MAPPING AND PREDICTING COMMUNITY VULNERABILITY TO HURRICANE FLORENCE IN COASTAL NORTH CAROLINA USING MACHINE LEARNING

by

Om Dahal

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Geoinformatics and Geospatial Intelligence

Committee:

____________________________________ Dr. Donglian Sun, Thesis Chair

____________________________________ Dr. John J. Qu, Committee Member

____________________________________ Dr. Arie Croitoru, Committee Member

____________________________________ Dr. Dieter Pfoser, Department Chairperson

____________________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

____________________________________ Dr. Ali Andalibi, Dean, College of Science

Date: ______________________________ Fall Semester 2019
George Mason University
Fairfax, VA

Mapping and Predicting Community Vulnerability to Hurricane Florence in Coastal North Carolina Using Machine Learning

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Om Dahal
Graduate Certificate of Geospatial Intelligence
George Mason University, 2016
Master of Natural Resources
Virginia Tech, 2013

Director: Donglian Sun, Professor
Department of Geography and Geoinformation Science
George Mason University

Fall Semester 2019
George Mason University
Fairfax, VA

DEDICATION

This is dedicated to my loving parents.

ACKNOWLEDGEMENTS

Most of all, I would like to thank my thesis adviser, Dr. Donglian Sun, for all of her invaluable guidance, encouragement, and support. I would also like to thank my committee members, Dr. John Qu and Dr. Arie Croitoru, for their valuable insight, encouragement, guidance, and feedback. Finally, my special thanks go to my wife and daughter for their unwavering encouragement and support during the entire time in this process.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF EQUATIONS

Equation Page

LIST OF ABBREVIATIONS

Adaptive Neuro-Fuzzy Inference System	ANFIS
American Community Survey	ACS
Artificial Neural Networks	ANN
Classification and Regression Trees	CART
Decision Trees	DT
Digital Elevation Model	DEM
Ensemble Prediction Systems	EPS
Land-use/land-cover	LULC
Matthews Correlation Coefficient	MCC
Mean-squared error	MSE
Multiple Perception	MLP
Normalized Difference Vegetation Index	NDVI
North Carolina Department of Transportation	NCDOT
Out-of-Bag	OOB
Random Forests	RF
Stream Power Index	SPI
Support Vector Machine	SVM
Wavelet Neural Network	WNN

ABSTRACT

MAPPING AND PREDICTING COMMUNITY VULNERABILITY TO HURRICANE FLORENCE IN COASTAL NORTH CAROLINA USING MACHINE LEARNING

Om Dahal, M.S.

George Mason University, 2019

Thesis Director: Dr. Donglian Sun

Extreme record breaking hurricanes followed by heavy rainfall and flooding claimed dozens of lives and damaged billions of dollar worth of property every year in the Atlantic coastal areas of the United States indicating that they are most vulnerable areas to hurricane hazards. Nevertheless, all the communities are not equally vulnerable due to their varying degrees of exposure and coping abilities. Thus, it is of vital importance to study the extent of vulnerability in different communities for the purpose of prevention, preparedness, response, and recovery efforts. This study attempted to predict and categorize vulnerability of communities to the hurricane Florence in the New Hanover County, North Carolina considering hurricane and subsequent disasters as a composite event. The Random Forests, a data driven machine learning method, was used to predict and categorize vulnerability of communities in census blocks. The explanatory variables were created from distance features and raster datasets. The training features were selected from crowdsourced data, disaster emergency evacuation locations, and satellite imagery collected during hurricane events. The regression results showed 0.93 percent R^2 value with tweets, roads, elevation, NDVI, and waterbodies as top five important variables. The classification results showed the accuracy per variable ranging from 0.96 to 1.00 with NDVI, roads, elevation, SPI, and tweets as top five important variables. The results demonstrated that the Random Forests ensemble learning method can be a valuable tool for categorical prediction and mapping of vulnerable communities from hurricanes. Furthermore, the results from both regression and classification models revealed that demographic variables are among the least important variables however they are not insignificant. It is recommended to combine all the three types of variables in prediction modeling for community vulnerability to hurricanes. The novel method used in this study may be used to identify the categories of vulnerable communities from various types of natural disasters in the other communities. It is also likely that predictions for vulnerability of buildings in the communities can be made using this method.

CHAPTER 1 INTRODUCTION

High wind and storm surge coupled with inundation are the major causes of infrastructure damage, loss of lives, and damage of property in the coastal United States (Helderop and Grubesic, 2019). The extreme weather events (e.g., hurricanes, floods, and fires) are increasing in number and intensity. The adverse effects on life and property are expected to only increase in the future as a consequence (Bouwer, 2019; Hoque et al., 2017), and that will make coastal human communities more vulnerable (Hoque et al., 2017).

Understanding demography of coastal areas is vital for hurricane vulnerability analysis. About 94.7 million (29.1 percent of the total U.S. population) live in coastline regions, of which about 44.4 million people live in the Atlantic coastline. The Atlantic coastline witnessed 13.2 percent population growth between 2000 and 2017. The percentage population of 85 and older is higher in the coastline counties compared to that of the United States (Cohen, 2019). The population in the Atlantic coastal areas is most vulnerable to hurricanes due to high frequency of devastating hurricanes in this region. It is evident from the fact that eight hurricanes made landfall in the Atlantic coastal areas between 2000 and 2017, each of which caused more than $10 billion worth of damages (Cohen, 2019).

The hurricane Florence is another disaster event of most devastating impact occurred in 2018 hurricane season. This hurricane lasted until September 18 since it made landfall in September 14 with the slow motion of about 3-4 miles per hour with a zone of tropical storm force winds nearly 400 miles wide (Feaster et al., 2018). This hurricane was at the intensity of category one along the southeastern coast of North Carolina. It caused a total of 52 fatalities, and estimated damage of approximately $24 billion of which a significant portion of loss was in North Carolina having to lose power in about one million households. Numerous trees were uprooted due to strong force of hurricane winds, but most of the damages to homes and commercial buildings were caused by freshwater flooding, with approximately 74,563 structures being flooded (Stewart and Berg, 2019). The loss of agricultural farm products and livestock from the hurricane Florence was accounted for at least $1.1 billion (Feaster et al., 2018).

The hurricane Florence produced 10 to 30 inches of rainfall in the New Hanover County and its surrounding areas due to slow movement and persistent rain bands before and after the hurricane made landfall that established a new highest record of rainfall in two decades. This extreme rain resulted in excessive low-land record breaking river floods across the New Hanover County. Eighteen record breaking peaks of streamflow were observed in North Carolina, and some of them were the highest since 1940 (Stewart and Berg, 2019).

The New Hanover County is a coastline county in the tidewater area in North Carolina (Figure 1a). The area of this county is approximately 329 square miles, of which approximately 192 sq. miles (58%) is land and 137 sq. miles (42%) is water (US Census

Bureau, 2019). This is one of the densely populated county in North Carolina with 232,274 population, 91,673 households, and 113,215 housing units according to US Census Bureau (2018).

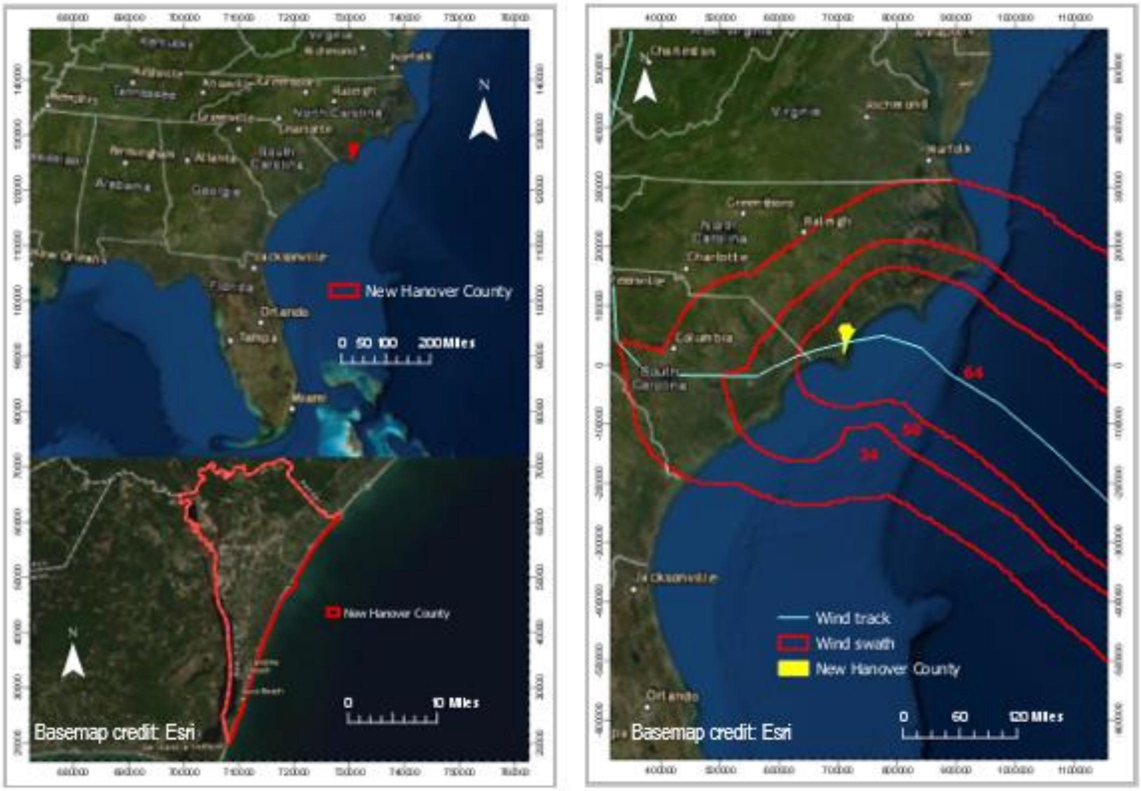


Figure 1 (a) The New Hanover County, North Carolina, (b) the hurricane Florence wind track and swath

The New Hanover County is one of the hardest hit area by the hurricane Florence in North Carolina coast where worst flash floods were experienced in the history of this locality. The Florence made landfall near Wrightsville Beach (Figure 1.b) causing up to three feet of flash flood that inundated Northchase, Writsboro, and Ogden neighborhoods. Similarly, downtown Wilmington was inundated by two feet of flood from the Cape Fear

River. As a result, the entire county was generally isolated from outside world due to access road closures for several days (Stewart and Berg, 2019).

Since the devastating hurricane, rainfall, and flooding events occur frequently in the coastal United States, it is vital to learn the levels of vulnerability of different communities for the purpose of damage prevention, preparedness, rescue, and recovery efforts. Moreover, it is necessary to understand what factors should be given higher priority in the efforts to deal with hurricanes. Similarly, whether the extensively used machine learning algorithm in prediction (mainly used to predict areas of potential landslides and flooding), Random Forests (RF), can be useful to predict vulnerable communities from hurricane hazards. Thus, the set of objectives for this study are as follows:

(a) To identify the level of vulnerability of different communities in the coastal New Hanover County, North Carolina from the hurricane Florence using geo-physical, socio-economic, and social media-generated explanatory variables.

(b) To examine the usefulness and applicability of the Random Forests algorithm to make categorical prediction of vulnerability in coastal communities from hurricane hazards.

CHAPTER 2 LITERATURE REVIEW

Disaster is an overall consequence of a hazard event (Klonner et al., 2016). Vulnerability is a function of exposure and coping ability, or it is an inability of people to deal with the hazards due to physical and social conditions of the place of their residence (Wu et al., 2002). Vulnerability of communities varies with their coping ability. Coping ability is combined derivative of resistance and resilience (Rygel et al., 2006). Also, levels of risks depend on the hazard intensity and levels of vulnerability. Hence, same hazard may have different impacts on different communities or places depending on their exposure and coping ability (Klonner et al., 2016). Vulnerability has been conceptualized as pre-existing conditions that potentially expose humans to hazards, e.g., human settled in hazardous areas. Loss of life and property is likely in the hazardous areas when there is a catastrophic natural event. This is the kind of vulnerability caused due to biophysical settings of the area of residence (Rygel et al., 2006). Second way of conceptualizing vulnerability is social vulnerability that stem from social marginalization due to age, race, disability, or income (Morrow, 199; Rygel et al., 2006). Assessment of social or community vulnerability needs to include select demographic data, essentially, disability, vulnerable age groups (children and aged population), and poverty (Morrow, 1999; Aubrecht et al., 2013). The third approach is the vulnerability of places which combines biophysical as well as social risk within a specific geographic area to evaluate vulnerability (Rygel et al., 2006). These are multiple frameworks to explain the root cause of vulnerability to natural disasters from social conditions inherent in the community (Morrow, 1999) to biophysical environment around the community, or combination of both. It is crucial to consider a coupled human-environmental system to identify the vulnerable communities (Cutter et al., 2008).

Identification and mapping of coastal communities at risk from hurricane hazards is crucial for every stage of disaster management consisting of prevention, preparedness, response and recovery. Use of remote sensing data analysis methods have been increasingly used for risk assessment from hurricane disasters (Hoque et al., 2017). This is promising due to the fact that large quantities of remotely sensed data are being collected having very high applicability in hurricane risk assessment (Zhang et al., 2019; Zhou et al., 2019).

The geotagged information from Twitter, Facebook, or Flicker also have been proved that they have high applicability in hurricane impact study because they provide valuable information regarding geometries, attributes, and semantic information. The social media-generated data can have spatial patterns. The social media posts during disaster strikes are at or closer to the affected areas. Therefore, they are likely to capture the information regarding the effects of disasters. For these reasons, social media-generated geographic information is a promising alternative geospatial data for natural hazard analysis. So, crowdsourced locations, messages, or images can be used as valuable complementary data for the natural disaster vulnerability models (Klonner et al., 2016).

Statistical, physical, and data-driven (e.g., machine learning) models were typically used in prediction of natural hazards risks. Even though the physical models have great capabilities of prediction of natural hazards risks, they require datasets collected from the ground, intensive computation, and high level of expertise (Mosavi et al., 2018). The most remarkable drawback of this modeling is that the prediction cannot be carried out in short time frame because of data collection efforts taking long time (Mosavi et al., 2018). Similarly, numerical prediction models could have systematic errors (Mosavi et al., 2018). In order to overcome the shortcomings of these models data-driven prediction modeling have been widely used. The strengths of machine learning models are that they do not require to have the knowledge of underlying physical processes, quicker to develop, allow fast training, validation, testing, and evaluation. Moreover, this approach has outperformed the conventional approaches with higher prediction accuracy, and data-driven algorithms can predict beyond the range of training datasets spatially and temporally (Mosavi et al., 2018). Artificial Neural Networks (ANNs), Multiple Perception (MLP), Adaptive Neuro-Fuzzy Inference System (ANFIS), Wavelet Neural Network (WNN), Support Vector Machine (SVM), Decision Tree (DT), and Ensemble Prediction Systems (EPSs) are the algorithms have highest favorability among natural hazards modeling community (Mosavi et al., 2016).

In the hurricane hazard risk analyses literature, apparently, socio-economic variables were preferred less than geo-physical variables to analyze vulnerability of coastal communities from hurricane hazards although it is a multi-variate non-linear problem.

CHAPTER 3 METHODOLOGY

This study was performed as shown in the workflow diagram in Figure 2 below. First, explanatory variables for input to the model were selected that likely could explain vulnerability from the hurricane Florence. The selected data for explanatory variables were then pre-processed and stored in a geodatabase. Features were collected and categorized for training and validation data for the model and stored in the same geodatabase. Likewise, polygon features to receive prediction were collected and stored. After building satisfactory RF classification and regression models separately, they were executed with input distance features and raster datasets to calculate explanatory variables that were later used to predict the vulnerability levels and generate two different predicted maps from these models.

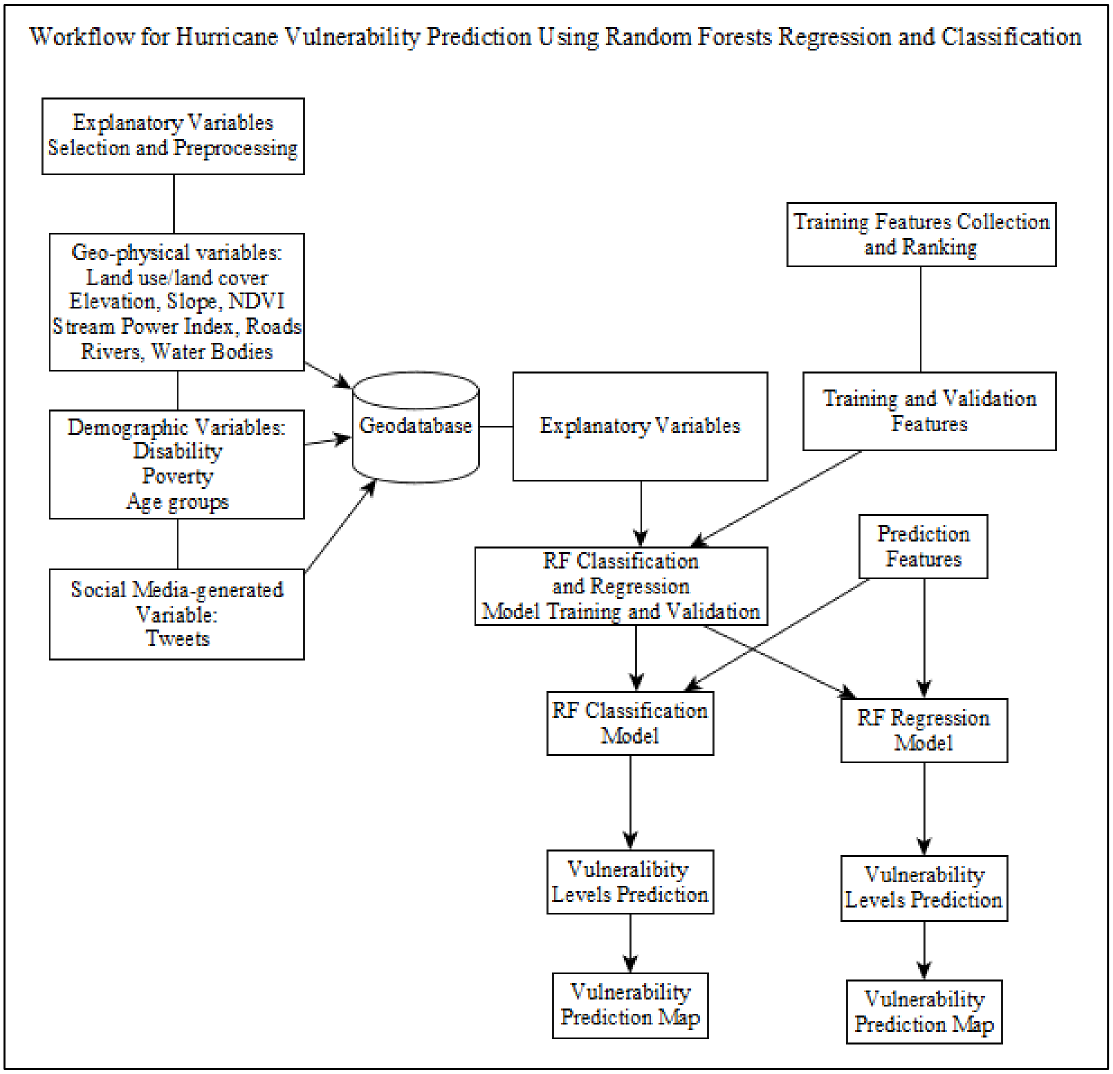

Workflow for Hurricane Vulnerability Prediction Using Random Forests Regression and Classification
Explanatory Variables Selection and Preprocessing
Geo-physical variables: Land use/land cover Elevation, Slope, NDVI Stream Power Index, Roads Rivers, Water Bodies
Demographic Variables: Disability Poverty Age groups
Social Media-generated Variable: Tweets
Geodatabase
Explanatory Variables
Training Features Collection and Ranking
Training and Validation Features
RF Classification and Regression Model Training and Validation
Prediction Features
RF Classification Model
RF Regression Model
Vulneralibity Levels Prediction
Vulnerability Levels Prediction
Vulnerability Prediction Map
Vulnerability Prediction Map

Figure 2 Workflow for hurricane vulnerability prediction modeling

Section 3.1 Training Features

NAPSG Foundation, GISCorps, and CEDR Digital maintained a Story Map displaying 2018 hurricane crowdsourced photos collected from Instagram, Twitter, Facebook and online news media (Figure 3). This is a collection of photos with the brief description of events by social media users illustrating the incidences (e.g., hurricane impact, hurricane intensity, damage, storm surge, flooding, and rescue efforts) before, during, and after the hurricane event (NAPSG Foundation et al., 2018). After careful observation of photos and their descriptions, they were classified into four different levels of severity as class vulnerability categories and were assigned the numbers from 1 to 4 (1 indicates the most at risk location, highest vulnerability, and 4 indicates the least at risk location, lowest or no vulnerability). Total of 99 locations within the study area were identified from the story map appropriate for training input. Similarly, emergency shelters (shelter locations designated by the New Hanover County to evacuate county residents during natural disaster emergencies including hurricanes and floods) were collected and they were considered as no risk or least risk locations. They were assigned to 5 and 6 in vulnerability categories. More locations were identified by observing satellite imagery and flood maps collected during the hurricane Florence and assigned numbers from 1 to 6 vulnerability categories depending on the severity of the impact observed. Total of 273 location points were identified for input as training features. These training features with their corresponding vulnerability labels are summarized in Table 1 and displayed on the map in Figure 4a below.

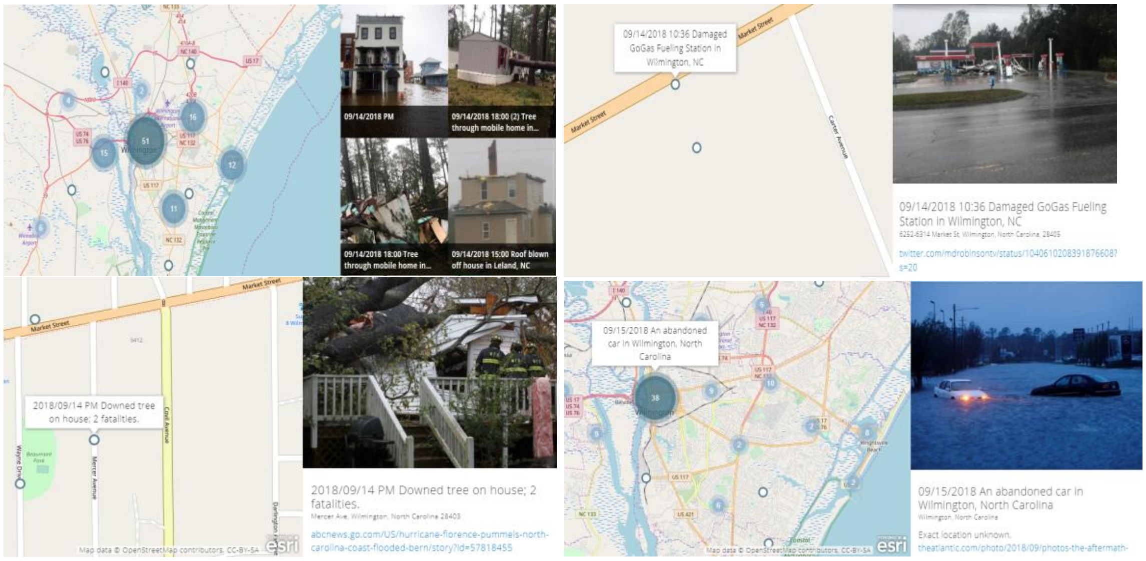

09/14/2018 PM
09/14/2018 18:00 (2) Tree through mobile home in...
09/14/2018 18:00 Tree through mobile home in...
09/14/2018 15:00 Roof blown off house in Leland, NC
09/14/2018 10:36 Damaged GoGas Fueling Station in Wilmington, NC
Market Street
Carter Avenue
09/14/2018 10:36 Damaged GoGas Fueling Station in Wilmington, NC
6252-6314 Market St, Wilmington, North Carolina, 28405
twitter.com/mdrobinsontv/status/1040610208391876608?s=20
2018/09/14 PM Downed tree on house; 2 fatalities.
Market Street
Mercer Avenue
Wayne Drive
Covil Avenue
Map data © OpenStreetMap contributors, CC-BY-SA
esri
2018/09/14 PM Downed tree on house; 2 fatalities.
Mercer Ave, Wilmington, North Carolina 28403
abcnews.go.com/US/hurricane-florence-pummels-north-carolina-coast-flooded-bern/story?id=57818455
09/15/2018 An abandoned car in Wilmington, North Carolina
38
Map data © OpenStreetMap contributors, CC-BY-SA
esri
09/15/2018 An abandoned car in Wilmington, North Carolina
Wilmington, North Carolina
Exact location unknown.
theatlantic.com/photo/2018/09/photos-the-aftermath-

Figure 3 2018 hurricane crowdsourced photos:
Photos above are powered by NAPSG Foundation, GISCorps, and CEDR Digital, a Story Map (a) upper left picture is a general map with cluster of locations with impacted locations; (b) upper right picture showing location of damaged gas station in Wilmington, NC on 9/14/2018, (c) lower left map showing location of a downed tree on a house on 9/14/2018; (d) lower right picture is the location on map and an abandoned car in Wilmington, NC on 9/15/2018.

Table 1 Vulnerability categories and number of locations for model training

Vulnerability Category	Number of Locations
1 highest	59
2	77
3	51
4	31
5	17
6 lowest	38
Total	273

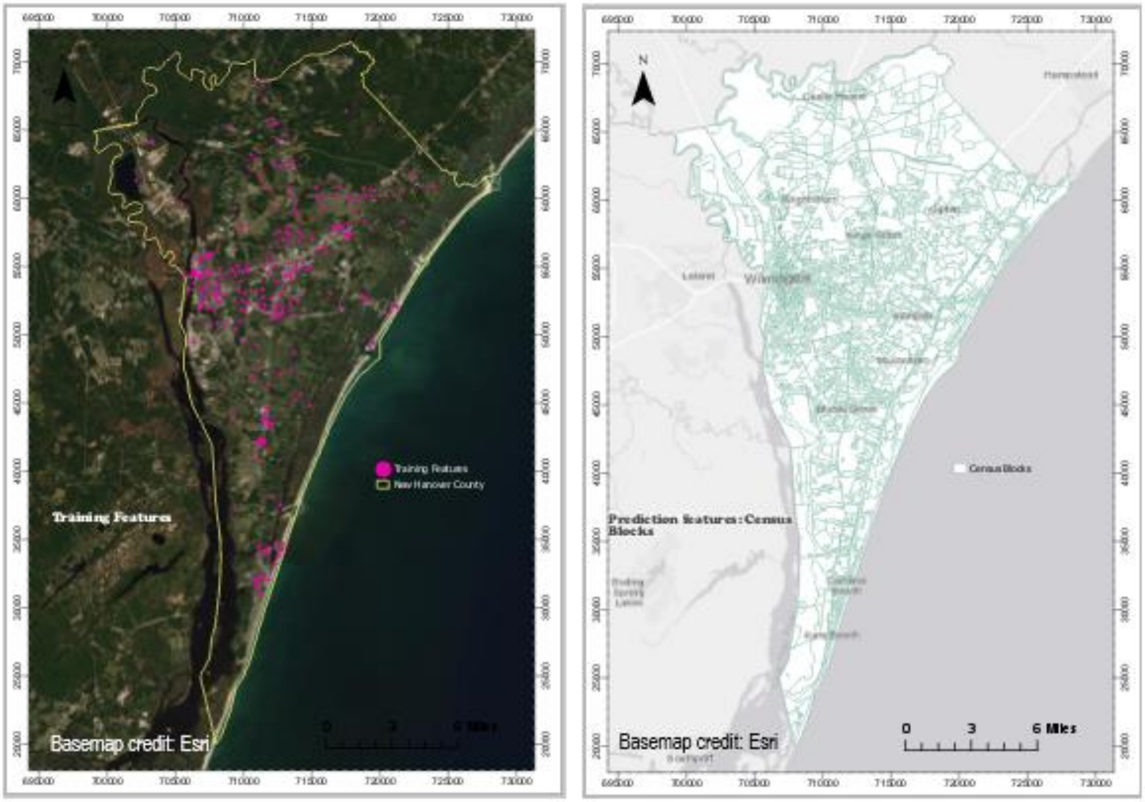


Figure 4 (a) left, training point features corresponding to Table 01, (b) right, prediction polygon features (census blocks)

Section 3.2 Prediction Polygon Features

Prediction polygon features are the features representing polygons to receive the results of the predictions made by the models. Since the goal of this work is to make prediction for vulnerability of communities, the census blocks would be ideal polygon features to predict on because census blocks are the areas that encompass small communities with distinctive geophysical and demographic similarities. The New Hanover County consists 5069 census blocks as delineated by the US Census Bureau in 2010 census (Figure 4b).

Section 3.3 Explanatory Variables

The vulnerability is due to combination of multiple geophysical, demographic, and socio-economic conditions of people and places where they live. These geophysical conditions, demographic conditions, information generated regarding these conditions, and information regarding hurricane itself can be defined as explanatory variables for hurricane disaster analysis. There is no consensus as to which factors should be given higher priority when categorizing vulnerability of communities from hurricanes in coastal areas (Bathi and Das, 2016). This work used combination of geo-physical, demographic, and social media-generated information as explanatory variables and found variable importance by a semi-automated process as discussed in the following sections.

Geophysical variables

1. ***Land use/land cover***

Sentinel-2 high resolution (10m) multispectral imagery for surface reflectance was used for land-use/land-cover (LULC) classification. The imagery was classified using Semi-automatic Classification Plugin for QGIS version 2.18. Out of 12 Sentinel-2 spectral bands, bands 1 (coastal aerosol), 9 (water vapor), and 10 (cirrus) were excluded from classification dataset. The imagery was classified into nine different land-use and land-cover classes: (a) forest, (b) ocean, (c) river, (d) lake/pond, (e) road, (f) residential, (g) agricultural, (h) commercial, and (i) marsh. The Maximum Likelihood algorithm was used to classify the imagery, which calculates the probability distribution for the classes, related to the Bayesian theorem to find a pixel that belongs to the land cover class in training (Richards and Jia, 2006). The classified output raster then resampled to 30m

(Figure 7c) to reduce the number of pixels to synchronize with the processing ability of ArcGIS Pro version 2.2, Forest-based Classification and Regression tool.

2. Elevation

3D Elevation Program (3DEP) (https://www.usgs.gov/core-science-systems/ngp/3dep), USGS, National Map Services collects 1/9 arc seconds (approximately 1m resolution) digital elevation model (DEM) (https://viewer.nationalmap.gov/basic/#productSearch) data. It was used as an elevation dataset for elevation explanatory variable. The DEM was resampled to 30m in order to overcome the computational limitation of the tool. The elevation of New Hanover County ranges from 0m to 30m (Figure 7b).

3. Slope

Slope tells steepness of a raster surface. Slope was calculated in degrees using DEM dataset discussed in previous section. Planar method parameter was used where slope is measured as maximum rate of change in value from a cell to its immediate neighbors. The following slope algorithm was used (Equation 1).

Equation 1 Algorithm to calculate slope in degrees

$$\text{Slope degrees} = \text{ATAN} \left(\sqrt{([dz/dx]^2 + [dz/dy]^2)} \right) * 180/\pi$$

Where, $\frac{dz}{dx}$ is rate of change in x-direction, and $\frac{dz}{dy}$ is rate of change in y-direction. Slope indicates the topographic change and variability of surface. Lower slope means flatter surface which has higher risk of flooding (Wang et al., 2015). Slope raster used as an input explanatory variable is shown in map (Figure 8)

4. *Stream Power Index (SPI)*

Stream Power Index (SPI) is a measure of power of flowing water on terrain surface. The higher the stream power index the more erosion it can cause downstream. Stream power is a hydrological factor that can condition or explain how damaging the flood could be (Wang et al., 2015; Lee et al., 2017). The SPI was calculated from slope and flow accumulation raster datasets obtained from terrain analysis of digital elevation models (Figure 7d). The percent rise slope was used to calculate SPI by the formula in ArcGIS raster calculator (Equation 2).

Equation 2 Percent rise slope

SPI = Ln (Flow accumulation raster + 0.001) * ((Slope raster /100) + 0.001)).

5. *Normalized Difference Vegetation Index (NDVI)*

Normalized Difference Vegetation Index (NDVI) measures the difference between near-infrared (NIR) and red values of wavelengths. NDVI values range from -1 to 1. Healthy vegetation has highest NDVI value, i.e., inclined towards 1 and water inclined towards -1. Other land cover values fall between these two extremes depending on the type, growth, soil moisture, and presence or absence of vegetation, snow, and soil roughness (Wang et al., 2015). NDVI of area of interest was computed from Sentinel-2 imagery bands, Band 4 (Red) and Band 8 (NIR), as given by the formula in Equation 3 below. NDVI explanatory variable used as an input is shown in map below (Figure 7c).

Equation 3 Normalized Difference Vegetation Index

$$\text{NDVI} = \frac{NIR\ (Band\ 8) - Red\ (Band\ 4)}{NIR\ (Band\ 8) + Red\ (Band\ 4)}.$$

6. *Major roads*

Major roads play a critical role before, during and after natural disasters from the perspective of evacuations, rescue and recovery needs. The wider roads available closer by a settlement the easier it will be to evacuate and provide post event assistance. As a result, the communities could become safer from the impacts of hurricanes and floods. Thus, road features can be considered as a remarkable variable to explain vulnerability prediction. Road features were obtained from North Carolina Department of Transportation (NCDOT) for explanatory variable (Figure 5b).

7. *Water features*

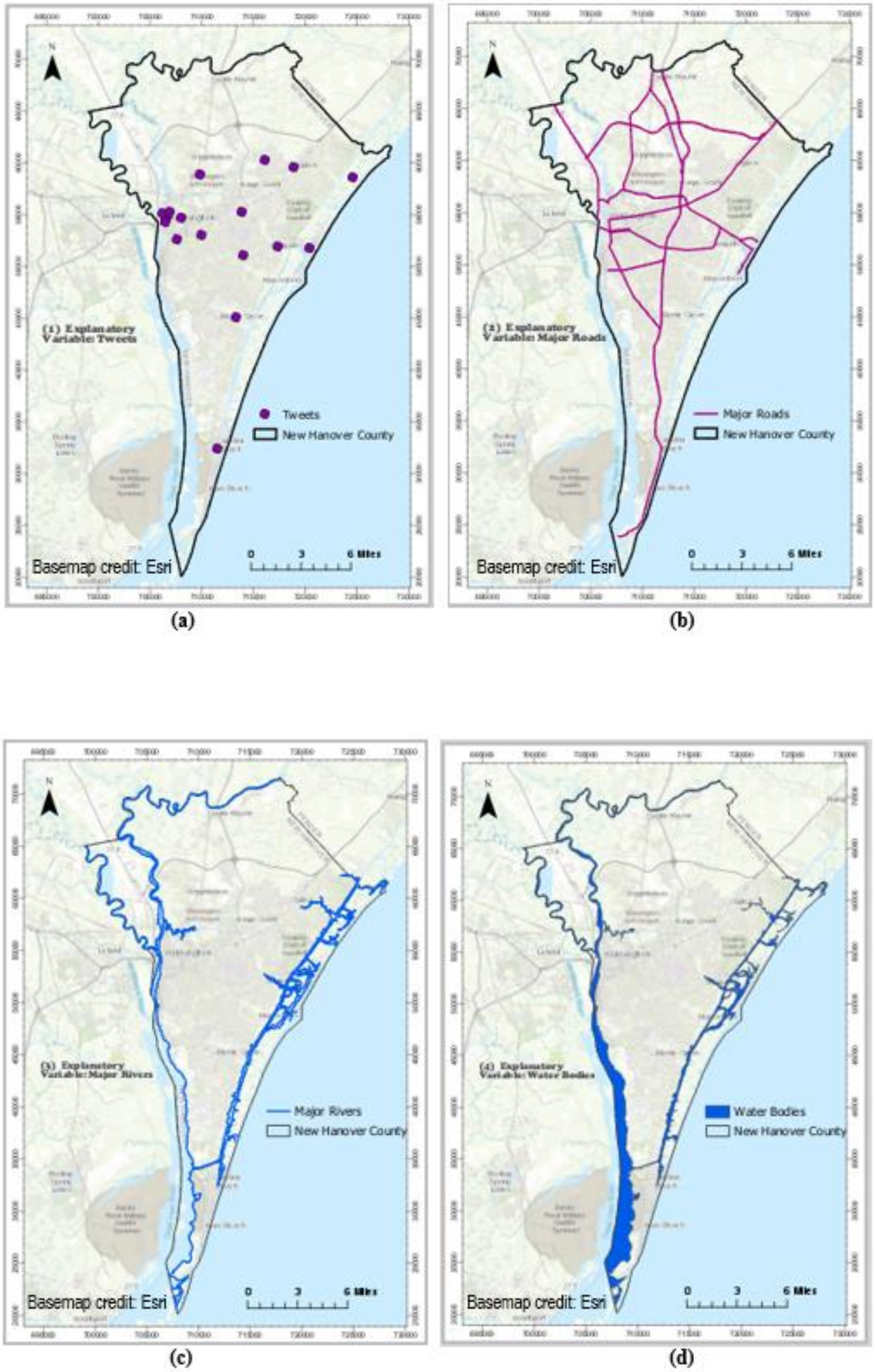
NC Center for Geographic Information and Analysis distributes the major hydrography data that include major rivers and water bodies (lakes, ponds, dams etc.) (Figures 5c and 5d). Rivers and other water features are the areas where floods surge during hurricane and heavy rainfall. People adjacent to water features could be in danger of being affected by flood. For this reason, this is an important addition to the list of explanatory variable of vulnerability prediction.

Demographic variables

Poverty, gender, race, ethnicity, age, and disability are demographic indicators of social vulnerability. Poor people, women, children, people with disability, and aged people are vulnerable because of their inability to have access to resources needed to protect themselves when disaster strikes, and recover in the aftermath of disaster (Rygel et al., 2006). Age groups 0-14 and 65 and older, population with disability, and population with poverty are considered more vulnerable than the rest of the population in the event of natural disasters such as hurricanes (Morrow, 1999). So, American Community Survey (ACS) 2017 data at block group level for age, disability and poverty were used for demographic explanatory variables (Figures 6a, 6b, 6c, and 6d).

Social media-generated variables

Social media is a fundamental tool for people to discriminate and consume real time information regarding storm intensity, routes, damages, safety, evacuation, rescue and recovery. As part of data collection, real time twitter stream was downloaded using "#Florence" as keyword during hurricane Florence, September 14 through September 19, 2018. Out of thousands of tweets with this hashtag from all over the world, 65 tweets were geo-enabled, and were posted from the New Hanover County (Figure 5a). Though the number is small, tweets are significant explanatory variable as they were tweeted real time and place as the hurricane event unfolded. Geographic information collected and disseminated online are crucial alternative of conventional data, and immensely useful for preparedness, response, and recovery in the event of natural disasters (Goodchild, 2010).



(a) (b) (c) (d)

Figure 5 Explanatory variables: (a) tweets, (b) major roads, (c) major rivers, and (d) water bodies

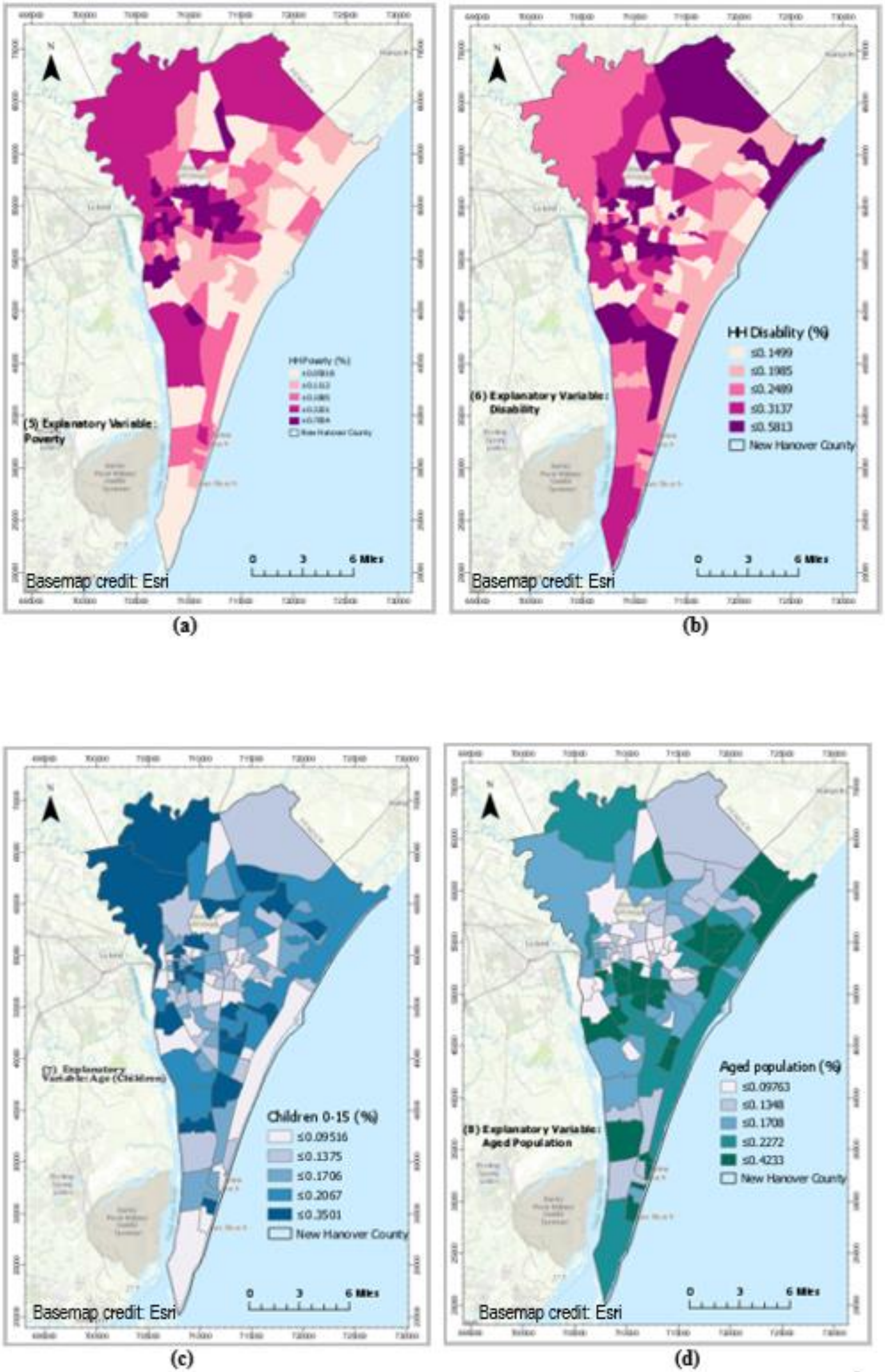
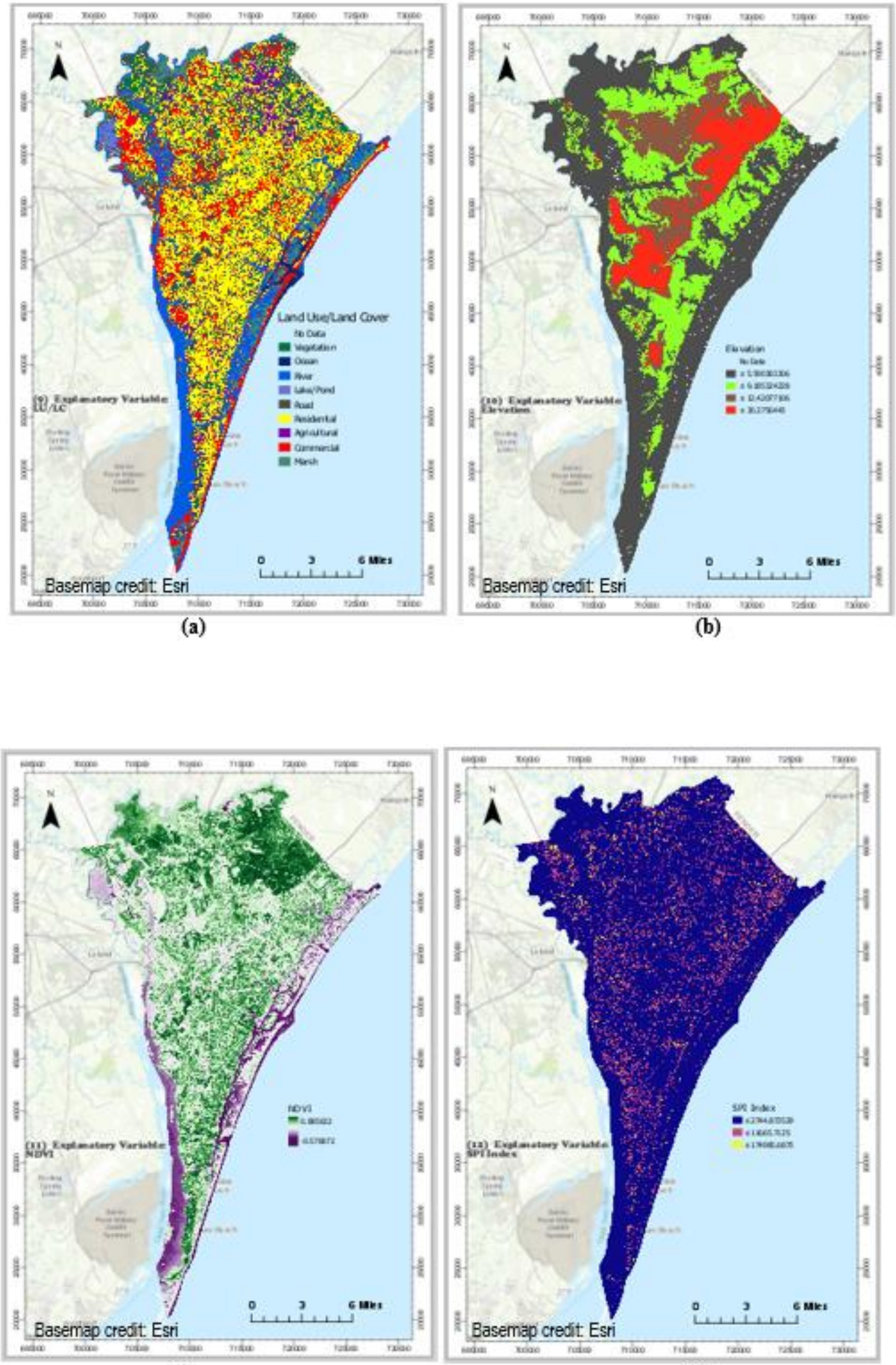


Figure 6 Explanatory variables: (a) poverty, (b) disability, (c) children, and (d) aged population



(a) (b)

(c) (d)

Figure 7 Explanatory variables: (a) land use/land cover, (b) elevation, (c) NDVI, and (d) SPI

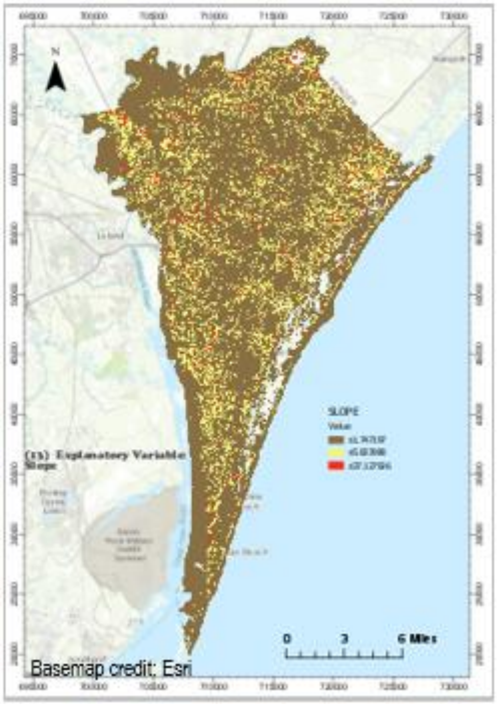


Figure 8 Explanatory variable: slope

Section 3.4 Random Forest Classification and Regression

Natural hazard risk prediction is multivariate and non-linear task (Wang et al., 2015) due to the combined role of a number of disaster-inducing factors. Several methods and machine learning algorithms have been employed to solve the predictive analysis such as the Support Vector Machine (SVM), the Artificial Neural Networks (ANN), and the Decision Trees (DT). The major weakness of these algorithms is their inability to estimate each conditioning factors contribution to the total risk (Wang et al., 2015). The Classification and Regression Trees (CART) algorithm decision trees are greedy. Even with bagging, the trees can have structural similarities that will result in high correlation in predictions. However, in the RF the trees are uncorrelated or least correlated because learning algorithms just select random sample of features from random sample of variables as specified in parameters (Storey, 2019). The RFs are modification of CART algorithms (Pourghasemi and Kerle, 2016). It is a supervised classification and regression method of modeling that allows growing an ensemble of trees and letting them vote for the most occurred class as the predicted class (Breiman, 2001). The RF is an algorithm capable of estimating the contribution of each factor to the total effect (Wang et al., 2015). The RF has high forecast accuracy, acceptable tolerance to outliers and noise, and has ability to easy avoidance of outfitting (Wang et al., 2015). The RF algorithm generates numerous binary trees which collectively called forests (Park and Kim, 2019). In the RF, trees grow based upon a bootstrap sample. For each node, random subsets of samples are selected. The "out-of-bag" error rate is calculated using samples out of the bootstrap sample (Park and Kim, 2019). Mean decrease in accuracy and mean in the Gini are calculated in the process, which then are used to calculate the variable importance scores (Park and Kim, 2019).

Given an observation for each tree in the model, the RF predicts outcomes using tree applied to an observation and store outcome as a list. If the model is classifier, it returns maximum count. If the model is regression, it returns average (Storey, 2018) (Figure 9).

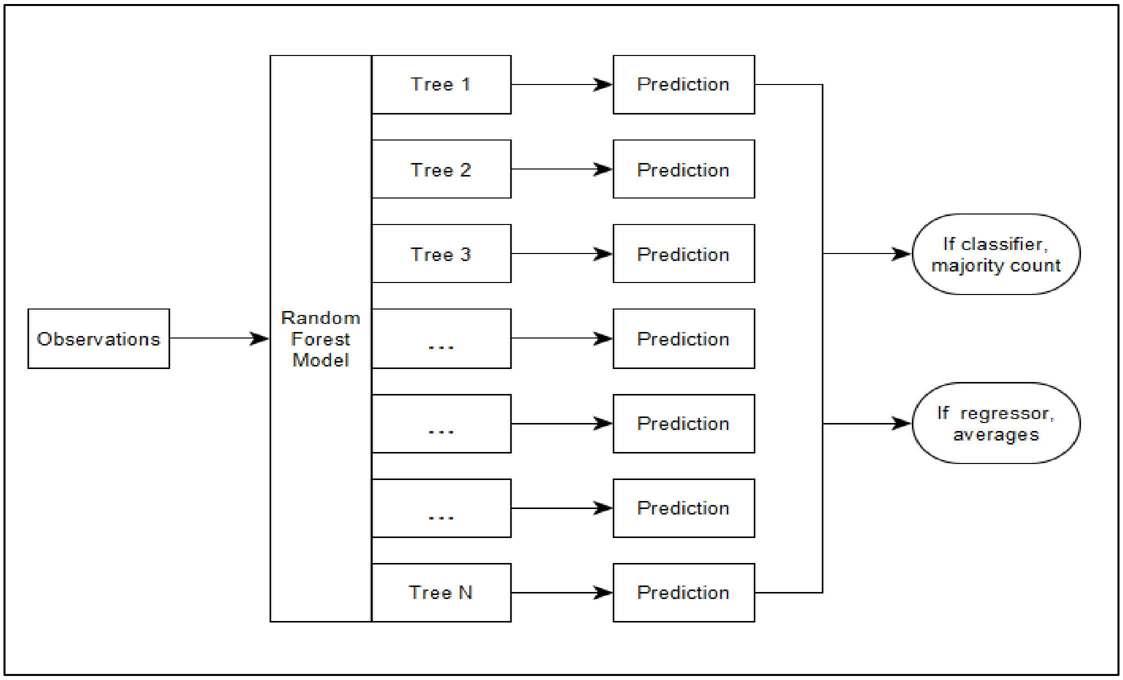


Figure 9 Random Forests process flow based on Storey (2018)

The RF algorithm relies on a parallel ensemble method known as "bagging" or bootstrap aggregation to generate classifiers. This is a method that averages multiple estimates that are measured from random subsamples of variables. A subset of observation are selected at random to form a subsample and used to train the model, and the process is repeated again to select the subset of samples from the original observation until the specified number of tree limit is reached. This process is known as bootstrapping (Storey, 2018). Random Forests is built by: specifying number of trees, specifying number of variables, specifying number of features (columns) to be used in each tree, Then, for each tree: number of samples are selected with replacement from all observations. Also, given number of features are selected randomly and a decision tree is trained with selected samples and features (Storey, 2018). Specified number of samples selected from original dataset is known as bootstrap samples. The RF process randomly selects variables from the sample for each node split. An unpruned classification tree is grown for each bootstrap sample. Finally, all the trees are aggregated and prediction for the new label is performed by majority votes (Ai et al., 2014).

Variable importance

Mean decrease accuracy and mean decrease Gini are widely used for measuring, ranking, and selecting variable importance (Park and Kim, 2019). Often in regression problems the drop in sum of squared errors, and in classification problems the Gini impurity score are calculated to estimate errors. The greater the impurity the greater the importance of variable (Brownlee, 2019). Gini impurity is computed by summing the probability of each item chosen multiplied by the probability of an error to classify that item into correct class (Ai et al., 2014). Gini impurity is obtained by the following equation (Equation 4).

Equation 4 Gini impurity

G(k) = $\sum_{i=1}^{n} P(i) \times (1 - P(i))$, where *P(i)* is the probability at node (i).

Gini impurity of parent node is higher than that of child node (Wang et al., 2017). The Gini decrease of each individual explanatory variable is combined to estimate the total contribution of it in the prediction of vulnerability (Wang et al., 2017). The variable importance is calculated by the given formula (Equation 5).

Equation 5 Variable importance

$$P_k = \frac{\sum_{i=1}^{n} \sum_{j=1}^{t} D_{Gkij}}{\sum_{k=1}^{m} \sum_{i=1}^{n} D_{Gkij}}$$

Where, P_k = the variable importance, m = total number of explanatory variables, n = total number of classification trees, t = total number of nodes, and D_{Gkij} = Gini decrease value of the j^{th} node in the i^{th} tree that belongs to the k^{th} vatiable. Mean-squared error (MSE) is obtained by the given equation (Equation 6).

Equation 6 Mean-squared error

$$\varepsilon = (V_{observed} - V_{response})^2$$

Where, ε is mean squared error, $V_{observed}$ is the variable from observed data, and $V_{response}$ is the variable from result (Lee et al., 2017)

Out-of-bag (OOB) error

Each tree in RF is constructed from a random sample of observations, usually called bootstrap samples. The observations that are left out from constructing a tree during classification process are called "out-of-bag" (OOB) observations, i.e., unseen data in classification (or out of bootstrap samples). Therefore, each tree is constructed from different samples from the whole dataset. The prediction for an observation is made from the trees for which the observations were not used to build them. The error rate estimated from these predictions is known as "out-of-bag" error (Ai et al., 2014; Janitza and Hornung, 2018).

CHAPTER 4 RESULTS AND DISCUSSION

The Random Forests (RF) regression and classification models were constructed and executed using “Forest-based Classification and Regression” tool in ArcGIS Pro 2.2. The explanatory distance variables and explanatory raster variables described in previous section were used to predict hurricane vulnerability by classification and regression methods separately. The model was constructed based on “vulnerability levels” that were the variables to predict. Variables to predict were classified into 6 categories from 1 to 6 (1 indicating the most vulnerable and 6 indicating the least vulnerable to hurricane hazards) as an attribute in training feature class. Thirteen vector and raster geospatial datasets that realistically would explain the vulnerability of communities in New Hanover County, North Carolina were used together as input variables in the analysis.

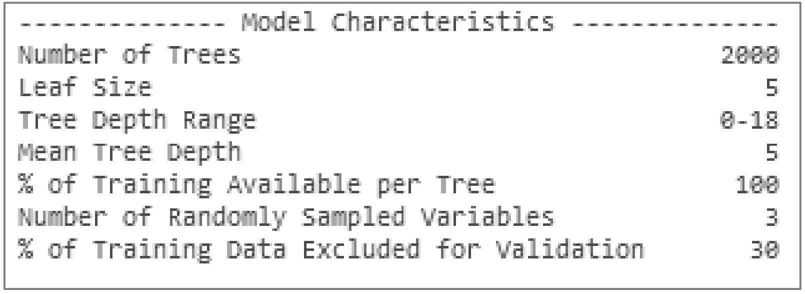
Explanatory variables from distance features were calculated by first finding distances from the nearest input distance features to each of the input training features. Likewise, explanatory variables were extracted from input raster dataset for each point location. The distance attributes were calculated from the training features to the closest segments of the polygons or lines of explanatory variables. The explanatory variables were then used in the constructed model and vulnerability of communities was predicted using census blocks as prediction areas.

Section 4.1 Regression Results and Analysis

Two thousand decision trees parameter was found to be optimal number of trees during model construction process. The prediction from regression model was made to census blocks to produce predicted vulnerability output corresponding to vulnerability levels in input training features. After the model is trained, the validation data were used to predict the values of the test data. The predicted values were then compared to the observed values to provide a measure of prediction accuracy based on data that were not included in the training process.

Leaf size parameter is the number of observations required to keep a terminal node without further split. Minimum leaf size parameter set for this regression model was 5, i.e., tree stopped growing after it has achieved minimum observation of 5 at its terminal node. Tree depth means number of nodes in each tree from root node to leaf. The tree depths in the forest ranged between 0-18, having 5 mean tree depth. Number of variables available to construct each tree was set to 100%, number of randomly sampled variables for each tree was 3 (square root of total number of variables, i.e., 12), and percent of data excluded for validation was set to 30 (Table 2).

Table 2 Regression model characteristics



```
-------------- Model Characteristics --------------
Number of Trees                                2000
Leaf Size                                         5
Tree Depth Range                               0-18
Mean Tree Depth                                   5
% of Training Available per Tree                100
Number of Randomly Sampled Variables              3
% of Training Data Excluded for Validation       30
```

Variable importance is a measure of how important a variable is in the prediction process. The RF determines variable importance by complex interactions among the variables by observing how much prediction error increased when data for that variable is permuted while all others are left unchanged. The calculations are carried out from each tree, and final variable importance score is obtained.

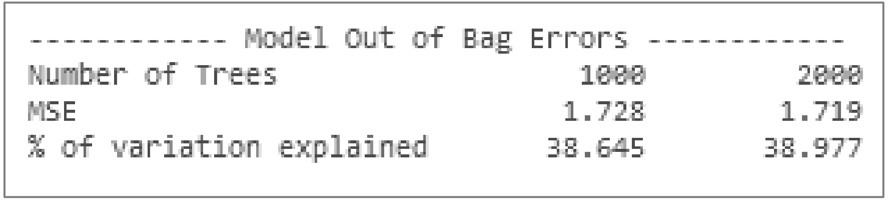
Mean decrease in accuracy is a measure to express the extent of contribution of a variable towards decreasing in accuracy of prediction during OOB error calculation. The variable with a large mean decrease in accuracy are more important for classification. The more the accuracy of a variable decreases due to the exclusion of a single variable the more important that variable is considered. The mean decrease in Gini is a measure of how each variable contributes to the homogeneity of the nodes. Each time a particular variable is used to split a node, the Gini for the child nodes are calculated and compared to that of the original node. Variables that result in nodes with higher purity have a higher decrease in Gini (San Diego State University, n.d.).

Mean squared error (MSE) is the average squared difference between the predicted values and the observed values. This is another measure of the quality of a model. The values closer to zero are better. In this model, the MSE for number of trees 1000 and 2000 were 1.728 and 1.729 respectively. While doubling the number of trees, the error decreased but not significantly in regression model.

Percent of variation explained is the determination of the degree of relationship in the patterns of variation, or how well the variation of one variable is explained by the variation of the other variable. The coefficient of determination (R^2) is the measure of the variation explained. The higher the value of R^2 the higher the predictive value of the regression may be. There can be situations that percent of variation explained and R^2 may be insignificant in case the number of data points is higher (Colby College, n.d.). The value of R^2 from this model for training data and validation data were 93% and 21% respectively (Tables 4 and 5). This R^2 value for validation data appears to be significantly lower than that training data. Even though this is the case, it should not be interpreted that the goodness of fit for validation data was insignificant and erroneous. The reason for that is: the quality of model should not be evaluated based solely on the value of R^2, besides P-values and standard errors are other measures that should be taken into account to evaluate the quality of model outcomes (Nau, 2019; Shalizi, 2015). Shalizi (2015) even demonstrated that R^2 can be low when the model is correct and claimed this alone most not be considered to evaluate the goodness of fit of any model. R^2 value may be low if data have high amount of noise or high variance. Even though there is no any threshold to call R^2 value good or bad for regression, it is always good to be in the position to have a higher value of R^2 (Nau, 2019).

Percent of variation explained may vary as the number of trees parameter is changed. In this case, percent of variation explained was 1.728 and 1.719 for 1000 and 2000 trees respectively. It slightly decreased when number of trees were doubled indicating that predictive ability of the model increased as the number of tree parameter increased from 1000 to 2000. This showed that it did not make a remarkable difference in the ability of model to predict (Table 3).

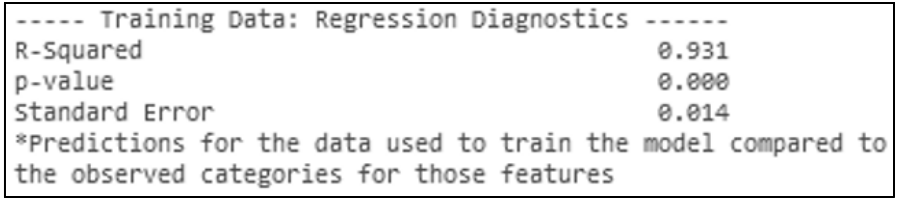
Table 3 Model out-of-bag errors



```
------------ Model Out of Bag Errors ------------
Number of Trees                     1000         2000
MSE                                1.728        1.719
% of variation explained          38.645       38.977
```

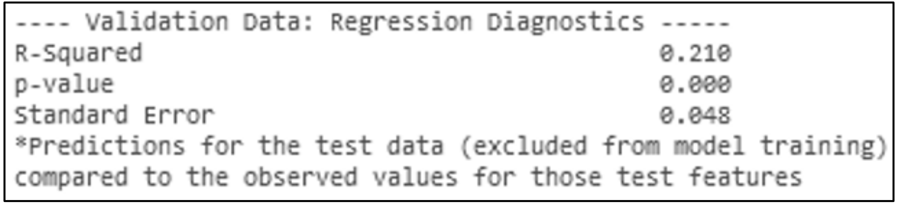
P-value in regression analysis measures the relationship between change in predictor and response variables. Higher P-values means the response variable is insignificant for prediction (Nau, 2019). P-value of 0.05 or lower indicates a significant relationship with predicted outcome (Minitab Blog, 2013). P-value in this analysis is zero (0) for both training and validation data. The standard errors for training and validation data are 0.014 and 0.048 respectively (Tables 4 and 5). These measures are alternatives to R^2 to evaluate the model performance. This is the evidence that the variables used in this analysis were statistically very significant having decent relationship with predicted outcome.

Table 4 Training data regression diagnostics



```
----- Training Data: Regression Diagnostics ------
R-Squared                                     0.931
p-value                                       0.000
Standard Error                                0.014
*Predictions for the data used to train the model compared to
the observed categories for those features
```

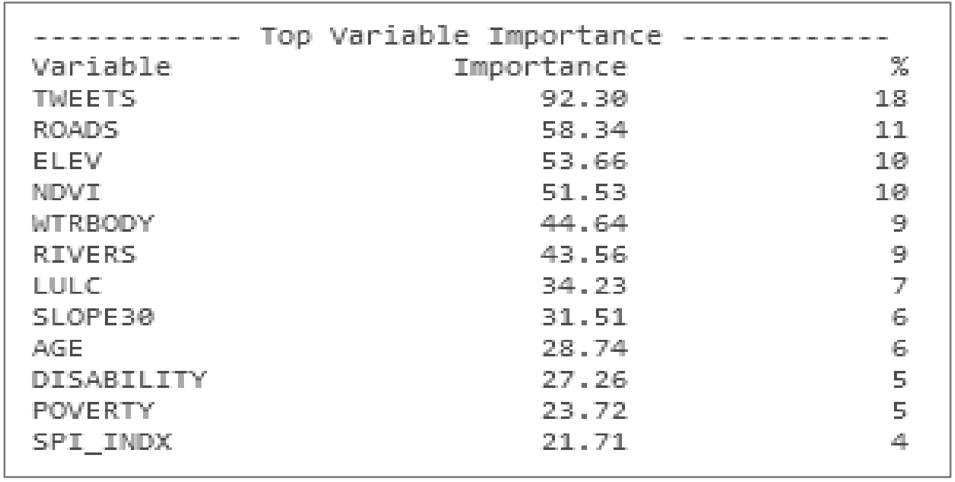
Table 5 Validation data regression diagnostics



```
---- Validation Data: Regression Diagnostics -----
R-Squared                                     0.210
p-value                                       0.000
Standard Error                                0.048
*Predictions for the test data (excluded from model training)
compared to the observed values for those test features
```

Variable importance ranked in Table 6 and Figure 10 demonstrate the contribution of each explanatory variable to predict the vulnerability situation of communities in the study area from hurricane Florence using regression model. Tweets, roads, elevation, and NDVI have highest contribution for predicting the vulnerable communities, whereas water body, land use/land cover, slope, demographic variables, and SPI have moderate contribution, yet not so much insignificant.

Table 6 Variable importance output from the RF regression model



```
------------ Top Variable Importance ------------
Variable                 Importance           %
TWEETS                        92.30          18
ROADS                         58.34          11
ELEV                          53.66          10
NDVI                          51.53          10
WTRBODY                       44.64           9
RIVERS                        43.56           9
LULC                          34.23           7
SLOPE30                       31.51           6
AGE                           28.74           6
DISABILITY                    27.26           5
POVERTY                       23.72           5
SPI_INDX                      21.71           4
```

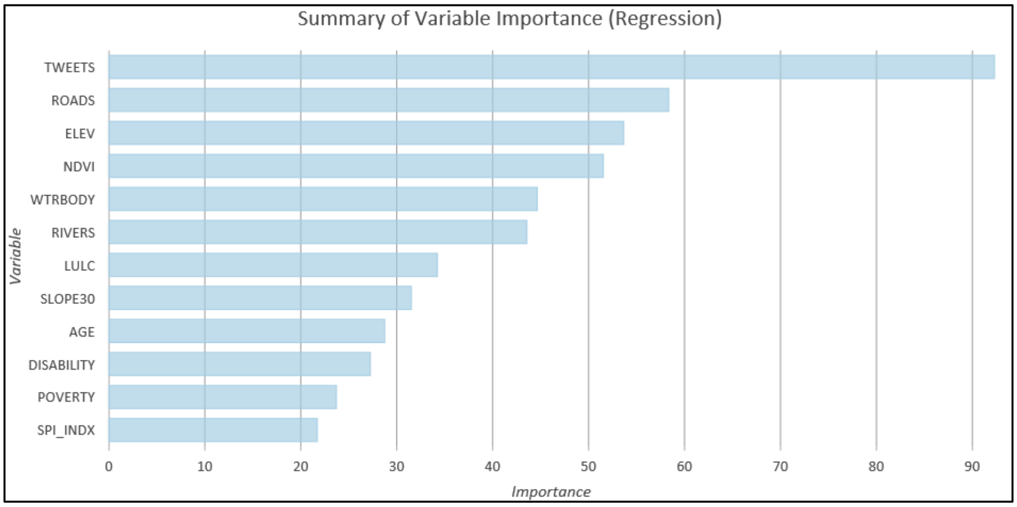


Figure 10 Summary of variable importance from RF regression model

The regression analysis predicted approximately 47 percent census blocks (2311) to category two, 31 percent (1538) to category 3, and the rest to category one, four, and five (Table 6 and Figure 11). Figure 12 shows vulnerability categories by explanatory variables indicating nearly 57 percent of the communities corresponding to the census blocks had highest level of vulnerability, 31 percent communities were moderately vulnerable, and the rest, 12 percent, had lower level of vulnerability to the risk associated to the hurricane Florence. Also, it is evident from the predicted map that generally the areas around the water features (ocean and rivers) and lowland areas have higher vulnerability than the areas away from these features (Figure 11).

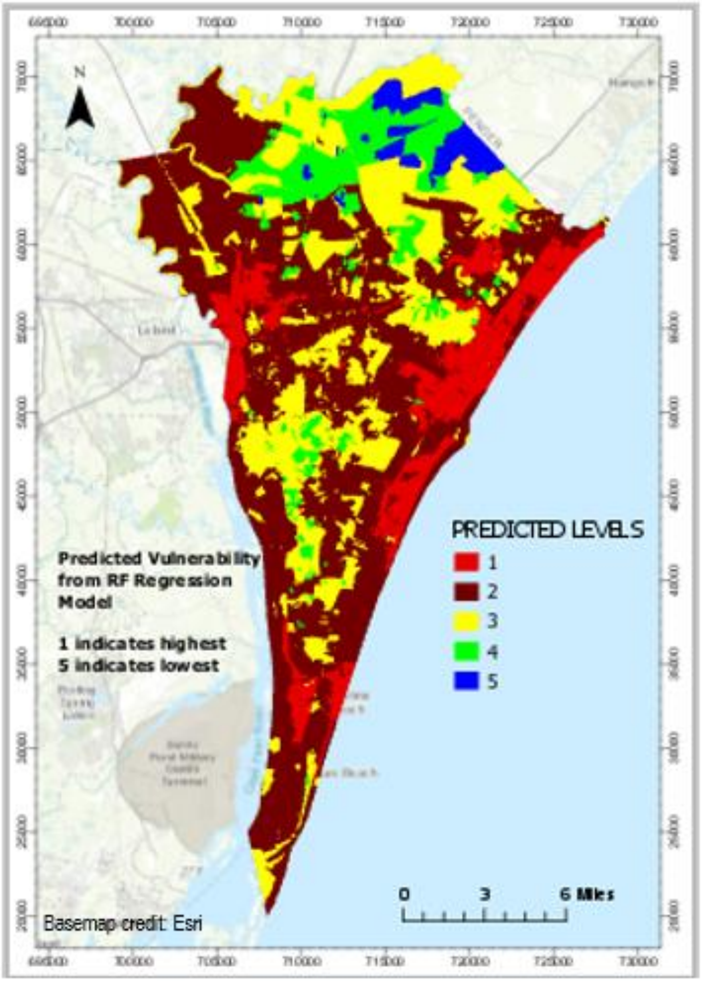
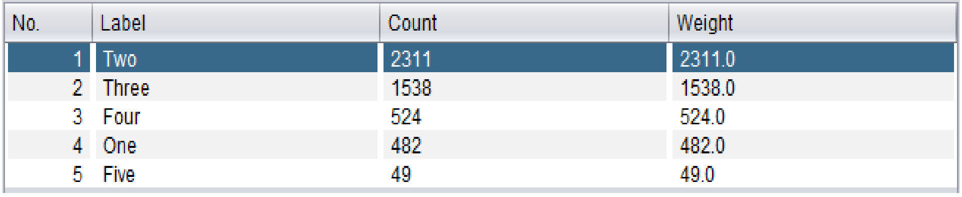


Figure 11 Predicted categories on census blocks from RF regression

Table 7 Predicted categories by number of census blocks from RF regression model



No.	Label	Count	Weight
1	Two	2311	2311.0
2	Three	1538	1538.0
3	Four	524	524.0
4	One	482	482.0
5	Five	49	49.0

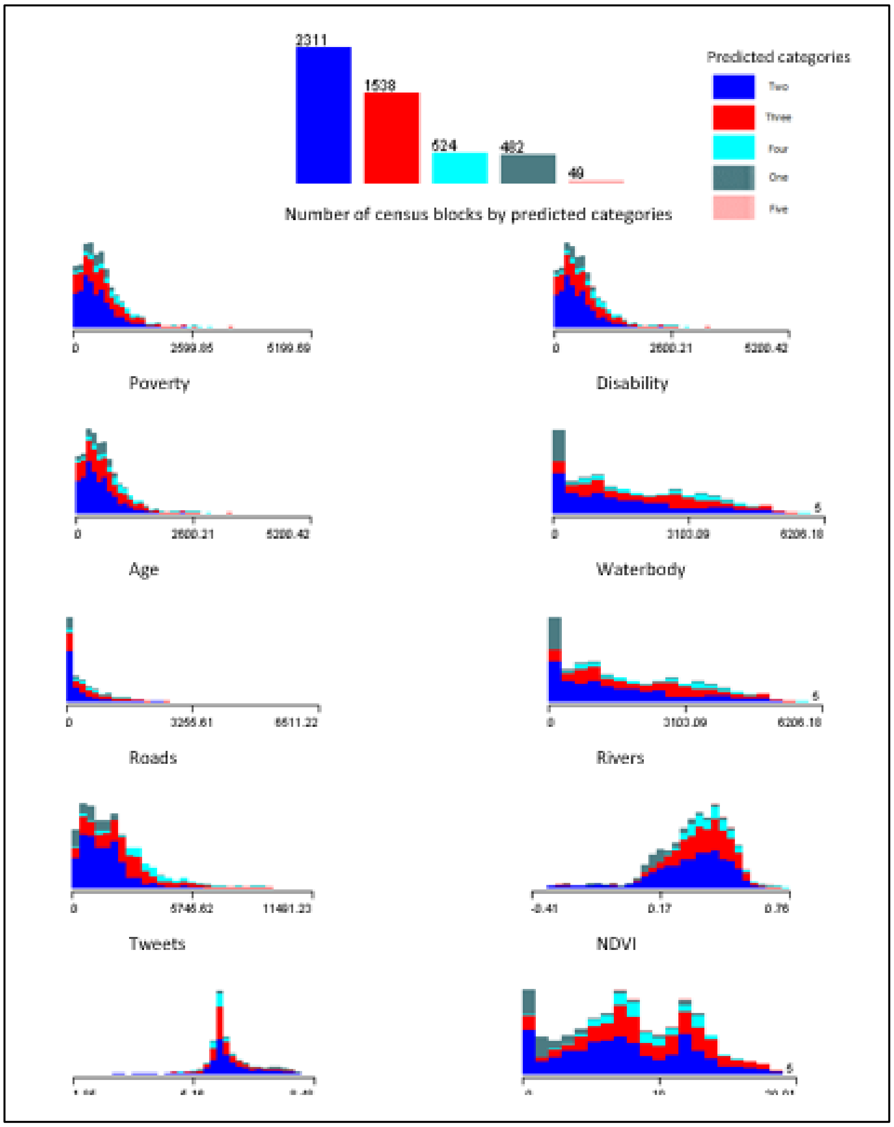


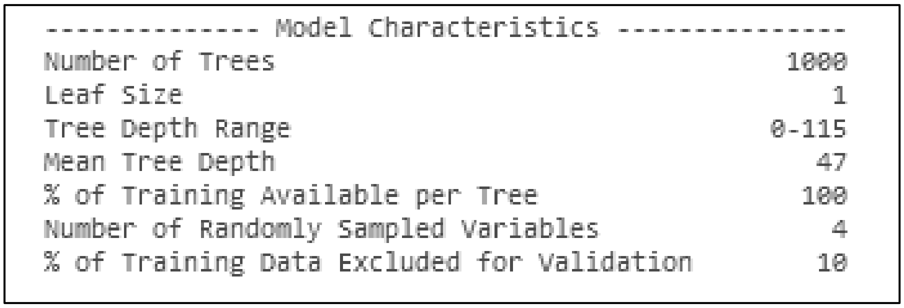
Figure 12 Predicted categories for different explanatory variables on census blocks from the RF regression model

Section 4.2 Classification Results and Analysis

One thousand was found to be optimal number during classification model construction process for decision trees parameter. The prediction from classification model was made to census blocks for predicted vulnerability output corresponding to vulnerability categories in the input training features. Explanatory variables were calculated from distance feature and raster datasets. Ten percent of training data were exclude from training the model for validation. The validation data were used to predict the values of the test data after the model was trained. The predicted values were then compared to the observed values to provide a measure of prediction accuracy based on the variables that were not included in the training process.

Leaf size parameter is the number of observations required to limit a terminal node from further split. Minimum leaf size parameter set for the classification model was 1, i.e., tree stopped growing after it achieved minimum observation of 1 at its terminal node. Tree depth means the number of splits from its root node to terminal node. The tree depths in the forest ranged between 0-115 with a mean tree depth of 47. Number of data available to construct each tree was set to 100%, and number of randomly sampled variables for each tree was 4 (approximately one third of the total number of variables, i.e., 12) (Table 8).

Table 8 Classification model characteristics



```
-------------- Model Characteristics ---------------
Number of Trees                                1000
Leaf Size                                         1
Tree Depth Range                              0-115
Mean Tree Depth                                  47
% of Training Available per Tree                100
Number of Randomly Sampled Variables              4
% of Training Data Excluded for Validation       10
```

Following measures are often used to measure the accuracy of a model and dependability of predicted output in supervised classification problem.

(a) Confidence: It is also referred to as “Precision.” It indicates the proportion of predicted positives that are real positives. This is a measure of accuracy of predicted positive rather than that of true positives (Powers, 2007). Confidence is given by Equation 7:

Equation 7 Confidence

$$Confidence = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

(b) Sensitivity: It also referred to as “Recall.” It is the proportion of true positives that are predicted as positive. It describes the effectiveness of model to predict positive cases as positive (Powers, 2007). Sensitivity is given by the Equation 8:

Equation 8 Sensitivity

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

(c) Accuracy: It measures systematic errors and statistical bias. It is a nearness of a predicted value to an observed value, or it measures how close the predicted values are to the actual values. The best accuracy value is 1. Accuracy is given by the formula in the Equation 9:

Equation 9 Accuracy

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + \ False\ Negative}$$

In this classification model, the accuracy of vulnerability categories for training data were near perfect (0.96) to perfect (1.00) indicating that the predicted outcomes were as expected. Similarly, the accuracy for validation data for vulnerability categories ranged between 0.70 and 0.85 which were not so much far off the training accuracy.

(d) F1 Score: it is the harmonic mean of precision and recall and is used to measure prediction accuracy. F1-score is given by the Equation 10:

Equation 10 F1-score

$$F1\text{-}score = 2.\ \frac{Precion\ .\ Recall}{Precision + Recall}$$

F1-score or harmonic mean of training data for all categories ranged between 0.94 and 1.00, whereas F1-score of validation data for categories 3, 4, and 6 were lower than expected (Table 09 and 10).

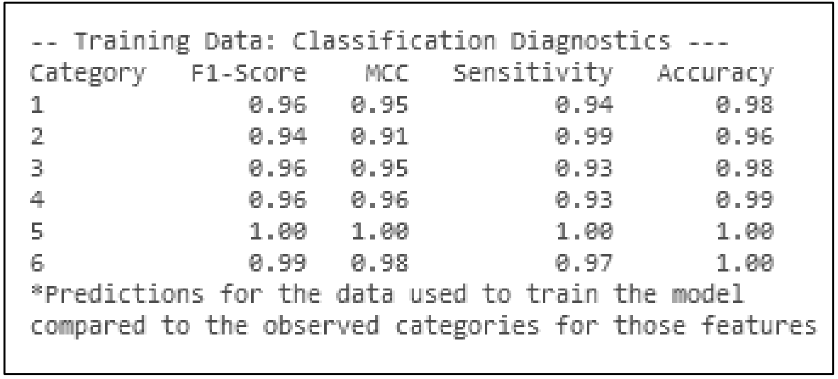
(e) Matthews Correlation Coefficient (MCC): It is a model performance measure of binary classification by taking true positive, true negative, false positive and false negative into account. This is a correlation coefficient between observed and predicted binary classification. This is an appropriate measure of prediction accuracy when there are very imbalanced data with different class sizes (Boughorbel et al., 2017). The values in this measure range between -1 and 1 (-1 indicates total disagreement and 1 perfect correlation) (Liu et al., 2015). The Matthews Correlation Coefficient is given by the Equation 11:

Equation 11 Mathews Correlation Coefficient (MCC)

$$MCC = \frac{TPx\,TN - FP\ x\ FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

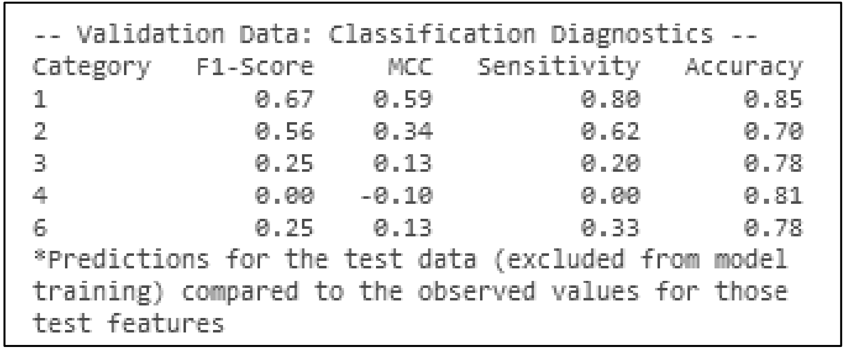
The Mathews Correlation Coefficient for training data ranged between 0.95 and 1.00 in this analysis confirming the high level of correlation between observed data and predicted outcomes for all vulnerability categories. Nevertheless, MCC for validation data for vulnerability for all categories were lower than MCC for training data. MCC for category 4 is -0.18. Despite one negative value, overall correlation of validation data and predicted outcome was high enough to indicate that there was better degree of correlation between observed data and predicted outcome (Tables 9 and 10).

Table 9 Classification model training data diagnostics



```
-- Training Data: Classification Diagnostics ---
Category   F1-Score    MCC   Sensitivity   Accuracy
1              0.96   0.95          0.94       0.98
2              0.94   0.91          0.99       0.96
3              0.96   0.95          0.93       0.98
4              0.96   0.96          0.93       0.99
5              1.00   1.00          1.00       1.00
6              0.99   0.98          0.97       1.00
*Predictions for the data used to train the model
compared to the observed categories for those features
```

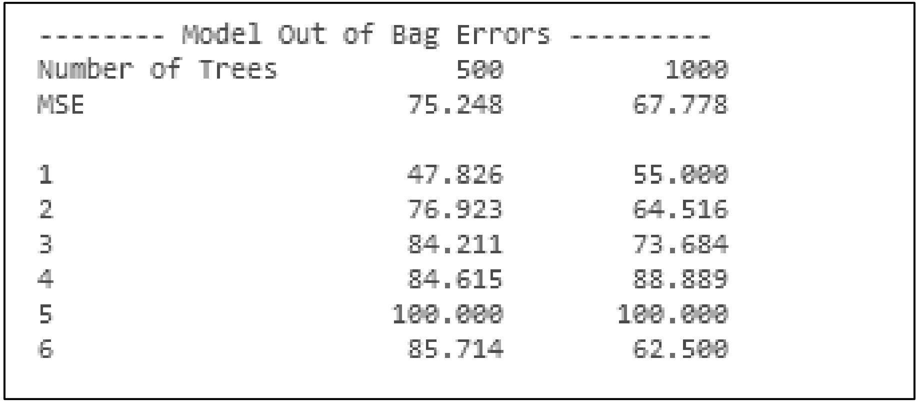
Table 10 Classification model validation data diagnostics



```
-- Validation Data: Classification Diagnostics --
Category   F1-Score     MCC   Sensitivity   Accuracy
1              0.67    0.59          0.80       0.85
2              0.56    0.34          0.62       0.70
3              0.25    0.13          0.20       0.78
4              0.00   -0.10          0.00       0.81
6              0.25    0.13          0.33       0.78
*Predictions for the test data (excluded from model
training) compared to the observed values for those
test features
```

Model “out-of-bag” (OOB) error (Table 11) shows average mean squared error (MSE) and MSE for each vulnerability category for the data that were excluded for trees construction. The average MSE decreased from 75.248 to 67.778 when number of trees were increased from 500 to 1000. The MSE increased for category 4 but MSE for category 5 did not change.

Table 11 Classification out of bag errors

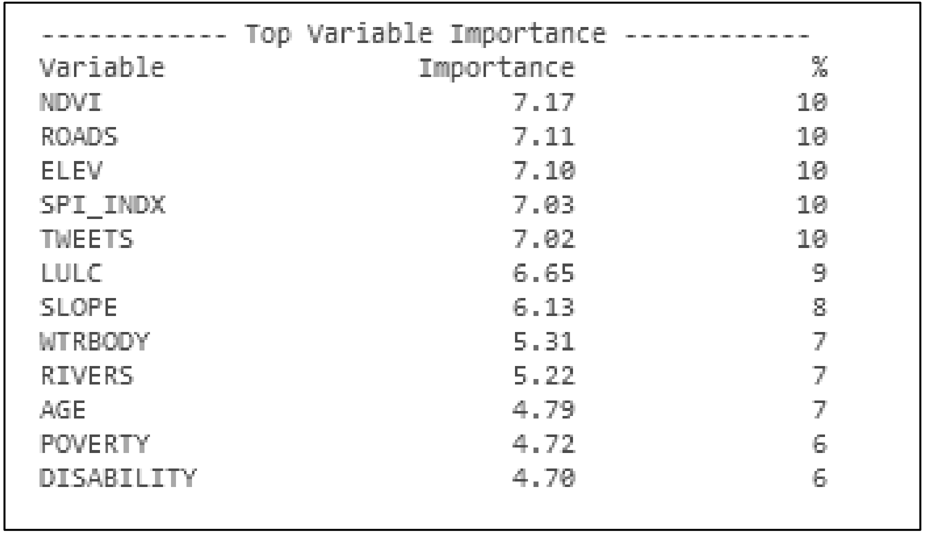


```
-------- Model Out of Bag Errors ---------
Number of Trees             500         1000
MSE                      75.248       67.778

1                        47.826       55.000
2                        76.923       64.516
3                        84.211       73.684
4                        84.615       88.889
5                       100.000      100.000
6                        85.714       62.500
```

Variable importance rank (Table 12 and Figure 13) shows the contribution of each explanatory variable to predict the vulnerability situation in the study area from hurricane Florence using the RF classification model. NDVI, roads, SPI, elevation, and tweets appeared have highest contribution in predicting the vulnerable communities, whereas water body, land use/land cover, slope, and demographic variables had moderate contribution. Despite low importance score, the contributions of demographic variables were not insignificant.

Table 12 Variable importance output from RF classification



```
------------ Top Variable Importance ------------
Variable                 Importance            %
NDVI                           7.17           10
ROADS                          7.11           10
ELEV                           7.10           10
SPI_INDX                       7.03           10
TWEETS                         7.02           10
LULC                           6.65            9
SLOPE                          6.13            8
WTRBODY                        5.31            7
RIVERS                         5.22            7
AGE                            4.79            7
POVERTY                        4.72            6
DISABILITY                     4.70            6
```

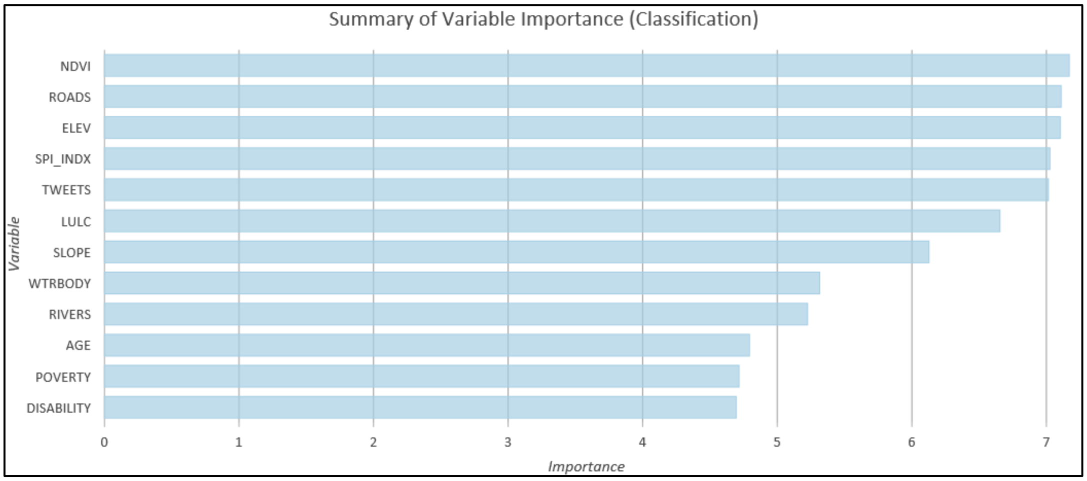


Figure 13 Summary of variable importance from classification model

The classification model predicted approximately 44 percent census blocks (2121) to category one, 27 percent (1339) to category two, and the rest to category three, four, five, and six (Table 13 and Figure 14). Figure 15 shows vulnerability categories by explanatory variables, and it indicates nearly 71 percent of the communities corresponding to the census blocks had highest level of vulnerability, nearly 16 percent communities were moderately vulnerable, and the rest (nearly 13 percent) had lower level of vulnerability to the risk attributed to the hurricane Florence. Also, it is evident from the map generated by prediction that the census blocks that are closer to the water bodies, and lowland areas appeared to have higher level of vulnerability than to the areas away from these physical features (Figure 14).

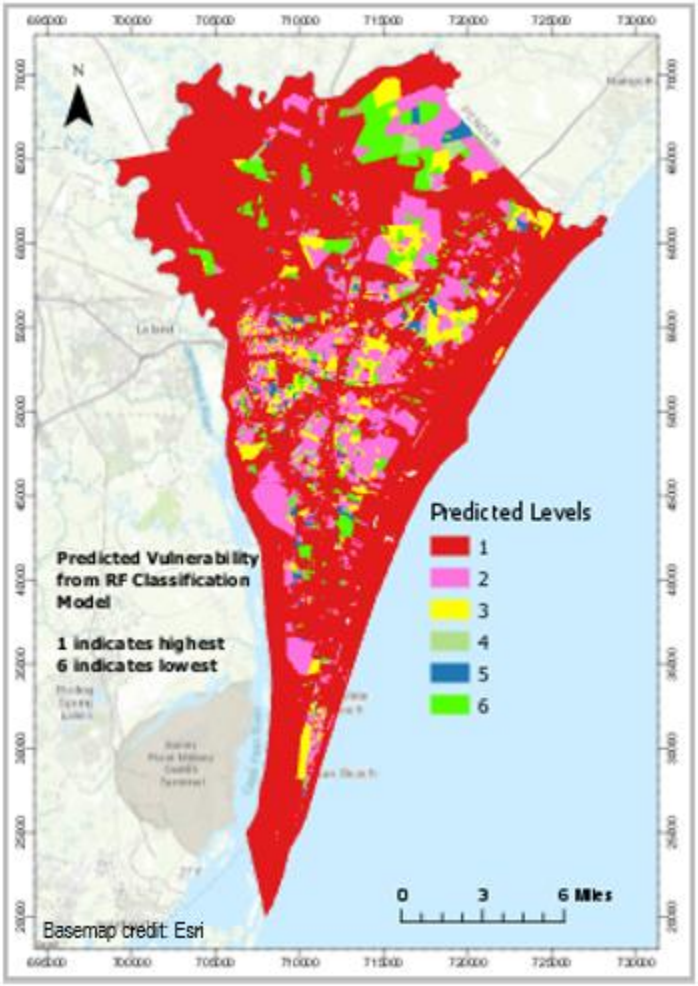
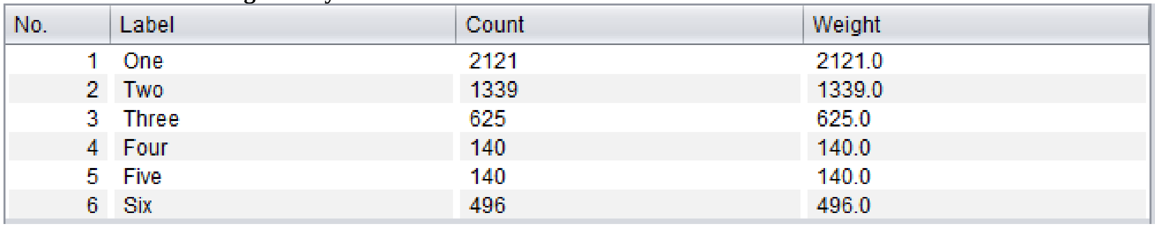


Figure 14 Predicted categories on census blocks from RF regression

Table 13 Predicted categories by number of census blocks from classification model



No.	Label	Count	Weight
1	One	2121	2121.0
2	Two	1339	1339.0
3	Three	625	625.0
4	Four	140	140.0
5	Five	140	140.0
6	Six	496	496.0

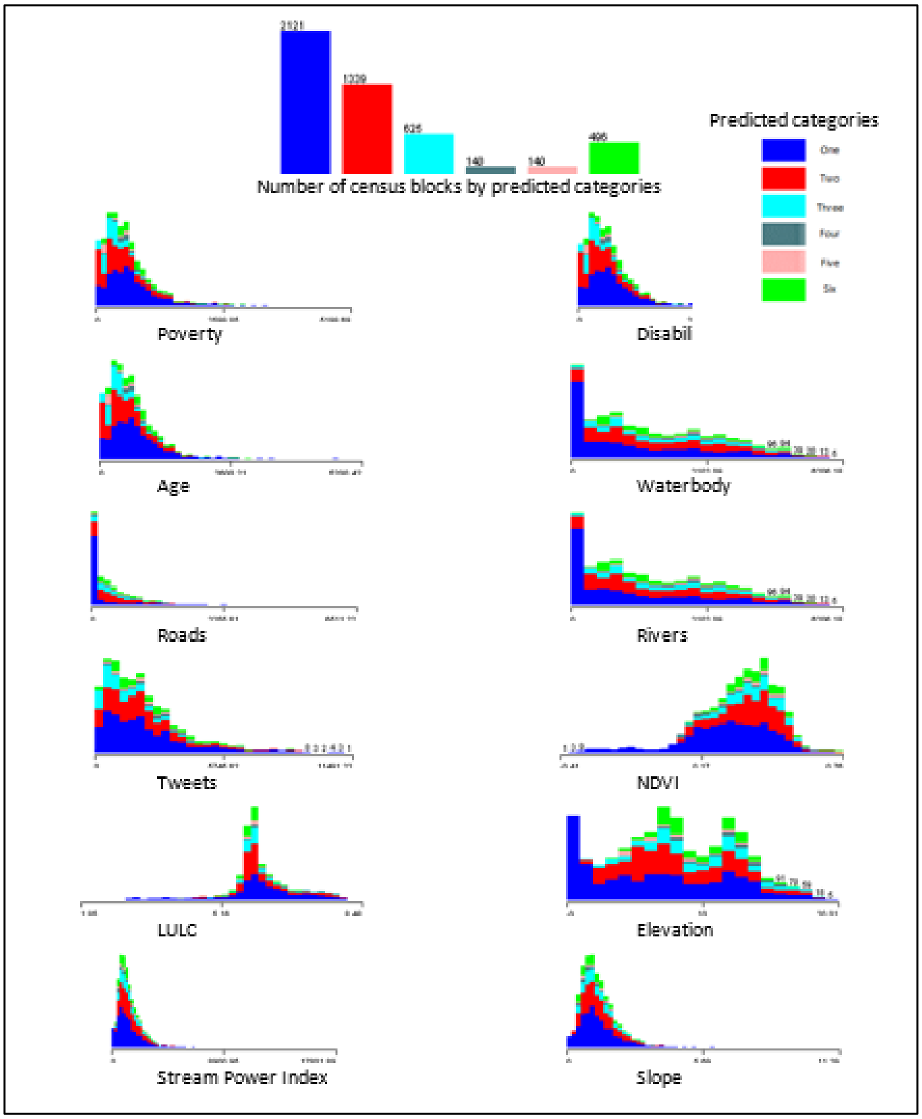


Figure 15 Predicted categories for different explanatory variables on census blocks from RF classification model

Section 4.3 Comparison between Regression and Classification Outputs

The Figure 16 below elucidates the differences between the prediction from RF regression and classification models. In order to compare the results a quantile method was used which distributes the observations equally across the class interval giving unequal class widths but it keeps the same frequency of observation per class.

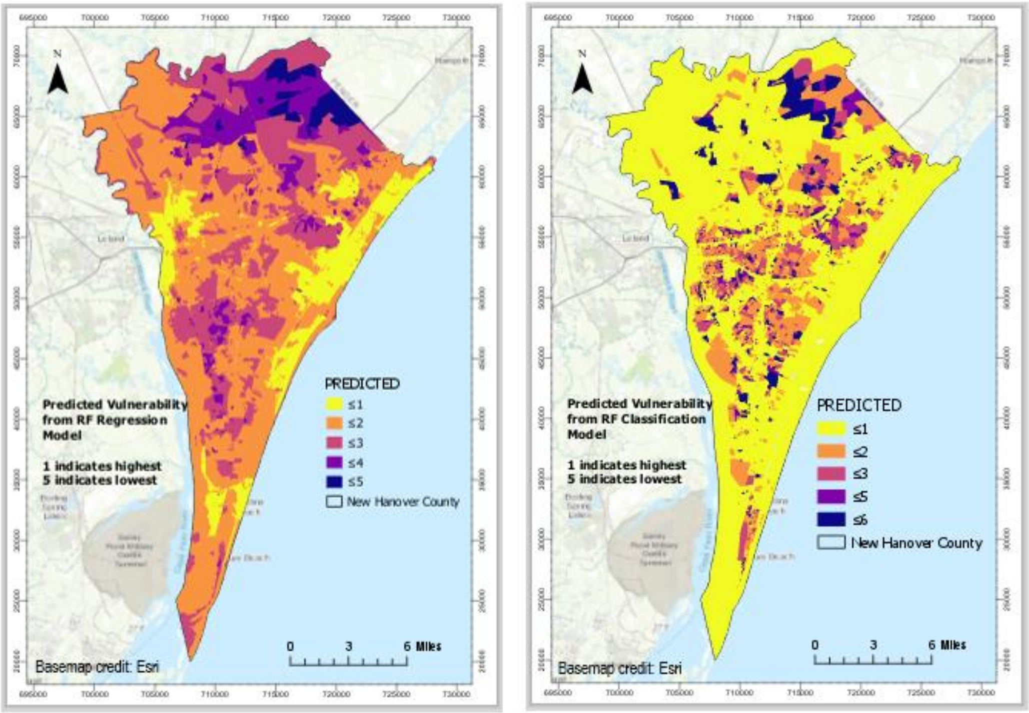


Figure 16 Predicted vulnerability categories: left from RF regression and right from RF classification

This comparison revealed that apparent tendency of both models is to predict the census blocks away from large water bodies and higher elevation to lesser vulnerability categories. The regression model did not predict any census blocks to category 6, it predicted low number of census blocks to category 1, and it predicted higher number of census blocks to category 2. On the other hand, regression model predicted very low number of census blocks to category 4 but higher number of census blocks to category 1. Despite these inconsistencies in the results, both models inclined to predict very high number of census blocks to the higher vulnerability categories and very low number of census blocks to the lower vulnerability categories. Further exhaustive investigation into the behaviors of these regression and classification models is required to understand this dissimilarity in the predicted outputs.

CHAPTER 5 CONCLUSIONS

The trend of extreme hurricane events and frequency are increasing in the Atlantic coastal areas making coastal communities more vulnerable every year. The population in the United States coastal areas is growing that increases the chance of causing loss of more lives and damage of more properties if a hurricane strikes the populated areas. The hurricane Florence made landfall in the New Hanover County, North Carolina as a category one storm and caused at least 24 billion dollar worth of property damage and loss of dozens of human lives. The damage on property and loss of human lives was mainly due to record breaking heavy rainfall and flooding. The area of this study, New Hanover County, is a coastal county comprising approximately 42% of water area. This is one of the main reasons New Hanover County witnessed most dangerous inundation flood resulting to be isolated from the rest of the world for several days.

Geospatial predictive analysis of vulnerability to hurricane hazards were only occasionally preformed using the RF classification and regression modeling in the studies thus far. Moreover, geophysical variables were preferably used rather than combined use of socio-demographic and social media-generated variables to carry out hurricane vulnerability modeling. Given the fairly lack of researches with the use of combination of variables that potentially can better explain vulnerability to hurricane, this work attempted to use demographic and social media-generated in addition to geophysical variables to initiate a new discourse in data modeling for hurricane vulnerability prediction. The objectives were to make categorical prediction and mapping vulnerable communities by the RF machine learning algorithm.

The vulnerability levels of communities vary with the variation in demographic, socio-economic, and physical-environmental conditions of the place, i.e., exposure and coping ability. It is indispensable to consider coupled human-environment system when mapping vulnerability from natural hazards.

Among statistical, physical, and data-driven models used to predict natural hazards, data-driven methods were proved to be the most useful. Thus, machine learning method with combination of geo-physical, demographic and social media-generated variables were used as explanatory variables for predicting vulnerability at the level of census blocks. Land use/land cover, elevation, NDVI, SPI, slope, major roads, major rivers, and water bodies were geo-physical variables; poverty, disability, and age were demographic variables; and tweets posted during hurricane event were social media-generated variable used to feed into the RF classification and regression models. Training data were collected from three different sources: (a) crowdsourced location features with photos from Instagram, Twitter, Facebook and online news media during the hurricane Florence; (b) the New Hanover County designated safe emergency shelter areas; and (c) imagery captured during hurricane event. Total of 273 point locations were used as labelled feature data for model training. The census blocks were used as prediction polygon features since they represented areas with geophysical and demographic similarities.

The RF is extensively used data modeling algorithm in natural hazard risk prediction such as landslides and floods. However, the uses of this modeling technique have been found to be very infrequent in hurricane vulnerability predictions. The RF is a supervised classification and regression method of modeling by growing ensemble of trees and selecting the predicted class by majority count or averaging. The trees grow based upon bootstrap samples, and the "out-of-bag" error rate is calculated using samples out of the bootstrap samples for validation. Variable importance is a fundamental output from the RF because it can be used to evaluate which variables are more useful than others to describe the vulnerability to the disaster event.

For prediction by the RF regression, two thousand decision trees was used as a number of tree parameter. Similarly, three randomly sampled variables for constructing each tree was allowed, and 30 percent data were excluded for model validation. The MSE for number of trees 1000 and 2000 were 1.728 and 1.729 respectively in regression model. It revealed that while doubling the number of trees the error decreased, but not significantly. Therefore, 2000 trees were considered an optimal number. However, the predictive ability did not appear to have increased remarkably by increasing the number of trees. Having R-squared value 0.931, P-value 0.000, and standard error 0.014 showed that variables used were statistically significant having good relationships with the predicted outcomes. Even though R-squared value (0.210) appeared lower than expected, and standard error (0.048) appeared higher for the validation data compared to the data used to train the model, P-value of 0 indicated there was still a better relationship between observed and predicted values. The variables, including tweets, roads, elevation, and NDVI appeared to have high importance for vulnerability prediction from hurricane using the RF regression model.

For classification model, 1000 decision trees was found to be an ideal number. Number of randomly sampled variables were 4 and percent of training data excluded for validation was 10. The classification accuracy of training data for different variables in this model ranged from 0.96 to 1.00, and that of validation data ranged from 0.70 to 0.85. The classification “out-of-bag” errors generally decreased from increasing number of decision trees from 500 to 1000 for most of the vulnerability categories. NDVI, roads, elevation, SPI, and tweets appeared to be the most important variables, whereas age, poverty, and disability are least important variables. Even though the demographic variables were least important, their percent importance values showed that they were not trivial either.

Both regression and classification results showed that geophysical and social media generated variables had higher weight in terms of importance than demographic variables. The communities in the majority of census blocks had highest level of vulnerability, whereas just around one tenth of the communities were least vulnerable in study area from the hurricane Florence. Despite some inconsistencies in results between regression and classification, both models inclined to predict very high number of census blocks to higher vulnerability categories and very low number of census blocks to lower vulnerability categories. Results from regression appeared to be more appealing than result from classification in terms of categorizing the communities to different vulnerability categories.

Conducting predictive analysis for vulnerability to hurricane risks using the RF algorithms for predicting the location of vulnerable communities is highly encouraged in the future works. Community vulnerability to hurricanes should be performed prior to hurricane strikes so that the findings help to reduce the loss. The novel method used in this study may be used to identify the categories of vulnerable communities from various types of natural disasters in the other communities. It is also highly likely that the prediction of vulnerability to hurricanes can be performed for each building in the hurricane affected or potentially affected communities.

REFERENCES

Ai, F., Bin, J., Zhand, Z., Huang, J., Wang, J., Liang, Y., You, L., and Yang, Z. (2014). Application of random forest to select premium quality vegetable oils by their fatty acid composition. *Food Chemistry*, 143:472-478.

Aubrecht, C., Ozceylan, D., Steinnocher, K., and Freire, S. (2013). Multi-level geospatial modelling of human exposure patterns and vulnerability indicators. *Natural Hazards*, 68:147-163.

Bathi, J. R., and Das, H. S. (2016). Vulnerability of coastal communities from storm surge and flood disasters. *International Journal of Environmental Research and Public Health*, 13(239):1-12.

Boughorbel, S., Jarray, F., and El-anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE*, 12(6):e0177678.

Bouwer, L.M. (2019). Observed and Projected Impacts from Extreme Weather Events: Implications for Loss and Damage. In: Mechler R., Bouwer L., Schinko T., Surminski S., Linnerooth-Bayer J. (eds.) *Loss and Damage from Climate Change. Climate Risk Management, Policy and Governance*. Springer, Cham.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5-32.

Brownlee, J. (2019). *Master Machine Learning Algorithms*, Edition v1.13. Retrieved from http://index-of.es/Varios-2/Master%20Machine%20Learning%20Algorithms.pdf (accessed 10/27/2019).

Cohen, D. (2019). About 60.2M Live in Areas Most Vulnerable to Hurricanes. U. S. Census Bureau [Blog Post]. Retrieved from https://www.census.gov/library/stories/2019/07/millions-of-americans-live-coastline-regions.html?CID=CBSM+AC (accessed 10/27/2019).

Colby College. (n.d.). Regression: Patterns of Variation [Blog Post]. Retrieved from https://www.colby.edu/biology/BI17x/regression.html (accessed 11/10/2019)

Cutter, S. L., Barnes, L., Berry, M., Burton, C., Evans, E., Tate, E., and Webb, J. (2008). A place-based model for understanding community resilience to natural disasters. *Global Environmental Change*, 18:598-606.

Feaster, T. D., Weaver, J. C., Gotvald, A. J. and Kolb, K. R. (2018). *Preliminary Peak Stage and Streamflow Data at Selected U.S. Geological Survey Streamgaging Stations in North and South Carolina for Flooding Following Hurricane Florence, September 2018.* U.S. Geological Survey Open File Report, 2018-1172, 36p.

Goodchild, M. F. and Glennon, J. A. (2010) Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3:231-241.

Helderop, E. and Grubesic, T. H. (2019). Hurricane storm surge in Volusia County, Florida: evidence of a tipping point for infrastructure damage. *Disasters,* 43(1):157-180.

Hoque, M. A., Phinn, S., and Roelfsema, C. (2017). A systematic review of tropical cyclone disaster management research using remote sensing and spatial analysis. *Ocean & Coastal Management* 146:109-120.

Hoque, M. A., Phinn, S., Roelfsema, C, and Childs, I. (2017). Tropical cyclone distaste management using remote sensing and spatial analysis: A review. *International Journal of Disaster Risk Reduction*, 22:345-354.

Janitza, S. and Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLOS ONE*, 13(8) e0201904.

Klonner, C., Marx, S., Usón, T., Albuquerque, J.P., and Höfle, B. (2016). Volunteered Geographic Information in Natural Hazard Analysis: A Systematic Literature Review of Current Approaches with a Focus on Preparedness and Mitigation. *International Journal of Geo-Information*, 5(103):1-20.

Lee, S., Kim, J., Jung, H., Lee, M., and Lee, S. (2017). Spatial prediction of flood susceptibility using random-forest and boosted-tree models in Seoul metropolitan city, Korea. *Geomatics, Natural Hazards and Risk*, 8(2):1885-1203.

Liu, Y., Cheng, J., Yan, C., Wu, X., and Chen, F. (2015). Research on the Matthews Correlation Coefficients Metrics of Personalized Recommendation Algorithm Evaluation. *International Journal of Hybrid Information Technology,* 8(1):163-172.

Minitab Blog (2013). How to Interpret Regression Analysis Results: P-values and Coefficients [Blog Post]. Retrieved from https://blog.minitab.com/blog/ adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients (accessed 10/27/2019)

Morrow, B. H. (1999). Identifying and Mapping Community Vulnerability. *Disasters,* 23(1):1-18.

Mosavi, A., Ozturk, P., and Chau, K. (2018). Flood Prediction Using Machine Learning Models: Literature Review. *Water,* 10(1536):1-40.

NAPSG Foundation, GIS Corps, and CEDR Digital. 2018. 2018 Hurricanes Crowdsourced Photos. National Alliance for Public Safety GIS. Retrieved from https://napsg.maps.arcgis.com/apps/StoryMapCrowdsource/index.html?appid=69b95886cf8e49a3a349c9d550174a91 (accessed 12/1/2019).

Nau, R. (2019). Statistical forecasting: notes on regression and time series analysis [Blog Post]. Duke University, North Carolina. Retrieved from http://people.duke.edu/~rnau/411home.htm (accessed 11/28/2019).

Park, S. and Kim, J. (2019). Landslide susceptibility mapping based on Random Forest and Boosted Regression tree models, and a comparison of their performance. *Applied Sciences*, 9(942):1-19

Pourghasemi, H. R. and Kerle, N. (2016). Random forests and evidential belief function-based landslide susceptibility assessment in Western Mazandaran Province, Iran. *Environmental Earth Sciences,* 75(185):1-17.

Powers, D. M. (2007). *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness*. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Richards, J. A. and Jia, X. (2006). *Remote Sensing Digital Image Analysis: An Introduction.* Berlin, Germany: Springer.

Rygel, L., O'Sullivan, D., and Yarnal, B. (2006). A method for constructing a social vulnerability index: an application to hurricane storm surges in a developed country. *Mitigation and Adaptation Strategies for Global Change*, 11(3):741-764.

San Diego State University. (n.d.). Metagenomics Statistics, Random Forests. Retrieved from https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html).

Shalizi, C. (2015). Lecture 10: F-Tests, R^2, and other distractions. Carnegie Mellon University. Retrieved from http://www.stat.cmu.edu/~cshalizi/mreg/15/

Stewart, S. R. and Berg, R. (2019). *Hurricane Florence (AL0620180), 31 August - 17 September 2018.* National Hurricane Center Tropical Cyclone Report, National Oceanic and Atmospheric Administration (NOAA).

Storey, D. (2018). Random Forests, Decision Trees, and Ensemble Methods Explained. Oracle Data Science Blog [Blog Post]. Retrieved from https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods (accessed 10/27/2019).

US Census Bureau. (2018). Quick Facts: New Hanover County North Carolina, Population estimates July 1, 2018 (V2018). Retrieved from https://www.census.gov/quickfacts/fact/table/newhanovercountynorthcarolina/PST045218 (accessed 11/24/2019)

US Census Bureau. (2019). 2019 Tiger/Line Shapefiles: Counties (and equivalent). US Census Bureau. Retrieved from https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2019&layergroup=Counties+%28and+equivalent%29 (accessed 11/24/2019).

Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., and Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527:1130-1141.

Wu, S., Yarnal, B., and Fisher, A. (2002). Vulnerability of coastal communities to sea-level rise: a case study of Cape May County, New Jersey, USA. *Climate Research*, 22:255-270.

Zhang, C., Durgan, S. D., and Lagomasino, D. (2019). Modeling risk of mangroves to tropical cyclones: a case study of hurricane Irma. *Estuarine, Coastal and Shelf Science,* 224:108-116

Zhou, Z., Gong, J., and Hu, X. (2019). Community-scale multi-level post-hurricane damage assessment of residential buildings using multi-temporal air-borne LiDAR data. *Automation in Construction*, 98:30-45.

BIOGRAPHY

Om Dahal received his Bachelor's degree majoring in anthropology from Tribhuvan University in Nepal. He received his Master of Natural Resources (MNR) degree from Virginia Tech in 2013, then he received his Graduate Certificate in Geointelligence (GEOINT) from George Mason University in 2016. He has been employed as a geographer in US Census Bureau for more than two years. Moreover, he is expecting to obtain his Master of Geoinformatics and Geospatial Intelligence degree from George Mason University in December 2019.