Wetland Prediction for Project Sites

Ben Shillington

Summary

This study builds on work completed by the North Carolina Department of Water Resources in their project *Automated Identification of Wetlands Using GIS in North Carolina*. In that project, they tested the accuracy of three models (Maximum Entropy, Overlay, and ESRI's Wetland Identification model) and their ability to predict wetlands throughout the state of North Carolina. They reported that the ESRI model failed to run, while the Maximum Entropy model produced the best results for predicting wetland areas.

The purpose of this study is to further compare additional machine learning models and their ability to accurately predict wetland areas, specifically near transportation project sites and roads. The models evaluated in this study include Random Forest, Linear Regression, and Gradient Boosting. The goal of this study is to compare how well each model predicts wetland presence, and to determine which model is most effective for use in project planning.

This work also addresses the fact that the National Wetlands Inventory is outdated and does not accurately reflect many current wetland areas. Improving wetland prediction models will help the South Carolina Department of Transportation identify potential wetland impacts earlier in the project development process and support better environmental planning and decision making.

Data

The dataset used to train and test the models consisted of 666,977 one-square-meter tiles, of which 111,158 were verified wetlands. Verified wetland data was collected from four South Carolina Department of Transportation (SCDOT) project sites located in Aiken, Dorchester, Sumter, and Ritter counties. Data for this project was stored in a geojson format

Digital Elevation Model (DEM) data was obtained from the U.S. Geological Survey (USGS) 2019 lidar survey of the Savannah Pee Dee region. This elevation data was used to calculate several topographic variables: Fill, Slope, Aspect, and Flow.

Additional environmental variables were sourced from Esri datasets. Soil Drainage Class was derived from the USA Soils Hydrologic Group dataset, and Average Annual Precipitation was obtained from the IMERG Precipitation dataset.

Data Sources

- **DEM Data:** USGS Savannah-Pee Dee 2019 Lidar Project
- **Soil Data:** Esri USA Soils Hydrologic Group ImageServer
- **Precipitation Data:** Esri IMERG Precipitation ImageServer

Variables

Short Name	Variable	Description	Source & Calculation Method
elv	Elevation	Raw digital elevation data	USGS DEM (2019)
fill	Filled Elevation	Depression-filled elevation data used to remove sinks	Derived from DEM using Sink geoprocessing tool, followed by raster math between filled and raw DEM
slope	Slope	Degree of surface incline	Calculated from DEM using slope geoprocessing tool
asp	Aspect	Direction of slope	Direction of maximum slope
flow	Flow	Flow accumulation	Calculated from DEM using D8 flow geoprocessing tool
soil	Soil Drainage Class	Classification of soil drainage capacity	Esri USA Soils Hydrologic Group dataset
precip	Average annual Rainfall	Mean annual precipitation (mm)	Esri Precipitation IMERG dataset
wet	wetland	Binary label (wetland / non-wetland)	SCDOT field recorded wetland files

Sample Data

{"type": "Feature", "geometry": {"type": "Point", "coordinates": [-81.60265067245076, 33.702653801788124]}, "properties": {"elv": 420.796875, "fill": 420.796875, "slope": 38.505069732666016, "asp": 154.13478088378906, "flow": 8.0, "soil": 3, "precip": 150.0, "wet": 0}}

Challenges

The most challenging aspect of this project was collecting and preparing the spatial data for modeling. While gathering soil type, precipitation, and wetland data was relatively straightforward, acquiring and processing the Digital Elevation Model (DEM) data proved significantly more complex. The DEM dataset comprised approximately 600 GB of high-resolution TIF files, which required the creation of a spatial index to reference elevation tiles relevant to the project areas.

Once the DEM data was organized, generating terrain variables (fill, slope, aspect, and flow) posed additional challenges. These calculations depend on the elevation values of neighboring cells to ensure accuracy, resulting in high computational demands and long processing times. Managing such large datasets required substantial memory and storage resources.

Initially, I aimed to compute all terrain variables for the entire state of South Carolina, storing the results in Zarr format for efficient access and scalability. While this approach worked on a per-tile basis, it failed to achieve the desired accuracy near tile boundaries, where edge cells lacked access to neighboring elevation data from adjacent tiles. I implemented a halo effect on the tiles to reduce this issue, but ultimately determined that processing data on a per-project basis provided the best balance between accuracy and computational feasibility.

Model Performance

Random Forest

F1 score: 0.8905780721265136

Classification Report:

	precision	**recall**	**f1-score**	**support**
0 (non-wet)	0.98	0.98	0.98	555819
1 (wet)	0.90	0.99	0.89	111158
accuracy			0.96	666977
Macro avg	0.94	0.93	0.93	666977
Weighted avg	0.96	0.96	0.96	666977

Confusion Matrix:

Actual/Predicted	**0**	**1**
0	544524	11295
1	12860	98298

Feature Importance:

Feature	**Importance**
elv	0.322254
fill	0.292179
slope	0.092807
precip	0.091484
asp	0.054114
flow	0.053484
soil_1	0.021105
soil_5	0.020025
soil_7	0.014841
soil_3	0.014241
soil_6	0.007801
soil_15	0.007781
soil_2	0.006624
soil_4	0.001263
soil_0	0.000000

Linear Regression

F1 score: 0.5195295174465935

Classification Report:

	precision	**recall**	**f1-score**	**support**
0 (non-wet)	0.96	0.73	0.82	555819
1 (wet)	0.38	0.83	0.52	111158
accuracy			0.74	666977
Macro avg	0.67	0.78	0.67	666977
Weighted avg	0.86	0.74	0.77	666977

Confusion Matrix:

Actual/Predicted	**0**	**1**
0	403283	152536
1	18622	92536

Feature Coefficients:

feature	coef	abs_coef
num__elv	-33.614158	33.614158
num__fill	30.637162	30.637162
cat__soil_5	3.049558	3.049558
cat__soil_15	-2.718454	2.718454
num__precip	-2.113412	2.113412
cat__soil_4	-1.969612	1.969612
cat__soil_2	-0.856262	0.856262
num__slope	-0.766562	0.766562
cat__soil_3	0.712873	0.712873
cat__soil_6	0.704048	0.704048
cat__soil_1	-0.596884	0.596884
cat__soil_7	0.315590	0.315590
num__flow	-0.223225	0.223225
num__asp	-0.088225	0.088225
cat__soil_0	-0.000876	0.000876

XGBoost

F1 score: 0.8823055564305997

Classification Report:

	precision	**recall**	**f1-score**	**support**
0 (non-wet)	0.98	0.98	0.98	555819
1 (wet)	0.88	0.88	0.88	111158
accuracy			0.96	666977
Macro avg	0.93	0.93	0.93	666977
Weighted avg	0.96	0.96	0.96	666977

Confusion Matrix:

Actual/Predicted	**0**	**1**
0	542795	13024
1	13129	98029

Feature Importance:

Feature	Importance
precip	0.313597
soil_5	0.308985
soil_15	0.093892
elv	0.086177
soil_3	0.045066
soil_7	0.035285
fill	0.032708
soil_6	0.021568
slope	0.020857
flow	0.011462
soil_2	0.011054
soil_1	0.009032
soil_4	0.006550
asp	0.003768
soil_0	0.000000

Model Comparison

Comparison Table

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.96	0.96	0.96	0.8906
XGBoost	0.96	0.96	0.96	0.8823
Linear Regression	0.74	0.86	0.74	0.5195

Analysis

Among the three models evaluated, the Random Forest model performed the best, achieving the highest accuracy (0.96) and F1-score (0.89) with a balanced precision and recall. The XGBoost model followed closely behind, matching Random Forest's accuracy, but with a slightly lower F1-score. In contrast, the Linear Regression model underperformed, with an accuracy of 0.74 and an F1-score of 0.52.

Looking at class-level performance, we see that both the Random Forest and XGBoost models had a strong predictive ability for non-wetland areas (F1 = 0.98), but performed less effectively for wetland areas (F1 = 0.89). This discrepancy can be attributed to the class imbalance in the dataset, where wetland samples were roughly 16% of all observations. This imbalance likely caused the models to favor the non-wetland class during training.

Overall, both Random Forest and XGBoost were far better suited for this task than Linear Regression, thanks to their ability to model nonlinearities. Although both achieved high accuracy, the Random Forest model demonstrated slightly better performance for the minority (wetland) class, suggesting it may be more robust for imbalanced datasets.

Challenges

In the early stages of model development, I experimented with referencing multiple tiles at once in an attempt to improve the Random Forest model's accuracy by incorporating data from neighboring areas. Unfortunately, the approach didn't work as expected and resulted in an F1 score of around 0.56. I spent some time troubleshooting the model but ultimately decided to process the data on a per-tile basis for the sake of simplicity and efficiency. Referencing multiple tiles simultaneously placed a heavy strain on my system and significantly increased processing time. With access to a more powerful computing setup, I would be interested in revisiting this idea to see if a spatially-aware approach could improve model performance.