Improvements to NCDOT's Wetland Prediction Model

Final Report

UNCC WAM Research Team **(for NCDOT RP 2013-13)**

P.I.: Sheng-Guo Wang, Professor, UNC Charlotte

Date: 02-25-2015

1. Report No. ***NCDOT RP 2013-13***	2. Government Accession No. … …	3. Recipient's Catalog No. … …

4. Title and Subtitle Improvements to NCDOT's Wetland Prediction Model	5. Report Date ***February 28, 2015*** 6. Performing Organization Code … …
7. Author(s) Sheng-Guo Wang (PI), Libin Bai, Jing Deng, Meijuan Jia, Mingzhi Chen, Shu Liang, Peng Wang, Sushmitha Yalla, Fangjian Huang, Shen-En Chen, Wenwu Tang	8. Performing Organization Report No. … …
9. Performing Organization Name and Address Dept. of Engineering Technology Dept. of Software & Information Systems Dept. of Geography and Earth Sciences Dept. of Civil & Environmental Engineering UNC – Charlotte Charlotte, NC 28223-0001	10. Work Unit No. (TRAIS) … … 11. Contract or Grant No. … …
12. Sponsoring Agency Name and Address NC Department of Transportation	13. Type of Report and Period Covered **Final Report** ***April 30, 2012 – February 15, 2015*** 14. Sponsoring Agency Code *NCDOT Project # 2013-13*

15. Supplementary Notes:

This project was supported by a grant from the U.S. Department of Transportation and the North Carolina Department of Transportation

16. Abstract

This Final Report is to summarize several main achievements of this project as follows:

(i) Automation Method and its Tools for the Wetland Identification and Analysis Process;
(ii) Method Development for Wetland Identification Process;
(iii) Reliability and Flexibility of Automation Tools and Methods; and
(iv) User Friendly deliverables.

These achievements fit the NCDOT research needs as: "while NCDOT has made significant advances with the concept, the process and tools of predicting wetlands using LiDAR is under-developed."

The goal of the project is to provide the improved LiDAR-based wetland prediction models with *highly automated, reliable, and user-friendly tools* based on ArcGIS for NCDOT.

The UNC Charlotte WAM Research Team has successfully completed a number of valuable research topics related to wetland prediction process, such as process automation, variables exploration, data mining, and statistical analysis.

The acclaimed results include the deliverable WAM Automation Process Tools and the Users' Guide to the Tools.

17. Key Words *Wetland, Automation, Modeling, Prediction, Analysis*	18. Distribution Statement … …

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages *48*	22. Price … …

DISCLAIMER

The contents of this report reflect the views of the authors and not necessarily the views of the University. The authors are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the North Carolina Department of Transportation or the Federal Highway Administration at the time of publication. This report does not constitute a standard, specification, or regulation.

ACKNOWLEDGEMENTS

The research team thanks the North Carolina Department of Transportation for supporting and funding this project. We extend special thanks to the project Steering and Implementation Committee members:

- Morgan Weatherford (Chair)
- Philip S. Harris III
- LeiLani Paugh
- David Johnson
- James Mason
- Colin Mellor
- Sarah Schwarzer
- Mrinmay Biswas
- John W. Kirby

The research team is indebted to the tremendous support provided by these committee members in helping advices and the scope of the project.

Special thanks are given to LeiLani Paugh, Morgan Weatherford, John Kirby and Dave Johnson at NCDOT who provided us with valuable information and expert advices about the WAM in North Carolina, and support to the project.

We also want to thank Alexander P. (Sandy) Smith and Scott Davis at Axiom Environmental for their expert support, especially, for wetland field visits and model testing.

Without the help of all of the above individuals, this project could not have such scientific results in such a successful manner, leading to our deliverable automation tools as a useful product.

EXECUTIVE SUMMARY

The goal of this project is to provide the improved LiDAR-based wetland prediction models with highly automated, reliable, and user-friendly tools based on ArcGIS for NCDOT as stated in the contracted proposal.

To follow the NCDOT Jan 15th (2014) Meeting and Guidance for Future Work (REVISED), we list the following Table 1 as a Task-and-Status Check list. The team is required to complete these task items by the end of the contract. We are pleased to report that all required tasks are successfully completed.

Table 1. Tasks and Status Check List

Tasks: Objective Items	1. Finish Lenoir County & Provide results to NCDOT	2. Finish refining RF model	3. Develop & apply models for Piedmont; & Assess accuracy	4. Finish automation programs and documentation & Provide demo to NCDOT	5. Automate identification/ mapping riparian areas	6. Develop a conceptual framework: how to identify wetland types
Key features	3 eco-regions 12 sampling areas for modeling	RF model	10 counties	V.3.1.1: Java 1.7 Sampling rate choice BDPrediction Logit & RF	Based on available software from web for Riparian	Discussion with experts L. Paugh, M. Weatherford, & S. Smith, S. Davis
Information need from NCDOT	Sampled wetland data in Lenoir		Sampled wetland data in Piedmont areas			
Team	Wang & All	Wang, Deng, M. Chen, Bai, Liang	Wang, Yalla, Bai, P. Wang, Liang, S. Chen	Wang, Bai, Yalla	Wang, Bai	Wang, Bai, P. Wang, Liang
Completed Time	Spring 2014 (Lenoir) Jan. 2015 (Piedmont)	Spring 2014	Field visit – Aug. 2014. Full sampling Jan. 2015	v3.0 Aug. 2014 v3.1 Dec. 2014 v3.1.1-Jan. 2015	Spring 2014	Spring 2014
Status Check	√	√	√	√	√	√

Recognized nationally in recent 2011 FHWA Environmental Excellence Awards (EEA) for the Excellence in Environmental Research, NCDOT and NCDENR developed and integrated airborne LiDAR, digital imagery technology and statistical modeling for wetland and stream measurement and inventory reporting. NCDOT has identified in [4] that “while NCDOT has made significant advances with the concept, the process and tools of predicting wetlands using LiDAR is under-developed. Additional research needs to be undertaken to improve the reliability and performance of the model across the various wetland types within the state.”

This project addresses the above mentioned NCDOT’s needs to enhance the NCDOT wetland prediction procedure via process automation and methods as well as automation tools.

The achievement of this project may be summarized as follows:

(i) Automation and its Tools of Wetland Identification and Analysis Process;

(ii) Systematic Methods of Wetland Identification Process;

(iii) Reliability and Flexibility of the Developed Tools and Methods; and

(iv) User Friendly Deliverables.

These achievements fit the NCDOT research needs through a number of valuable research topics related to wetland prediction process, such as process automation, variables exploration, data mining, post-treatment, and statistical analysis.

In addition, early September 2014, the NCDOT IT made a new request on Java 1.7 issue, even though ArcGIS v10.1 at both NCDOT and UNCC only supports Java 1.6 and does not support Java 1.7. In order to satisfy the NCDOT IT requirement, the PI proposed a new idea to run our Java files outside ArcGIS, and use Java 1.7. Thus, this approach has led to develop our WAMAT (WAM Automation Tools) v3.1 and the current v3.1.1. The initial results show this method works very well to not only run in Java 1.7, but it also increases the speed and the memory benefit. Based on NCDOT expert Morgan Weatherford’s much helpful feedback, the team has also updated the Users’ Guide for WAMAT v3.1.1.

The acclaimed results include the deliverable WAM Automation Process Tools and the Users’ Guide to the Tools, which can benefit NCDOT and EPA by innovative automation and models to have labor saving in the NEPA process. In view of the 2011 FHWA Environmental Excellence Awards (EEA) to NCDOT and NCDENR, we believe that our work and deliverable product will be valued on a national level.

CONTENTS

DISCLAIMER.. 3

ACKNOWLEDGEMENTS 4

EXECUTIVE SUMMARY 5

CONTENTS.. 7

1. INTRODUCTION.. 11

2. WETLAND PREDICTION MODELS 13

2.1 WETLAND PREDICTOR VARIABLES 13

2.2 LOGISTIC REGRESSION (LOGIT) MODEL 14

2.3 RANDOM FOREST (RF)...16

2.3.1 RF model for project areas in three eco-regions 17

2.3.2 Prediction results 18

2.4 RANDOM FOREST SMART VERSION (RF-S)...23

3. AUTOMATION PROCESS.. 24

3.1 OVERALL ARCHITECTURE OF THE AUTOMATION TOOLSET 25

3.2. DISPLAY 25

3.3. ADVANTAGES OF THE SOFTWARE 27

4. CASE STUDY AND FIELD VALIDATION 28

4.1 WETLAND PREDICTION OF LENOIR COUNTY 28

4.1.1 Model construction...28

4.1.2 Model validation 29

4.2 WETLAND PREDICTION OF SOUTHERN OUTER PIEDMONT.......................................32

4.3 FIELD VALIDATION 34

5. AUXILIARY VARIABLE AND DATA 39

5.1 Riparian variable 39

5.2 Soil data 40

5.3 Land cover...40

6. VALUABLE ADDITIONAL METHODOLOGY 41

6.1. SCAD..41

6.2 Balance treatment of input data 41

7. CONCLUSION 44

8. REFERENCES... 45

9. APPENDIX.. 48

[A] WAMAT – WAM Automation Tools v.3.1.1, UNCC WAM Team, Patent Pending, 01-20-2015. ...48

[B] WAM Automation Tools (WAMAT) – QUICK START GUIDE, Version 3.1.1, Sheng-Guo Wang (PI), Libin Bai and Sushmitha Yalla, 01-23-2015...48

[C] Installation Guide for Computer with Old Version of WAMAT, UNCC WAM Team, 8-14-2014. 48

[D] Random Forest Classification and Automation for Wetland Identification based on DEM Derivatives, S.-G. Wang, J. Deng, M. Chen, M. Weatherford, and L. Paugh, 2015. 48

LIST OF FIGURES

Figure 1. Project Goal 12
Figure 2. Random Forest Method 16
Figure 3. Study area 17
Figure 4. An example of relative importance of variables (FW project B4085) 18
Figure 5. Summary of model comparison by error rate... 22
Figure 6. Scheme for RF improvements as Smart RF 23
Figure 7. Automation Process.. 25
Figure 8. Display of wetland prediction result: (a) Logistic Regression (left), (b) RF (right) 26
Figure 9. Prediction for Project U3826 (SFLT) (a) Logistic Regression (left), (b) RF (right)..... 27
Figure 10. Sample data of 13 project areas.. 28
Figure 11. Study area of field trip for Lenoir County.. 35
Figure 12. Study area of field trip for Southern Outer Piedmont region 36
Figure 13. Sample to expand the minority type.. 42

LIST OF TABLES

Table 1. Tasks and Status Check List 5

Table 2. Predictor Variables used to build wetland models.. 14

Table 3. Results comparison between RF and Logit for FW projects 19

Table 4. Results comparison between RF and Logit model for RCP projects 20

Table 5. Results comparison between RF and Logit model for SFLT projects............................ 21

Table 6. Performance comparison of Logit models for Lenoir County....................................... 30

Table 7. Performance comparison of RF models for Lenoir County 31

Table 8. Performance comparison of models for Southern Outer Piedmont region.................... 33

Table 9. Summary validation results in Southern Outer Piedmont Region.................................. 37

Table 10. Comparison of balancing sampling function for Logit method................................... 43

Table 11. Comparison of balancing sampling function for RF method 43

1. Introduction

This Final Report is for the NCDOT Project RP 2013-13, titled “Improvements to NCDOT’s Wetland Prediction Model” during 04-30-2012 through 02-15-2015. It concludes several main achievements of this project as follows:

(i) Automation and its Tools of Wetland Identification and Analysis Process;

(ii) Systematic Methods of Wetland Identification Process;

(iii) Reliability and Flexibility of the Developed Tools and Methods; and

(iv) User Friendly Deliverables.

Recently, the FHWA presented the 2011 Environmental Excellence Awards (EEA) to the NCDOT and NCDENR for the Excellence in Environmental Research in “GIS-based Wetland and Stream Predictive Models” [1]. Recognized nationally and internationally in [1-4, 12-13], as a necessary future trend in North America to develop and integrate airborne LiDAR, digital imagery and pattern recognition technology [4-5, 9-14] for 21st century transportation and environment monitoring, the NCDOT and NCDENR achievement in integrating LiDAR imagery measurement and GIS coupled with stream/wetland prediction is not trivial.

This project addresses the NCDOT research needs defined previously in [4] as: “while NCDOT has made significant advances with the concept, the process and tools of predicting wetlands using LiDAR is under-developed. Additional research needs to be undertaken to improve the reliability and performance of the model across the various wetland types within the state.” Further, procedures and models to identify wetland types and characteristics do not exist but would be highly valuable information to obtain [4].

The need definition of the NCDOT addressed by this project is to enhance the NCDOT wetland prediction procedure via procedure automation and further mathematical modeling. The above mentioned achievements fit the NCDOT research needs.

To further enable reliable identification of wetland locations by improving the NCDOT/NCDENR model, this project does research in process automation and systematic methods development. The benefits to NCDOT include significantly reducing the time and cost of field delineations and providing early awareness of potential wetland impact areas in NC [1].

The significance of LiDAR implementation into wetland identification and modeling, as stated by the FHWA is to exemplify "how innovative technologies can be used to speed the environmental assessment process and ultimately advance transportation projects while protecting the environment" [1]. Therefore, this project research [4, 5] is important and highly needed. In addition, it contributes to NCDOT by keeping the leading status in this important area of research, which can benefit NCDOT by innovative models and significant labor saving in the NEPA process [2].

This project includes a number of valuable research topics related to wetland prediction, such as process automation, variables exploration, data mining, and statistical analysis. According to the project proposal [5], our goal for this project is to provide improved NCDOT LiDAR-based wetland prediction models with *highly automated, reliable, and user-friendly tools* for NCDOT based on ArcGIS as shown in Figure 1. Therefore, we mainly concentrate on the topics of process automation and modeling and prediction methods for this project.

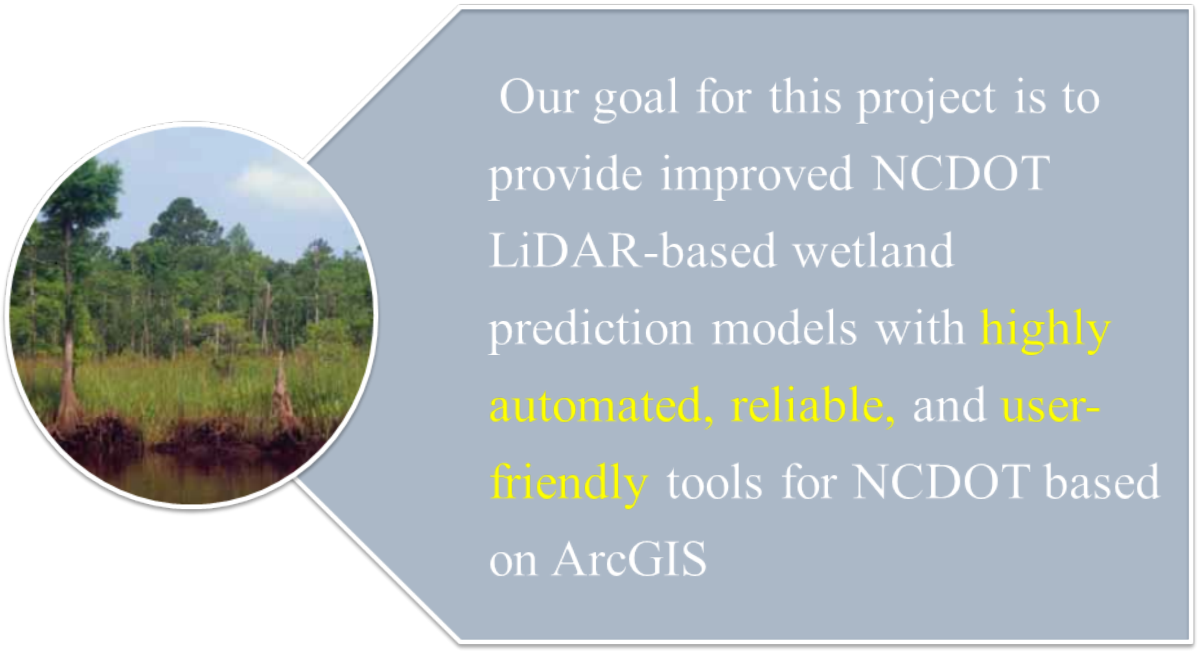

Our goal for this project is to provide improved NCDOT LiDAR-based wetland prediction models with highly automated, reliable, and user-friendly tools for NCDOT based on ArcGIS

Figure 1. Project Goal

In this report, we conclude our work for these research topics, and at the same time we focus on the NCDOT Guidance Memo of 01-17-2014.

The rest of this report is organized in the following manner. Section 2 is to illustrate the prediction models we applied in this research, including the original models and refined models. Section 3 describes the process automation with different tasks in wetland prediction. In Section 4, two case studies are conducted by applying our refined models and automation process to Lenoir County and Southern Outer Piedmont Region. The prediction results are reported and the models are evaluated in this section. Section 5 is about the research data and variables we conducted in addition to the DEM (Digital Elevation Model), e.g., soil and land cover, as well as riparian variable. In Section 6, we describe an additional methodology for WAM variable selection via SCAD (Smoothly Clipped Absolute Deviation) penalty, and input data balance method via balanced sampling. Section 7 provides summary remarks of the project. This final report also includes the attached deliverable WAM Automation Tools (WAMAT) for the Wetland Prediction Process Automation, and the Users' Guide to the Tools.

2. Wetland Prediction Models

In this section, we summarize the wetland prediction models and their methods we applied and developed with their performances as follows:

(1) Logistic Regression model,

(2) Random Forest model,

(3) Random Forest Smart model as RF-s, and

(4) Logistic Regression Smart model as Logit-s.

2.1 Wetland Predictor Variables

For wetland prediction, we take the following predictor variables as in Table 2.

Table 2. Predictor Variables used to build wetland models

Variable Name	Full Name	Formula and Illustrations
elev	Elevation	Elevation of each cell: z(x,y).
asp	Aspect	asp = 57.29578 * atan2 ([dz/dy], -[dz/dx])
slp	Slope	$\mathrm{slp} = 57.29578 \times \mathrm{atan}(\sqrt{{}^{dz}/_{dx}{}^{2} + {}^{dz}/_{dy}{}^{2}})$
cv	Curvature	Slope of the slope: $\mathrm{cv} = 57.29578 \times \mathrm{atan}(\sqrt{{}^{ds}/_{dx}{}^{2} + {}^{ds}/_{dy}{}^{2}})$
prcv	Profile Curvature	Curvature on vertical (y) direction
plcv	Plan Curvature	Curvature on horizontal (x) direction
batwi	Ratio of Slope by Drainage Area	batwi = slp / drainage contributing area (calculated with breach-all DEM)
wei	Wetness Elevation Index	Series of increasingly larger neighborhoods used to determine the relative landscape position of each cell.
weiRe	Reclassification of wei	Wei value of each cell will be reclassified as 0 if original value is bigger than a predefined threshold, else is reclassified as 1 (default threshold is 0.5).
mdec	Maximum Downslope Elevation Change	Maximum difference of z(x,y) between the cell and its neighbor cells.
rawda	Stochastic Depression Analysis	Stochastic depression analysis based on raw DEM.
curv5	Smooth Curvature	Each cell gets mean value of curvature from its 5*5 neighbors. $Cur5 = \sum_{i=i1}^{i25} \frac{cv(i)}{25}$
depan	Stochastic Depression Analysis	Stochastic depression analysis based on breach-all DEM.
gap	Land Cover Data	Categorized land use types.
soil	Soil Data	Reclassified as 1 or 0 to indicate hydric or non hydric soil type.

Table 2 lists these variables with an illustration of variables calculation. These predictor variables are selected by NCDOT models, most of them are calculated based on DEM data.

2.2 Logistic Regression (Logit) Model

Firstly, we have applied the logistic regression model to classify the landscape into two categories of wetland and non-wetland. Before we describe the logistic regression model, let's first describe a linear regression model as in (1), that predicts the occurrence of wetland as a function $y(x)$ of the selected explanatory variable vector $\boldsymbol{x}$ at a data point as

$$y = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \tag{1}$$

where $\boldsymbol{x}$ is the wetland variables vector $\boldsymbol{x} = [x_1, x_2, \cdots, x_m]^T$, y is a response variable as the prediction result, $\boldsymbol{\beta}$ is the coefficient vector as a "weighting factor" for the variable vector, and ε is an estimator/noise error or adjustment of this linear estimator. In a prediction area, each point (e.g., 20×20 feet as a point), the variable vector $\boldsymbol{x}$ can be arranged in a matrix X, and the corresponding response variable y can be presented as a vector $\boldsymbol{y}$, where each row represents a data point. Then we have the following linear regression model in a matrix-vector format as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

Because the response vector should be a binary-valued vector, i.e., the prediction model is a two-category classification, therefore a binary-valued model is used with a logistic function transform to (1) and called logistic regression. Logistic regression is just to take a transform on the continuous-valued response variable to predict a binary response with a "probability" value in [0, 1]. In statistics the probability describing the possible outcomes of a single trial is modeled as a function of predictor variables, using a logistic function

$$p = F(t) = \frac{e^t}{1+e^t} = \frac{1}{1+e^{-t}} \tag{3}$$

where $t = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$, i.e., to transform a continuous response y in (1) to a binary response. After the logistic function transform, we may have a generalized linear model for binary response in probability as

$$logit\ (E[y|\boldsymbol{x}\ ]) = logit\ (p) = \ln\left(\frac{p}{1-p}\right) = t = \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \tag{4}$$

$$p = E[y|\boldsymbol{x}\ ] = \frac{1}{1+e^{-\boldsymbol{\beta}^T x - \varepsilon}} \tag{5}$$

Sometimes, it is simply written as a new response variable y as follows

$$y = \frac{1}{1+e^{-\boldsymbol{\beta}^T x - \varepsilon}} \tag{6}$$

2.3 Random Forest (RF)

In order to reduce the sensitivity to data noise, we have applied a tree-based classification method Random Forest (RF) with 15 derivative variables to identify wetlands. Random trees are built by a set of rules with random and optimization, that uses bagging technique to randomly select sub-datasets and optimization technique to determine best decision tree nodes from a randomly selected sub-set of variables [41, D]. Thus, it leads to a random forest. Then, in the prediction process, RF can recursively partition the data into categories.

The classification tree analysis (CTA), also referred to classification and regression trees (CART), is a typical tree-based classification method. RF aims at improving predictive ability by taking the majority vote result from the prediction results of multiple trees in classification mode, or taking the average result of the prediction results of multiple trees in regression mode. Thus, this method is not sensitive to noise or overtraining, as resampling is not based on weighting. Furthermore, it is computationally much lighter than methods based on boosting and somewhat lighter than simple bagging. In the literature, it is used for land cover classification [34].

Here we have developed the RF for the wetland classification [41, D], especially using the above listed variables and creating RF-smart for modeling and prediction. For prediction, each tree in the forest generates a class result based on input data, and then the method collects the voting results from those trees. Thus, we evaluate the modeling method by testing cross-validation overall accuracies. It is described in the following Figure 2.

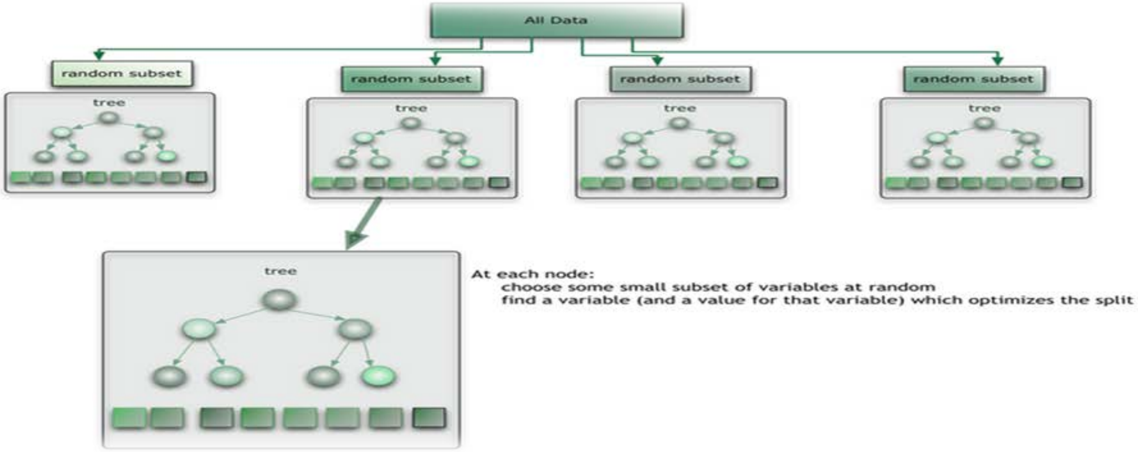

All Data
random subset
tree
random subset
tree
random subset
tree
random subset
tree
tree
At each node:
choose some small subset of variables at random
find a variable (and a value for that variable) which optimizes the split

Figure 2. Random Forest Method

2.3.1 RF model for project areas in three eco-regions

We first built RF models for 13 different wetland delineation projects in three eco-regions in North Carolina. The data and delineation results are provided by NCDOT. See Figure 3 for the whole study area. For each project, we randomly split the data into two parts, where 80% of the data are used for training purpose, while the remaining 20% of the data are used for validation. Using RF we can also obtain the relative importance of different variables, see Figure 4 as an example.

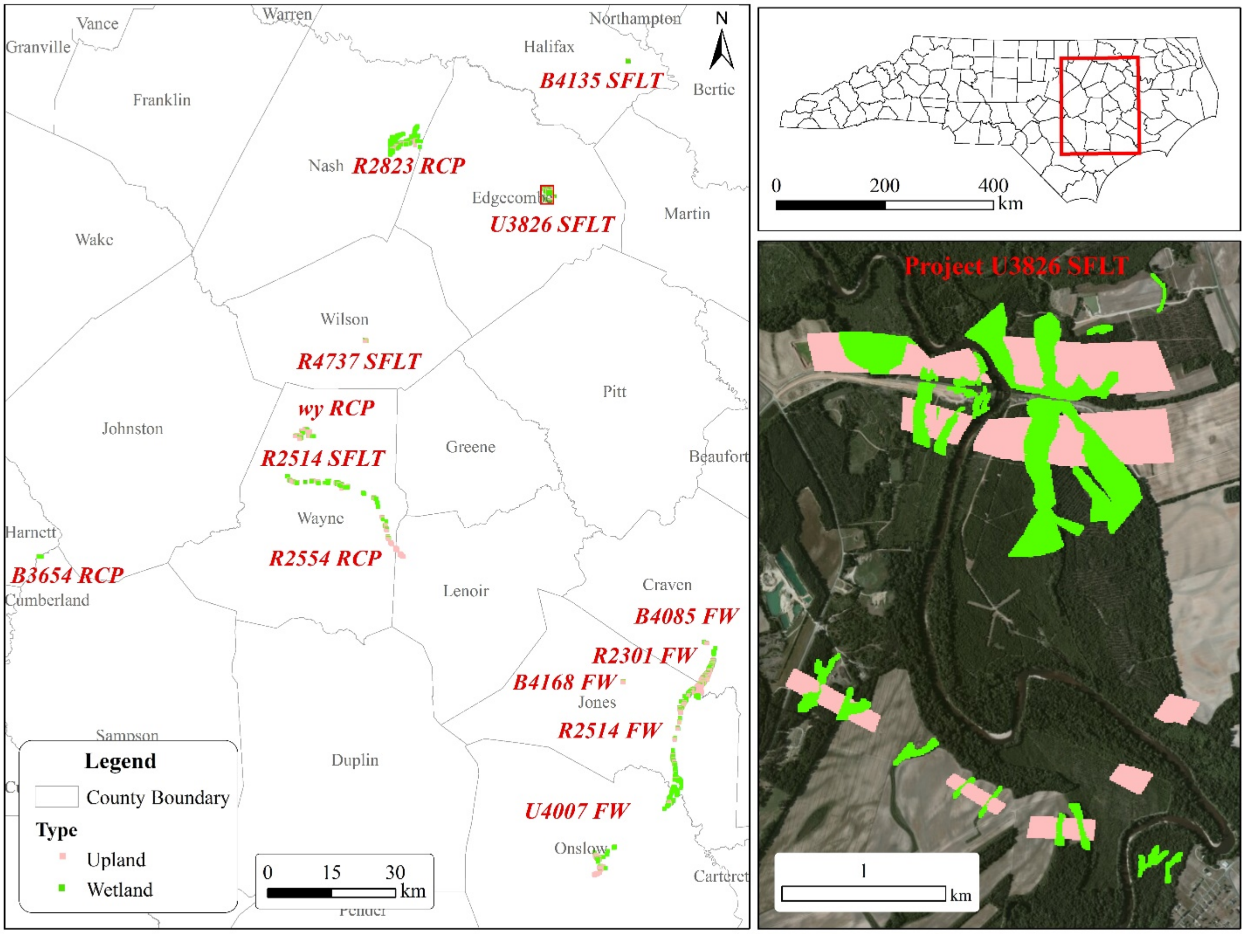


Figure 3. Study area

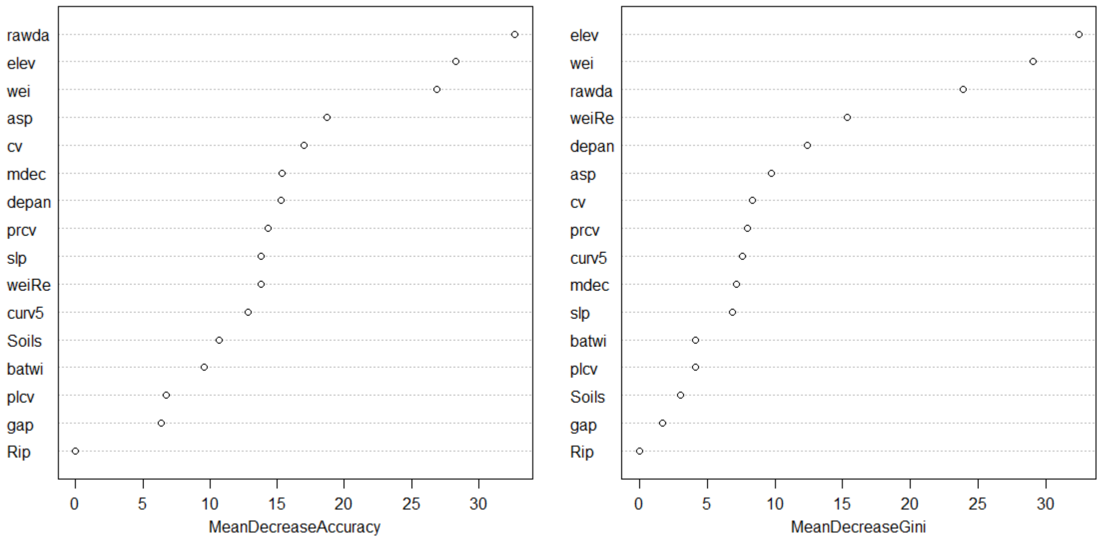


Figure 4. An example of relative importance of variables (FW project B4085)

As shown in Fig. 4, these are two common used methods to quantify and rank the importance of the variables. One is to calculate the decrease of Gini index (an impurity measure, the right figure) and another is to calculate the decrease of accuracy error (the left figure) every time introducing a new variables. These two figures here give readers some idea about the relative importance of these variables by the Gini index and the accuracy as a measure respectively. The top group of variables in the figures is important by these measures for the RF method.

For Logistic method, we apply SCAD method [9, 11] to effectively identify the important variables in the regression method, which we shall describe in Section 6.1.

2.3.2 Prediction results

In order to validate the accuracy for the RF models and to test whether the method can improve the prediction result, we compare the accuracy generated by our created Logit regression models (standard form) and RF models, where we do use the Xs (variables) and Y (wetland/non-wetland) data from NCDOT, but do remove variable Crews out of the Xs. Thus, our Logit model is different from NCDOT Logit model. Since the project data include three types of ecoregions, Flatwoods (FW), Rolling Coastal Plain (RCP), and Southeastern Floodplains and Low Terraces (SFLT), we group the data based on the ecoregion. The results are illustrated in Table 3, Table 4, and Table 5. We treat the data into two groups: the first group contains 80% records for training models (Logit and RF), and the second group contains the remaining 20% records for testing/checking the generated models. In these tables, the evaluation data (Y) are provided by NCDOT, which are verified wetland and non-wetland data.

Table 3. Results comparison between RF and Logit for FW projects

				Random Forest			Logistic Regression	Comparison
Type	**Project**	**Data**	**Total Records**	**0-1 Error Rate**	**1-0 Error Rate**	**Total Error Rate**	**Total Error Rate**	**Total Improvement**
FW	B4085	80%	780	0.00%	0.00%	0.00%	6.41%	100.00%
		20%	195	2.38%	3.60%	3.08%	5.64%	45.45%
		100%	975	0.44%	0.76%	0.62%	6.26%	90.16%
	B4168	80%	812	0.00%	0.00%	0.00%	3.69%	100.00%
		20%	203	3.67%	2.13%	2.96%	2.46%	-20.00%
		100%	1015	0.77%	0.40%	0.59%	3.45%	82.86%
	R2301	80%	89787	0.00%	0.00%	0.00%	18.65%	100.00%
		20%	22447	3.23%	36.11%	10.94%	18.50%	40.86%
		100%	112234	0.64%	7.30%	2.19%	18.62%	88.24%
	R2514	80%	32827	0.00%	0.00%	0.00%	32.36%	100.00%
		20%	8207	22.00%	8.47%	14.09%	32.75%	56.99%
		100%	41034	4.43%	1.69%	2.82%	32.44%	91.32%
	U4007	80%	23203	0.00%	0.00%	0.00%	13.02%	100.00%
		20%	5801	0.99%	36.95%	6.36%	13.38%	52.45%
		100%	29004	0.20%	7.61%	1.28%	13.09%	90.23%
Group			**184262**	**1.06%**	**4.75%**	**2.17%**	**20.67%**	**89.51%**

From the results as listed in above table, it can be observed for all FW projects that RF models perform better than the corresponding logistic regression models. The total improvement for each project for all 100% of all the datasets is around 80% ~ 90%.

Table 4. Results comparison between RF and Logit model for RCP projects

				Random Forest			Logistic Regression	Comparison
Type	**Project**	**Data**	**Total Records**	**0-1 Error Rate**	**1-0 Error Rate**	**Total Error Rate**	**Total Error Rate**	**Total Improvement**
RCP	R2554	80%	31469	0.00%	0.00%	0.00%	12.29%	100.00%
		20%	7868	2.33%	6.41%	3.69%	11.64%	68.34%
		100%	39337	0.47%	1.29%	0.74%	12.16%	93.91%
	B3654	80%	626	0.00%	0.00%	0.00%	7.19%	100.00%
		20%	157	17.39%	1.49%	3.82%	3.82%	0.00%
		100%	783	2.56%	0.32%	0.77%	6.51%	88.24%
	R2823	80%	37094	0.00%	0.00%	0.00%	8.68%	100.00%
		20%	9274	5.35%	1.34%	2.89%	8.65%	66.58%
		100%	46368	1.06%	0.27%	0.58%	8.67%	93.31%
	wy	80%	14594	0.00%	0.00%	0.00%	11.04%	100.00%
		20%	3649	3.68%	6.88%	4.66%	10.06%	53.68%
		100%	18243	0.73%	1.39%	0.93%	10.84%	91.41%
Group			**104731**	**0.72%**	**0.68%**	**0.70%**	**10.34%**	**93.21%**

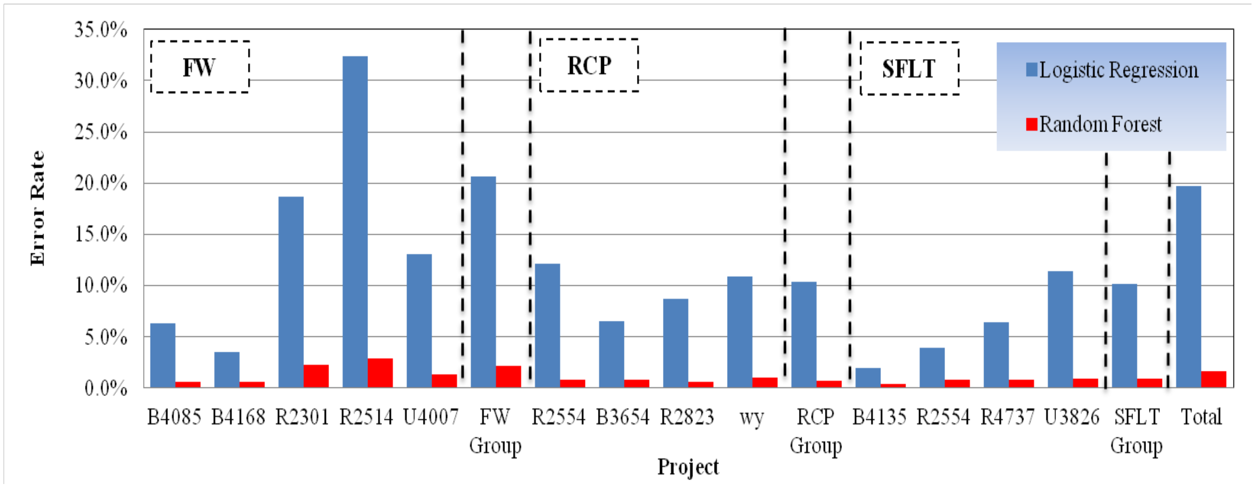
Table 5. Results comparison between RF and Logit model for SFLT projects

				Random Forest			**Logistic Regression**	**Comparison**
Type	**Project**	**Data**	**Total Records**	**0-1 Error Rate**	**1-0 Error Rate**	**Total Error Rate**	**Total Error Rate**	**Total Improvement**
SFLT	B4135	80%	1769	0.00%	0.00%	0.00%	1.87%	100.00%
		20%	443	1.89%	2.11%	2.03%	2.26%	10.00%
		100%	2212	0.37%	0.43%	0.41%	1.94%	79.07%
	R2554	80%	416	0.00%	0.00%	0.00%	3.85%	100.00%
		20%	104	5.48%	0.00%	3.85%	3.85%	0.00%
		100%	520	1.15%	0.00%	0.77%	3.85%	80.00%
	R4737	80%	2112	0.00%	0.00%	0.00%	6.06%	100.00%
		20%	528	6.67%	1.98%	3.98%	7.95%	50.00%
		100%	2640	1.32%	0.40%	0.80%	6.44%	87.65%
	U3826	80%	20172	0.00%	0.00%	0.00%	11.45%	100.00%
		20%	5044	3.51%	5.69%	4.48%	11.30%	60.35%
		100%	25216	0.71%	1.13%	0.90%	11.42%	92.12%
Group			**30588**	**0.74%**	**0.97%**	**0.85%**	**10.17%**	**91.61%**

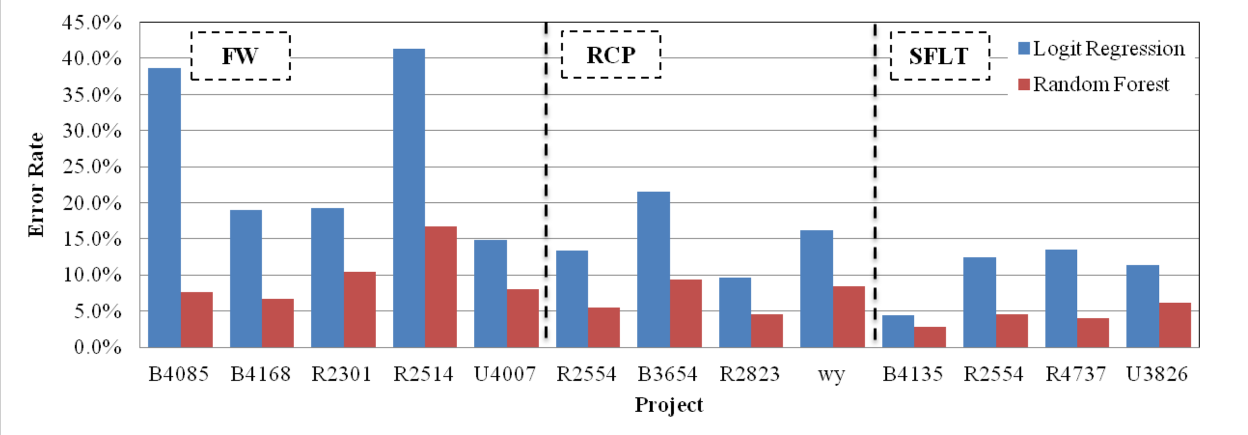
For all the SFLT projects, RF model outperforms the logistic regression model, and the total improvements comparing to the regression models are around 80%. Among the different groups, FW type has the most number of data records, while the RCP Group has the most improvement, 93.21%.

Figure 5 shows a summary of comparison between RF model and Logit model for each project area in three eco-regions, based on the prediction error rate as reported in Table 4.

Remark: Here, the validation data are from NCDOT, which are verified by expert field delineation. The validation is run on the RF models and Logit models for two groups of data, i.e., (i) to verify the Y values of the first group records data (preselected 80% of total data) which are used in the modeling process as the modeling error, and (ii) to verify the Y values of the second group records data (preselected 20% of total data) which are not used in the modeling process as the prediction error as shown in Fig. 5(b). Based on that, we also get the combination error rate on all total 100% data as shown in Fig. 5(a).



(a) Error rate on all data



(b) Error rate on test data

Figure 5. Summary of model comparison by error rate

2.4 Random Forest Smart version (RF-s)

According to the results, the prediction accuracy varies among RF models for different project areas and regions (see Figure 5). Some data noise may exist that affect the prediction in some sampling areas and the uncertainty of input data will significantly affect the quality of training model. Therefore, we have further conducted some experiments to improve the RF models as the PI proposes as a smart RF. The idea is to smartly select project area models for a combination model based on their individual model error rate and/or other information, thus to identify the most correct and efficient configuration. The overall scheme is shown in Figure 6. Currently, it is not included in our WAMAT tools. But it can be run by preselecting training area data as input data into our WAMAT for modeling.

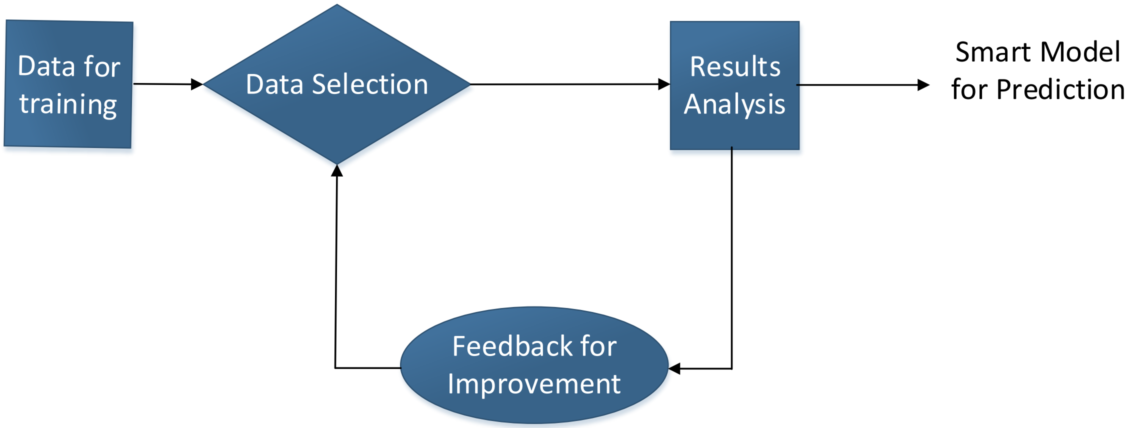


Figure 6. Scheme for RF improvements as Smart RF

Considering the scheme above, our models for wetland prediction in Lenoir County have the following setting:

(1) The configuration of RF is: to use 80% of the dataset as training data; to build 100 trees in the RF; to set the maxnodes index $= 7$ ($2^7 = 128$ maximum number).

(2) For the data selection, the smart RF model selects different sub-regions for its construction in three ecoregions as follows:

- RCP: *B3654, R2554, and R2823* (from B3654, R2554, R2823, wy)

- FW: *B4168 and R2514* (from B4085, B4168, R2301, R2514, U4007)

- SFLT: *B4135 and R2554* (from B4135, R2554, R4737, U3826)

The smart selection is logical and smart in view of Figure 5 (a) and area knowledge.

3. Automation Process

In this section, we summarize the automation process and the tools we developed [6, 7]. The detail of WAMAT (WAM Automation Tools) can be found in the Users' Guide as Appendix [B].

These tools can be flexibly organized as GIS workflows to implement several tasks related to wetland prediction. The main tasks of wetland prediction include data pre-processing, model training, predicting, wetland mapping, post-treatment processing, model performance evaluation, and wetland map display. The automation tools are developed based on ArcGIS 10.1. Users can easily assemble these tools together through graphic user interface (Model Builder) to automate the entire process, or in a piecemeal fashion. We have developed the following individual processes:

(1) Automation process for generating DEM derivatives and extracting values from multiple raster layers.

(2) Automation process for sampling data based on users' preferred strategy.

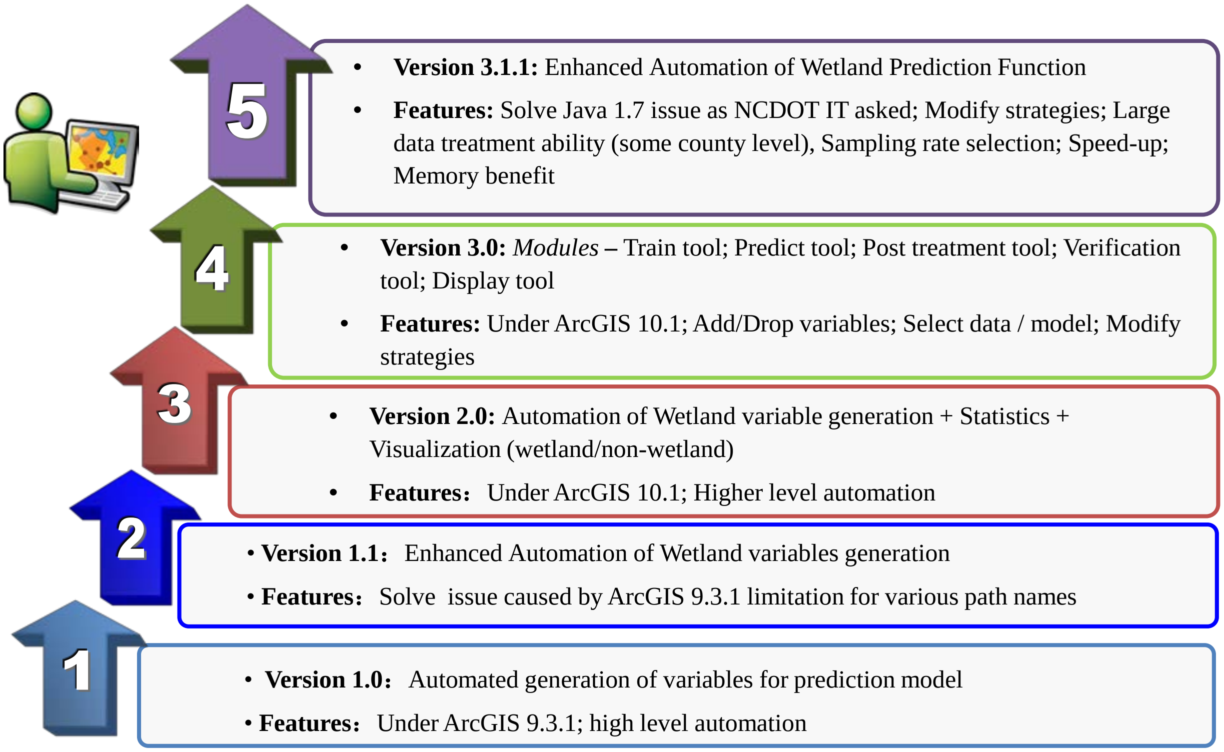
(3) Automation process for building model using training dataset.

(4) Automation process for wetland prediction by the selected prediction model generated from (3).

(5) Automation process for post-treatment on the prediction results with assistance of other auxiliary data (land use and/or river/water body).

(6) Automation process for accuracy evaluation of the prediction results and/or post-treatment results based on the evaluation of field data.

During the project period, we have provided NCDOT three new major versions of our automation process tools for wetland prediction [A, B, C]. It can be summarized as in following Figure 7.



- **Version 3.1.1:** Enhanced Automation of Wetland Prediction Function
- **Features:** Solve Java 1.7 issue as NCDOT IT asked; Modify strategies; Large data treatment ability (some county level), Sampling rate selection; Speed-up; Memory benefit

- **Version 3.0:** *Modules* **–** Train tool; Predict tool; Post treatment tool; Verification tool; Display tool
- **Features:** Under ArcGIS 10.1; Add/Drop variables; Select data / model; Modify strategies

- **Version 2.0:** Automation of Wetland variable generation + Statistics + Visualization (wetland/non-wetland)
- **Features：** Under ArcGIS 10.1; Higher level automation

- **Version 1.1：** Enhanced Automation of Wetland variables generation
- **Features：** Solve issue caused by ArcGIS 9.3.1 limitation for various path names

- **Version 1.0：** Automated generation of variables for prediction model
- **Features：** Under ArcGIS 9.3.1; high level automation

Figure 7. Automation Process

3.1 Overall architecture of the automation toolset

We organize the programs into the following sub toolsets: (1) *Train (Modeling)*, (2) *Predict*, (3) *Post-treatment*, (4) *Verification - Analysis*, and (5) *Display*.

Each tool is an independent component that can be applied individually, or logically combined for full automation. Users can connect WAMAT in ArcGIS Model Builder as other regular build-in tools. In addition, this project also provides an automation tool to generate wetland variables from the DEM data and combines them and other data (soil and land-cover) into a table for modeling and/or prediction. That automation tool is from our automation version 1. Of course, this function tool is included as one module in version 2 and version 3.

3.2. Display

The *Display* tool is to display the prediction results using a pre-defined color scheme. The output is a raster file of wetland prediction result, where 1 means wetland and 0 means non-wetland. In addition, the map generated by this software can highlight both correct and error prediction areas, as 1-1, 0-0, 1-0 error, and 0-1 error by four different colors respectively. Figure 8 and Figure 9 show the examples of displaying the prediction results.

(1) We use 4 different colors for the following cases, where 1 for wetland and 0 for non-wetland, $x - \hat{x}$ as "true – estimate".

1 – 1: Green color; 1 – 0 error: Red; 0 – 0: Grey; 0 – 1 error: Yellow.

(2) This dynamic delineation was presented to NCDOT during our annual progress meeting on 8-30-2013.

(3) Our WAMAT tools can automatically delineate and mark the wetland and non-wetland areas on ArcGIS map after the run of our WAMAT prediction tool.

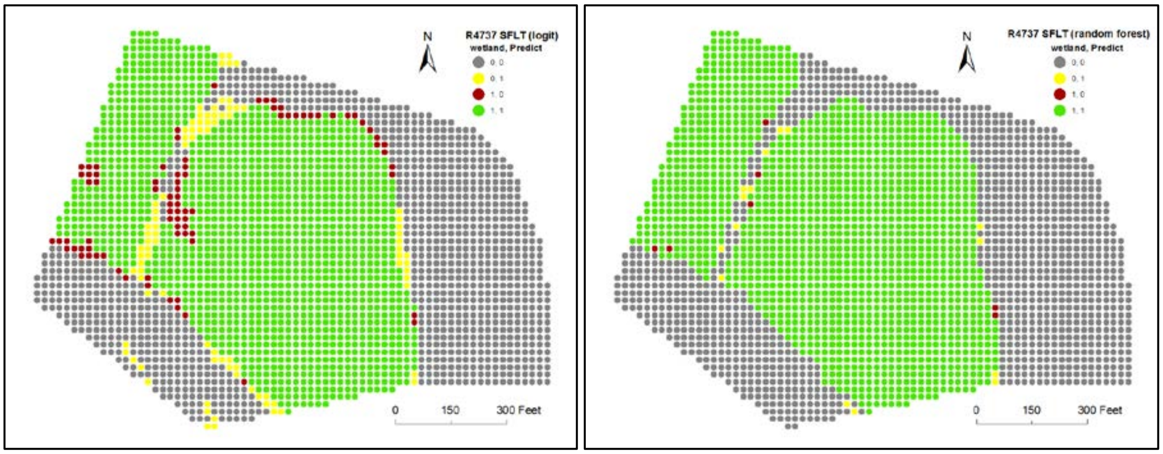


Figure 8. Display of wetland prediction result: (a) Logistic Regression (left), (b) RF (right)

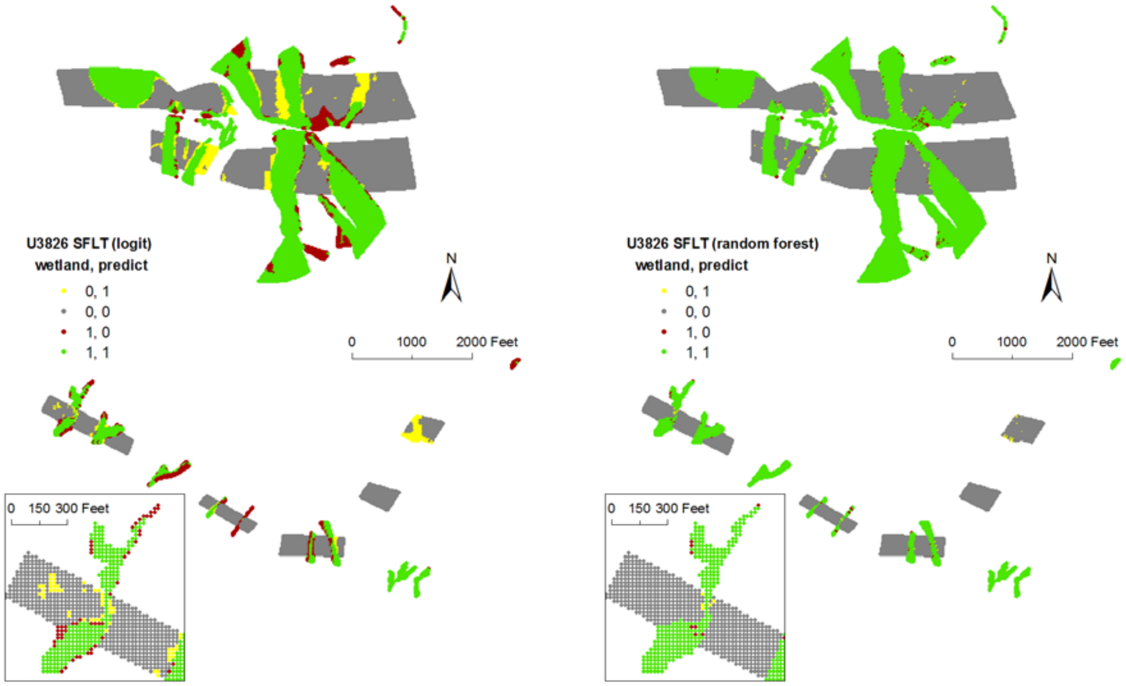


Figure 9. Prediction for Project U3826 (SFLT) (a) Logistic Regression (left), (b) RF (right)

3.3. Advantages of the software

There are some important features of this software of WAMAT.

(1) Flexible

- Users can use individual tools to conduct certain tasks, or they can also assemble these tools within the model builder to conduct more complex tasks.
- Another is the operation of add/drop explanatory variables for modeling.

(2) Efficient

The algorithm and software are both enhanced to be able to process large areas. We efficiently compressed the data for the algorithm, thus it can be quickly calculated.

(3) User friendly

The graphic interface is more straightforward and applicable. Users can easily change parameters and settings, such as linking different tools through “drag and drop” operations.

4. Case Study and Field Validation

This section describes two case studies and field visits. One is for Lenoir County, February 2014, and the other is for the Piedmont Region, early August 2014.

4.1 Wetland prediction of Lenoir County

We have implemented the automation process of wetland prediction for Lenoir County. Our prediction models are built by the sampled training data from 13 projects areas provided by NCDOT. The sample data included three eco-regions (RCP, FW and SFLT) within the 13 project areas, with only one project in Lenoir County. (Figure 10).

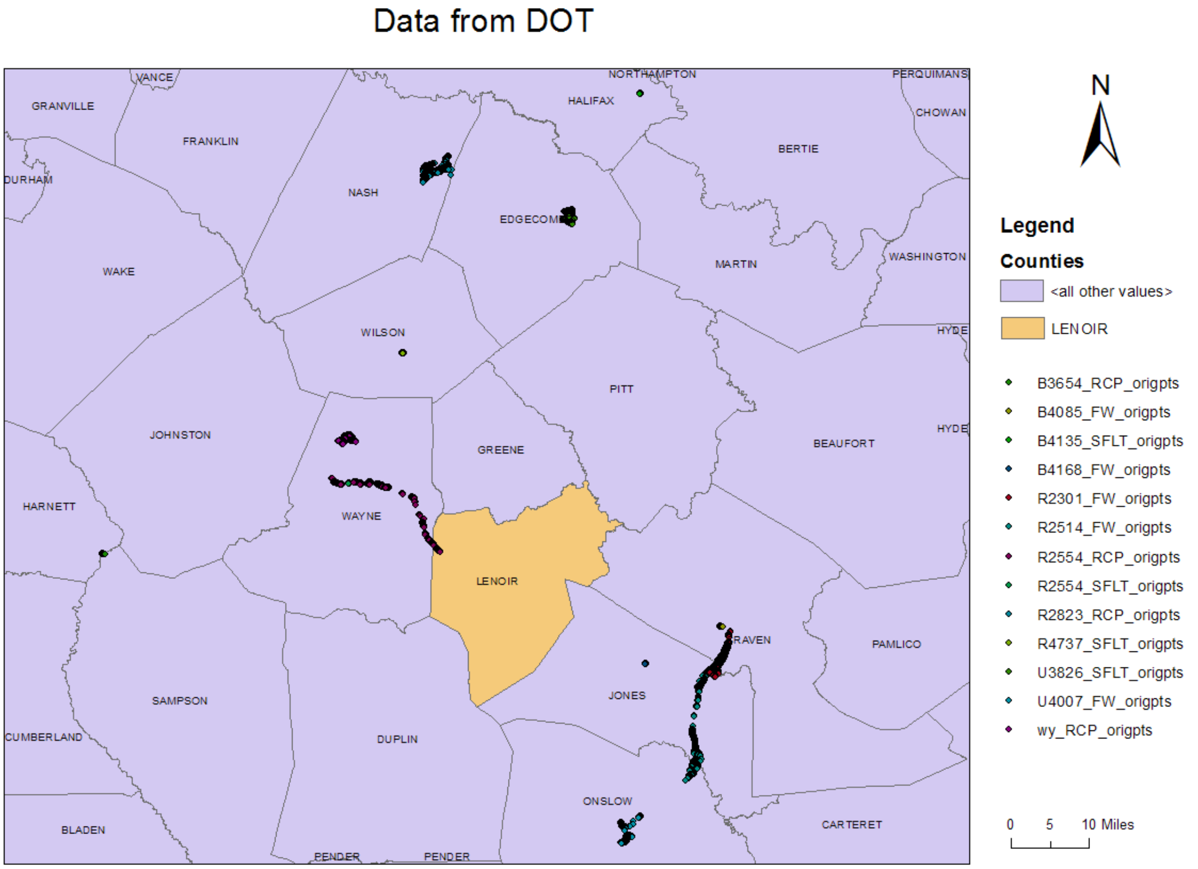


Figure 10. Sample data of 13 project areas

4.1.1 Model construction

We have run the following process of building our models and predicting wetlands as follows:

(1) To build Logistic Regression models for 3 eco-regions from the corresponding eco-regions in 13 sampling areas, respectively;

(2) To build RF models for 3 eco-regions from the corresponding eco-regions in 13 sampling areas, respectively;

(3) To predict wetland areas in Lenoir County using Logistic Regression models and RF models based on their 3 corresponding eco-regions.

(4) To do a statistical analysis of the prediction results with respect to the provided evaluation area data in Lenoir County.

4.1.2 Model validation

Using provided validation data in Lenoir County, we have evaluated the prediction results in Lenoir County, as shown in Table 6 and Table 7.

A smart modeling method has been proposed to both RF and Logit models to improve the prediction accuracy.

(1) We applied both standard and smart versions of the models.

- The extension –"a" means using all 13 sampling areas for modeling;
- The extension –"s" stands for smart use of modeling areas.

(2) We applied different strategies related to riparian data.

- The extension –1 stands for the method without riparian variable;
- The extension –2 is to use riparian variable for eco-region sub-division;

(3) We sampled different proportion of data for model training (100% or 80%).

Table 6. Performance comparison of Logit models for Lenoir County

		NCDOT model results		Logit-a-1 100%		Logit-a-2 100%		Logit-s-1 80%	
	Error Type	Pixels	Error rate	Pixels	Error rate	Pixels	Error rate	Pixels	Error rate
BR	1-1	2137	-	2204	-	2221	-	2132	
	1-0	642	23.10%	575	20.69%	558	20.08%	647	23.28%
	0-0	9113	-	9389	-	9074	-	9594	
	0-1	1086	10.65%	810	7.94%	1125	11.03%	605	5.93%
	Total		**13.31%**		**10.67%**		**12.97%**		**9.65%**
CF	1-1	7592	-	7133	-	12490	-	9387	
	1-0	4898	39.22%	5357	42.89%	0	0.00%	3103	24.84%
	0-0	761	-	792	-	0	-	554	
	0-1	50	6.17%	19	2.34%	811	100.00%	257	31.69%
	Total		**37.20%**		**40.42%**		**6.10%**		**25.26%**
Hornpipe	1-1	1538	-	1076	-	1468	-	954	
	1-0	234	13.21%	696	39.28%	304	17.16%	818	46.16%
	0-0	377	-	430	-	393	-	437	
	0-1	69	15.47%	16	3.59%	53	11.88%	9	2.02%
	Total		**13.66%**		**32.10%**		**16.10%**		**37.29%**
R2719	1-1	10383	-	10264	-	10505	-	9001	
	1-0	2826	21.39%	2945	22.30%	2704	20.47%	4208	31.86%
	0-0	17674	-	17672	-	17484	-	19050	
	0-1	5832	24.81%	5834	24.82%	6022	25.62%	4456	18.96%
	Total		**23.58%**		**23.91%**		**23.77%**		**23.60%**
SFLT	1-1	11468	-	10452	-	10493	-	10034	
	1-0	802	6.54%	1818	14.82%	1777	14.48%	2236	18.22%
	0-0	1012	-	1124	-	1089	-	961	
	0-1	257	20.25%	145	11.43%	180	14.18%	308	24.27%
	Total		**7.82%**		**14.50%**		**14.45%**		**18.79%**
Total Sample Areas	1-1	33118	-	31129	-	37177	-	31508	
	1-0	9402	22.11%	11391	26.79%	5343	12.57%	11012	25.90%
	0-0	28937	-	29407	-	28040	-	30596	
	0-1	7294	20.13%	6824	18.83%	8191	22.61%	5635	15.55%
	Total		**21.20%**		**23.13%**		**17.19%**		**21.14%**

Table 7. Performance comparison of RF models for Lenoir County

		NCDOT model results		**Logit-s-2 80%**		**RF-s-1 80%**		**RF-s-2 80%**	
	Error Type	Pixels	Error rate	Pixels	Error rate	Pixels	Error rate	Pixels	Error rate
BR	**1-1**	2137	-	2087		1182		1312	
	1-0	642	23.10%	692	24.90%	1597	57.47%	1467	52.79%
	0-0	9113	-	9497		9513		9039	
	0-1	1086	10.65%	702	6.88%	686	6.73%	1160	11.37%
	Total	12978	**13.31%**		**10.74%**		**17.59%**		**20.24%**
CF	**1-1**	7592	-	12490		10531		10224	
	1-0	4898	39.22%	0	0.00%	1959	15.68%	2266	18.14%
	0-0	761	-	0		522		564	
	0-1	50	6.17%	811	100.00%	289	35.64%	247	30.46%
	Total		**37.20%**		**6.10%**		**16.90%**		**18.89%**
Hornpipe	**1-1**	1538	-	1318		1533		1716	
	1-0	234	13.21%	454	25.62%	239	13.49%	56	3.16%
	0-0	377	-	415		397		311	
	0-1	69	15.47%	31	6.95%	49	10.99%	135	30.27%
	Total		**13.66%**		**21.87%**		**12.98%**		**8.61%**
R2719	**1-1**	10383	-	9211		7603		7818	
	1-0	2826	21.39%	3998	30.27%	5606	42.44%	5391	40.81%
	0-0	17674	-	18892		19641		18792	
	0-1	5832	24.81%	4614	19.63%	3865	16.44%	4714	20.05%
	Total		**23.58%**		**23.46%**		**25.80%**		**27.52%**
SFLT	**1-1**	11468	-	9946		10160		10149	
	1-0	802	6.54%	2324	18.94%	2110	17.20%	2121	17.29%
	0-0	1012	-	1039		1024		1048	
	0-1	257	20.25%	230	18.12%	245	19.31%	221	17.42%
	Total		**7.82%**		**18.86%**		**17.39%**		**17.30%**
Total Sample Areas	**1-1**	33118	-	35052		31009		31219	
	1-0	9402	22.11%	7468	17.56%	11511	27.07%	11301	26.58%
	0-0	28937	-	29843		31097		29754	
	0-1	7294	20.13%	6388	17.63%	5134	14.17%	6477	17.88%
	Total		**21.20%**		**17.59%**		**21.14%**		**22.57%**

A summary for wetland prediction of three eco-regions in Lenoir County shows that by comparing NCDOT model, Logit-s and RF-s respectively have an improvement of 34.76% and 4.54% in total RCP (BR+Hornpip), FW (CF) and SFLT in Lenoir evaluation areas. Here we do not include a region R2719 because it is across 2 eco-regions.

4.2 Wetland prediction of Southern Outer Piedmont

The second study region is Southern Outer Piedmont region, which includes ten counties. We use regular Logistic Regression model and RF model to predict wetland distribution of this region. And two models are conducted based on the training datasets.

Table 8 shows the prediction results. According to the results, the overall prediction accuracy is higher for models built based on the RF method than the Logit models, but this current RF model greatly under-predicts wetland occurrence. From the theoretical analysis, the RF should have potential ability to dramatically reduce this under-prediction rate for the modeling data. Currently, the PI has guided the team to change the RF model to be built from R library, not from the Python library, and with ability to set adjustable parameters such that we may further improve the whole prediction accuracy in near future.

Notes:

(1) Logit: Logistic Regression model constructed based on the training dataset of each individual county.

(2) RF: Random Forest model constructed based on the training dataset of each individual county.

Table 8. Performance comparison of models for Southern Outer Piedmont region

		Logit		**RF** **N = 100**	
County	**Type**	**Pixels**	**Error rate**	**Pixels**	**Error rate**
Cabarrus	**1--1**	457		337	
	1--0	45	8.96%	165	32.87%
	0--0	73764		86041	
	0--1	12396	14.39%	119	0.14%
	Total	**86662**	**14.36%**	**86662**	**0.38%**
Catawba	**1--1**	491		457	
	1--0	92	15.78%	235	40.31%
	0--0	21573		23187	
	0--1	1864	7.95%	250	1.07%
	Total	**24020**	**8.14%**	**24020**	**2.02%**
Cleveland	**1--1**	2047		1608	
	1--0	193	8.62%	632	28.21%
	0--0	182669		213226	
	0--1	31145	14.57%	588	0.28%
	Total	**216054**	**14.50%**	**216054**	**0.56%**
Davidson	**1--1**	965		718	
	1--0	133	12.11%	380	34.61%
	0--0	97134		116710	
	0--1	20060	17.12%	484	0.41%
	Total	**118292**	**17.07%**	**118292**	**0.73%**
Davie	**1--1**	4914		4462	
	1--0	914	15.68%	1366	23.44%
	0--0	138531		173164	
	0--1	36679	20.93%	2046	1.17%
	Total	**181038**	**20.77%**	**181038**	**1.88%**

Table 8 (b): Performance comparison of models for Southern Outer Piedmont region

		Logit		**RF** **N = 100**	
County	**Type**	**Pixels**	**Error rate**	**Pixels**	**Error rate**
Forsyth	**1--1**				
	1--0	142	9.17%	533	34.43%
	0--0	108792		133078	
	0--1	24901	18.63%	615	0.46%
	Total	**135241**	**18.52%**	**135241**	**0.85%**
Gaston	**1--1**	485		408	
	1--0	57	10.52%	134	24.72%
	0--0	28890		35469	
	0--1	6667	18.75%	88	0.25%
	Total	**36099**	**18.63%**	**36099**	**0.61%**
Iredell	**1--1**	2621		2412	
	1--0	238	8.32%	447	15.63%
	0--0	38275		45070	
	0--1	7728	16.80%	933	2.03%
	Total	**48862**	**16.30%**	**48862**	**2.82%**
Lincoln	**1--1**	3430		2926	
	1--0	217	5.95%	721	19.77%
	0--0	81154		91517	
	0--1	11483	12.40%	1120	1.21%
	Total	**96284**	**12.15%**	**96284**	**1.91%**
Rowan	**1--1**	**1451**		**1341**	
	1--0	313	17.74%	423	23.98%
	0--0	31782		38622	
	0--1	7249	18.57%	409	1.05%
	Total	**40795**	**18.54%**	**40795**	**2.04%**

4.3 Field validation

With an expert Sandy Smith at Axiom Environmental, our team executed a validation trip to Lenoir County on February 7th, 2014 and Piedmont Region (10 Counties) on August 8th – 9th, 2014.

The goals of the trip in Lenoir are: (1) to validate two automated wetland identification digital maps generated using Logistic Regression (Logit) and Random Forest (RF) regression models; (2) to differentiate the wetland types (riparian vs. non-riparian).

The goals of the Piedmont visit are to check two kinds of models and make comparison among them: Logit and RF for newly selected areas which are not used in the training process.

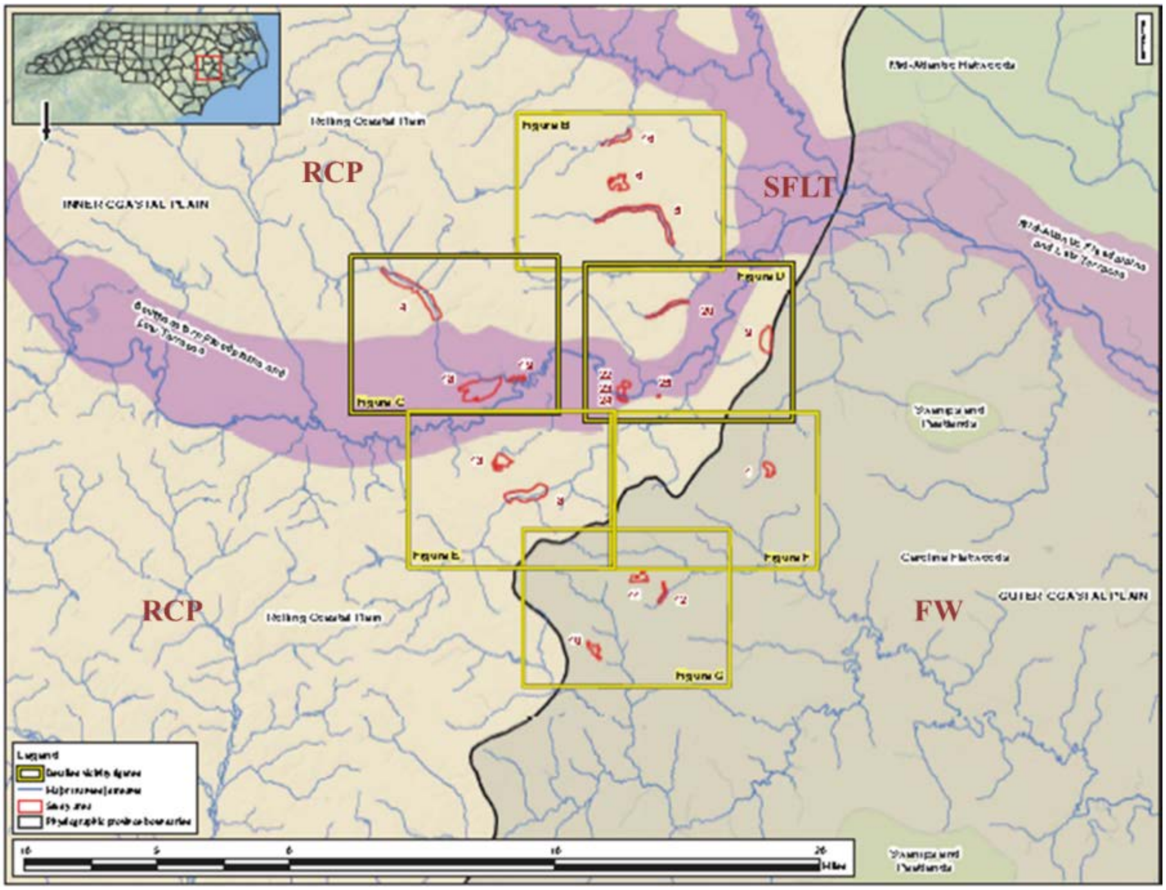
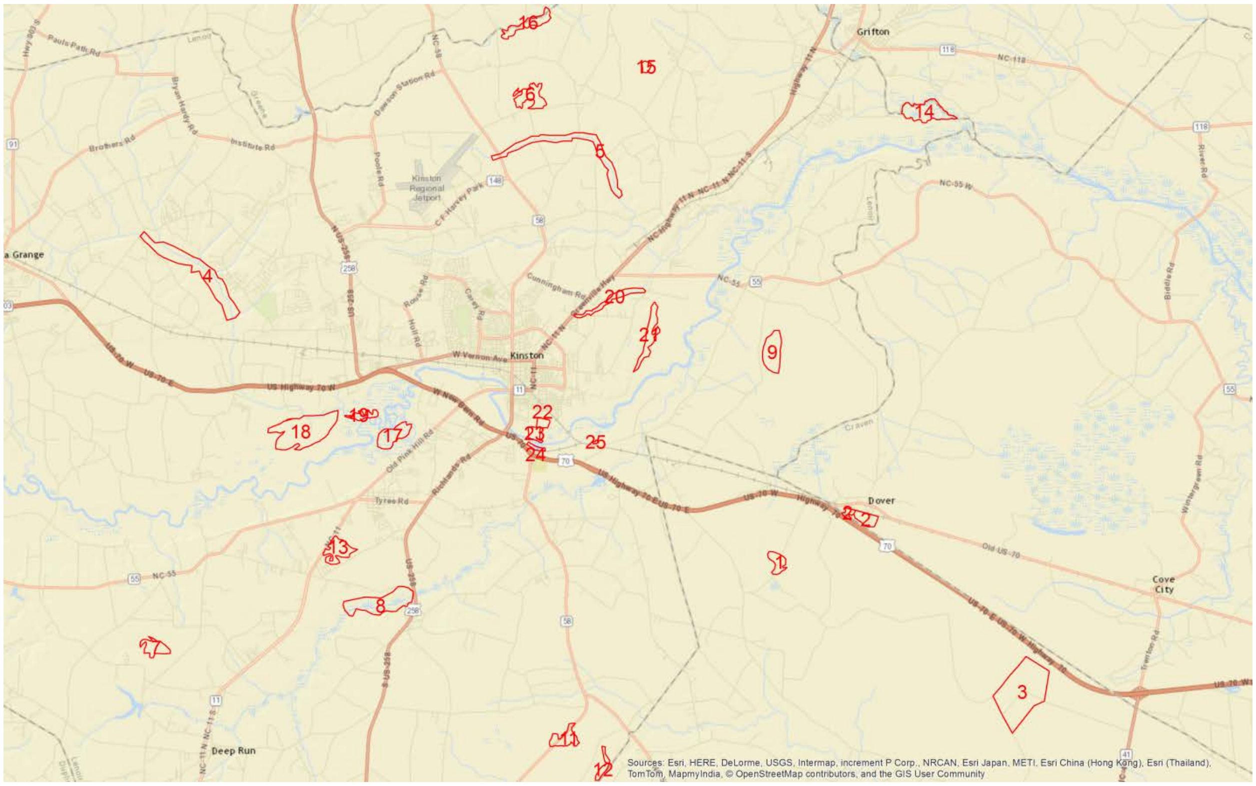


Figure 11. Study area of field trip for Lenoir County

In Lenoir, we selected out twenty five locations (see Figure 11) and were able to visit sixteen (ten on the ground and six drive-by) of them. There are sites where observations indicate challenging differentiation between wetland and non-wetland, e.g., locations 13 and 24 labeled as “uncertain”. The results can be found in our previous report. Also refer to the detail comments and summary from the Axiom Environmental report as submitted.

It is important to point out that the training data are almost totally out of Lenoir County. From the statistics view point, it is suggested that the best way to develop future models is to have sample data from Lenoir County only for training the models. Although the model training data are from the same ecoregions including Lenoir County, our used training data from these samples are majority out of Lenoir County and may not include all wetlands features in Lenoir County.

For Piedmont region, the sample locations are in 10 counties. See Figure 12. The detail maps of each location were submitted separately in a Visiting Report of Piedmont Region. A summary figure and summary table are as follows.



Sources: Esri, HERE, DeLorme, USGS, Intermap, increment P Corp., NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, MapmyIndia, © OpenStreetMap contributors, and the GIS User Community

Figure 12. Study area of field trip for Southern Outer Piedmont region

Table 9. Summary validation results in Southern Outer Piedmont Region

County	Coordinate	Riparian	Logit	RF	Wetland	REMARK
Cabarrus						
Location 1	35°14'10"N, 80°35'47"W	Yes	Yes	No	No	Pond
Location 2	35°14'07"N, 80°35'39"W	Yes	Yes	No	No	
Location 3	35°14'10"N, 80°35'34"W	Yes	No	No	No	
Catawba						
Location 1	35°36'28"N, 80°58'40"W	Yes	Yes	No	No	on boundary maybe wet
Location 2	35°36'32"N, 80°58'40"W	Yes	Yes	No	No	on boundary maybe wet
Cleveland						
Location 1	35°14'44"N, 81°34'37"W	Yes	No	Yes	No	
Location 2	35°14'45"N, 81°34'27"W	Yes	No	No	No	
Location 3	35°14'37"N, 81°34'26"W	Yes	No	Yes	No	
Gaston						
Location 1	35°17'40"N, 81°12'05"W	Yes	Yes	No	yes	
Location 2	35°14'52"N, 81°12'03"W	Yes	Yes	No	Yes	
Location 3	35°14'46"N, 81°12'19"W	Yes	Yes	No	No	
Lincoln						
Location 1	35°26'53"N, 81°16'13"W	Yes	Partly Yes/No	Partly Yes/No	Yes	
Location 2	35°27'08"N, 81°16'19"W	Yes	Partly Yes/No	Partly Yes/No	No	

County	Coordinate	Riparian	Logit	RF	Wetland	REMARK
Davidson						
Location 1	35°44'21"N, 80°21'44"W	Yes	Yes	No	No	
Location 2	35°44'12"N, 80°21'52"W	Yes	No	No	No	
Location 3	35°44'12"N, 80°23'13"W	Yes	Yes	No	Yes	Beaver Dam Found
Location 4	35°44'19"N, 80°23'28"W	Yes	No	No	Partly Wet	Caused by Beaver Dam, Flood area
Davie						
Location 1	35°57'07"N, 80°32'07"W	Yes	Yes	No	No	
Location 2	35°56'58"N, 80°32'10"W	Yes	No	No	No	
Location 3	35°56'57"N, 80°32'12"W	Yes	Yes	No		Not visit
Forsyth						
Location 1	36°00'53"N, 80°24'53"W	Yes	No	No	No	
Location 2	36°00'46"N, 80°24'51"W	Yes	Yes	No	Yes	
Location 3	36°00'42"N, 80°24'49"W	Yes	Yes	No	No	
Iredell						
Location 1	35°51'40"N, 80°42'45"W	Yes	Yes	Yes	Yes	
Location 2	35°51'36"N, 80°43'.03"W	Yes	No	No	No	
Location 3	35°51'32"N, 80°42'33"W	Yes	Yes	No	No	
Rowan						
Location 1	35°31'16"N, 80°34'43"W	Yes	Yes	No	No	
Location 2	35°31'06"N, 80°34'26"W	Yes	Yes	No	No	
Valid #	Clear count # = 24		13 (54.17%)	18 (75.00%)		Not considering “partly”
	Count # =27		16 (59.26%)	21 (77.78%)		Considering “partly”

In Table 9, we select 28 locations for a comparison of our two modeling methods, i.e., Logit and RF. They all had no verified wetland values. In these 28 locations there are 19 locations where the resultant predictions from these two methods are different. Thus, Axiom expert helps the team to run a field visit for the wetland verification of these locations. However, one location is not visited because it is a special restrict area, making 27 visiting locations. It is noticed that 2 locations have partly wet-predictions from both methods, and one location has a partly wet verification from the field visit. Thus, we make two summary lines: one is for the 24 clear count locations, and another is for the 27 count locations.

It should be pointed out that in both Lenoir County and Piedmont regions all the above models are without post-treatment. That is a new task that the PI suggests for future to improve the models further by using post-treatment on lakes and rivers as well as some kind of land cover. However, the automation tool for it has been developed recently.

5. Auxiliary variable and data

Besides the variable data that can be derived from the DEM data, we use some variables that cannot be simply derived by DEM data, such as the riparian variable, land cover data and soil data. In this project, they serve as ancillary data to provide important information to identify wetlands and filter information.

5.1 Riparian variable

Since the original regression model considers riparian as a predictor variable, we applied two strategies based on how to utilize this variable. One is to use riparian variable as a normal predictor variable. The second way is to use riparian variable to further divide the region into two sub-regions, where one is a riparian sub-region, and another one is a non-riparian sub-region. We use Lenoir County as an example where originally it has three eco-regions. Now, by using riparian variable data, we have a total six sub-regions. The general performance of this riparian division plus smart models are better overall than regular models for Logit, but it is not obvious for RF, see Table 6 and Table 7.

Remark:

- The strategy “1” (with extension -1) means that we do not utilize riparian variable.
- The strategy “2” (with extension -2) means that we use the riparian variable to further divide the data into two categories (sub-regions) for the model training and prediction.
- For the extension, “-s” means that the training area data are smartly selected to train both Logit and RF model, in contrast to “-a”, which means no selection has been applied in the model. Selection rule is mainly based on the individual training model errors.

We observe the results for the RF models in Lenoir County where the training data are from 13 training project areas and 12 of 13 are out of Lenoir. The experiment results show that the smart models (-s) is better than non-smart model (-a). We also see that using riparian variable does not provide further benefit to RF models. On the other hand, from above Table 6 and Table 7, we see that using riparian variable to further divide regions makes the Logit model (-2) better. For using riparian variable as a normal predictor variable, we did not see any enhancement to model accuracy from experiments and some related papers.

5.2 Soil data

In wetland areas, soil has changed to adapt to the saturated environment. Therefore, with study of soil data, we are able to help prediction of the probability for an area to be wetland. Along this research direction, we study the soil tabular data and analyze our soil data downloaded from the web-link suggested by NCDOT. We find that we need to further study this important soil variable and utilize it in future.

5.3 Land cover

Land cover is another very useful variable for wetland prediction. We use it as a regular predictor variable for modeling and prediction. Furthermore, we use it for post-treatment in the wetland prediction to remove areas or spots which are surely not wetland as indicated by the land cover variable classification.

Land cover data include six categories in Lenoir: “Agricultural Vegetation”, “Developed & Other Human Use”, “Forest &Woodland”, “Open Water”, “Recently Disturbed or Modified”, and “Shrubland & Grassland”. We reclassify these six categories to “0” for post-treatment.

6. Valuable additional methodology

6.1. SCAD

We have performed a SCAD (Smoothly Clipped Absolute Deviation penalty) analysis to identify key effective variables and to know how to improve the modeling method. By using SCAD analysis on the logistic regression model with 16 variables, we find that the following 6 variables are important: batwi, dem, land cover, soil, slope_brdem, and weir. This key variable identification for the logit regression is independent from the WAMAT automation tools.

The SCAD method is designed to scientifically find a minimal set of variables with almost the same accuracy of the model with the full set of variables. It is noticed that the Nagelkerke's R^2 is only slightly reduced by 0.007 from 0.517 to 0.510, but the variable number is reduced from 16 to 6. It clearly shows the power of SCAD method for model method improvement. For the SCAD method, we referred to Fan and Li [11]. Recently, Yang, Wang and Bao also applied it with a stable algorithm to AADT estimation [9]. For mathematical background of SCAD, please see [11] or [9]. A good penalty function should merit the following three properties, e.g., unbiasedness for the large true unknown estimator, sparsity that can set small estimator to be zero automatically, and continuity of the resulting estimator to avoid instability in model prediction [11, 9].

6.2 Balance treatment of input data

Because the Logit method normally generates models favor to identify the class (type) which has majority training data in the whole training dataset, thus we discuss and develop some technique to solve this problem, i.e., to reduce the error rate for identifying the class (type) which has minority training data in the whole training dataset.

In this project, we have also developed a group of techniques to improve the wetland estimation. They are pre-treatment and post-treatment to study the solvability to the following problems:

(1) How to sample when the project area for model training has majority points in non-wetland (or wetland) and minority points in wetland (or non-wetland)? It means that we have biased data in one type over another. It is a training problem. For that, we propose a pre-treatment of balanced sampling technique to study the possible effect.

(2) How to deal with the regression results that have some non-regulated mistakes/errors, e.g., very small holes and strange boundary? We propose a post-treatment technique to deal with the regression results in geographical shape. That is a Dilation–Erosion post-processing to smooth the wetland areas in order to remove noises in the data and/or errors in regression prediction.

In this sub-section we describe a balance treatment when the input data for model training is unbalanced between wetland and non-wetland sample data, which leads the regression result to yield to the majority type. For example, because the number of non-wetland samples is much greater than the number of the wetland samples in project area U4007_FW, the nominal logistic regression leads to a 1-0 error (missing wetland error) 78.9374%, and a 0-1 error (mismarked wetland error) 1.9445%. To balance these two-type errors and evenly to reduce the total error rate, we propose two sampling methods for the modeling.

(1) To Enhance Minority Sampling. This first method as in Figure 13 is to randomly sample the minority type data repeatedly with replacement (i.e., one observation of minority type may be used more than once) such that its total sampling number is close to the number of the majority type samples. Thus, the model training data set contains three parts: one from majority type, one from minority, and one from randomly repeat minority samples. By using this sampling technique, we have run the logistic regression and help to reduce the error rate.

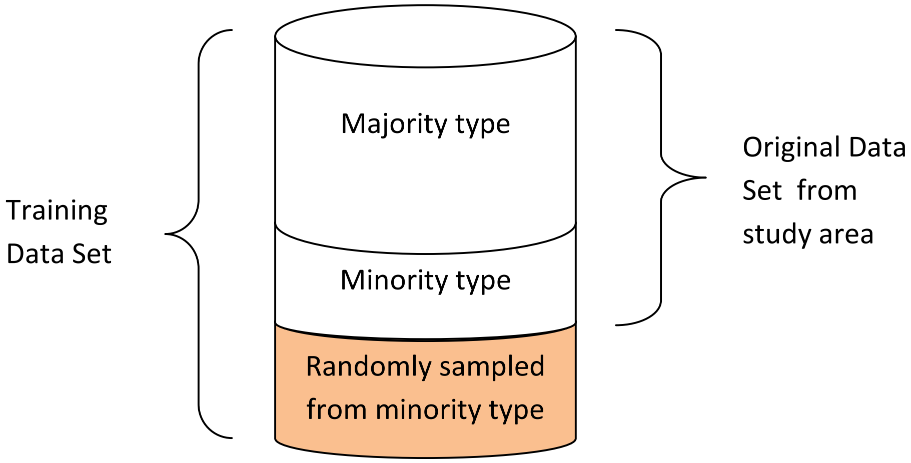


Figure 13. Sample to expand the minority type

(2) To Lessen Majority Sampling. This second method is to randomly sample a smaller proportion of the majority type data, instead of using all samples of this majority type data, such that the sample number in the majority type is close to the sample number in the minority type.

The theoretical analysis and experimental results show that the first method to enhance minority sampling is better than the second method to lessen majority sampling. Here Table 10 lists our recent test results on enhancing minority sampling by balancing function in our WAMAT. The test county is Rowan County.

Table 10. Comparison of balancing sampling function for Logit method

County		Logit with balancing sampling		Logit without balancing sampling	
Rowan	**1--1**	**1451**		**1426**	
	1--0	313	17.74%	338	19.16%
	0--0	31782		32011	
	0--1	7249	18.57%	7020	17.99%
	Total	**40795**	**18.54%**	**40795**	**18.04%**

Table 11. Comparison of balancing sampling function for RF method

County		RF with balancing sampling		RF without balancing sampling	
Rowan	**1--1**	**1341**		**1347**	
	1--0	423	23.98%	417	23.64%
	0--0	38622		38608	
	0--1	409	1.05%	423	1.08%
	Total	**40795**	**2.04%**	**40795**	**2.06%**

7. Conclusion

This project mainly focuses on two major objectives: to improve the prediction model and to develop automation process. As for the model improvement, we have conducted literature review to select and develop modeling methods that are suitable for our prediction problem. According to the results, we summarize as follows:

(1) We have successfully completed this important project for the NCDOT needs and for wetland protection requirement.

(2) We have developed wetland prediction automation tool, WAMAT, as a deliverable product for NCDOT to use internally. Users' Guide of WAMAT is also provided. Our WAMAT (patent pending) is easy to install and user-friendly to use with a full process automation and/or a module process automation as user's choice.

(3) Two systematic models are presented and developed with the automation. They are logistic regression model and Random Forest (RF) model. Their corresponding smart version models are also developed.

(4) The models with automation have been applied to predict wetland for Lenoir County and Piedmont Regions (10 Counties). The resultant data and digital maps have been delivered to NCDOT.

(5) Two main field visits to Lenoir County and Piedmont Regions (10 Counties) have been conducted with Axiom Environmental support. Our prediction results are mainly based on the terrain data with soil and land cover data, which may change over time, especially due to human activities.

(6) The deliverable product includes:

  (i) WAMAT tool,

  (ii) WAMAT Users' Guide,

  (iii) Systematic Logit model and RF model for wetland prediction in automation tools,

  (iv) Digital wetland maps from the above models for Lenoir County and Southern Outer Piedmont region (10 Counties).

(7) A further research and study in this important research area and direction is needed to advance our developed system and NCDOT's excellent WAM work to continue leading in the Nation.

8. References

A. General

[1] FHWA, *2011 Environmental Excellence Awards*, GIS-based Wetland and Stream Predictive Models – For Excellence in Environmental Research, http://environment.fhwa.dot.gov/eea2011/environment_research.htm, 2011.

[2] National Environmental Policy Act (NEPA), *US Environmental Law*, 1970.

[3] N.C. Wetland Functional Assessment Team, *Wetland Assessment Method (NC WAM) User Manual*, version 4, Oct. 2010.

[4] Morgan Weatherford and Phil Harris, "Improvements to NCDOT's Wetland Prediction Model," *Call for New Research Needs*, 3115, NC DOT, 2011.

[5] Sheng-Guo Wang, "Project Proposal for Improvements to NCDOT's Wetland Prediction Model," *NCDOT RP 2013-13*, 2012.

[6] Sheng-Guo Wang, Libin Bai, Jing Deng, Meijuan Jia, Morgan Weatherford, LeiLani Paugh, Wenwu Tang, Mingzhi Chen and Shen-En Chen, "Automation Process Method of Generating Wetland Predictive Variables", Invention Documents, UNC Charlotte, April 18, 2014. (USPTO 62/003,869, 05/28/2014)

[7] Sheng-Guo Wang, Libin Bai, LeiLani Paugh and Morgan Weatherford, "Automation Process Method of Wetland Modeling and Prediction", Invention Documents, UNC Charlotte, April 18, 2014. (USPTO 62/003,887, 05/28/2014)

[8] S.-G. Wang and B. Wang, "Modeling of Distributed RLC Interconnect and Transmission Line via Closed Forms and Recursive Algorithms," *IEEE Trans. VLSI*, 18(1), pp.119-130, 2010. (regular paper) (This research was supported in part by the NSF)

[9] B. Yang, S.-G. Wang and Y. Bao, "New Efficient Regression Method for Local AADT Estimation via SCAD Variable Selection", *IEEE Transactions on Intelligent Transportation Systems*, vol.15, no.6, pp.2726-2731, 2014.

[10] Z. Cai, J. Fan and R. Li, "Efficient estimation and inferences for varying-coefficient models," *Journal of American Statistical Association*, 95, pp.888-902, 2000.

[11] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oralce

properties,"J. Am. Statist.Ass.,vol. 96, pp.1348-1360, 2001.

[12] Minnesota Geospatial Information Office, "How is LiDAR Data Used to Protect Water Quality in Minnesota?" www.mngeo.state.mn.us/chouse/.../lidar_uses_waterquality.html.

[13] "An investigation into the role of Integrated LiDAR and Multiband Orthophotography in the production of Enhanced Forest Inventories in the Great Lakes St. Lawrence Forest," http://www.mrnf.gouv.qc.ca/ecanusa/documents/affiches/woods_pineau_courville.pdf.

[14] E. Kreyszig, *Advanced Engineering Mathematics*, 7th Ed., Wiley, N.Y. 1993.

B. Wetland and wetland inventory:

[15] EPA, U. S. "Section 404 of the Clean Water Act: how wetlands are defined and identified", Retrieved 0529, 2014, from http://water.epa.gov/type/wetlands/outreach/fact11.cfm

[16] J. Boyd, "Compensating for Wetland Losses under the Clean Water Act", *Environment: Science and Policy for Sustainable Development, 44*(9), 43-44, 2002.

[17] North Carolina Floodplain Mapping Program. Retrieved 0606, 2014, from http://www.ncfloodmaps.com/

[18] C. Castañeda & D. Ducrot, "Land cover mapping of wetland areas in an agricultural landscape using SAR and Landsat imagery", *Journal of Environmental Management, 90*(7), 2270-2277, 2009.

[19] C. Baker, R. Lawrence, C. Montagne, & D. Patten, "Mapping wetlands and riparian areas using Landsat ETM+ imagery and decision-tree-based models", *Wetlands*, *26*(2), 465-474, 2006.

[20] F. M. Henderson & A. J. Lewis, "Radar detection of wetland ecosystems: a review", *International Journal of Remote Sensing*, *29*(20), 5809-5835, 2008.

[21] L.-M. Rebelo, C. Finlayson, & N. Nagabhatla, "Remote sensing and GIS for wetland inventory, mapping and change analysis", *Journal of Environmental Management, 90*(7), 2144-2153, 2009.

[22] L. L. Hess, J. M. Melack, E. M. Novo, C. C. Barbosa & M. Gastil, "Dual-season mapping of wetland inundation and vegetation for the central Amazon basin", *Remote Sensing of Environment, 87*(4), 404-428, 2003.

[23] C. Wright & A. Gallant, "Improved wetland remote sensing in Yellowstone National Park using classification trees to combine TM imagery and ancillary environmental data", *Remote Sensing of Environment, 107*(4), 582-605, 2007.

[24] R. W. Tiner Jr, “Use of high-altitude aerial photography for inventorying forested wetlands in the United States”, *Forest Ecology and Management, 33*, 593-604, 1990.

C. Terrain derivatives for wetland prediction:

[25] A. Hogg & K. Todd, “Automated discrimination of upland and wetland using terrain derivatives. *Canadian Journal of Remote Sensing, 33*(S1), S68-S83, 2007.

[26] J. Li & W. Chen, “A rule-based method for mapping Canada's wetlands using optical, radar and DEM data”, *International Journal of Remote Sensing, 26*(22), 5051-5069, 2005.

D. Random Forest:

[27] L. Breiman, "Random forests," Machine learning, vol. 45, pp. 5-32, 2001.

[28] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," Machine learning, vol. 40, pp. 139-157, 2000.

[29] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 832-844, 1998.

[30] G. Biau, L. Devroye, and G. Lugosi, "Consistency of random forests and other averaging classifiers," Journal of Machine Learning Research, vol. 9, pp. 2015-2033, 2008.

[31] J. Peters, N. E. Verhoest, R. Samson, M. Van Meirvenne, L. Cockx, and B. De Baets, "Uncertainty propagation in vegetation distribution models based on ensemble classifiers," Ecological Modelling, vol. 220, pp. 791-804, 2009.

[32] J. Peters, B. D. Baets, N. E. Verhoest, R. Samson, S. Degroeve, P. D. Becker, et al., "Random forests as a tool for ecohydrological distribution modelling," Ecological Modelling, vol. 207, pp. 304-318, 2007.

[33] A. Verikas, A. Gelzinis, and M. Bacauskiene, "Mining data with random forests: A survey and results of new tests," Pattern Recognition, vol. 44, pp. 330-349, 2011.

[34] P.O. Gislason, J.A. Benediktsson, & J. R. Sveinsson, "Random forests for land cover classification," Pattern Recognition Letters, 27(4), 294-300, 2006.

[35] R. Lawrence, A. Bunn, S. Powell, & M. Zambon, "Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis," Remote Sensing of Environment, 90(3), 331-336, 2004.

[36] R. Lawrence, & A. Wright, "Rule-based classification systems using classification and regression tree (CART) analysis," Photogrammetric Engineering and Remote Sensing,

67(10), 1137-1142, 2001.

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

[38] E. Olhan, S. Gun, Y. Ataseven, & H. Arisoy, "Effects of agricultural activities in Seyfe Wetland," Scientific Research and Essays, 5(1), 9-14, 2010.

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

[40] M. Toner, & P. Keddy, "River hydrology and riparian wetlands: a predictive model for ecological assembly," Ecological applications, 7(1), 236-246, 1997.

[41] S.-G. Wang, J. Deng, M.-Z. Chen, M. Weatherford, and L. Paugh, "Random Forest Classification and Automation for Wetland Identification based on DEM Derivatives", 2015.

E. Soil:

[42] Scott Davis and Alexander P. (Sandy) Smith, Axiom Soils Tables and Notes, March 2014.

[43] J. Pomeroy et al., "Prairie Hydrological Model Study," Final Report, Center for Hydrology, University of Saskatchewan, http://www.usask.ca/hydrology/reports/CHRpt07_PHMS-Final-Report_Jan10.pdf, 2010.

9. Appendix

[A] WAMAT – WAM Automation Tools v.3.1.1, UNCC WAM Team, Patent Pending, 01-20-2015.

[B] WAM Automation Tools (WAMAT) – QUICK START GUIDE, Version 3.1.1, Sheng-Guo Wang (PI), Libin Bai and Sushmitha Yalla, 01-23-2015.

[C] Installation Guide for Computer with Old Version of WAMAT, UNCC WAM Team, 8-14-2014.

[D] Random Forest Classification and Automation for Wetland Identification based on DEM Derivatives, S.-G. Wang, J. Deng, M. Chen, M. Weatherford, and L. Paugh, 2015.