further apply spatial rules to refine the prediction result, such as applying a wetland type contiguity matrix or using “focal statistics” to generate a smoother result.

4.3 Calibration

Calibration represents a critical phase to adjust model parameters to ensure that the prediction results are satisfactory. However, calibration is a challenging task for this integrated expert system due to the complicated interactions among different modules, which contributes to the non-linear relationship between model parameters and results. The intensity of calibration also relies on the number of parameters and the number of parameter combinations, which forms the calibration space. In this section, Formula (7) shows the targeted parameters set of $\{ w_1, w_2, w_3,\ w_c, w_o, t\}$. Formula (9) and Formula (10) illustrate the constraints and relationships for the parameters. Based on the constraints, we can further reduce the calibration space to a subset of $\{w_c, w_2, w_3\}$. In this paper, I used an iterative random searching method. Figure 16 shows the calibration process using two parameters (v1 and v2) as an example. The first step is a general grid search, in which I applied a grid to regularly divide the parameter space evenly. The red dot represents the parameter combination that returns the best performance among all the sampling parameter sets (blue dots). Second, perform another search to identify the best parameter combination from the first round. It aims at identifying the direction that gives the best performance improvement. I randomly select the parameter combinations that locate on the radius circle of previously selected parameter set. If the performance is improved for any parameter set, the process will continue with the new best result as the centroid to draw another radius for the next search. Repeat this process until no better results can be identified on the circle.

Table 8 shows the pseudo code for the calibration.

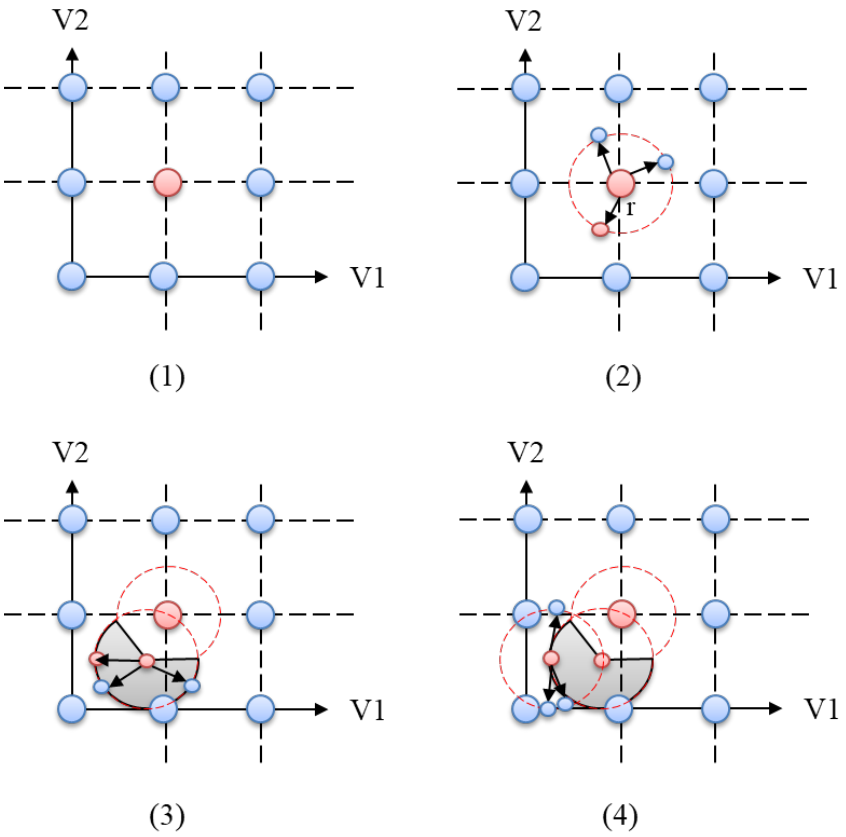


Figure 16. The calibration process for the model parameter search (an example of two-dimensional search space)

Table 8. Calibration process implemented for the wetland expert system

Algorithm:

Input: searching grid size $\boldsymbol{s_t}$, searching radius $\boldsymbol{s_r}$ (usually half size of $\boldsymbol{s_t}$), {(minimum value $\boldsymbol{minVal_i}$, **maximum value** $\boldsymbol{maxVal_i}$)| $\boldsymbol{i = 1, 2, 3}$} where $\boldsymbol{i}$ is the index for 3 parameters list [$\boldsymbol{w_c, w_2, w_3}$]

Output: optimal parameters set $\boldsymbol{O}$ denoting optimal value for [$\boldsymbol{w_c, w_2, w_3}$]

```
Initialization:
SET highest accuracy of the model as $A_{model} \leftarrow 0$
SET controller to stop the searching as $counter \leftarrow 5$
LOOP PROCESS FOR THE FIRST ROUND AS GRID SEARCH:
    For $v_1 \leftarrow minVal_1$ to $maxVal_1$ with a step $s_t$ DO
        For $v_2 \leftarrow minVal_2$ to $maxVal_2$ with a step $s_t$ DO
            For $v_3 \leftarrow minVal_3$ to $maxVal_3$ with a step $s_t$ DO
                $A_{\{v\}} \leftarrow$ model accuracy using $\{v_1, v_2, v_3\}$
                IF $A_{\{v\}} > A_{model}$ THEN
                    $A_{model} \leftarrow A_{\{v\}}$
                    $O \leftarrow [v_1, v_2, v_3]$
                END IF
            END FOR
        END FOR
    END FOR
THEN LOOP PROCESS FOR THE SECOND ROUND:
    WHILE $counter > 0$ DO
            FOR $i \leftarrow 0$ to 3 DO
                $\beta_1 \leftarrow$ random number selected from 0 to $2\pi$
                $\beta_2 \leftarrow$ random number selected from 0 to $2\pi$
                $v_{1i} \leftarrow v_1 + S_r * \cos\beta_1 * \sin\beta_2$
                $v_{2i} \leftarrow v_2 + S_r * \cos\beta_1 * \cos\beta_2$
                $v_{3i} \leftarrow v_3 + S_r * \sin\beta_1$
                $A_{\{v_i\}} \leftarrow$ model accuracy using $\{v_{1i}, v_{2i}, v_{3i}\}$
                IF $A_{\{v_i\}} > A_{model}$ DO
                    $A_{model} \leftarrow A_{\{v_i\}}$
                    $O \leftarrow [v_{1i}, v_{2i}, v_{3i}]$
                    $counter \leftarrow 5$ // reset counter
                ELSE
                    $\boldsymbol{counter \leftarrow counter - 1}$
                END IF
            END FOR
    END WHILE
    RETURN $\boldsymbol{O}$
```

CHAPTER 5: EXPERIMENTS

In this research, I designed three experiments to compare the performance of models in different situations. The study area in this paper refers to the Kinston bypass project corridor and nearby sites described in Chapter 4 in which NCDOT conducted wetland delineations. In the first experiment, I trained the closed-loop models using data samples randomly extracted from the dataset of the entire study area. I adopted a stratified random sampling strategy (Jensen and Lulla 1987) to maintain the same percentage of samples and spatial distribution for all wetland types. In this experiment, I used all the variables for model training and prediction. The objective of this experiment is to test the performance variation of the models by applying different sample sizes. In the second experiment, I conducted variable selection based on the performance of different variables in the ML training process demonstrated in Experiment I. The data for each treatment in Experiment II are consistent with the data from the first experiment, only without selected variables. The objective of this experiment is to examine whether the variable deduction process will affect ML model performance. I conducted Experiment III to test the performance of the integrated model by comparing classification results between the conventional ML models and the integrated models.

5.1. Experiment I

In this experiment, I conducted a stratified sampling strategy to select a subset of all the data samples for each wetland type as the training dataset. After the training process, I applied the models generated from this process to predict the wetland types for the entire study area. This experiment includes six treatments, with corresponding sampling percentages of 80%, 60%, 40%, 20%, 10%, and 5%. For RF and GBM models, the model parameters are set as 100 trees and the maximum depth of a tree is 5.

5.1.1 Performance Variation

Table 9 shows the training performance of the closed-loop models based on different sample datasets. GBM models always reveal the best performance by completely fitting the training dataset. Kappa for RF models are all above 0.9. The GLM model achieves higher kappa with smaller sample size comparing to the larger sample size.

Table 10 and Figure 17 show the performance of the models for predicting wetland types for the entire study area. Similarly,

Table 11 and Figure 18 show the results evaluated in the form of accuracy. The kappa and accuracy for all the ML models reveal the same pattern—they decrease as the sample size decreases. GBM models reveal the best performance among all the models, followed by RF and GLM models. Furthermore, GBM tends to reveal higher dependency on the sample size—its performance dropped more rapidly as the sampling size decreases. GLM tends to be least sensitive to the data sample size according to the standard deviation of its prediction performance.

Table 9. Training performance for the closed-loop models (kappa)

Treatment	Sample Percentage	GLM	RF	GBM	Number of Samples
T1	80%	0.88	0.90	1.00	26,478
T2	60%	0.88	0.90	1.00	19,860
T3	40%	0.88	0.91	1.00	13,238
T4	20%	0.89	0.91	1.00	6,620
T5	10%	0.89	0.93	1.00	3,309
T6	5%	0.91	0.93	1.00	1,655

Table 10. Prediction performance for the study area using different sampling datasets (kappa)

Treatments	GLM	RF	GBM
T1 (80%)	0.880	0.900	0.994
T2 (60%)	0.882	0.896	0.990
T3 (40%)	0.879	0.898	0.983
T4 (20%)	0.880	0.892	0.971
T5 (10%)	0.875	0.894	0.953
T6 (5%)	0.876	0.882	0.933
Mean	0.879	0.894	0.971
Std.	0.002	0.006	0.022

Table 11. Prediction performance for the study area using different sampling datasets (accuracy)

Treatments	GLM	RF	GBM
T1 (80%)	0.915	0.929	0.996
T2 (60%)	0.916	0.926	0.993
T3 (40%)	0.914	0.927	0.988
T4 (20%)	0.915	0.923	0.980
T5 (10%)	0.912	0.925	0.966
T6 (5%)	0.912	0.917	0.953
Mean	0.914	0.925	0.979
Std.	0.002	0.004	0.015

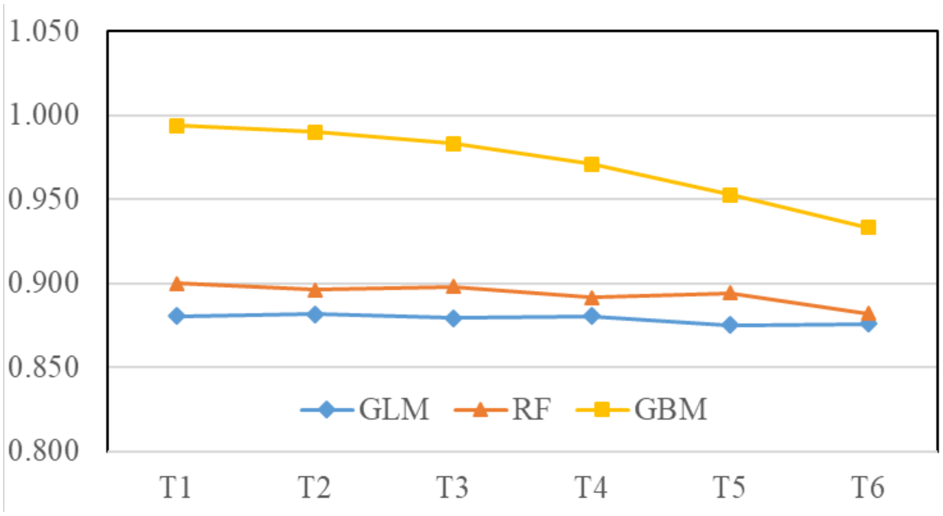


Figure 17. Prediction result of the study area based on the closed-loop models (kappa)

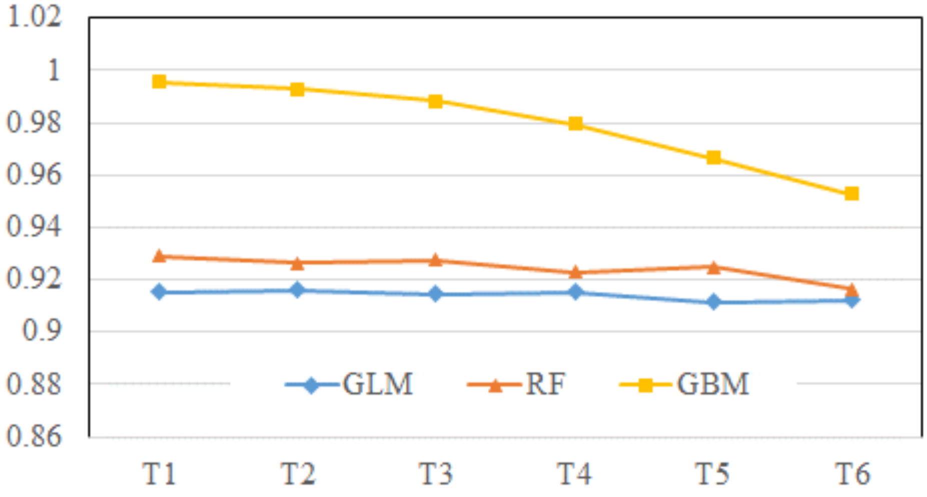


Figure 18. Prediction result of the study area based on the closed-loop models (accuracy)

5.1.2 Variable Importance

During the model training process of RF and GBM, the algorithm also calculates the variable importance basing on the decrease of squared error over all trees. Figure 19 to Figure 24 show the rank of variable importance for each treatment. For each treatment, the two models reveal very similar rank. Among all the treatments, “soil” is the most important variable. Other important variables include “elvpcd,” “land cover,” soil index, and vegetation indexes, such as “msavi,” “msavi2,” “evi,” “nvdi,” and “ndwi”. Some LiDAR-based variables also have very high rank, such as “highi,” “highp,” and “lowi”. However, most of the elevation derivatives tend to be less important in the training process. Based on the variable rank, the ten most important variables are: “soil,” “elvpcd,” “ndwi,” “ndvi,” “evi,” “lc,” “dropd,” “msavi,” “msavi2,” and “weid”.

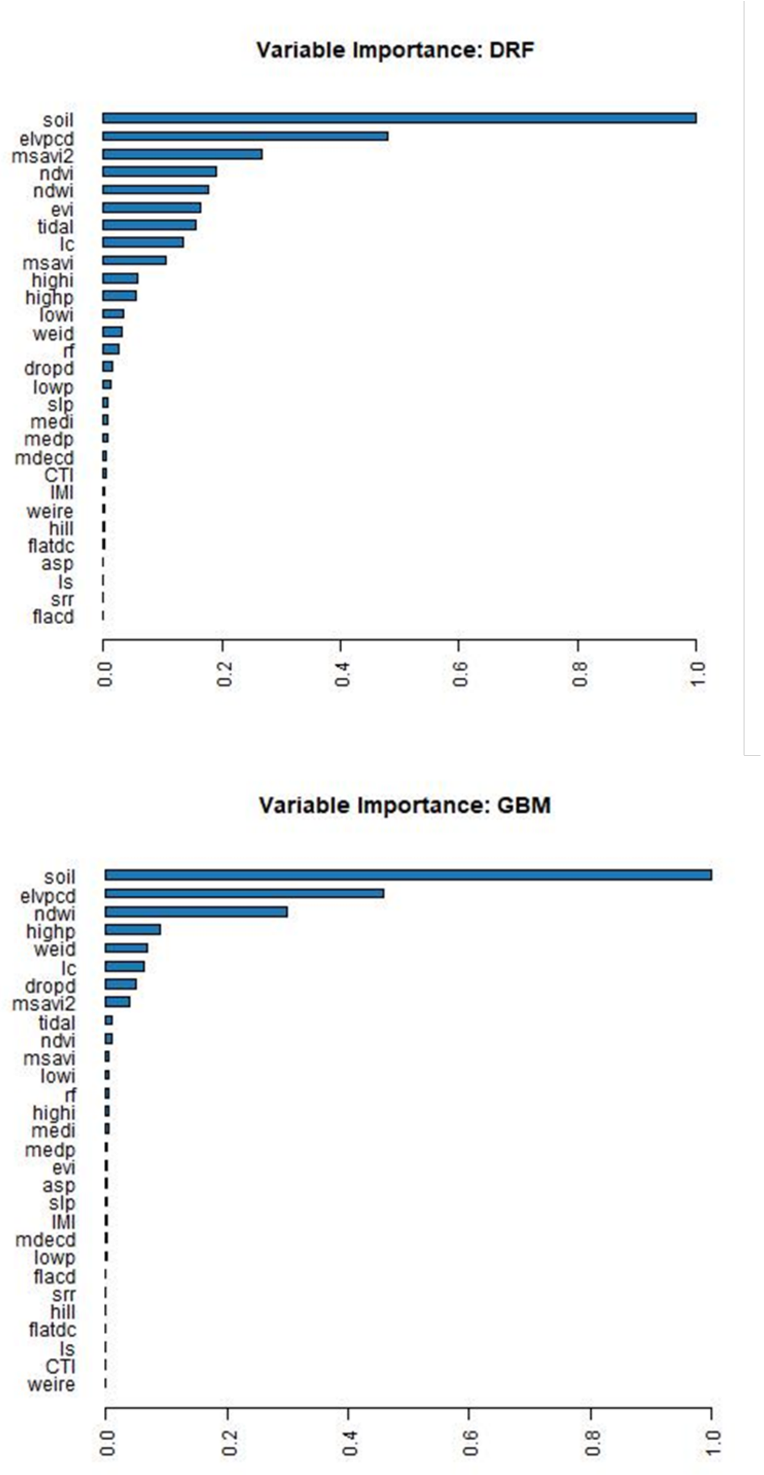


Figure 19. Variable importance rank for *T*1

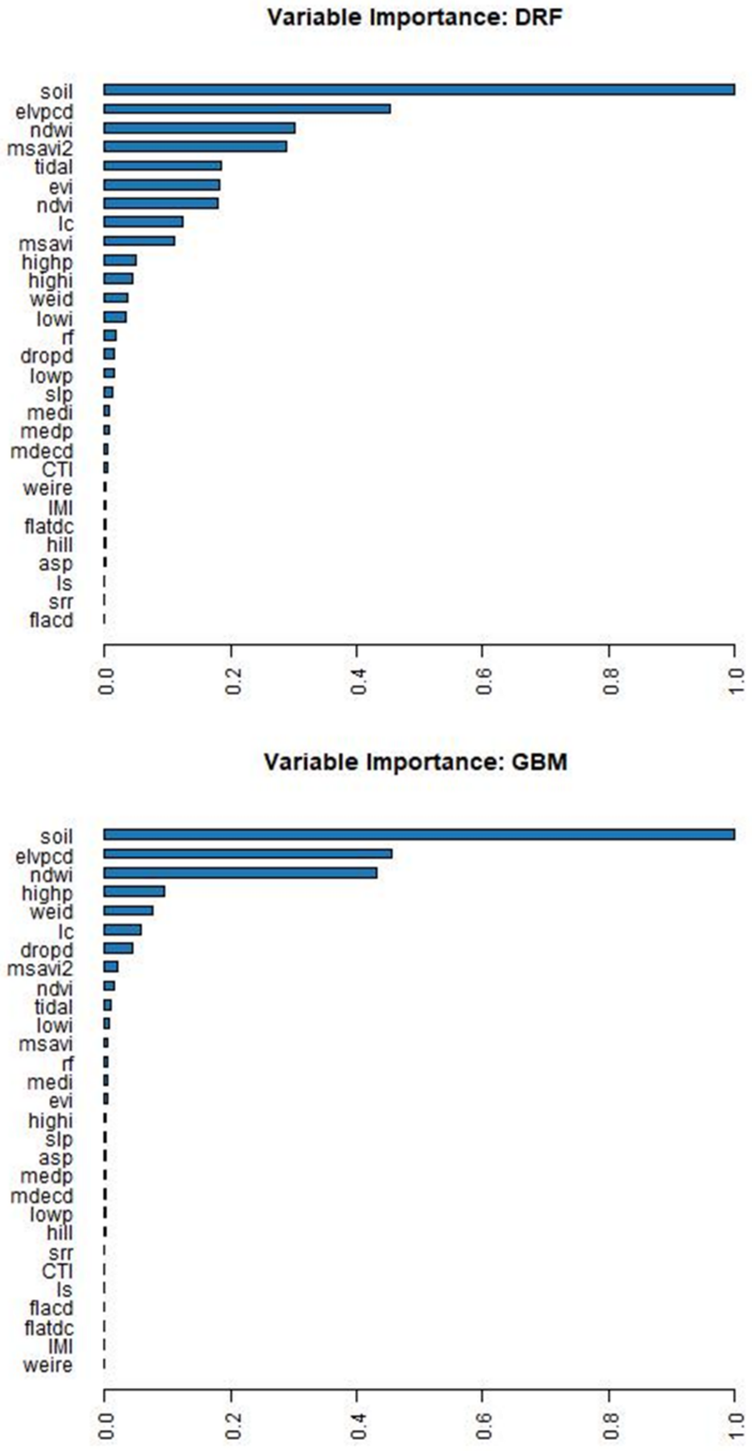
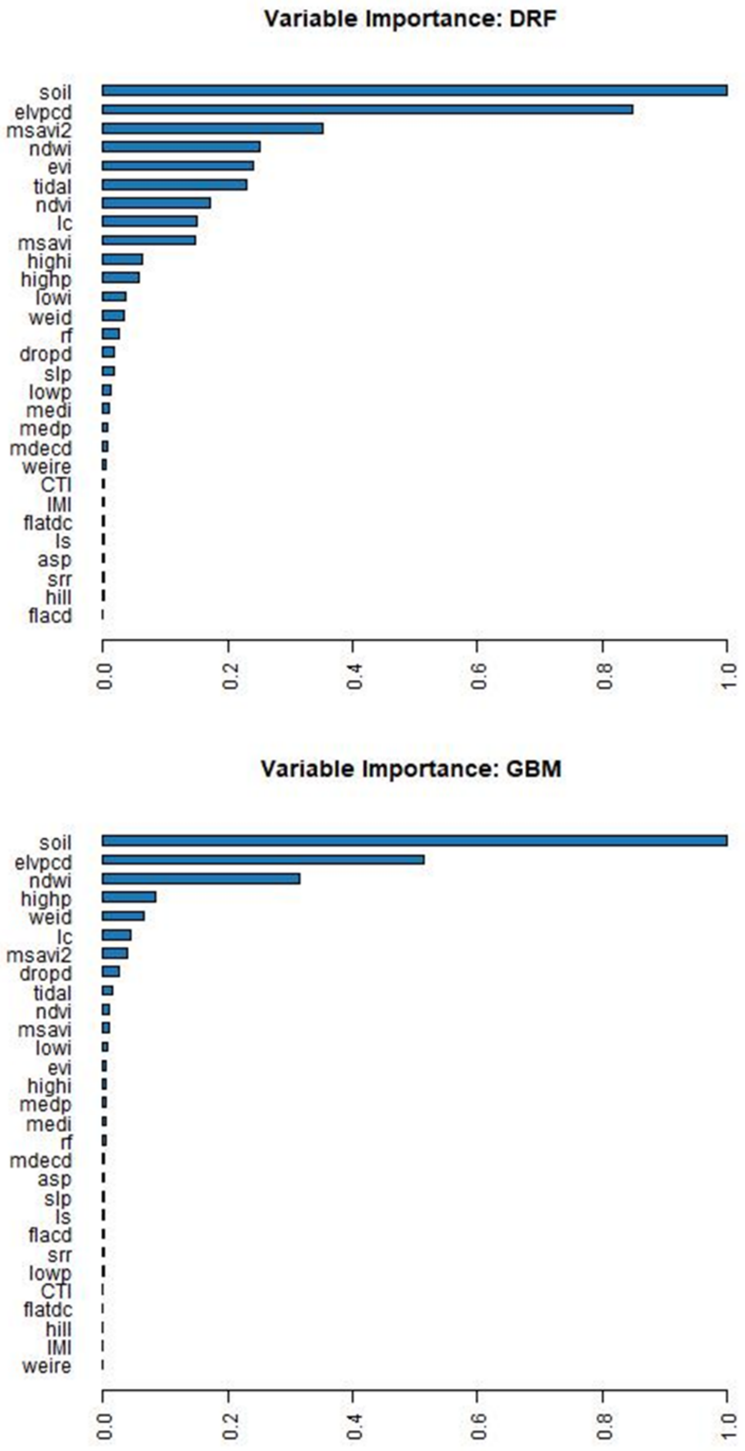


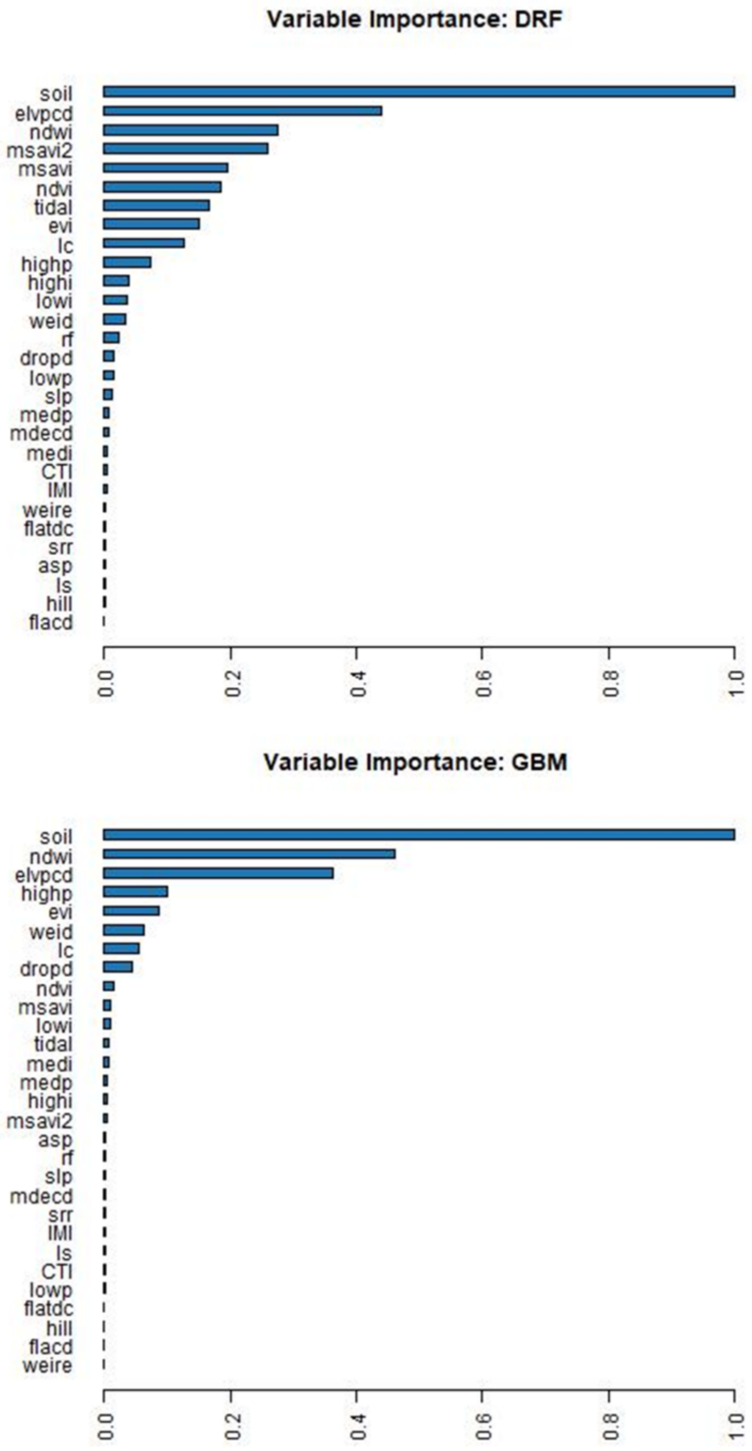
Figure 20. Variable importance rank for *T*2



Variable Importance: DRF

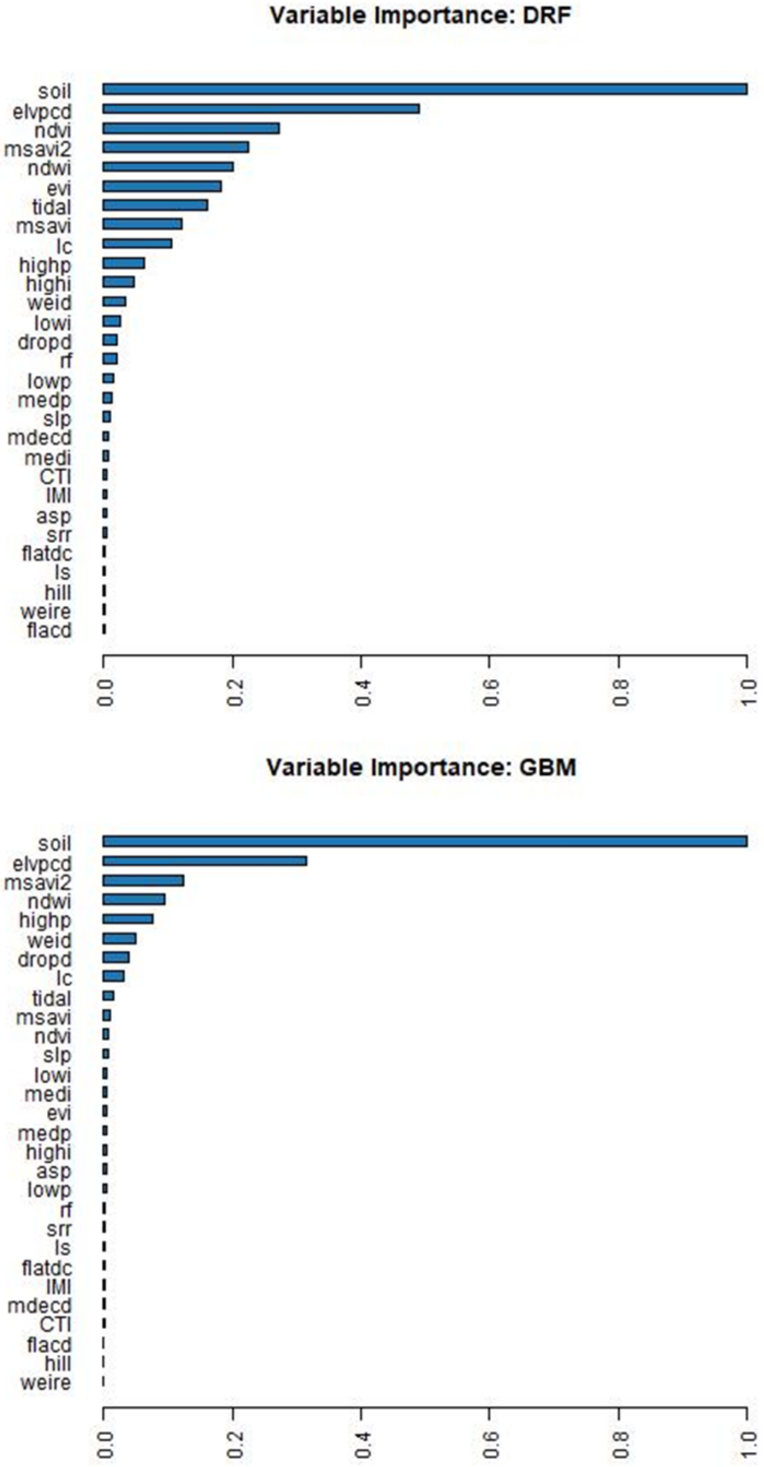
Variable Importance: GBM

Figure 21. Variable importance rank for *T*3


Variable Importance: DRF

Variable Importance: GBM

Figure 22. Variable importance rank for *T*4


Variable Importance: DRF

Variable Importance: GBM

Figure 23. Variable importance rank for *T*5

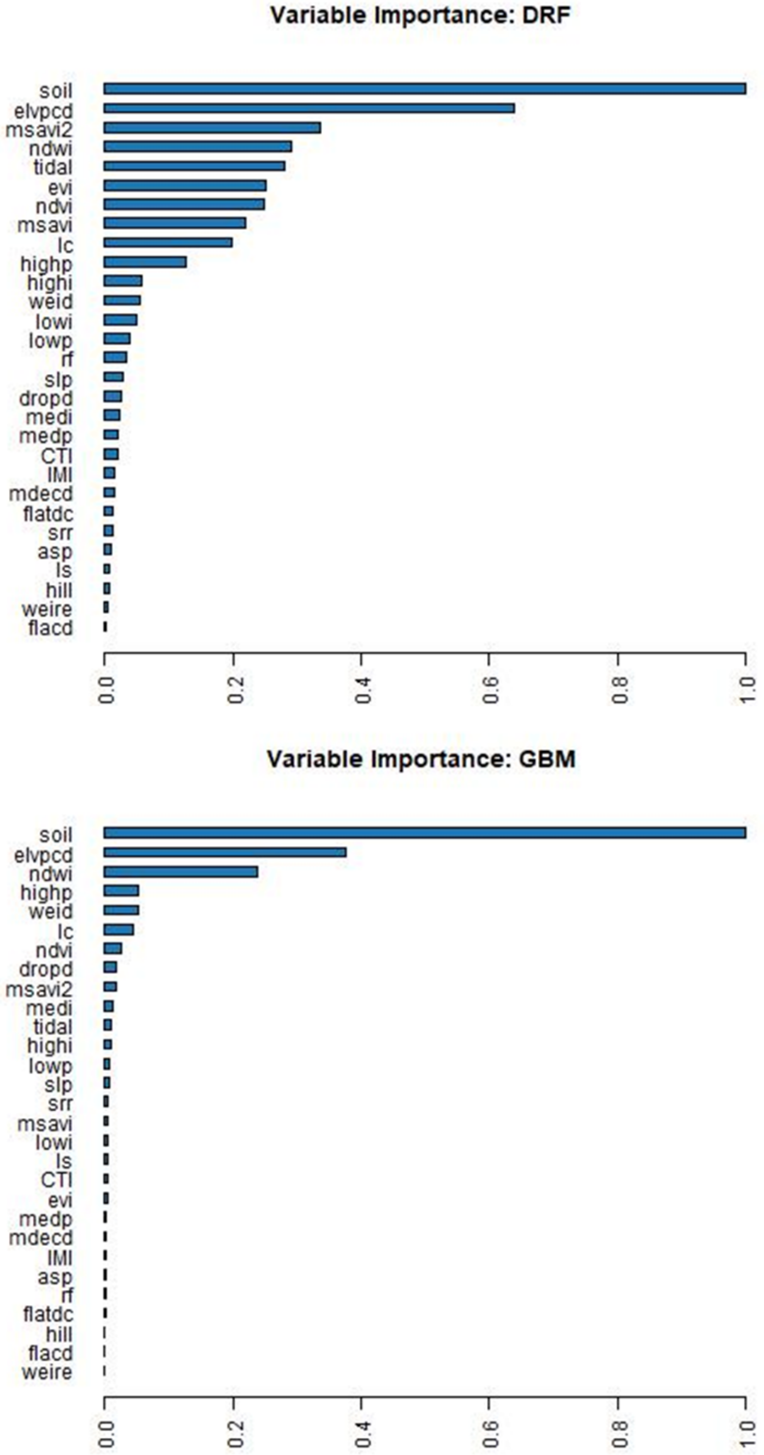


Figure 24. Variable importance rank for *T*6

5.1.3 Confusion Matrix

Table 12, Table 13, and Table 14 illustrate the confusion matrix for each ML model in Treatment 6. For GLM, type 10 (pine flat) has the highest recall followed by type 8 (pocosin); and type 8 has the highest precision followed by type 10. For the RF model, type 8 has the highest recall and type 11 (basin) has the highest precision. For GBM, type 8 has the highest recall and precision. Among all the models, type 5 (seep) and type 6 (hardwood flat) have the worst prediction.

Table 12. Confusion matrix of GLM in Treatment 6

	Classification									
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	9979	0	0	0	1	0	42	0	1176	11198	0.891
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	0	0	0	0	1	1	0.000
8	0	0	0	2833	78	0	0	0	0	2911	0.973
10	9	0	0	52	4445	2	0	29	8	4545	0.978
11	0	0	0	15	33	2	0	3	0	53	0.038
13	45	0	0	0	6	0	514	0	132	697	0.737
15	6	0	0	3	98	2	0	781	151	1041	0.750
16	927	0	0	1	2	0	31	49	11638	12648	0.920
Total	10966	0	0	2904	4663	6	587	862	13110		
Precision	0.910	NA	NA	0.976	0.953	0.333	0.876	0.906	0.888		

Table 13. Confusion matrix of RF in Treatment 6

	Classification									
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	10183	0	0	0	0	0	21	0	994	11198	0.909
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	0	0	0	1	0	1	0.000
8	0	0	0	2910	1	0	0	0	0	2911	1.000
10	0	0	0	10	4511	0	0	2	22	4545	0.993

11	0	0	0	6	44	3	0	0	0	53	0.057
13	71	0	0	0	0	0	377	0	249	697	0.541
15	5	0	0	18	78	0	0	764	176	1041	0.734
16	1012	0	0	2	13	0	19	13	11589	12648	0.916
Total	11271	0	0	2946	4647	3	417	780	13034		
Precision	0.903	NA	NA	0.988	0.971	1.000	0.904	0.979	0.889		

Table 14. Confusion matrix of GBM in Treatment 6

	Classification										
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	10598	0	0	0	0	0	44	0	556	11198	0.946
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	0	0	0	1	0	1	0.000
8	0	0	0	2911	0	0	0	0	0	2911	1.000
10	0	0	0	0	4521	0	0	5	19	4545	0.995
11	0	0	0	1	33	17	0	1	1	53	0.321
13	92	0	0	0	0	0	585	0	20	697	0.839
15	4	0	0	4	53	25	0	879	76	1041	0.844
16	520	0	0	0	9	0	80	23	12016	12648	0.950
Total	11214	0	0	2916	4616	42	709	909	12692		
Precision	0.945	NA	NA	0.998	0.979	0.405	0.825	0.967	0.947		

5.2. Experiment II

In this experiment, I used the same datasets from corresponding treatments in Experiment I but only keeping ten variables with the highest importance rank according to RF and GBM.

5.2.1 Variation of Model Training Performance

Table 15 lists the classification training results. Among the three ML models, GBM has the best performance, the kappa index for RF is also above 0.9. According to the training performance from Experiment I, kappa for GLM drops the most as the sample size decreases; kappa for GBM is slightly lower for only the first two treatments. However, the performance of RF is slightly improved comparing to the first experiment.

Table 15. Training performance for the closed-loop models with selected variables (kappa)

Treatment	GLM	RF	GBM	Number of Samples
T1 (80%)	0.87	0.91	0.99	26,478
T2 (60%)	0.87	0.92	0.99	19,860
T3 (40%)	0.87	0.92	1.00	13,238
T4 (20%)	0.88	0.93	1.00	6,620
T5 (10%)	0.88	0.94	1.00	3,309
T6 (5%)	0.88	0.93	1.00	1,655

5.2.2 Variation of Model Prediction Performance

Table 16 and Table 17 summarize the prediction results of applying the trained ML models to the entire study area. The accuracy of all the models is above 90%. GBM has the best performance, and the prediction accuracy can reach as high as 99%. Even for the treatment with the lowest sampling percentage, the GBM model can predict 98% of data samples correctly.

Figure 25 and Figure 26 also demonstrate the performance variation. Among all three models, GBM has the best performance and biggest variation. Comparing the results with the first experiment, GLM reveals degradation in performance while RF and GBM both have better performance. RF and GBM both reveal higher average accuracy and kappa comparing to the first experiment. Meanwhile, the variation for all three models are smaller in this experiment, it denotes that the models become slightly less sensitive to the change of sample size.

Table 16. Prediction performance for the study area using different sampling datasets with selected variables (kappa)

Treatments	GLM	RF	GBM
T1	0.867	0.911	0.990
T2	0.867	0.917	0.987
T3	0.867	0.913	0.982
T4	0.869	0.913	0.972
T5	0.863	0.906	0.954
T6	0.866	0.901	0.946
Mean	0.867	0.910	0.972
Std.	0.002	0.005	0.017

Table 17. Prediction performance for the study area using different sampling datasets with selected variables (accuracy)

Treatments	GLM	RF	GBM
T1	0.906	0.937	0.993
T2	0.906	0.941	0.991
T3	0.905	0.938	0.987
T4	0.907	0.938	0.980
T5	0.903	0.934	0.967
T6	0.905	0.929	0.961
Mean	0.905	0.936	0.980
Std.	0.001	0.004	0.012

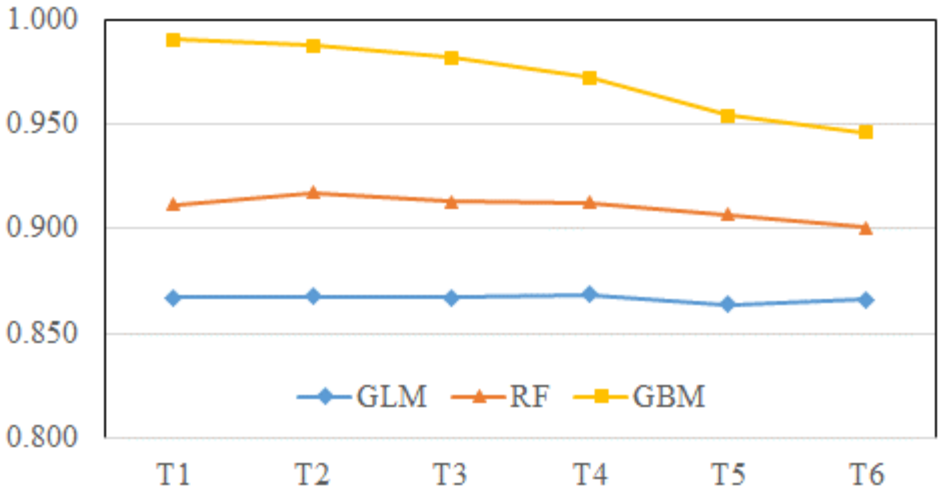


Figure 25. Prediction result of the study area based on the closed-loop models with selected variables (kappa)

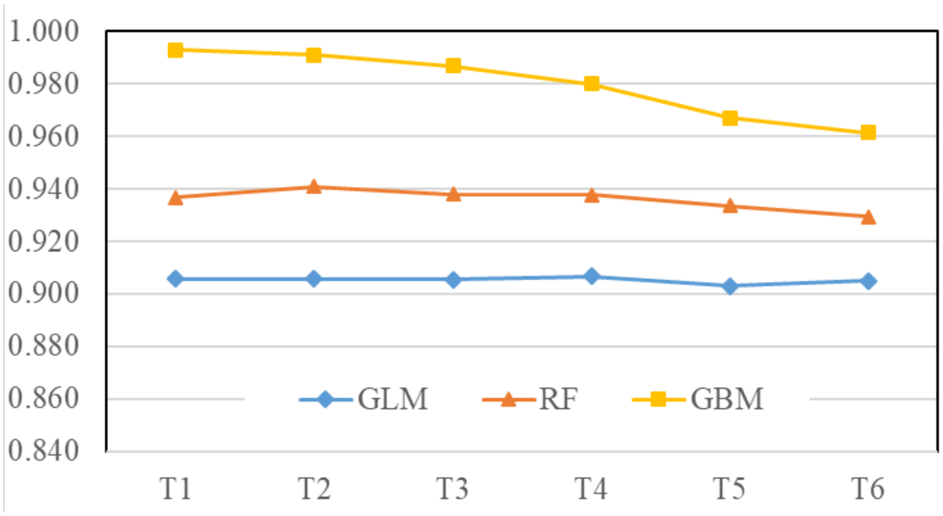


Figure 26. Prediction result of the study area based on the closed-loop models with selected variables (accuracy)

5.2.3 Confusion Matrix

Similar to the first experiment, Table 18, Table 19, and Table 20 list the confusion matrix for the three models applied in the last treatment. For all three models, type 8

(Pocosin) has the highest recall and precision; type 5 (Seep) and 6 (Hardwood Flat) have the worst prediction results. This result is consistent with the result of treatment 6 in Experiment I.

Table 18. Confusion matrix of GLM in Treatment 6 with selected variables

	Classification									
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	9913	0	0	0	0	0	93	0	1192	11198	0.885
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	0	0	0	0	1	1	0.000
8	0	0	0	2833	77	0	0	0	1	2911	0.973
10	0	0	0	115	4333	0	19	76	2	4545	0.953
11	0	0	0	8	39	0	0	6	0	53	0.000
13	30	0	0	0	0	0	550	0	117	697	0.789
15	1	0	0	7	117	0	0	766	150	1041	0.736
16	1020	0	0	0	3	0	15	54	11556	12648	0.914
Total	10964	0	0	2963	4569	0	677	902	13023		
Precision	0.904	NA	NA	0.956	0.948	NA	0.812	0.849	0.887		

Table 19. Confusion matrix of RF in Treatment 6 with selected variables

	Classification									
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	10340	0	0	0	0	0	81	0	777	11198	0.923
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	1	0	0	0	0	1	0.000
8	0	0	0	2911	0	0	0	0	0	2911	1.000
10	0	0	0	0	4516	0	0	10	19	4545	0.994
11	0	0	0	1	48	4	0	0	0	53	0.075
13	9	0	0	0	0	0	602	0	86	697	0.864
15	3	0	0	9	87	4	0	815	123	1041	0.783
16	1024	0	0	0	0	0	18	34	11572	12648	0.915
Total	11376	0	0	2921	4652	8	701	859	12581		
Precision	0.909	NA	NA	0.997	0.971	0.500	0.859	0.949	0.920		

Table 20. Confusion matrix of GBM in Treatment 6 with selected variables

	Classification									
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	10646	0	0	0	0	0	78	0	474	11198	0.951
5	0	0	0	0	0	0	0	0	4	4	0.000
6	0	0	0	0	1	0	0	0	0	1	0.000
8	0	0	0	2911	0	0	0	0	0	2911	1.000
10	3	0	0	0	4542	0	0	0	0	4545	0.999
11	0	0	0	1	6	32	0	14	0	53	0.604
13	21	0	0	0	0	0	651	0	25	697	0.934
15	5	0	0	1	41	36	0	917	41	1041	0.881
16	394	0	0	0	0	0	101	29	12124	12648	0.959
Total	11069	0	0	2913	4590	68	830	960	12668		
Precision	0.962	NA	NA	0.999	0.990	0.471	0.784	0.955	0.957		

5.3. Experiment III

The first two experiments are both based on the scenario that the samples for model training are extracted from the entire dataset while maintaining the same wetland type proportion and spatial distribution. It proves to be ideal for the models to capture the general characteristics of the study area and maintain stable prediction results. However, in real applications, it can be a challenge to implement a balanced sampling strategy. First, without knowing the spatial distribution of existing wetland types, it is difficult to conduct the survey and collect a similar amount of data samples for each wetland type. Second, some regions will have better accessibility than others, therefore the less easily accessed areas will possibly end up with less survey and data samples. Third, when the targeted area is large, it is time-consuming to collect spatially balanced data across the entire study area.

In this experiment, I divided the study area into two regions to simulate the scenario that the training data are collected from one area and the trained models are applied to a completely different region. These two regions generally cover the eastern and western portions of the sampling data along the corridor. I trained the closed-loop models in the two regions individually. Models constructed based on the datasets of the first (western) region are called R1 models, while models trained based on the second (eastern) region are R2 models. Each region contains around 50% of the total data samples.

5.3.1 Sample Data Distribution

Figure 27 illustrates the spatial distribution of the sample points in the study area. In the figure, green dots represent R1 data samples while orange dots represent R2 data samples. Table 21 lists the numbers of samples for different wetland types within each region. The entire dataset for the study area contains 33,098 data samples representing 9 wetland types. The first region (R1) contains 17,113 sample points, covering 6 wetland types. The second region (R2) includes 15,985 sample points, containing 9 wetland types. For R1, the majority of wetland types are type 16 (bottomland hardwood forest), followed by type 4 (riverine swamp forest), and type 10 (pine flat), which is around 20%, while the other types are less than 1%. For R2, the dominant wetland types are also categories 4 and 16. For type 8 (pocosin), almost all the data samples fall within R2. Similarly, type 5 (seep), type 6 (hardwood Flat), and type 13 (non-tidal freshwater marsh) all fall within R2.

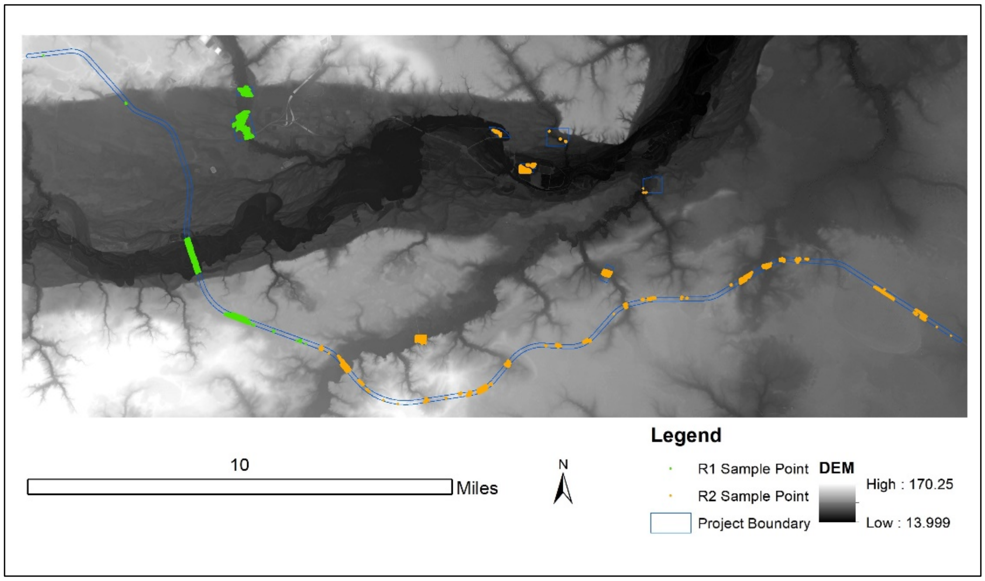


Figure 27. Two study regions for Experiment III

Table 21. Wetland sample data distribution for Experiment III

Wetland Code	Wetland Type	Number (percentage) of data sample in R1	Number (percentage) of data sample in R2
4	Riverine Swamp Forest	5,024	6,174
		29.36%	38.62%
5	Seep	-	4
		-	0.03%
6	Hardwood Flat	-	1
		-	0.01%
8	Pocosin	1	2,910
		0.01%	18.20%
10	Pine Flat	3,317	1,228
		19.38%	7.68%
11	Basin Wetland	3	50
		0.02%	0.31%
13	Non-Tidal Freshwater Marsh	-	697
		-	4.36%
15	Headwater Forest	20	1,021
		0.12%	6.39%
16	Bottomland Hardwood Forest	8,748	3,900
		51.12%	24.40%
Total		17,113	15,985

5.3.2 Performance of Model Training

For the model training process, I used 100 trees with the max depth of 5 for both RF and GBM. Table 22 compares the training performance among different models. Overall, the training performance of R1 models is slightly better than R2 models due to fewer wetland types and more data samples for the majority types. For both study regions, GBM reveals the best performance, and it can fit the training dataset 100%.

Table 22. Model performance in the training area (overall accuracy and kappa)

Model	R1		R2	
	Accuracy	Kappa	Accuracy	Kappa
GLM	0.998	1.00	0.874	0.83
RF	1.000	1.00	0.895	0.86
GBM	1.000	1.00	0.999	1.00

5.3.3 Performance of Model Prediction

I applied R1 and R2 ML models to predict the entire study area and compared the prediction results with the ground truth data, see results in Table 23. For the open-loop model, it does not have the training process as the closed-loop models, so the performance for R1 and R2 regions reveal the same. According to the result, prediction results based on R2 models are slightly better than R1 models (except for GLM), although R1 can better fit the training dataset. All the models can predict the entire study area with the accuracy of around 0.7. Among the R1 models, the order of models in terms of decreasing prediction accuracy is GLM, RF, Rule-based, and GBM. Among R2 models, the corresponding order is GBM, RF, Rule-based and GLM.

Table 23. Comparison of model performance based on different training area (overall accuracy and kappa)

Model	**Prediction based on R1 models**		**Prediction based on R2 models**	
	Accuracy	Kappa	Accuracy	Kappa
GLM	0.695	0.549	0.647	0.511
RF	0.692	0.541	0.737	0.636
GBM	0.674	0.519	0.766	0.677
Rule-based	0.675	0.584	0.675	0.584

5.3.4 Performance of the Integrated Model

In order to further compare the individual models with the integrated method proposed in this research, I ran the calibration process for model integration based on ML models trained in R1 and R2 respectively. Table 24 and Table 25 summarize a subset of model calibration results. These results are in the 20 runs with the highest overall accuracy in terms of predicting the entire study area. Figure 30 and Figure 31 illustrate the prediction results for all the rounds in ascending order.

I used 0.2 as the interval in the parameter searching space, which means that in the first round of calibration, it will conduct the grid search by dividing the entire parameter space into fine grids using 0.2 as the resolution for each dimension. In terms of choosing the interval, the smaller value will lead to a larger number of parameter combinations and increase computational intensity for the first round of exploration. Meanwhile, it can reduce the chance for the search process to be trapped in local optima. It is noted that, due to the stochastic nature of the calibration algorithm, to repeat the calibration process will not produce the same results. Furthermore, the best result in the calibration does not represent the global optimal solution.

Table 24. Calibration results for R1 (top 20 calibration results)

Parameter					**Performance**	
w_1	w_2	w_3	t	w_c	Accuracy	Kappa
1	0	0	1	0.2	0.799	0.734
0.82	0.06	0.12	0.94	0.21	0.796	0.731
0.82	0.06	0.12	0.88	0.21	0.796	0.731
0.82	0.06	0.12	0.18	0.21	0.796	0.731
0.82	0.06	0.12	0.06	0.21	0.796	0.731
0.82	0.06	0.12	0.82	0.21	0.796	0.731
0	0	1	1	0.2	0.795	0.731
0.4	0	0.6	0.6	0.2	0.795	0.730
0.4	0	0.6	0.4	0.2	0.795	0.730
0.2	0	0.8	1	0.2	0.795	0.730
0.2	0	0.8	0.2	0.2	0.795	0.730
0.2	0	0.8	0.8	0.2	0.795	0.730
0.4	0	0.6	1	0.2	0.795	0.730
0.6	0	0.4	1	0.2	0.795	0.730
0.6	0	0.4	0.6	0.2	0.795	0.730
0.6	0	0.4	0.4	0.2	0.795	0.730
0.8	0	0.2	1	0.2	0.795	0.730
0.8	0	0.2	0.8	0.2	0.795	0.730
0.82	0.03	0.15	0.85	0.2	0.795	0.731
0.82	0.03	0.15	0.97	0.2	0.795	0.731

Table 25. Calibration results for R2 (top 20 calibration results)

Parameter					**Performance**	
w_1	w_2	w_3	t	w_c	Accuracy	Kappa
0.4	0.6	0	0	0.6	0.775	0.689
0	0.6	0.4	0	0.6	0.774	0.688
0.2	0.6	0.2	0	0.6	0.772	0.686
0	0.2	0.8	1	1	0.770	0.688
0	0.4	0.6	1	1	0.770	0.688
0	0.6	0.4	1	1	0.770	0.688
0	0.8	0.2	1	1	0.770	0.688
0.2	0.4	0.4	0.8	1	0.770	0.688
0	0.6	0.4	0	0.8	0.766	0.677
0	0.6	0.4	0	1	0.766	0.677
0	0.8	0.2	0	0.6	0.766	0.677

0	0.8	0.2	0	0.8	0.766	0.677
0	0.8	0.2	0	1	0.766	0.677
0	1	0	0	0.6	0.766	0.677
0	1	0	0	0.8	0.766	0.677
0	1	0	0	1	0.766	0.677
0	1	0	1	1	0.766	0.677
0.2	0.6	0.2	0	0.8	0.766	0.677
0.2	0.6	0.2	0	1	0.766	0.677
0.2	0.8	0	0	0.6	0.766	0.677

*Note: w_c means the weight for the closed-loop result; lower value denotes lower level of impacts; *t* represents the threshold for integrating closed-loop models; lower *t* means it is easier for the closed-loop models to be introduced to affect the final result.

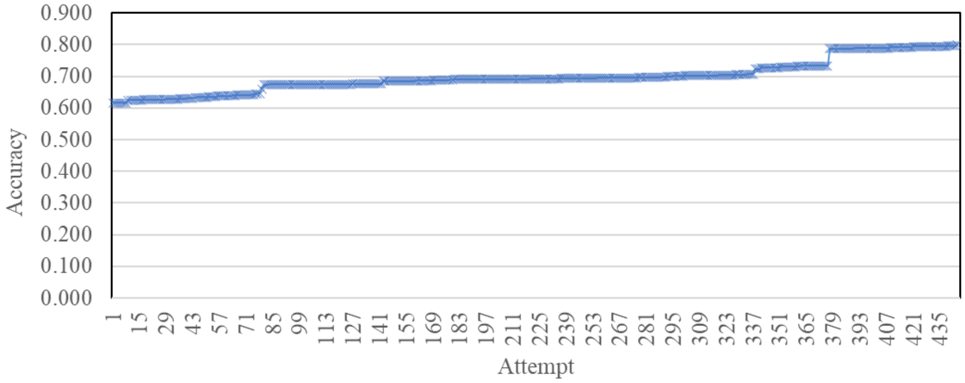


Figure 28. Prediction variation based on R1 models

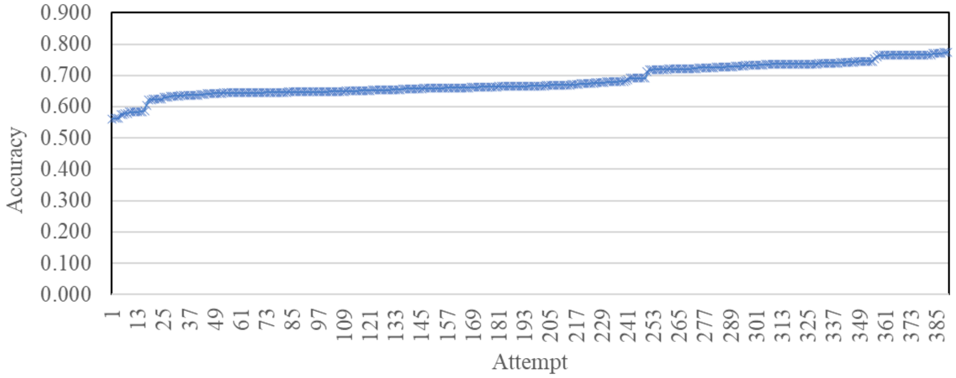


Figure 29. Prediction variation based on R2 models

According to the calibration results, R1 has a slightly better prediction result than R2—the highest accuracy for R1 is 0.799 while for R2 it is 0.775. The average accuracy for R1 among all the rounds is 0.698, and average kappa is 0.578. These two metrics are higher than the prediction result based on any individual models in R1. The average accuracy and kappa for all calibration rounds based on R2 models are 0.683 and 0.580, respectively, which are not as good as the best single model performance from R2. However, the best calibration result based on the integration model performs better than the best individual models trained in R2.

According to the parameters among the 20 best prediction results, R1 integration models reveal high value for the integration threshold and low weight for the closed-loop models. This means the open-loop models have a higher contribution in the final integration, and it is stricter to apply the results from the closed-loop methods. For R2, the best results come from the parameter combinations—the weights of the closed-loop method are relatively high.

5.3.5 Confusion Matrix

To further check the prediction results for detailed wetland types, Table 26 and Table 27 present the confusion matrix for the two best prediction results based on the integrated models from R1 and R2. Using R1-trained integration models, type 8 (pocosin), type 13 (non-tidal freshwater marsh), and type 15 (headwater forest) have the highest precision, while type 5 (seep) and type 13 have the highest recall. However, types 5, 8, 13, and 15 all have very few representations in R1 training data samples, so these wetland types are predicted by the open-loop method. This can be explained by the high contribution of the open-loop method shown in the pattern of parameter combinations.

Table 26. Confusion matrix of the best prediction result based on R1 integration model

	Prediction											
Ground Truth	0	4	5	6	8	10	11	13	14	15	16	Total	Recall
0	0	0	0	0	0	0	0	0	0	0	0	0	NA
4	11	8452	29	0	0	0	0	0	2593	0	113	11198	0.755
5	0	0	4	0	0	0	0	0	0	0	0	4	1.000
6	0	0	1	0	0	0	0	0	0	0	0	1	0.000
8	16	0	664	0	1148	1083	0	0	0	0	0	2911	0.394
10	135	0	114	0	0	4130	146	0	0	0	20	4545	0.909
11	0	0	4	0	0	25	24	0	0	0	0	53	0.453
13	0	0	0	0	0	0	0	697	0	0	0	697	1.000
14	0	0	0	0	0	0	0	0	0	0	0	0	NA
15	0	0	0	0	0	0	0	0	89	952	0	1041	0.915
16	16	7	252	0	0	0	2	0	1349	0	11022	12648	0.871
Total	178	8459	1068	0	1148	5238	172	697	4031	952	11155		
Precision	0.000	0.999	0.004	NA	1.000	0.788	0.140	1.000	0.000	1.000	0.988		

Table 27. Confusion matrix of the best prediction result based on R2 integration model

	Classification										
Ground Truth	4	5	6	8	10	11	13	15	16	Total	Recall
4	10845	0	0	0	0	0	18	0	335	11198	0.968
5	0	2	0	0	0	0	0	0	2	4	0.500
6	0	0	1	0	0	0	0	0	0	1	1.000
8	0	0	0	2911	0	0	0	0	0	2911	1.000
10	0	0	0	0	4545	0	0	0	0	4545	1.000
11	0	0	0	0	4	46	0	2	1	53	0.868
13	0	0	0	0	0	0	687	0	10	697	0.986
15	0	0	0	0	20	0	0	1021	0	1041	0.981
16	6549	0	0	0	0	0	5	507	5587	12648	0.442
Total	17394	2	1	2911	4569	46	710	1530	5935		
Precision	0.623	1.000	1.000	1.000	0.995	1.000	0.968	0.667	0.941		

According to the confusion matrix of the R2 integration model, type 6 (hardwood flat), type 8 (pocosin), and type 10 (pine flat) have the highest recall, while type 5 (seep), type 6 (hardwood flat), type 8, and type 11 (basin) have the highest precision. However, among these wetland types, except for type 8, all other wetland types are rarely represented by R2 training data. However, the overall accuracy and kappa for the R2 integration model are lower than the R1 integration model. According to the confusion matrix, the R2 model is able to maintain high precision and recall for the majority of the wetland types except for type 16 (Bottomland Hardwood Forest). Besides, the R1 model also generated wrong prediction results as the non-wetland type and type 14 (Floodplain Pool), which never occurred in the training data samples. Some regions are classified as non-wetland because the area has conflicts among multiple wetland types—more than three wetland types are calculated as the same probability.

5.3.6 Prediction Map

I generated the prediction results based on the best integration parameters calibrated from the two regions. Figure 30 shows the prediction based on the best R1 integration model; Figure 31 shows the classification result based on R2 integration model; Figure 32 illustrates wetland type distribution based on the open-loop method.

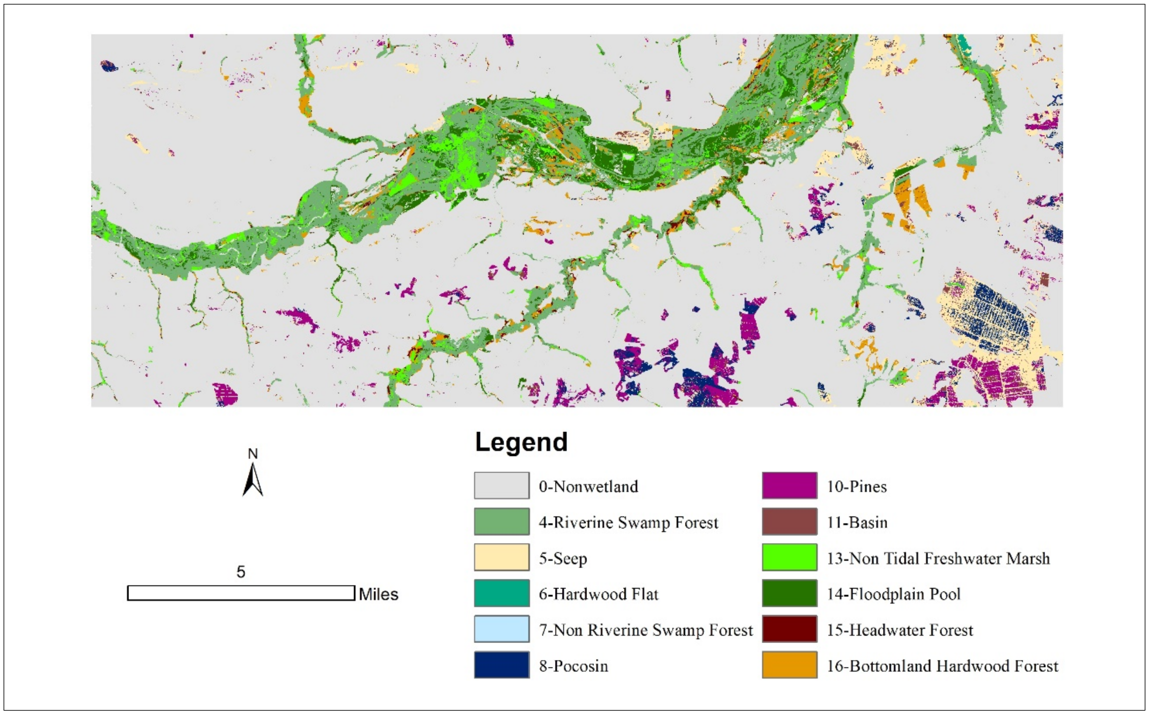


Figure 30. Prediction result based on the integrated method in R1

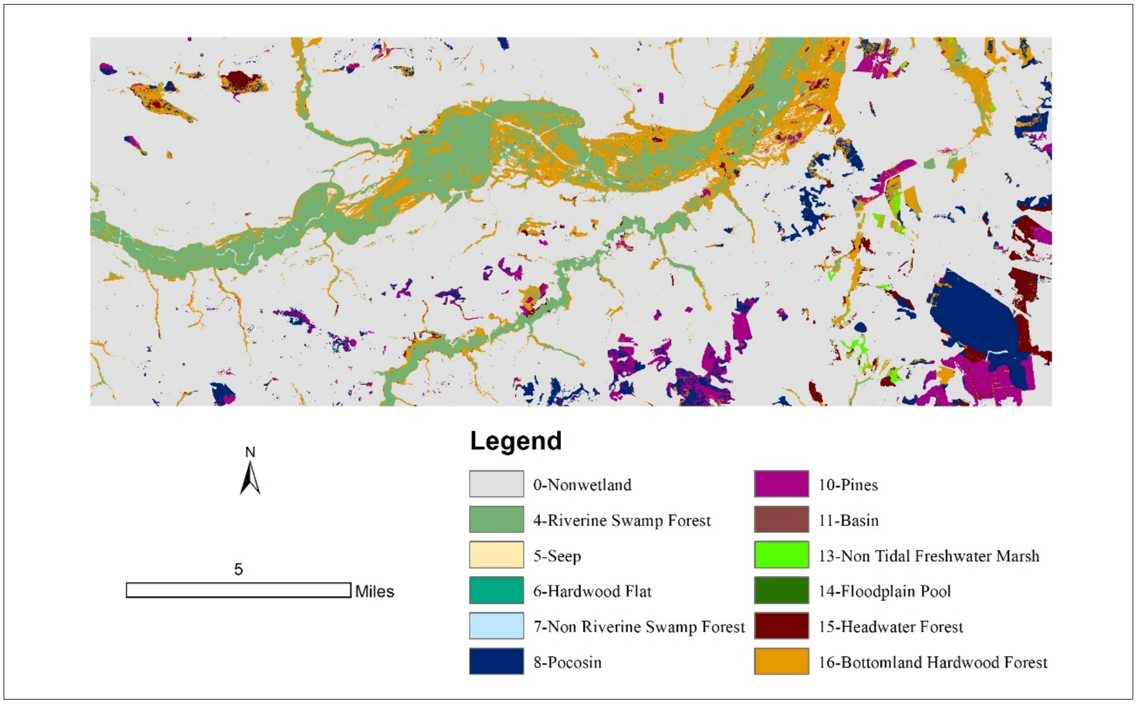


Figure 31. Prediction result based on the integrated method in R2

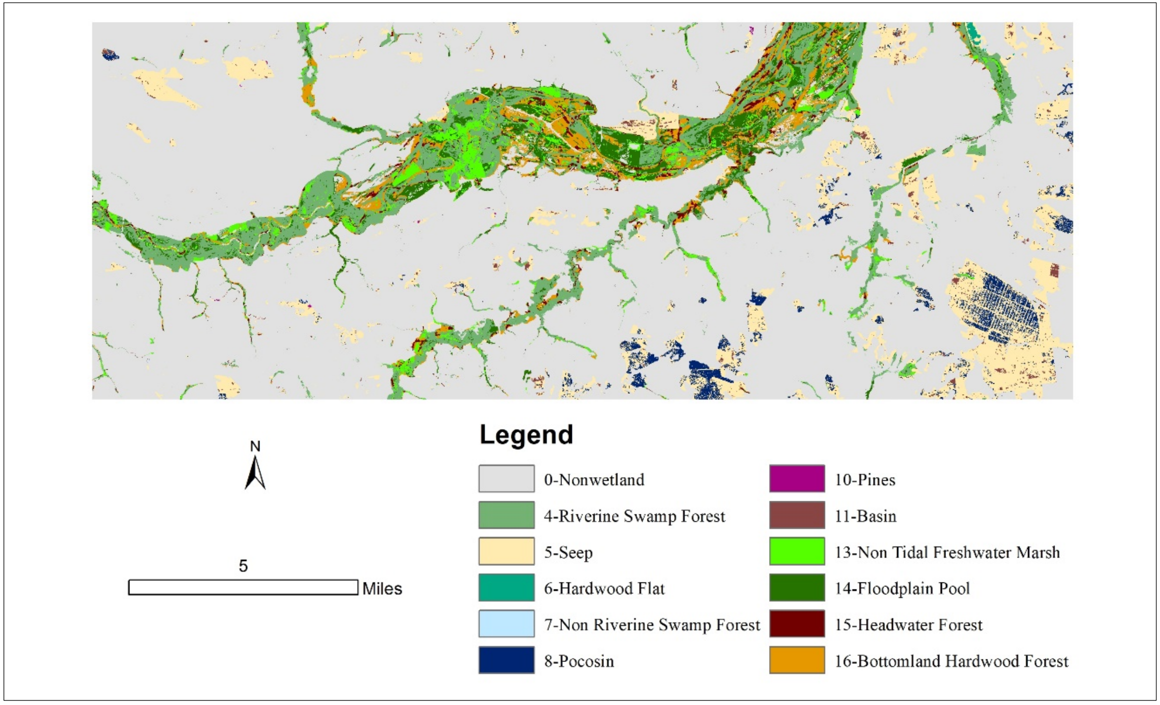


Figure 32. Prediction result based on the open-loop method

The wetlands identified by the open-loop method reveal the most fragmented pattern within the riverine area. The classification result of the R1 integration model is more similar to the open-loop prediction than the R2 model, because the R2 model gives higher weight to the closed-loop method. Referring to the R2 model classification result, we can observe the general pattern of the closed-loop classification result. Overall, the closed-loop models predict more pine and pocosin wetlands than the open-loop model. However, the closed-loop models tend to be constrained by the wetland types occurring in the samples. The open-loop model tends to predict more seep, which has very few data samples for the training process of the closed-loop models. Furthermore, the open-loop model can predict wetland types such as type 13 (non-tidal freshwater marsh) and type 14 (floodplain pool)—also see the R1 classification result. On the other hand, the R2 integration model tends to predict the riverine area as type 4 (Riverine Swamp Forest) and type 16 (Bottomland Hardwood Forest), which has a larger percentage in the training data sample.

The conflicting results within riparian areas between R1 and R2 integration models are likely due to difficulties in remotely classifying variations in localized hydrology. Therefore, the transition area adjacent to rivers create challenges for classification, and discrepancies among the three ML models will cause the reducing contribution from the closed-loop method.

In non-riparian regions, the elliptical Carolina Bay located in the southeast portion of the project area shows an example of a major conflict, characterized by intensive silvicultural activities, including ditching, vegetation removal, and maintenance of a monoculture community of loblolly pine. These activities result in localized changes in hydrology and vegetative composition that are difficult to quantify. The open-loop method and R1 model tend to emphasize the details, possibly ditches and sloppy surface within the large patch of pocosin area and classify them as seep.

Figure 33 shows the probability maps produced during the construction of the open-loop method. It lists the examples of non-riparian wetland types in the left column (including basin, pine, and pocosin), and examples of riparian wetland types in the right column (including riverine swamp forest, bottomland hardwood forest, and non-tidal freshwater marsh). All the probability values are in the range of 0 to 1. Red color denotes a higher probability for the wetland type to occur.

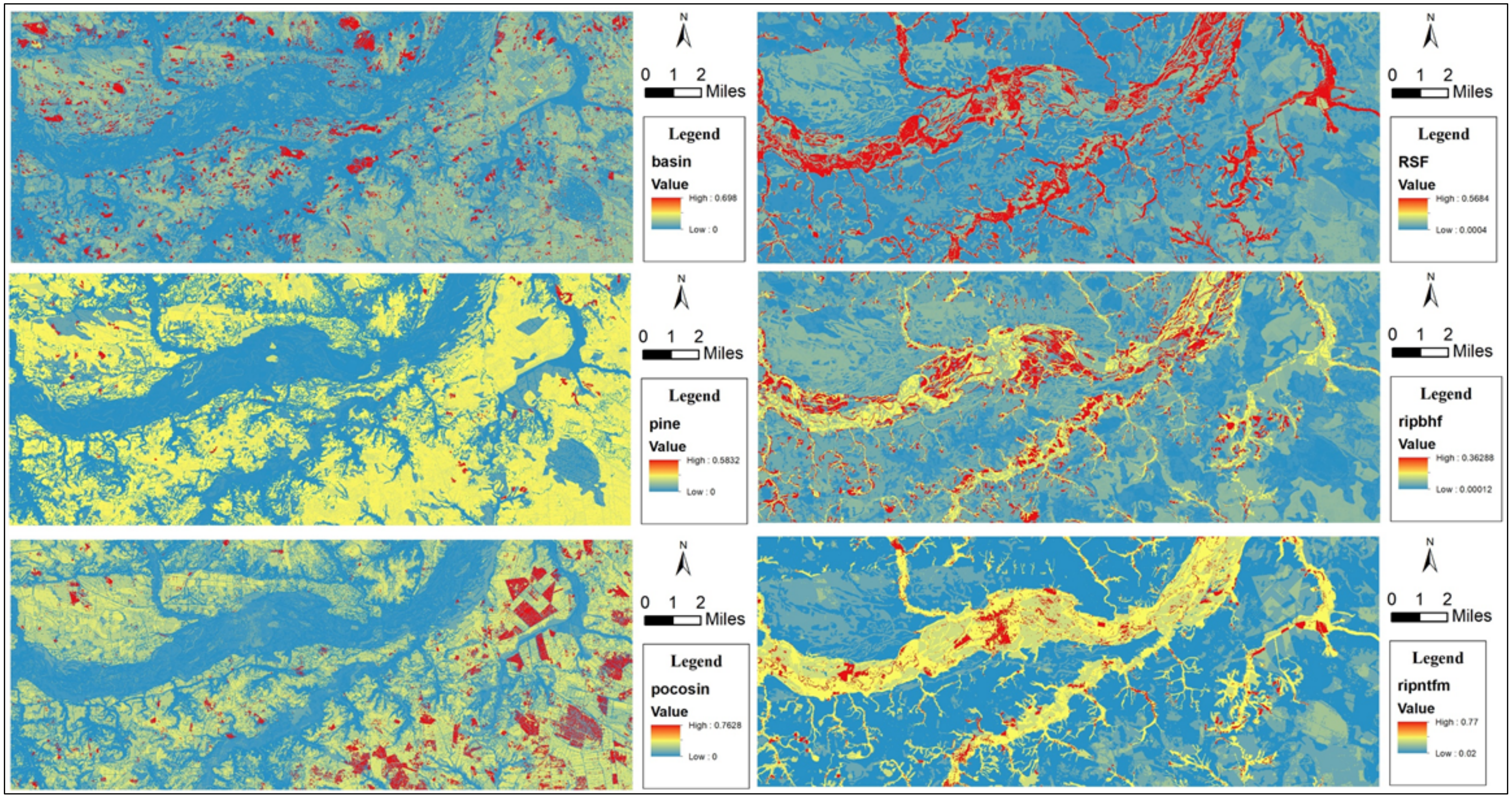


Figure 33. Probability maps generated by the rule-based model

CHAPTER 6: DISCUSSIONS

6.1. Summary

Wetland inventory mapping represents an important step for us to gain a better understanding of wetland distribution and to support wetland studies and management. Various studies have applied RS-based supervised classification methods to serve the objective of wetland mapping. However, it remains a challenge to build an accurate and robust classification model for detailed wetland types. In recent years, machine learning models gained popularity in wetland classification applications, due to their advantages of less constraints and fewer prior rules required, especially their excellent prediction performance.

However, several limitations exist when merely applying the machine-learning methods in wetland classification. First, the classification rules and strategies generated by the methods do not present in an intuitive and straightforward fashion. This situation hinders users from understanding the rules and evaluating them—more specifically, to evaluate the application range for these rules and the applicability in another study area. Second, the objective of machine-learning-based models is to best fit the training data samples, which can also lead to over-fit issues wherein the noise in the training data can disturb and mislead the construction of these rules. The rule-based classification model, on the other hand, requires predefined classification rules, but it suffers from mediocre performance.

To tackle those issues, I proposed an expert system to build a new integration model combining two different modeling perspectives—closed-loop modeling and open-loop modeling—which are represented by three machine learning models and a rule-based model, respectively. The integration aims to leverage the advantages of both methods since they complement each other in many ways. Another contribution of this research is to make the classification process highly automated and flexible so that researchers can easily transfer the workflow to predict wetland types in any location in North Carolina or even a broader region with a given classification system.

I first conducted a literature review to summarize the applications of current methods in wetland classification. Based on the limitation s of current studies, I introduced the proposed methodology and implementation. In the experiment section, I conducted three experiments to test the performance and limitations of conventional machine learning methods applying to the study case. In the end, I compared the results of the conventional method with the new integration method. In this final chapter, I summarized the findings and lessons learned in the research.

6.2. Performance Under Different Scenarios

Assuming the fieldwork data we collected are the source of truth for the entire study area, the first two experiments simulate the scenario with good sampling data. Through the stratified sampling process, we can maintain the same proportion of samples from each wetland type for the entire study area. Under this scenario, the closed-loop models revealed excellent performance even with greatly reduced sampling size. All the closed-loop models can maintain a high prediction accuracy above 90%. The first experiment shows that when the samples are representative, the sample size does not significantly affect the performance of the closed-loop models.

The second experiment added another variation on top of the sample size. It selected a subset of 10 variables based on the variable importance rank in the first experiment, generated by RF and GBM models. Surprisingly, the variable reduction only impacted the performance of GLM significantly. RF even showed better prediction accuracy with fewer variables. The results of this experiment illustrated the importance of data quality over quantity for the close-loop models. On the other hand, poor data quality, such as a variable with too much noise may mislead the tree building process and result in bad performance.

The scenario setup for the first two experiments both represent a relatively ideal situation in which the training samples can represent the entire population. However, data samples in most real cases could not completely capture the distribution of the entire population in advance. Furthermore, the sample quantify representing different wetland types varies, which can affect the performance of the classification model. To simulate this scenario, I designed the third experiment to use the unbalanced training data samples by dividing the study area into two spatial regions. The results show that all the machine learning-based models experienced performance degradation. In the third experiment, I applied the integrated method to both regions, and it generated better prediction results than any individual closed-loop or open-loop models. I also explored different parameter combinations for the integrated model to gain insights into the prediction pattern. The two regions have slightly different wetland sample distributions: R1 has fewer wetland types and the accuracy achieved in model training is higher, indicating that it is easier for the model to fit in R1 sample data than in R2. In terms of the integration model, R1 put higher weights on the open-loop method while the R2 integration model values the closed-loop method more. The overall prediction accuracy and kappa based on the R1 integration model are higher than the R2 integration model.

To evaluate the prediction performance for different wetland types, I used two metrics: recall and precision to reflect the false negative and false positive cases for the prediction of specific wetland types. The higher recall denotes less false negative prediction for this wetland type; similarly, higher precision means fewer false positive cases. The R2 integration model maintain high precision and recall for the majority of the wetland types except for type 16 (Bottomland Hardwood Forest), which is well represented by the data samples. This is the major reason why the overall accuracy of the R2 integration model is worse than R1.

According to the experiment results, the quality of the sample data should guide the determination of integration weights for different methods. If the data samples are imbalanced among types, or spatially concentrated in a small area, the training results from the closed-loop models are then less reliable. Users should then consider increasing the weight of the open-loop method and the threshold for the closed-loop models for the integration. Otherwise, if the data are collected across the entire study area with a spatially balanced sampling strategy, users should assign relatively high weights to the closed-loop model. If it is already challenging to determine the wetland type for the sites during the fieldwork, the possibility of generating highly conflicting prediction results are then expected. Under this scenario, it is suggested to lower the threshold for integrating different prediction models while building the integration model.

In terms of computational cost, the ML training procedure and the model integration step consume the most time. The time cost of model training mainly depends on data sample size. In the first two experiments, the computational time decreases as the sample size decreases. The first treatment took approximately 5 hours to complete data training, while the last treatment took about 20 minutes to finish. The wetland type prediction procedure is also memory-intensive due to the large number of pixels for the high-resolution raster datasets. To solve this issue, I applied decomposition strategy on the raster data layer to divide the data into partitions each containing 200 rows. I then applied the trained models to each partition, I collected all the classification results and proceeded to mosaic the pieces together into one raster layer.

6.3. Data Resolution and Scale Analysis

In this research, I collected spatial datasets with different resolutions from different sources. For instance, LiDAR data is two points per square meter, the Sentinel-2 imagery pixel is 10 meters, and land cover raster data is 30 meters. I processed all the spatial datasets towards a certain scale level (the same spatial resolution) for data representation and spatial analysis. This research is using 20 feet (about 6 meters) as the resolution. I down-sampled datasets with higher resolution by using average value to merge neighboring pixels. I also up-sampled the datasets with lower resolution by dividing a grid cell to multiple ones with the same pixel value.

In spatial modeling, scale selection or scale analysis is an important topic. In general, when the resolution is coarse, multiple classification objects may coexist within a single pixel and detailed information can be lost. In our case, we will easily omit the wetland patches with smaller size in the map representation. However, if the resolution is too high, the map may reveal a higher level of heterogeneity, which introduces a lot of noises to the classification models. For wetland types that occur as large-size patches, fine pixels may increase the level of data variation within the type and increase the difficulty for classification.

Therefore, there is a trade-off between a generic pattern and detail representation. The selection of scale should be based on the specific study cases, such as the general size of the objects to be classified and the resolution of data sources being used. In wetland-related studies, researchers have explored the scale effect by conducting experiments to evaluate the change of prediction performance by varying the spatial resolution of input data (Powers et al. 2012). I have also conducted related research to explore the relationship between classification model performance and scale variation, described in (Deng et al. 2018). According to the experiment results in previous studies, for the same study area, 10 feet and 20 feet can generate the best classification prediction accuracy based on the LiDAR DEM derivatives. In this research, I chose 20 feet as the spatial resolution since most of the spatial datasets are collected under similar resolution.

Although this research did not elaborate direct experiment for scale analysis, the first two experiments regarding the variation of sampling percentage can also shed light on the same topic. The results show that when more data samples are included, it is not necessary to generate better classification results since noise and unnecessary information could misguide the classification model. Furthermore, to use a higher resolution for classification will lead to an up-sampling process for the spatial data, which will not bring in more information but create duplication records in the sample data.

6.4. Application Cases for the Research

This study proposed and implemented a framework to generate and integrate rules from both the expert knowledge and machine learning training process. It reveals several advantages according to the experiment results. First, it provides reliable classification results by combining the advantages of both open-loop and closed-loop methods. Second, the application of this framework is less constrained by the training data. It can identify wetland types that did not occur in the training samples. Third, it allows human interactions through assigning weights to different methods for realistic scenarios. Furthermore, it is a fully automatic workflow that can be flexibly applied to any location in the state of North Carolina or even broader region with a given classification system.

One can use this system to generate a wetland type distribution map to guide wetland management or wetland protection activities. For instance, we can use the product of this system—the potential wetland type distribution map—to guide the development plan. Under the context of a construction project (e.g., bridge, highway, urban expansion), the wetland distribution map can help with evaluating the environmental impact of the project and minimizing the risk of damaging wetlands. Another useful application is to monitor wetland change, by applying the wetland classification model on spatial data (e.g., remote sensing data) obtained at different time.

The prediction results based on different temporal data can help us gain a better understanding of the dynamics of the wetland system. This system can also provide important function to support high accuracy wetland inventory mapping. The distribution map can guide us to collect sample data in the field work. As illustrated in the previous experiments, the quality of the training data samples is critical for classification models. With the wetland spatial distribution map, we can better plan the path for the field work and collect more balanced data samples for each wetland type. If the dataset itself is more representative and collected in a more balanced manner, the prediction results will be more reliable. Furthermore, based on different versions of classification maps generated by different methods, we can analyze the area with a higher level of wetland type conflicts, and specifically visit that area for data samples.

6.5. Limitations and Future Work

The major limitation of this research relates to data unavailability. For example, for the open-loop method, the procedure of translating each step in the decision-making process to measurable data variables is critical for good prediction results. However, this process may suffer from a lack of data and errors introduced by prior simplifications and assumptions. For example, a key step in distinguishing pocosin wetlands from pine flat wetlands is a determination of the presence of dense shrub species. This step can be visually judged by an expert in the field, but the data layer that represents “dominated by shrubs” may not be acquirable in a digital and remote fashion. This study applied QL2 LiDAR point cloud data to calculate the percentage of medium-height vegetation points relative to the total vegetative composition of each pixel defined as a wetland area.

However, the classification could benefit more if accurate shrub distribution survey map is available.

The system introduced a threshold system in the rule-based model to determine whether each pixel meets these criteria and can, therefore, be labeled as “dominated by shrubs” or not. This type of variable generation process introduces complexities in the calibration of the thresholds and parameters. To resolve these problems and improve the performance of rule-based models, I applied probability factors on binary variables while constructing the decision tree. I calculated an error rate for each data variable according to which wetland types are supposed to be derived from the given data layer. I then converted the binary (0, 1) values to segmented values based on the error rate for calculating overall probabilities. Basically, the higher error rate corresponds to two probability values that are less distant from each other to represent a higher level of uncertainty. In this way, I can better control the quality of the condition data layers in the open-loop method and recheck the data generating process if the error rate is unreasonably high. We can consider this process as a “training” or “calibration” step for the rule-based model.

In pursuit of future improvement, I believe this expert system should be more dynamic. The database and model base should keep evolving by introducing more study cases and domain knowledge. For example, we should collect more data samples for minority wetland types as well as more field testing data corresponding to the highly-conflicted regions predicted by different models. Meanwhile, we can summarize more domain knowledge and keep refining the rules based on current prediction and validation process. For example, the open-loop method tends to classify more “seep” type. One reason is that the rules defined to distinguish seep are not sufficient. Domain experts can interact with the classification results and further investigate the false positive cases, thus to formulate more rules to limit areas being falsely classified as seep.

Furthermore, we can apply some post process and spatial analyses to further refine the classification results. When the classification map reveals “salt and pepper effect,” it means individual pixels are classified into different wetland types from the neighboring pixels. This situation occurs frequently for small wetland sites. We can apply neighborhood rules and spatial constraints to fix such a situation. For example, the development of context rules for each wetland type in terms of the possible size, shape, and adjacency constraints with other wetland types limits the areas in which various wetlands may occur and is expected to provide positive feedback for additional modeling efforts.

REFERENCES

Adam, E., O. Mutanga & D. Rugege (2010) Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecology and Management,* 18**,** 281-296.

Ahmad, R. (2001) Expert Systems: Principles and Programming. *Scalable Computing: Practice and Experience,* 7.

Allen, T., Y. Wang, B. Gore, J. Swords & D. Newcomb. 2011. Coastal Wetland mapping Using Time Series SAR Imagery and LiDAR: Alligator River National Wildlife Refuge, North Carolina. In *Proceedings, Pecora 18 symposium, Herndon, Virginia*, 14-17.

Anderson, J. E. & J. E. Perry (1996) Characterization of wetland plant stress using leaf spectral reflectance: implications for wetland remote sensing. *Wetlands,* 16**,** 477-487.

Anderson, R. R. & F. J. Wobber. 1973. Wetlands mapping in New Jersey. In *American Society of Photogrammetry and American Congress of Surveying and Mapping, Annual Convention, Washington, D. C., Mar. 12-17, 1972.) Photogrammetric Engineering*, 353-358.

Argialas, D. & C. Harlow (1990) Computational image interpretation models: an overview and a perspective.

Aschbacher, J., P. Tiangco, C. Giri, R. Ofren, D. Paudyal & Y. Ang. 1995. Comparison of different sensors and analysis techniques for tropical mangrove forest mapping. In *Geoscience and Remote Sensing Symposium, 1995. IGARSS'95.'Quantitative Remote Sensing for Science and Applications', International*, 2109-2111. IEEE.

Aspinall, R. (1992) An Inductive Modeling Procedure Based on Bayes Theorem for Analysis of Pattern in Spatial Data. *International Journal of Geographical Information Systems,* 6**,** 105-121.

Augusteijn, M. & C. Warrender (1998) Wetland classification using optical and radar data and neural network classification. *International Journal of Remote Sensing,* 19**,** 1545-1560.

Baker, C., R. Lawrence, C. Montagne & D. Patten (2006) Mapping Wetlands and Riparian Areas Using Landsat ETM+ Imagery and Decision Tree Based Models. *Wetlands,* 26**,** 465-474.

Baker, C., R. L. Lawrence, C. Montagne & D. Patten (2007) Change Detection of Wetland Ecosystems Using Landsat Imagery and Change Vector Analysis. *Wetlands,* 27**,** 610-619.

Ball, G. H. & D. J. Hall. 1965. ISODATA, a novel method of data analysis and pattern classification. DTIC Document.

Bolstad, P. V. & T. Lillesand (1992) Rule-based classification models: flexible integration of satellite imagery and thematic spatial data. *Photogrammetric Engineering and Remote Sensing,* 58**,** 965-971.

Boyd, J. (2002) Compensating for Wetland Losses under the Clean Water Act. *Environment: Science and Policy for Sustainable Development,* 44**,** 43-44.

Bratko, I. 2001. *Prolog programming for artificial intelligence*. Pearson education.

Breiman, L. (2001) Random forests. *Machine learning,* 45**,** 5-32.

Brisco, B., M. Kapfer, T. Hirose, B. Tedford & J. Liu (2011) Evaluation of C-band polarization diversity and polarimetry for wetland mapping. *Canadian Journal of Remote Sensing,* 37**,** 82-92.

Bronge, L. B. (1999) Mapping boreal vegetation using Landsat-TM and topographic map data in a stratified approach. *Canadian Journal of Remote Sensing,* 25**,** 460-474.

Bronge, L. B. & B. Näslund-Landenmark (2002) Wetland classification for Swedish CORINE Land Cover adopting a semi-automatic interactive approach. *Canadian journal of remote sensing,* 28**,** 139-155.

Butera, M. K. (1983) Remote sensing of wetlands. *IEEE Transactions on Geoscience and Remote Sensing***,** 383-392.

Castañeda, C. & D. Ducrot (2009) Land cover mapping of wetland areas in an agricultural landscape using SAR and Landsat imagery. *Journal of environmental management,* 90**,** 2270-2277.

Chopra, R., V. Verma & P. Sharma (2001) Mapping, monitoring and conservation of Harike wetland ecosystem, Punjab, India, through remote sensing. *International Journal of Remote Sensing,* 22**,** 89-98.

Chust, G., D. Ducrot & J. L. Pretus (2004) Land cover discrimination potential of radar multitemporal series and optical multispectral images in a Mediterranean cultural landscape. *International Journal of Remote Sensing,* 25**,** 3513-3528.

Cihlar, J., C. Tarnocai, A. Jano, I. Kettles, B. Lacelle, J. Liu, L. Maynard, T. Moore, S. Robinson & N. Roulet (2000) Wetlands of Canada and Climate Change: Observation Strategy and Baseline Data.

Click, C., M. Malohlava, A. Candel, H. Roark & V. Parmar (2016) Gradient Boosted Models with H2O.

Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement,* 20**,** 10.

--- (1968) Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin,* 70**,** 213.

Corcoran, J. M., J. F. Knight & A. L. Gallant (2013) Influence of Multi-Source and Multi-Temporal Remotely Sensed and Ancillary Data on the Accuracy of Random Forest Classification of Wetlands in Northern Minnesota. *Remote Sensing,* 5**,** 3212-3238.

Cowardin, L. M., V. Carter, F. C. Golet & E. T. LaRoe (1979) Classification of wetlands and deepwater habitats of the United States. *US Fish and Wildlife Service FWS/OBS,* 79**,** 131.

Crawford, M. M., J. Ham, Y. Chen & J. Ghosh. 2003. Random forests of binary hierarchical classifiers for analysis of hyperspectral data. In *Advances in Techniques for Analysis of Remotely Sensed Data, 2003 IEEE Workshop on*, 337-345. IEEE.

Crawford, M. M., S. Kumar, M. R. Ricard, J. C. Gibeaut & A. Neuenschwander (1999) Fusion of airborne polarimetric and interferometric SAR for classification of coastal environments. *IEEE Transactions on Geoscience and Remote Sensing,* 37**,** 1306-1315.

Crist, E. P. & R. C. Cicone (1984) A physically-based transformation of Thematic Mapper data---The TM Tasseled Cap. *IEEE Transactions on Geoscience and Remote sensing***,** 256-263.

Deng, J., S.-G. Wang, A. S. Smith, S. Davis, M. Weatherford, L. Paugh & S. Jiang. 2018. Scale Analysis of a Wetland Classification Model Based on Lidar Data and Machine Learning Methodology. In *Transportation Research Board (TRB) Annual Meeting*. Washington, D.C.

Drusch, M., U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti & P. Martimort (2012) Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment,* 120**,** 25-36.

EPA, U. S. (2012) Section 404 of the Clean Water Act: how wetlands are defined and identified. 2014. http://water.epa.gov/type/wetlands/outreach/fact11.cfm (last accessed Octorber 05, 2012).

Forgette, T. A. & J. A. Shuey (1997) A comparison of wetland mapping using SPOT satellite imagery and national wetland inventory data for a watershed in northern Michigan. *Northern Forested Wetlands: Ecology and Management***,** 61-70.

Franklin, S. (1991) Satellite remote sensing of mountain geomorphic surfaces. *Canadian Journal of Remote Sensing,* 17**,** 218-230.

Franklin, S., R. Gillespie, B. Titus & D. Pike (1994) Aerial and satellite sensor detection of Kalmia angustifolia at forest regeneration sites in central Newfoundland. *International Journal of Remote Sensing,* 15**,** 2553-2557.

Franklin, S. E. & J. E. Moulton (1990) Variability and classification of Landsat Thematic Mapper spectral response in southwest Yukon. *Canadian Journal of Remote Sensing,* 16**,** 2-13.

Freund, Y. & R. E. Schapire. 1996. Experiments with A New Boosting Algorithm. In *ICML*, 148-156.

Geerling, G., M. Labrador-Garcia, J. Clevers, A. Ragas & A. Smits (2007) Classification of floodplain vegetation by data fusion of spectral (CASI) and LiDAR data. *International Journal of Remote Sensing,* 28**,** 4263-4284.

Gessler, P., O. Chadwick, F. Chamran, L. Althouse & K. Holmes (2000) Modeling Soil–Landscape and Ecosystem Properties Using Terrain Attributes. *Soil Science Society of America Journal,* 64**,** 2046-2056.

Ghedira, H., M. Bernier & T. Ouarda. 2000. Application of neural networks for wetland classification in Radarsat SAR imagery. In *Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. IEEE 2000 International*, 675-677. IEEE.

Gislason, P. O., J. A. Benediktsson & J. R. Sveinsson (2006) Random forests for land cover classification. *Pattern Recognition Letters,* 27**,** 294-300.

Gluck, M. J., R. S. Rempel & P. Uhlig. 1996. *An evaluation of remote sensing for regional wetland mapping applications*. Sault Ste. Marie: Ontario Forest Research Institute.

Goetz, S. J. (2006) Remote sensing of riparian buffers: past progress and future prospects. *Journal of the American Water Resources Association***,** 133-143.

Gorham, E. (1991) Northern peatlands: role in the carbon cycle and probable responses to climatic warming. *Ecological applications,* 1**,** 182-195.

Gosselin, G., R. Touzi & F. Cavayas (2014) Polarimetric Radarsat-2 wetland classification using the Touzi decomposition: case of the Lac Saint-Pierre Ramsar wetland. *Canadian Journal of Remote Sensing,* 39**,** 491-506.

Grenier, M., A.-M. Demers, S. Labrecque, M. Benoit, R. A. Fournier & B. Drolet (2007) An object-based method to map wetland using RADARSAT-1 and Landsat ETM images: test case on two sites in Quebec, Canada. *Canadian Journal of Remote Sensing,* 33**,** S28-S45.

Hardisky, M., M. Gross & V. Klemas (1986) Remote sensing of coastal wetlands. *BioScience,* 36**,** 453-460.

Harper, J. & G. Ross (1983) Digital analysis of Landsat data in the Athabasca Delta. *Environmental assessment and resource management***,** 319-327.

Heidemann, H. K. 2012. Lidar Base Specification.

Henderson, F. M. & A. J. Lewis (2008) Radar detection of wetland ecosystems: a review. *International Journal of Remote Sensing,* 29**,** 5809-5835.

Hess, L. L., J. M. Melack, E. M. Novo, C. C. Barbosa & M. Gastil (2003) Dual-season mapping of wetland inundation and vegetation for the central Amazon basin. *Remote Sensing of Environment,* 87**,** 404-428.

Hess, L. L., J. M. Melack & D. S. Simonett (1990) Radar detection of flooding beneath the forest canopy: a review. *International Journal of Remote Sensing,* 11**,** 1313-1325.

Hill, R. & A. Thomson (2005) Mapping woodland species composition and structure using airborne spectral and LiDAR data. *International Journal of Remote Sensing,* 26**,** 3763-3779.

Hines, M. E., R. E. Pelletier & P. M. Crill (1992) Emissions of sulfur gases from marine and freshwater wetlands of the Florida Everglades: Rates and extrapolation using remote sensing.

Hodgson, M., J. Jensen, H. Mackey Jr & M. Coulter (1987) Remote sensing of wetland habitat: a wood stork example. *Photogrammetric Engineering and Remote Sensing,* 53**,** 1075-1080.

Hofle, B., M. Vetter, N. Pfeifer, G. Mandlburger & J. Stotter (2009) Water surface mapping from airborne laser scanning using signal intensity and elevation data. *Earth Surface Processes and Landforms,* 34**,** 1635.

Huang, C., L. Davis & J. Townshend (2002) An assessment of support vector machines for land cover classification. *International Journal of remote sensing,* 23**,** 725-749.

Huang, S., C. Potter, R. L. Crabtree, S. Hager & P. Gross (2010) Fusing optical and radar data to estimate sagebrush, herbaceous, and bare ground cover in Yellowstone. *Remote Sensing of Environment,* 114**,** 251-264.

Huang, X. & J. R. Jensen (1997) A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. *Photogrammetric engineering and remote sensing,* 63**,** 1185-1193.

Hughes, G. (1968) On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory,* 14**,** 55-63.

Huguenin, R. L., M. A. Karaska, D. Van Blaricom & J. R. Jensen (1997) Subpixel classification of bald cypress and tupelo gum trees in Thematic Mapper imagery. *Photogrammetric Engineering and Remote Sensing,* 63**,** 717-724.

Hui, Y., Z. Rongqun & L. Xianwen (2009) Classification of wetland from TM imageries based on decision tree. *WSEAS Trans. Infor. Sci. Appl,* 6**,** 1790-0832.

Hupp, C. R. & W. Osterkamp (1996) Riparian vegetation and fluvial geomorphic processes. *Geomorphology,* 14**,** 277-295.

Hurd, J. D., D. L. Civco, M. S. Gilmore, S. Prisloe & E. H. Wilson. 2006. Tidal wetland classification from Landsat imagery using an integrated pixel-based and object-based classification approach. In *American Society for Photogrammetry and Remote Sensing, 2006 Annual Conference, Reno, Nevada*.

Iverson, L. R., M. E. Dale, C. T. Scott & A. Prasad (1997) A GIS-derived integrated moisture index to predict forest composition and productivity of Ohio forests (USA). *Landscape Ecology,* 12**,** 331-348.

Jacek, S. (1997) Landform characterization with geographic information systems. *Photogrammetric Engineering & Remote Sensing,* 63**,** 183-191.

Jean, M. & A. Bouchard (1991) Temporal changes in wetland landscapes of a section of the St. Lawrence River, Canada. *Environmental Management,* 15**,** 241-250.

Jensen, J. R., E. J. Christensen & R. Sharitz (1984) Nontidal wetland mapping in South Carolina using airborne multispectral scanner data. *Remote Sensing of Environment,* 16**,** 1-12.

Jensen, J. R. & K. Lulla (1987) Introductory digital image processing: a remote sensing perspective.

Jensen, J. R., K. Rutchey, M. S. Koch & S. Narumalani (1995) Inland wetland change detection in the Everglades Water Conservation Area 2A using a time series of normalized remotely sensed data. *Photogrammetric Engineering and Remote Sensing,* 61**,** 199-209.

Johnston, R. M. & M. M. Barson (1993) Remote sensing of Australian wetlands: An evaluation of Landsat TM data for inventory and classification. *Marine and Freshwater Research,* 44**,** 235-252.

Kaplan, G. & U. Avdan (2018) Sentinel-1 and Sentinel-2 data fusion for wetlands mapping: Balikdami, Turkey. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences,* 42.

Kasischke, E. S. & L. L. Bourgeau-Chavez (1997) Monitoring South Florida wetlands using ERS-1 SAR imagery. *Photogrammetric Engineering and Remote Sensing,* 63**,** 281-291.

Keddy, P. A. 2010. *Wetland ecology: principles and conservation*. Cambridge University Press.

Kennedy, G. & T. Mayer (2002) Natural and constructed wetlands in Canada: An overview. *Water Quality Research Journal of Canada,* 37**,** 295-325.

Kindscher, K., A. Fraser, M. Jakubauskas & D. Debinski (1997) Identifying wetland meadows in Grand Teton National Park using remote sensing and average wetland values. *Wetlands Ecology and Management,* 5**,** 265-273.

Klemas, V., F. Daiber, D. Bartlett, O. Crichton & A. Fornes (1974) Inventory of Delaware's wetlands. *Photogrammetric Engineering,* 40.

Kontoes, C., G. Wilkinson, A. Burrill, S. Goffredo & J. Megier (1993) An experimental system for the integration of GIS data in knowledge-based image analysis for remote sensing of agriculture. *International Journal of Geographical Information Systems,* 7**,** 247-262.

Kushwaha, S., R. Dwivedi & B. Rao (2000) Evaluation of various digital image processing techniques for detection of coastal wetlands using ERS-1 SAR data. *International Journal of Remote Sensing,* 21**,** 565-579.

Lang, M., J. Awl, B. Wilen, G. McCarty & J. Galbraith (2009) Improved Wetland Mapping. *National Wetlands Newsletter,* 31.

Lang, M. W., E. S. Kasischke, S. D. Prince & K. W. Pittman (2008) Assessment of C-band synthetic aperture radar data for mapping and monitoring Coastal Plain forested wetlands in the Mid-Atlantic Region, USA. *Remote Sensing of Environment,* 112**,** 4120-4130.

Lauver, C. L. & J. Whistler (1993) A hierarchical classification of Landsat TM imagery to identify natural grassland areas and rare species habitat. *Photogrammetric engineering and remote sensing,* 59**,** 627-634.

Lefsky, M. A., W. B. Cohen, G. G. Parker & D. J. Harding (2002) Lidar Remote Sensing for Ecosystem Studies Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. *BioScience,* 52**,** 19-30.

Li, J. & W. Chen (2005) A rule-based method for mapping Canada's wetlands using optical, radar and DEM data. *International Journal of Remote Sensing,* 26**,** 5051-5069.

Li, J., W. Chen & R. Touzi (2007) Optimum RADARSAT-1 configurations for wetlands discrimination: a case study of the Mer Bleue peat bog. *Canadian Journal of Remote Sensing,* 33**,** S46-S55.

Lichvar, R. W., D. C. Finnegan, S. Newman & W. Ochs. 2006. Delineating and Evaluating Vegetation Conditions of Vernal Pools Using Spaceborne and Airborne Remote Sensing Techniques, Beale Air Force Base, CA. DTIC Document.

Lim, K., P. Treitz, M. Wulder, B. St-Onge & M. Flood (2003) LiDAR remote sensing of forest structure. *Progress in physical geography,* 27**,** 88-106.

Llewellyn, D. W., G. P. Shaffer, N. J. Craig, L. Creasman, D. Pashley, M. Swan & C. Brown (1996) A Decision-Support System for Prioritizing Restoration Sites on the Mississippi River Alluvial Plain. *Conservation Biology,* 10**,** 1446-1455.

Loh, W. Y. (2011) Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* 1**,** 14-23.

Lunetta, R. S. & M. E. Balogh (1999) Application of multi-temporal Landsat 5 TM imagery for wetland identification. *Photogrammetric Engineering and Remote Sensing,* 65**,** 1303-1310.

Lyon, J. G., R. D. Lopez, L. K. Lyon & D. K. Lopez. 2001. *Wetland landscape characterization: GIS, remote sensing and image analysis*. CRC Press.

MacAlister, C. & M. Mahaxay (2009) Mapping wetlands in the Lower Mekong Basin for wetland resource and conservation management using Landsat ETM images and field survey data. *Journal of Environmental Management,* 90**,** 2130-2137.

Marceau, D. J., P. J. Howarth, J.-M. M. Dubois & D. J. Gratton (1990) Evaluation of the grey-level co-occurrence matrix method for land-cover classification using SPOT imagery. *IEEE Transactions on Geoscience and Remote Sensing,* 28**,** 513-519.

McCullagh, P. & J. A. Nelder. 1989. *Generalized Linear Models, Second Edition*. Taylor & Francis.

Melgani, F. & L. Bruzzone (2004) Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing,* 42**,** 1778-1790.

Melloh, R. A., C. C. Racine, S. Sprecher, N. H. Greeley & P. B. Weyrick. 1999. Comparisons of digital terrain data for wetland inventory on two Alaskan Army bases. DTIC Document.

Milne, A., G. Horn & M. Finlayson (2000) Monitoring wetlands inundation patterns using RADARSAT multitemporal data. *Canadian Journal of Remote Sensing,* 26**,** 133-141.

Milton, G. & R. Hélie. 2003. Wetland inventory and monitoring: partnering to provide a national coverage. In *Wetland Stewardship in Canada: Contributed Papers from the Conference on Canadian Wetlands Stewardship*, 3-5.

Moreau, S., R. Bosseno, X. F. Gu & F. Baret (2003) Assessing the biomass dynamics of Andean bofedal and totora high-protein wetland grasses from NOAA/AVHRR. *Remote Sensing of Environment,* 85**,** 516-529.

Moreno-Seco, F., J. M. Inesta, P. J. P. De León & L. Micó. 2006. Comparison of classifier fusion methods for classification in pattern recognition tasks. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 705-713. Springer.

Mountrakis, G., J. Im & C. Ogole (2011) Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing,* 66**,** 247-259.

Munyati, C. (2000) Wetland change detection on the Kafue Flats, Zambia, by classification of a multitemporal remote sensing image dataset. *International Journal of Remote Sensing,* 21**,** 1787-1806.

N.C. Wetland Functional Assessment Team. 2010. N.C. Wetland Assessment Method (NC WAM) User Manual.

Naidoo, L., M. Cho, R. Mathieu & G. Asner (2012) Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing,* 69**,** 167-179.

Nykodym, T., T. Kraljevic, N. Hussami, A. Rao & A. Wang. 2016. Generalized Linear Modeling with H2O.

Olaya, V. (2009) Basic Land-Surface Parameters. *Developments in Soil Science,* 33**,** 141-169.

Olhan, E., S. Gun, Y. Ataseven & H. Arisoy (2010) Effects of agricultural activities in Seyfe Wetland. *Scientific Research and Essays,* 5**,** 9-14.

Ozesmi, S. L. & M. E. Bauer (2002) Satellite remote sensing of wetlands. *Wetlands ecology and management,* 10**,** 381-402.

Pakhale, G. & P. Gupta (2010) Comparison of advanced pixel based (ANN and SVM) and object-oriented classification approaches using landsat-7 Etm+ data. *International Journal of Engineering and Technology,* 2**,** 245-251.

Pal, M. & P. M. Mather (2003) An Assessment of the Effectiveness of Decision Tree Methods for Land Cover Classification. *Remote Sensing of Environment,* 86**,** 554-565.

Pike, R. J. & S. E. Wilson (1971) Elevation-relief ratio, hypsometric integral, and geomorphic area-altitude analysis. *Geological Society of America Bulletin,* 82**,** 1079-1084.

Pope, K. O., E. Rejmankova, H. M. Savage, J. I. Arredondo-Jimenez, M. H. Rodriguez & D. R. Roberts (1994) Remote sensing of tropical wetlands for malaria control in Chiapas, Mexico. *Ecological Applications,* 4**,** 81-90.

Powers, R. P., G. J. Hay & G. Chen (2012) How wetland type and area differ through scale: A GEOBIA case study in Alberta's Boreal Plains. *Remote Sensing of Environment,* 117**,** 135-145.

Ramsey III, E. (1998) Radar remote sensing of wetlands. *Remote sensing change detection: environmental monitoring methods and applications***,** 211-243.

Ramsey III, E. W. & S. C. Laine (1997) Comparison of Landsat Thematic Mapper and high resolution photography to identify change in complex coastal wetlands. *Journal of coastal research***,** 281-292.

Rebelo, L.-M., C. Finlayson & N. Nagabhatla (2009) Remote sensing and GIS for wetland inventory, mapping and change analysis. *Journal of Environmental Management,* 90**,** 2144-2153.

Reese, H. M., T. M. Lillesand, D. E. Nagel, J. S. Stewart, R. A. Goldmann, T. E. Simmons, J. W. Chipman & P. A. Tessar (2002) Statewide land cover derived from multiseasonal Landsat TM data: a retrospective of the WISCLAND project. *Remote Sensing of Environment,* 82**,** 224-237.

Richards, J. A. & J. Richards. 1999. *Remote sensing digital image analysis*. Springer.

Rodríguez-Galiano, V., F. Abarca-Hernández, B. Ghimire, M. Chica-Olmo, P. Atkinson & C. Jeganathan (2011) Incorporating spatial variability measures in land-cover classification using Random Forest. *Procedia Environmental Sciences,* 3**,** 44-49.

Rogers, A. & M. Kearney (2004) Reducing signature variability in unmixing coastal marsh Thematic Mapper scenes using spectral indices. *International Journal of Remote Sensing,* 25**,** 2317-2335.

Running, S. W., T. R. Loveland, L. L. Pierce, R. Nemani & E. Hunt (1995) A remote sensing based vegetation classification logic for global land cover analysis. *Remote sensing of Environment,* 51**,** 39-48.

Sader, S. A., D. Ahl & W.-S. Liou (1995) Accuracy of Landsat-TM and GIS rule-based methods for forest wetland classification in Maine. *Remote Sensing of Environment,* 53**,** 133-144.

Safavian, S. R. & D. Landgrebe (1991) A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics,* 21**,** 660-674.

Schmullius, C. & D. Evans (1997) Review article Synthetic aperture radar (SAR) frequency and polarization requirements for applications in ecology, geology, hydrology, and oceanography: A tabular status quo after SIR-C/X-SAR. *International Journal of Remote Sensing,* 18**,** 2713-2722.

Shaeffer, D. L. 2008. Characterizing Jurisdictional Wetlands Using Aerial LiDAR. East Carolina University.

Shang, J. 1996. Evaluation of multi-spectral scanner and radar satellite data for wetland detection and classification in the Great Lakes Basin. Masters thesis, University of Windsor, Ontario.

Short, N. (2010) The remote sensing tutorial.

Skidmore, A. K. (1989) An expert system classifies eucalypt forest types using thematic mapper data ans a digital terrain model. *Photogrammetric Engineering and Remote Sensing,* 55**,** 1449-1464.

Sohl, T. L., A. L. Gallant & T. R. Loveland (2004) The characteristics and interpretability of land surface change and implications for project design. *Photogrammetric Engineering & Remote Sensing,* 70**,** 439-448.

Sokol, J., T. Pultz & V. Bulzgis. 2000. Investigating Labrador fens and bogs using multi-temporal ERS-2 and RADARSAT data. In *Proceedings of the 22nd Canadian Symposium on Remote Sensing*, 357-364.

Song, J.-H., S.-H. Han, K. Yu & Y.-I. Kim (2002) Assessing the possibility of land-cover classification using lidar intensity data. *International archives of photogrammetry remote sensing and spatial information sciences,* 34**,** 259-262.

Spanglet, H. J., S. L. Ustin & E. Rejmankova (1998) Spectral reflectance characteristics of California subalpine marsh plant communities. *Wetlands,* 18**,** 307-319.

Töyrä, J. & A. Pietroniro (2005) Towards operational monitoring of a northern wetland using geomatics-based techniques. *Remote Sensing of Environment,* 97**,** 174-191.

Töyrä, J., A. Pietroniro & L. W. Martz (2001) Multisensor hydrologic assessment of a freshwater wetland. *Remote sensing of Environment,* 75**,** 162-173.

Tenenbaum, D. E., L. E. Band, S. Kenworthy & C. Tague (2006) Analysis of soil moisture patterns in forested and suburban catchments in Baltimore, Maryland, using high-resolution photogrammetric and LIDAR digital elevation datasets. *Hydrological Processes,* 20**,** 219-240.

Thenkabail, P. S., J. G. Lyon & A. Huete. 2016. *Hyperspectral remote sensing of vegetation*. CRC Press.

Tiner, R. W. (1990) Use of high-altitude aerial photography for inventorying forested wetlands in the United States. *Forest Ecology and Management,* 33**,** 593-604.

Toner, M. & P. Keddy (1997) River hydrology and riparian wetlands: a predictive model for ecological assembly. *Ecological Applications,* 7**,** 236-246.

Townsend, P. A. (2001) Mapping seasonal flooding in forested wetlands using multi-temporal Radarsat SAR. *Photogrammetric engineering and remote sensing,* 67**,** 857-864.

Townsend, P. A. & S. J. Walsh (1998) Modeling floodplain inundation using an integrated GIS with radar and optical remote sensing. *Geomorphology,* 21**,** 295-312.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer Science & Business Media.

Vierling, K. T., L. A. Vierling, W. A. Gould, S. Martinuzzi & R. M. Clawges (2008) Lidar: shedding new light on habitat characterization and modeling. *Frontiers in Ecology and the Environment,* 6**,** 90-98.

Waite, W. P. & H. C. MacDonald (1971) " Vegetation Penetration" with K-Band Imaging Radars. *IEEE Transactions on Geoscience Electronics,* 9**,** 147-155.

Walker, P. A. & D. Moore (1988) SIMPLE An inductive modelling and mapping tool for spatially-oriented data. *International Journal of Geographical Information System,* 2**,** 347-363.

Wang, J., J. Shang, B. Brisco & R. Brown (1998) Evaluation of multidate ERS-1 and multispectral Landsat imagery for wetland detection in southern Ontario. *Canadian journal of remote sensing,* 24**,** 60-68.

Wang, S.-G., J. Deng, M. Chen, M. Weatherford & L. Paugh. 2015. Random Forest Classification and Automation for Wetland Identification Based on DEM Derivatives. In *ICOET*. Raleigh, NC, USA.

Wang, S.-G., A. S. Smith & S. Davis. 2014. Improvements to NCDOT's Wetland Prediction Model (Phase II). In *NCDOT Proposal 2016-16*.

Wei, W., X. Zhang, X. Chen, J. Tang & M. Jiang (2008) Wetland mapping using subpixel analysis and decision tree classification in the Yellow River delta area. *ISPRS Archives,* 38**,** 667-670.

Wright, C. & A. Gallant (2007) Improved wetland remote sensing in Yellowstone National Park using classification trees to combine TM imagery and ancillary environmental data. *Remote Sensing of Environment,* 107**,** 582-605.

Xu, X. 2014. *A knowledge-based approach of satellite image classification for urban wetland detection*.

Yang, J., F. J. Artigas & J. Wang. 2009. *Mapping salt marsh vegetation by integrating hyperspectral and LiDAR remote sensing*. CRC Press: Boca Raton, FL, USA.

Zhang, C. (2014) Combining hyperspectral and LiDAR data for vegetation mapping in the Florida Everglades. *Photogrammetric Engineering & Remote Sensing,* 80**,** 733-743.

Zhang, C. & Z. Xie (2012) Combining object-based texture measures with a neural network for vegetation mapping in the Everglades from hyperspectral imagery. *Remote Sensing of Environment,* 124**,** 310-320.

--- (2014) Data fusion and classifier ensemble techniques for vegetation mapping in the coastal Everglades. *Geocarto International,* 29**,** 228-243.

Zhang, L., Y. Zhong, B. Huang, J. Gong & P. Li (2007) Dimensionality reduction based on clonal selection for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing,* 45**,** 4172-4186.

Zhang, Y., D. Lu, B. Yang, C. Sun & M. Sun (2011) Coastal wetland vegetation classification with a Landsat Thematic Mapper image. *International Journal of Remote Sensing,* 32**,** 545-561.