Hurricane Strikes Again! Forecasting Power Outages for Tropical Cyclones

Daisy Arokiasamy\textsuperscript{1}, Luis Damiano\textsuperscript{2}, Mai Dao\textsuperscript{3}, Samuel Gailliot\textsuperscript{4}, Akira Horiguchi\textsuperscript{5}, Ramesh Kesawan\textsuperscript{3}, Yiming Xu\textsuperscript{6}

Problem Presenters: Kimberly Kaufeld\textsuperscript{7}, Mary Frances Dorn\textsuperscript{7}, Faculty Mentors: Brian Reich\textsuperscript{8}, Yawen Guan\textsuperscript{8}

Abstract

Severe weather conditions such as hurricanes or tropical storms can damage critical infrastructure and power supplies. Power outages during natural disasters are a major concern to policy makers and responders, who make resource allocation decisions to alleviate the aftermath of such events. It is important when making informed real-time decisions before the storm that highly accurate forecasts of the severity and area of effect of power outages are available and updated as the storms move through regions. Real time forecasting at the county level will allow decision makers and emergency responders to focus their efforts in optimal locations. Forecasting one day ahead provides sufficient time while minimizing error. Many statistical methods have previously been applied to this problem including parametric (GLMs, GAMs) and nonparametric (BART, CARTs, Random Forest) using a wide range of predictors such as tree trimming practices and number of transformers per spatial element. Previous works have focused on a limited number of states and has often made use of proprietary information. Our approach expands on this by increasing the area of focus to include the entire eastern seaboard. In this article, we forecast one day ahead in near-real-time the number of tropical cyclone-induced county-level electrical outages in the Atlantic coast of the United States, using only publicly available data. We provide informative maps for predicting severity of a particular storm and conclude that the log ratio of average daily outages to historical median, as a response, provides the best predictive power among various output transformations.

1 Introduction

Power has become an indispensable part of human life in the modern age. A continuous year-round supply of electricity is often interrupted by natural disasters such as hurricanes, storms and floods. Many utility companies report these power outages, prompting the need to accurately understand and forecast such disruptions. These insights help authorities better

\textsuperscript{1}Department of Mathematical Sciences, Kent State University
\textsuperscript{2}Department of Statistics, Iowa State University
\textsuperscript{3}Department of Mathematics and Statistics, Texas Tech University
\textsuperscript{4}Department of Mathematical and Statistical Sciences, Marquette University
\textsuperscript{5}Department of Statistics, The Ohio State University
\textsuperscript{6}Department of Mathematics, The University of Utah
\textsuperscript{7}Los Alamos National Laboratory
\textsuperscript{8}Department of Statistics, North Carolina State University
prepare to mitigate the impact of these storms. The goal of this project is to forecast one day ahead in near-real-time the number of tropical cyclone-induced county-level electrical outages in the Atlantic coast of the United States, using only publicly available data.

The problem of predicting power outages due to extreme weather relies on choosing a spatial resolution. For example, a city government may only care about handling power outages in their city while the state government or a large utility company will need to focus their efforts over an entire state. Beyond this, emergency responders at the federal level such as the Federal Emergency Management Administration (FEMA) will make allocation decisions over a much wider spatial range and may not need information at the town/city level. Thus, it is important to choose the right tool for the job.

Attempts to predict outages due to extreme weather have historically fallen into two categories: engineering fragility models and statistical models. Engineering fragility models model the probability of failure of a single component as a function of a single demand parameter. This approach was originally used only to predict damage but was later expanded to estimate the number of customers who would be left without power. This approach offers a high level of specificity and works well on high resolution problems but is based on only a single measure of demand on the system. Thus, while city engineers may be the best people to contact in their specific case it is difficult to expand the scope of these methods to the county level. Statistical models, on the other hand, capture a much larger set of driving factors than engineering fragility models due to their top down nature and ability to take in many potential driving factors.

Many statistical models have been developed for estimating the number of outages caused by hurricanes. These models have differed in specifics but have typically been focused on individual states (Connecticut, North Carolina, and South Carolina) with spatial resolutions ranging from 1-10 km². The first published model is Liu et al. Their approach used a negative binomial generalized linear model (GLM) to predict outages in North and South Carolina. Their data consisted of outage numbers, power system inventories, hurricane wind speed, rainfall, types of trees, land cover, and soil drainage levels per geographical area. However, their models included hurricane and company indicator variables, making the model specific to their data set and limiting their model’s ability to make predictions for future hurricanes. Guikema et al. focused on the effects of tree trimming practices using a generalized linear mixed model (GLMM) and the same dataset Liu et al. used previously. Liu et al. attempted to use a spatial GLMM for better inference on variables but did not achieve improved prediction accuracies. Advances were made with better variable selection methods such as the generalized additive models (GAMs) and random forests. Quiring et al. used classification and regression trees (CARTs) to show that some land cover variables are proxies for the power system. Later, Guikema et al. revisited their Outage Prediction Model (OPM) to create the Spatially Generalized Hurricane Outage Prediction Model (SGHOPM) which combined elevation, land cover, soil and vegetation information with the wind characteristics used in their original version. Guikema et al. also studied the possibility of first using classification methods to highlight areas of high risk and then using GLMs to predict outages in those areas. Their hybrid tree/regression approach worked well for simulated data but did not differ significantly from a zero-inflated logistic-GLM method when applied to real world data. Most recently, Wanik et al. have added tree trimming and leaf cover radar (LIDAR) information into their OPM and used a multi-model optimization
approach [5].

This paper extends previous work by (1) forecasting on a finer spatial resolution and (2) using information aggregated at the county level as predictors. This expansion raises new problems. Counties in the lower 48 states range in size from 59.13 km$^2$ for New York County (NY) to 16,040 km$^2$ for Brewster County (TX) with an average of 2,911 km$^2$. Thus, instead of the uniformly proportioned spatial elements used in previous work, we face heterogeneous aerial units. Additionally, the size of the average county is two orders of magnitude larger than the spatial elements previously studied. Thus, we are unable to pinpoint exactly where within a county the majority of outages will occur. Similarly, we are unaware of the true cause of the power outage. Power outages have many causes, including vehicular accidents, which are completely unrelated to weather activity. This means we are looking for patterns in a signal with a low signal to noise ratio. Another difficulty is that our weather data (wind speed, precipitation, and temperature) comes from weather station data stations. Stations are not in every county in the US which means that for many counties we lack first hand accounts. Instead, we rely on simulations obtained with the stormwind R package [2] alongside data obtained by interpolating weather data at the county level using a Gaussian Processes, a spatial interpolation method. Finally, we are predicting the effects of fairly rare events. Between 2015 and 2017 there were only 13 hurricanes and sufficiently powerful tropical storms which made landfall.

In this paper we present two forecasting models: a classification approach and a regression approach. This provides the benefit that multiple different questions are answered. The classification approach works well as a results-oriented indicator of risk wherein counties are split into quantiles based on the predicted severity of the number of outages. The regression approach, on the other hand, is meant to provide information about how important different variables are with respect to the variability of the response as well as providing estimates of the actual number of power outages which will occur in a specific county. In this paper these two approaches will be described in detail with regard to both their similarities and differences, as well as the statistical methods used to implement them. Detailed overviews of the problem, the data, and the feature engineering performed on them are given in section two. This is followed by a discussion of the two approaches, then by section detailing the results of our trials.

2 The Problem Statement

2.1 Description of Problem

The goal of this paper is to forecast, using only publicly available data, one-day-ahead in near-real-time the number of tropical cyclone-induced electrical outages in the counties of the eastern United States.

Of the variety of natural disasters that impact the United States, we chose to study hurricanes and tropical storms due to their well-defined nature, acute damage impact, and abundance of established research in the field. The National Weather Service defines hurricanes and tropical storms as special cases of tropical cyclones. Tropical storms are a rapidly rotating storm system characterized by a low-pressure center, closed low-level atmospheric circulation,
strong winds, and a spiral arrangement of thunderstorms that produce heavy rains. Tropical storms are tropical cyclones with maximum sustained surface wind speeds of 39 mph or higher, where hurricanes are more severe with maximum sustained surface wind speeds reaching 74 mph. Many hurricanes originate in the Atlantic basin from June 1 through November 30. Some, however, occur as early as May.

We also chose to narrow our forecasts to the following states: Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, Pennsylvania, New Jersey, Delaware, Maryland, West Virginia, Virginia, North Carolina, South Carolina, Georgia, Florida, Alabama, Mississippi, Louisiana, Texas and would include Vermont but their data was unavailable. Some inland states are omitted to fully focus on coastal states, where the effects are observed to be most severe in recent years.

2.2 The Data

A large part of the predictive challenge lies in constructing a database with enough information which encompasses all hypothesized factors that are even remotely related to the distribution of outages during hurricanes and tropical storms. In fact, the data engineering stage accounted for 50% of the time invested in this project. This includes combining data from several sources with different reporting standards, such as reporting frequency, aerial definitions, periodicity and data quality.

2.2.1 Data sources

The data for this report has been obtained from various sources. Later on all the data are combined to form a single dataset to analyse. Following are the brief information about the data sources.

- FIPS: Federal Information Processing [18] Standard Publications 6-4 is a 5 digit code which uniquely identifies each county and county equivalents in the United States. The first two digits correspond to the state code and last 3 digits pertains to a county within the state. For example, 01001 corresponds to 01 - Alabama and 003 - Autauga county. We have used FIPS for the lower 48 states (3108 counties) but limit our analysis to only the eastern states (1243 counties).

- Weather: The National Weather Service [23] operates 122 weather forecast offices in total, each of which has a county warning area over which it is responsible for issuing local public, marine, aviation, fire, and hydrology forecasts. The stations’ main tasks are to provide severe weather warnings and to gather frequent weather observations of the local areas. The offices often communicate their broadcasts of weather information with the National Oceanic and Atmospheric Administration Weather Radio All Hazards stations. Because there are not weather stations in every county, we used a Gaussian process with Matérn kernel function to interpolate weather information for counties without known information.

- HURDAT: HURDAT stands for Hurricane Database which contains information about the tracks of the Atlantic and North East & North Central Pacific tropical cyclones
managed by the National Hurricane Center. The database of our interest is the Atlantic Hurricane database referred to as HURDAT2 which includes the best tracks data of the Atlantic Zone dated 1851 to 2018. It is a comma delimitied text format file with six-hourly information on the location, maximum winds, central pressure, and size of known tropical and subtropical cyclones.

- **Population density:** According to the United States Census Bureau, the population density is expressed as the number of people per square mile of land area, calculated by dividing the total US population by the total US land area. As of 2013, the average US population densities vary greatly across the country since the states have different areas of rural and city parts.

- **Tree species:** Trees are important factors to take into account since certain trees are more fragile than others, which makes them fall easily to strong winds, others have large limbs which can break off and take out power lines. In the case of winter storms the leaf cover a single tree provides may work as a good proxy for the amount of snow that it could accumulate in its branches. Fallen trees can immensely disturb electric poles and households causing power outages. Information regarding the different species of trees and their unweighted proportional distribution per total trees in a county. The most common types of trees include Alder, Ash, Birch, Buckeye, Cherry, Cottonwood, Cypress, Dogwood, Douglas Fir, Elm, Gum, Hackberry, Hemlock, Hickory, Juniper, Larch, Laurel, Magnolia, Mangrove, Maple, Oak, Pea, Persimmon, Pine, Spruce, Sycamore, Walnut, Willow, and others.

- **Land Cover Data:** The land cover data is a key component in modeling the outages as certain landforms are susceptible to hurricanes and other natural disasters and serve as a good proxy for the electrical system in certain areas. The weighted proportional distribution of different forms of land namely cultivated areas, forest areas, developed areas, pasture lands, wetlands, barren lands, among others. The data comes from National Land Cover Database (NLCD).

- **Outage data (LANL):** The daily max and average outages data for 3108 counties (continental US) were simulated outages based upon storm events. The data from which the averages were computed were recorded as number of outages per county per 15 minute interval. However, this data set does not include the outages information for all the counties on all the days. It has 2,132,925 records within the data range Nov 1, 2014 to September 16, 2018. We are only interested in the data pertaining to Atlantic hurricane season which is from May through November of every year. 36% of the daily average outages and max daily outages data in this area were missing.

- **Stormwindmodel Package:** We have used the stormwindmodel [2] package in R [21] developed by Willoughby and coauthors (2006), which allows users to model wind speeds at grid points in the US based on best tracks hurricane tracking data (HURDAT2). The package includes functions for interpolating hurricane tracks and for modelling and mapping wind speeds during the storm. Since the package includes population mean center locations for all US counties, we can use it to map winds by counties. We remained
cognizant of the caveat that the accuracy of the storm wind model estimates diminishes as distance from the storm increases. By limiting usage only to the eastern seaboard we aimed to diminish inaccuracies. The overview of the modeling process implemented by this package is as follows [2]:

1. Impute location and maximum wind speeds from the hurricane track data (every 6 hours) to more frequent intervals. The default is to impute to every 15 minutes.

2. For each storm track location, calculate all the inputs needed for the Willoughby wind speed model: forward speed, direction of forward motion of the storm, gradient-level wind speed, radius of maximum winds, parameters for decay of winds away from the storms center for Willoughby model.

3. For each county center, estimate surface-level sustained wind and 3-second wind gusts at all storm observation points. This step includes: measuring distance to county from storm center (radius); calculating tangential gradient wind components at that grid point; calculating gradient wind direction at that grid point; calculating surface wind speed; calculating surface wind direction, adding storm forward motion back into surface wind estimate.

4. Determine for each county: the maximum sustained winds and wind gust speeds at any point on the storms track; the duration of sustained and gust winds over a certain speed (i.e., how many minutes winds were above a cutoff).

5. The model ultimately uses the maximum sustained wind speed as an input to map winds based on counties.

2.3 Feature Engineering Common to Both Approaches

With the advent of big data, which forced a shift in the data analysis paradigm from data scarcity to data flooding, experts have been increasingly placing more importance on feature engineering. In the machine learning literature, a feature is an attribute that describes some aspect of individual data objects. Features are also known as inputs, predictors, regressors, and covariates. Depending on the goal of the model, feature extraction may be driven by their interpretability or their predictive power. Different data types require different kinds of analysis due to structural differences. However, regardless of their types, feature engineering is one of the most essential first steps in data analysis. Feature engineering is a general term encompassing data analysis processes such as feature transformation, generation, extraction, selection, analysis, and evaluation. More information can be found in [7]. This step often involves transformation of collected data such as changes in scales, mathematical transformations, projections, dimension reduction, examination of correlation and interaction, and clustering among many other techniques. For this project, we rely on spatial interpolation, clustering, and principal component analysis.

Each of the classification and regression approaches requires certain separate feature engineering to fit its respective tasks, but some of the performed feature engineering is common to both approaches.
2.3.1 Spatial Interpolation of Weather Data

Environmental datasets typically include geographically referenced and temporally correlated measurements. Daily weather data published by the National Climate and Data Center [1] is collected from 9,135 stations which are located in 1,145 out of the 1,243 (92%) counties conforming our region of interest. Their irregular spatial distribution over the Atlantic area is positively correlated with the population density. The number and the quality of measurements vary across the observational units, with up to 55 different weather variables recorded in total. The number of variables reported by one single station may change across days. Precipitation (tenths of mm), wind speeds (m/s), and temperatures (tenths of degree C) records are reported by 92.3%, 5.1%, and 40.1% of the stations respectively. Data also display errors due to lack of calibration, mis-operation, sensor errors, and recording errors, among others. We use spatial interpolation to account for unreported variables in cases of counties without stations, to deal with counties with multiple stations, and to smooth noisy measurements. In this way, we associate with each missing weather variable of interest at each day with exactly one smoothed value and its measure of uncertainty.

The weather dataset is point-referenced. Let $W(s)$ be a random variable denoting the spatial process at a location $s$, where $s$ varies continuously over the region of interest $D$. We set up a Gaussian process, a nonparametric approach that finds a distribution over all smoothed functions that are consistent with the observed data. In details, it defines a prior over functions which can be converted into a posterior over functions given the observed data. If the observed data used as inputs are similar, for some valid definition of similarity, the output of functions at those points are expected to be similar as well. In this spatial case, the similarity measure has the intuitive definition of spatial distance.

Assume that any two vectors $x, x' \in \mathbb{R}^n$ jointly follow a multivariate Gaussian distribution with some mean and covariance function. Further assume that the latter depends on the distance between them. This relationship is most frequently modeled via a parametric covariance function, which defines the similarity or nearness between data points. This function is also sometimes known as the kernel function, which is in fact a more general definition for a function with two arguments mapping a pair of inputs into the reals. Note that not any arbitrary function with input pairs will, in general, be a valid covariance function; it must be: (1) positive semidefinite, i.e. $\int k(x, x')f(x)f(x')d\mu(x)d\mu(x') \geq 0$ where $\mu$ denotes a measure; and (2) symmetric, i.e. $k(x, x') = k(x', x)$. Commonly used definitions include the exponential, squared exponential, and the Matérn family of covariance functions. The latter is a rather flexible and general family that encompasses many of the simpler options, thus becoming a frequent choice among spatial statisticians. Its formulation is given by

$$C_\nu(d) = \sigma^2 \ Gamma(\nu) \left( \sqrt{\frac{2d}{\rho}} \right)^\nu K_\nu \left( \sqrt{\frac{2d}{\rho}} \right)$$

where $\Gamma$ is the gamma function, $K_\nu$ is the modified Bessel function of the second kind, and $\rho$ and $\nu$ are non-negative parameters of the covariance. It allows for varying degrees of smoothness via the $\nu$ parameter: a Gaussian process with the Matérn covariance function is $\lceil \nu \rceil - 1$ times differentiable, thus matching physical processes realistically. The Matérn kernel simplifies to the exponential kernel when $\nu = 1/2$ and converges to the squared exponential (also known as Gaussian Radial Basis Function, or simply RBF) as $\nu \to \infty$. Analytically
interesting cases include $\nu$ being integer multiples of halves such as $\nu = \frac{3}{2}$ and $\nu = \frac{5}{2}$, where again it simplifies into nice expressions.

Utilizing the properties discussed above, for each weather variable individually at each day, we estimate the parameters of a Gaussian Process with a Matérn covariance function using the available station data. The parameters are estimated via weighted least squares fit with weights equal to $N_j/h_j^2$, where $N_j$ is the number of pairs and $h_j^2$ is the square of the bin spatial distance in the $j$-th spatial bin, using the R package gstat [20, 9]. All the variables are left in the original scale, except for precipitation which is interpolated in the log scale. We estimate the mean predicted value along with the prediction variance at the centroid of every county. Predictive values of precipitation are transformed back to the original scale using the well-known relationship between the first two central moments of the lognormal and normal distributions, namely:

$$\hat{\mu}_y = e^{\hat{\mu}_x + \frac{1}{2}\hat{\sigma}_x^2}, \quad \hat{\sigma}_y^2 = e^{2\hat{\mu}_x + \hat{\sigma}_x^2} \left( e^{\hat{\sigma}_x^2} - 1 \right),$$

where $y \sim \log\text{Normal}(\mu_y, \sigma_y^2)$ is the precipitation and $x = \log(y) \sim N(\mu_x, \sigma_x^2)$ is the log precipitation.

The transformation decisions were made based on visual inspection of empirical semivariograms, some of which are reported for a single day in Figure 1. The semivariance is an autocorrelation statistic defined as half the average squared difference between the values of two points $y_h$ and $y_i$ separated at distance $h$,

$$\gamma(h) = \frac{1}{2N} \sum_{h=1}^{N} \sum_{i=1}^{N} (y_h - y_i)^2$$

where $N$ is the number of points with distance equal to $h$.

The typical pattern is that of an increasing distance until a plateau, which signifies the idea of closer observations having more similar values (high correlation or less variance) than those further apart [3]. Some patterns can be read from the presented semivariograms 1: sill, range, and nugget. The nugget is the variance of any single spatial point and captures the geological microstructure or measurement error. The range, visualized as the distance at which the variogram function reaches the sill, represents the distance at which the spatial autocorrelation dies off. Finally, the sill is the sum of micro and macrostructure error.

In general, from Figure 1 we note that the Matérn covariance function proves flexible enough to accommodate for the different correlation patterns.

The resulting map is shown in Figure 2 which includes both the smoothed predictions as well as the irregularly distributed observations produced by the weather stations.

Here we fit the Matérn model and obtain the semivariogram plot in figure ???. From this plot, we can see that the sill is obtained at around 16, where the range is about 900. The fit is quite adequate in capturing most points on its curve. From the output of the Matérn covariance function, we can see the estimated values of nugget effect being 2.5, the sill 16, and the range 164.8.
## 3 The Classification Model Approach

### 3.1 Description of Classification Problem

When a hurricane or tropical storm hits the continental US, the Department of Homeland Security must decide where to allocate resources to best reduce the storms impact on civilians. These allocation decisions should adapt to the storms path and impact over time. As such, we wish to classify counties based on their probable severity of impact of a storm one day in the future based on the current days weather and land data. Counties in the training set are classified based on their average daily outages, which we use as a measure of impact severity. Classification cutoffs are determined by quantiles of average daily outages across all Atlantic coast states during days which a hurricane or tropical storm hit the continental US. During these dates, all Atlantic coast states are included, regardless of whether they were hit by a storm. Classification cutoffs are summarized in Table 1.

### 3.2 Construction of Test, Validation, and Training Sets

Storms were partitioned into training, validation, and test sets based on their severity. Each set was chosen to have roughly the same distribution of weak and strong storms. Thirteen storms hit the continental US during the hurricane season (May-November) between 2015 and 2017. Hence, we assigned seven storms into the training set, three storms into the test set, and three storms into the validation set. Specifics of storm distribution are in Figure 5.

### 3.3 Models

We train several classification parametric and non-parametric models due to the complexity of our data and to severe class imbalance. In addition, the cost of misclassifying a Very High observation as either Medium or Low is larger than the cost of any other misclassification. Thus, we are interested in classification accuracy as well as where misclassified observations are placed.

In the following section, we briefly introduce the different methods used in the classification model.

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentile Range</th>
<th>Outage Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Below 85th Percentile</td>
<td>Below 75</td>
</tr>
<tr>
<td>Medium</td>
<td>Between 85th and 95th Percentile</td>
<td>Between 75 and 251</td>
</tr>
<tr>
<td>High</td>
<td>Between 95th and 99th Percentile</td>
<td>Between 251 and 998</td>
</tr>
<tr>
<td>Very High</td>
<td>Above 99th Percentile</td>
<td>Above 998</td>
</tr>
</tbody>
</table>

Table 1: Table of classification cutoffs based on average daily outages determined by the percentiles.
3.3.1 Multinomial Logistic Regression

Multinomial Logistic Regression is a method for modeling and classifying counts data into ordered categories, say 1, ..., 4, with increasing severity of outages [25]. By introducing these categories, the original prediction problem turns into a multinomial classification problem, which falls into the realm of Generalize Linear Models (GLM) [19]. The GLM, assumes that the natural parameter $H = (h_1, ..., h_k)^T \in \mathbb{R}^k$ of a parametric family in which the response variable resides, is related to the covariate $X \in \mathbb{R}^m$ in a linear manner: For every $1 \leq i \leq k$,

$$h_i = X^T \beta_i$$

for some $\beta_i \in \mathbb{R}^m$. The $\hat{\beta} = (\beta_1, \cdots, \beta_m)^T$ are estimated using the maximum likelihood estimates from the training data, which is denoted by $\hat{\beta}$. Hypothesis tests on $\hat{\beta}$ as well as its variants can be done using the large sample properties of the maximum likelihood estimate. The prediction of the response variable based on a new covariate $\tilde{X}$ can be obtained from the estimated distribution in the parametric family. Particularly, in our case, the predicted probability for $\tilde{X}$ falling in category $i$ ($2 \leq i \leq 4$) is given by

$$p_i(\tilde{X}) = \frac{e^{h_i(\tilde{X})}}{1 + \sum_{i=1}^{3} e^{h_i(\tilde{X})}}$$

and for category 1 is given by

$$p_1(\tilde{X}) = \frac{1}{1 + \sum_{i=1}^{3} e^{h_i(\tilde{X})}},$$

where $h(\tilde{X}) = (h_1(\tilde{X}), h_2(\tilde{X}), h_3(\tilde{X}))^T$ is the natural parameter in the multinomial model and estimated by $(\tilde{X} \hat{\beta}_1, \tilde{X} \hat{\beta}_2, \tilde{X} \hat{\beta}_3)^T$.

3.3.2 Linear Discriminant Analysis

Linear Discriminant Analysis is a classification method originally introduced by Fisher in [8]. Compared to MLR, LDA has the advantage of being more robust to noise when categories are well-separated. LDA uses an indirect estimation following the Bayes Discriminant Rule. In our case, this means that the probability of observation $X = x \in \mathbb{R}^m$ being assigned to class $i$, $1 \leq i \leq 4$, is given by

$$\mathbb{P}(Y = i|X = x) = \frac{\mathbb{P}(Y = i) f_{X|Y=i}(x)}{f(x)} \propto \mathbb{P}(Y = i) f_{X|Y=i}(x), \quad (1)$$

where $f(x)$ is the density of $X$ and $f_{X|Y=i}(x)$ is the conditional density of $X$ on $Y = i$. The classification for $X$ is therefore given by

$$\arg \min_{i=1,2,3,4} = \mathbb{P}(Y = i|X = x).$$

Note that $Y$ is categorical. Thus, the expression $P(Y = i)$ in Equation (2) corresponds to the probability mass function (or prior distribution) of class $i$, which can be estimated by
the proportion of class $i$'s observations. In LDA, observation $X$ given class $i$ is assumed to be Normally distributed with mean $\mu_i$ and common variance $\sigma^2$. Hence, the expression $f_{X|Y=i}(x)$ is Equation (2) is assumed to be the density function of the distribution $N(\mu_i, \sigma^2)$, where the mean $\mu_i$ can be estimated by the sample mean of the observations in class $i$. The variance $\sigma^2$ is assumed the same for all the categories.

### 3.3.3 $k$-Nearest Neighbors

$k$-Nearest Neighbors [6] is a non-parametric method mainly used for classification. Here the inputs are the $k$ closest training points in the feature space under a given metric. In $k$-NN classification, the output is a class membership by a majority of votes of its neighbors, with the object being assigned to the class most common among its $k$ nearest neighbors where $k$ is a positive integer. If $k = 1$, then the object is simply assigned to the class of that single nearest neighbor. The simplicity and the versatility of this method is the primary motivator for us to consider it in the first place, although the accuracy may not be as efficient as the other supervised learning methods.

The $k$-NN takes 4 inputs namely the training sample, test data, predefined class labels along with the number of nearest neighbors $k$. Ideally the test data set is categorized into one of the four previously defined risk categories. The model is evaluated using the confusion matrix and its attributes with special focus on the third and fourth categories as our goal is to help the authorities in identifying high risk counties to provide immediate assistance and support. After training the model for several $k$’s, the highest prediction accuracy is obtained when $k = 10$ with 52 cases being classified in to the category 4 out of the 69 Very High risk cases. Table 2 below shows the prediction accuracy for various $k$ values.

<table>
<thead>
<tr>
<th>$k$-values</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>73.17%</td>
</tr>
<tr>
<td>8</td>
<td>73.11%</td>
</tr>
<tr>
<td>9</td>
<td>73.77%</td>
</tr>
<tr>
<td>10</td>
<td>74.00%</td>
</tr>
<tr>
<td>11</td>
<td>73.59%</td>
</tr>
<tr>
<td>20</td>
<td>72.99%</td>
</tr>
</tbody>
</table>

Table 2: $k$-NN Classifier Prediction Accuracy for Different $k$

### 3.3.4 Random Forest

Given the possible limitation of the parametric methods, non-parametric techniques such as RF become promising alternatives which can hopefully lead to better results. Briefly speaking, the implementation of RF [13] consists of three steps [4]. The first step is to independently break the training data into smaller subsets using bootstrap. This process is known as the bootstrap aggregation, or ‘bagging’ for short. Each subset gives a partitioning of covariates’ range based on the threshold values of some independently randomly selected covariates.
Points in the same area of partitioning are considered of the same category as the most common category (average in the case of regression) of the training data within it. When a new input arrives, each subset will return a predicted category for it according to where it falls into the partition and RF will choose the one that most subsets return (average in the case of regression). Intuitively, the extra randomness introduced in the first two steps allows RF algorithm to explore the extreme points more often, and the averaging step under independence makes the output predictive in nature. $k$-NN is similar to RF in the sense that test data is predicted by taking local averages, but in a deterministic manner how the locality is chosen. Even though it is known that RF and $k$-NN do not usually work well with time series data when a certain trend is obvious (this can be easily seen from the averaging mechanism of RF and the $k$-NN), it is not quite a big concern for us since our data is categorical rather than numerical.

### 3.3.5 Blind Classifier

The “Blind” classifier, which predicts all observations to be in the Low category, is also included. The blind classifier will correctly classify a large percent of the test set observations simply because of the large number of observations categorized as Low. Obviously, the blind classifier is not a desirable classifier due to it not taking into account predictor values and because it misclassifies every Very High observation. However, it is useful in providing a baseline for the overall performance of the other classifiers.

### 3.4 Results

Figure 6 summarizes the five classifiers’ performance against the validation set.

As shown in Figure 6, the random forest classifier performs best in all categories among all performed methods. In particular, it correctly classifies the Very High observations more often than the other classifiers do, but it is important to see how it assigns the misclassified observations. Now looking at Figure 7, we see that the random forest classifier assigns every misclassified Very High observation to the High category. That is, no misclassified Very High observation is assigned to either the Medium or Low category. Hence, the misclassified Very High observations are at most one category away from the true classification. Other classifiers, such as the $k$-NN classifier ($k = 10$) as shown in Figure 7b, misclassify some Very High observations as Medium or Low.

- The confusion matrices in Figure 7 show how well the two models (random forest and $k$-NN where $k = 10$) have predicted. Non-parametric methods such as RF and $k$-NN tend to perform better than the parametric ones when predicting “Very High” observations. This is consistent with our discussion at the beginning of this section.

- Overall, RF outperforms the other three methods in terms of the average accuracy. This can be roughly explained by the fact that RF is less sensitive to the training data size as long as it is reasonably large, as opposed to the other methods such as MLR and LDA, which strictly rely on some large sample assumptions when the underlying dimension is high. In our case, the dimension of the parameter space is more than 200. Another advantage of RF prediction is that the confusion matrix is tridiagonal, implying that
misclassified data only fall into neighbouring categories of the true one. In practice, this can help policy makers take pre-measures against unexpected events of outage in time.

- One problem concerning the performance of MLR is that overfitting of covariates causes the default algorithm to fail to converge in \( R \) [21]. To resolve this problem, we conducted a greedy-algorithm type variable selection process, based on the magnitude of the relative standard deviation of the estimated parameters. This validity of this method is due to the asymptotic normality of the maximum likelihood estimator. After setting a threshold, it has been found that the accuracy of the model does significantly improve. However, such variable selection process is highly sensitive to the training sets, in the sense that variables selected from training sets of different sizes can vary greatly. This in some sense still implies that MLR may not be the best parametric model to use in this particular case.

4 The Regression Model Approach

4.1 Description of Regression Problem

A complementary approach to the classification methods, regression methods are also tested to predict the magnitude of the impact of a hurricane on the number of outages. We focus on one regression model, random forest, considering different predictor configurations and predicted variables. As the first step in the modeling process, careful consideration is given to the relationship among predictors and different feature engineering techniques aiming at improving the predictive power of the final model. Here we consider mostly decorrelation and unsupervised clustering techniques.

4.2 Construction of Test, Validation, and Training Sets

Storms considered for the regression approach were partitioned into training, validation, and test sets based on their severity levels. Each set was chosen such that each has roughly the same proportions of weak and strong storms. Records showed thirteen storms hit the continental US during the hurricane season (May-November) between 2015 and 2017. Hence, we assigned seven storms into the training set, three storms into the validation set, and three storms into the test set. Specifics of storm distribution are in Figure 5.

4.3 Regression Feature Engineering

4.3.1 Winsorization and dimension reduction

Weather stations record three types measurements related to wind speed: average daily wind speed (AWND), fastest two-minute sustained wind speed (WSF2), and fastest five-second sustained wind speed (WSF5). We winsorize the four most extreme observations. With pairwise correlations greater than 0.80, we apply principal component analysis and retain the first principal component, which explains 97.3% of the variability. Similarly, minimum
and maximum daily temperatures have a sample correlation of 0.89. We apply the same winsorizing technique and retain the two components.

### 4.3.2 Clustering of land usage and tree characteristics data

Here we process county-level data for land usage and tree characteristics resourcing to unsupervised learning for feature extraction. The former is reported by the US Bureau of the Census, reported in 1000 of acres, and tree types data is reported as proportions per species within each county. In both cases, we fill missing data with zeroes. For the tree proportions only, which is recorded as a simplex, we first apply the logarithm transformation (the inverse of the Softmax transformation). We scale to zero mean and unit variance, and run K-means for $K \in \{1, \ldots, 50\}$. We explore the intra-class sum of squared error (elbow method) and compute the BIC as justified below. This methodology was applied independently for trees and land usage.

K-means clustering is performed through the minimization of the intra-class variation, commonly known as K-means total within sum of squares. This quantity is defined as the sum of squared Euclidean distances among data points and their corresponding centroids, i.e. the Euclidean norm. Each observation is then assigned to the cluster corresponding to the closest centroid. Because the target of the optimization problem is a distance function, scaling the observations to zero mean and unit variance is required to avoid different measurement units having an implicit weighing effect. No distributional assumptions are made at this point, but the use of the Euclidean norm implies that each cluster is spherical. The most widely used numerical algorithm is the Hartigan-Wong, which we run using the R programming language [22].

Further assume that the observation vector follows a mixture of Gaussian with $K$ fixed components, that is that all data points are generated from a mixture of finite number of Gaussian distributions. This can be seen as a generalization of $k$-means clustering that incorporates information about the covariance structure of the data. Being a probabilistic clustering method, we can estimate the value of the likelihood function and compute the Bayesian Information Criterion (BIC) to select the number of components. Let $n$ being the sample size, $K$ the numbers of parameters, and $\hat{L}$ the maximized likelihood function value, then the BIC can be computed as

$$\text{BIC} = \ln(n) \times K - 2 \ln(\hat{L}).$$

Figure 1 displays the intra-class variation as a function of the number of components for both variables (also known as the elbow plot). Although increasing the number of components decreases the within-class total sum of squared error, we found no evident cut-off. The BIC is ever increasing in the defined range (not show), which means that the fitting improvements associated with each additional mixture component more than compensates the additional parameters. We then arbitrarily set $K = 6$ in both cases, thus reducing the dimension of 21 different tree types and 14 different land usages to two sets with 5 dummy variables each.
4.4 Results

The purpose of the regression model is twofold. First, we compare different definitions of the random variable associated with the problem of interest to assess its predictability. Second, we compare different sets of inputs to assess their relevance for prediction. We then set up a Random Forest model for each combination and estimate the in-sample and out-of-sample coefficient of determination between the predicted and actual values.

As output variables, we first consider the daily average number of outages $y_{ti} \in \{1, 2, \ldots\}$ at time step $t \in \{1, \ldots, T_i\}$ for the $i$-th county with $i \in \{1, \ldots, I\}$. We estimate the typical number of outages per county $\tilde{y}_i$ by the median of the daily average number of outages computed over the historical time series excluding the days with in-land hurricanes. Finally, we construct the difference $y_{ti}^{(d)} = y_{ti} - \tilde{y}_i$, the ratio $y_{ti}^{(r)} = y_{ti} / \tilde{y}_i$, and the log ratio $y_{ti}^{(l)} = \ln(y_{ti} / \tilde{y}_i)$.

The predictors considered for the random forest fits are as below:

- **Distance**: the distance between the county centroid and the center of the storm, measured in kilometers.
- **Maximum gust speed**: the maximum 10-meter 1-minute gust wind experienced in the county, measured in minutes per second.
- **Gust duration**: the duration of time a certain gust wind was experienced in the county, measured in minutes.
- **Sustained wind duration**: the duration of time a certain sustained wind was experienced in the county, measured in minutes.
- **Tree cluster**: five binary variables identifying the six different tree type based clusters.
- **Land usage cluster**: five binary variables identifying the six different land usage based clusters.
- **Precipitation**: the predicted precipitation obtained via Gaussian process interpolation, measured in tenths of mm.
- **Wind speed principal component**: the first principal component of average wind speed, fastest two-minute wind speed, and fastest five-second wind speed.
- **Temperature principal component**: both principal component of minimum and maximum temperatures.
- **Longitude and latitude**: corresponding to the county centroid.

Based on these, we propose seven different input configurations:

1. **Baseline**: contains all the aforementioned predictors, except for longitude and latitude.
2. **Baseline without tree clustering**: all predictors included in baseline, except for the tree clusters.
3. Baseline without land usage clustering: all predictors included in baseline, except for the land clusters.

4. Baseline with forced interaction: all predictors included in baseline, except for gust duration and sustained wind duration, plus a new variable equal to the product of maximum gust speed and distance to the storm center.

5. Baseline with a dummy: all predictors included in baseline, except for gust duration and sustained wind duration, plus a new binary variable equal to one when gust duration and sustained wind duration is greater than zero.

6. Baseline with forced interaction only: all predictors included in baseline, except for maximum gust speed, gust duration, sustained wind duration, plus a new variable equal to the product of maximum gust speed and distance to the storm center.

7. Baseline with long/lat: all predictors included in baseline, plus the longitude and latitude of the county centroids.

Configurations 2. and 3. are meant to assess the partial effect of the tree and land usage clusters respectively. Configurations 4., 5. and 6. were designed to evaluate possible non-linear effects among storm-specific characteristics. Finally, the last configuration was proposed to evaluate the potential impact of considering the geographical locations.

Tables 3 and 4 respectively summarize the in-sample and out-of-sample estimated coefficients of determination. First, we notice large element-wise discrepancies between these two tables, a strong indication of overfitting. By inspecting the column average on the second table, we observe that most input configurations have approximately similar predicting power on average. By analyzing the mean coefficient across different input configurations, we find that the log ratio is the most predicted variable within the score of our study. Since baseline with longitudes and latitudes performs best as predictors of the log ratio (the best performing model), it would seem that baseline features do not capture all the spatial effects of the data.

Two measures of variables importance are the percentage increase in Mean Squared Error (MSE) and the increase node purity as measured by the Gini index, which are reported in Table 5 for the best performing model. The former is the mean increase in out-of-bag samples predicted MSE when a given variable is excluded from the fitted model, while the latter corresponds to the total increase in node purity resulting from splits over a given variable averaged over all fitted trees. [14] We note that the most important predictors are windPCA and tempPCA1 with approximately equal % increase in MSE, each is about 35%.

In Figure 8, we observe the partial effects of each variable in the best model on the log ratio. The plots show the partial dependence of log ratio based on each of the individual predictors. There are some interesting trends that can be observed. The relationships between the tree and land clusters are linearly related to log ration, either positive or negative. The log ratio has a steady increasing trend as the principal component increases while it decreases exponentially with the distance of storm. There are relationship with large increases at first but soon reach to a ceiling relationships like those of log ratio and precipitation, gust duration, and sustained

---

9Because no estimates of uncertainty were computed, we do not dispose of enough information to make a probabilistic statement about the coefficients being equal or not.
<table>
<thead>
<tr>
<th></th>
<th>Log Ratio</th>
<th>Daily Avg</th>
<th>Ratio</th>
<th>Difference</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>0.43</td>
<td>0.72</td>
<td>0.70</td>
<td>0.51</td>
<td>0.59</td>
</tr>
<tr>
<td>2. Baseline - Tree</td>
<td>0.42</td>
<td>0.71</td>
<td>0.72</td>
<td>0.48</td>
<td>0.58</td>
</tr>
<tr>
<td>3. Baseline - Land</td>
<td>0.42</td>
<td>0.70</td>
<td>0.70</td>
<td>0.47</td>
<td>0.57</td>
</tr>
<tr>
<td>4. Baseline + Interaction</td>
<td>0.43</td>
<td>0.72</td>
<td>0.73</td>
<td>0.50</td>
<td>0.60</td>
</tr>
<tr>
<td>5. Baseline + Dummy</td>
<td>0.43</td>
<td>0.74</td>
<td>0.73</td>
<td>0.47</td>
<td>0.59</td>
</tr>
<tr>
<td>6. Baseline + Int only</td>
<td>0.42</td>
<td>0.72</td>
<td>0.69</td>
<td>0.52</td>
<td>0.59</td>
</tr>
<tr>
<td>7. Baseline + Long + Lat</td>
<td>0.43</td>
<td>0.69</td>
<td>0.74</td>
<td>0.49</td>
<td>0.59</td>
</tr>
<tr>
<td>Avg across inputs</td>
<td>0.42</td>
<td>0.71</td>
<td>0.72</td>
<td>0.59</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 3: Adjusted $R^2$ for the random forest fits for seven configurations and four transformations of the response variable on training set.

<table>
<thead>
<tr>
<th></th>
<th>Daily Avg</th>
<th>Difference</th>
<th>Ratio</th>
<th>Log Ratio</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Baseline</td>
<td>0.10</td>
<td>0.13</td>
<td>0.00</td>
<td>0.12</td>
<td>0.09</td>
</tr>
<tr>
<td>2. Baseline - Tree</td>
<td>0.07</td>
<td>0.12</td>
<td>0.01</td>
<td>0.13</td>
<td>0.08</td>
</tr>
<tr>
<td>3. Baseline - Land</td>
<td>0.09</td>
<td>0.13</td>
<td>0.01</td>
<td>0.09</td>
<td>0.08</td>
</tr>
<tr>
<td>4. Baseline + Interaction</td>
<td>0.03</td>
<td>0.11</td>
<td>0.01</td>
<td>0.07</td>
<td>0.06</td>
</tr>
<tr>
<td>5. Baseline + Dummy</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td>6. Baseline + Int only</td>
<td>0.15</td>
<td>0.07</td>
<td>0.00</td>
<td>0.12</td>
<td>0.09</td>
</tr>
<tr>
<td>7. Baseline + Long + Lat</td>
<td>0.07</td>
<td>0.02</td>
<td>0.01</td>
<td>0.17</td>
<td>0.07</td>
</tr>
<tr>
<td>Avg across inputs</td>
<td>0.08</td>
<td>0.08</td>
<td>0.01</td>
<td>0.12</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 4: Adjusted $R^2$ for the random forest fits for seven configurations and four transformations of the response variable on validation set.

wind duration. There is some unpredictable fluctuating relationship like that of log ratio and the first principal component of temperature. For more details, examine closer at the Figure 8.

5 Results

Overall, we approached the problem in a scale different than what was done before. We expanded the scope from individual states which are focused on in the literature to the whole of the Atlantic coast. Additionally we had the constraint of using only publicly available data as predictors. Sometimes power outages can be viewed in a different angle like risk assessment and transformations of power outages. Realizing that, we approached the problems in two inter-related and complementary procedures: classification and regression. Each method can be expanded on separately depending on the goals and purposes of problem presenters.

The goal of the risk map is to identify counties that should be prioritized. Clever manipulation of logarithm of the outages helps to pinpoint the good separation percentile levels into four different categories. This allows full attention to the very high risk areas to be analyzed.
<table>
<thead>
<tr>
<th></th>
<th>% Increase in MSE</th>
<th>Increase in Node Purity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature PC1</td>
<td>35.67</td>
<td>1482.70</td>
</tr>
<tr>
<td>Wind PC</td>
<td>35.51</td>
<td>2831.21</td>
</tr>
<tr>
<td>Precipitation</td>
<td>30.42</td>
<td>2917.36</td>
</tr>
<tr>
<td>Temperature PC2</td>
<td>28.71</td>
<td>2252.78</td>
</tr>
<tr>
<td>Longitude</td>
<td>27.68</td>
<td>1293.33</td>
</tr>
<tr>
<td>Distance</td>
<td>27.65</td>
<td>1875.12</td>
</tr>
<tr>
<td>Latitude</td>
<td>26.64</td>
<td>1196.67</td>
</tr>
<tr>
<td>Max Gust Speed</td>
<td>24.55</td>
<td>2111.25</td>
</tr>
<tr>
<td>Gust duration</td>
<td>17.54</td>
<td>1211.20</td>
</tr>
<tr>
<td>Tree Cluster 4</td>
<td>16.07</td>
<td>157.59</td>
</tr>
<tr>
<td>Land Cluster 4</td>
<td>16.00</td>
<td>136.94</td>
</tr>
<tr>
<td>Tree Cluster 5</td>
<td>15.63</td>
<td>91.01</td>
</tr>
<tr>
<td>Tree Cluster 6</td>
<td>14.43</td>
<td>183.94</td>
</tr>
<tr>
<td>Land Cluster 5</td>
<td>13.84</td>
<td>136.18</td>
</tr>
<tr>
<td>Land Cluster 3</td>
<td>13.72</td>
<td>102.71</td>
</tr>
<tr>
<td>Tree Cluster 3</td>
<td>12.63</td>
<td>167.05</td>
</tr>
<tr>
<td>Tree Cluster 2</td>
<td>12.54</td>
<td>78.32</td>
</tr>
<tr>
<td>Land Cluster 6</td>
<td>11.89</td>
<td>86.05</td>
</tr>
<tr>
<td>Sustained Wind Duration</td>
<td>10.81</td>
<td>465.93</td>
</tr>
<tr>
<td>Land Cluster 2</td>
<td>10.45</td>
<td>83.41</td>
</tr>
</tbody>
</table>

Table 5: Variable importance of the baseline with longitude and latitude configuration with respect to the log ratio response variable. The percentage increase in MSE and increase in node purity are arranged in decreasing order, with the most important variables at the top of the table.

and classified. Among methods considered, random forest outperforms all the others with not only low misclassification rates, but also provides a conservative misclassified class just one class away of the high class. Based on this result, it can be very helpful to concentrate on regression for the identified counties to quantify the predicted counts, thus gives out best advisory forecasts to resources allocation.

The regression model is the next important step in answering the quantification of the risk and filling in the details needed for decision making. Though predictive powers across the regression fits are weaker compared to risk maps, it is important to realize this is a challenging problems with other things to consider such as low signal to noise ratio. One important result found is that how the output variable is characterized can greatly impact the predictive powers, while different combinations of covariates or configurations may not necessarily have that great effect. This can be viewed even more clearly in the differences in adjusted $R^2$ values for in-sample prediction, giving basis for inference research in effort of improving out-sample forecast accuracy. From the results, we will pay closer attention to the model with all covariates and geographical inputs regressed against the log ratio of daily average outages.

Both approaches provide meaningful insights for what can be improved in answering this
complex problem. It will be important to tie the two methods’ results as improvement and expansion are made thus potentially assessing the risks better.

6 Summary and Future Work

To conclude, both methods of classification and regression have their own advantages and can be used for different purposes. Classification can quickly and accurately give information about which counties are at greatest risk, helping decision makers respond quickly to the situation. Regression, on the other hand, can quantify the estimated number of power outages, instead of simply applying a risk assessment label, and gives information about which variables have the greatest effect on the response, though this task remains challenging. The regression fits can be great for trying to understand the data and the underlying relationships among predictors or patterns of the problems based on good in-sample metrics. However, it performs poorly on the out-of-sample forecast at such low resolution during the one-day ahead window and needs more tuning as well as more informed modeling approach. One goal for future work is to combine the classification and regression approaches into a hybrid model which would first identify counties at greatest risk using classification methods and then forecast the number of outages which would occur in that county. Other goals are including lower intensity storms into the forecasting model. This would require expanding the reference frame both temporally, to capture the weather effects which occur during fall, winter and spring, and spatially, to areas in the United States where winters are more severe and different weather patterns appear. This expansion clearly poses a significant problem as the definition of what counts as an event can be quite nebulous, containing everything from a northeaster to a snowstorm covering only a few counties. Additionally, the numbers of outages would be lower than in the case of a hurricane which would lower the signal to noise ratio even further.
7 Appendix

7.1 Description of the predictors

$AWND_{pred}$: Average daily wind speed (tenths of meters per second)
$AWND_{var}$: Average daily wind speed variance (tenths of meters per second)
$PRCP_{pred}$: Predicted precipitation (tenths of mm)
$PRCP_{var}$: Precipitation Variance (tenths of mm)
$TMAX_{pred}$: Predicted Maximum temperature (tenths of degrees C)
$TMAX_{var}$: Variance of Maximum temperature (tenths of degrees C)
$TMIN_{pred}$: Predicted Minimum temperature (tenths of degrees C)
$TMIN_{var}$: Variance of Minimum temperature (tenths of degrees C)
$WSF2_{pred}$: Predicted Fastest 2-minute wind speed (tenths of meters per second)
$WSF2_{var}$: Variance of Fastest 2-minute wind speed (tenths of meters per second)
$WSF5_{pred}$: Predicted Fastest 5-second wind speed (tenths of meters per second)
$WSF5_{var}$: Variance of Fastest predicted 5-second wind speed (tenths of meters per second)
$popDensity$: Number of people per square mile of land area
$lBARREN$: Ratio of barren land
$lCULTIVATED_{CROPS}$: Ratio of cultivated crops
$lDECIDUOUS_{FOREST}$: Ratio of deciduous forest
$lDEVELOPED_{HIGH}$: Ratio of highly developed
$lDEVELOPED_{LOW}$: Ratio of low developed land
$lDEVELOPED_{MEDIUM}$: Ratio of medium developed land
$lDEVELOPED_{OPEN}$: Ratio of open land
$lEVERGREEN_{FOREST}$: Ratio of evergreen forest
$lGRASSLAND$: Ratio of grassland
$lMIXED_{FOREST}$: Ratio of mixed forest
$lPASTURE$: Ratio of pasture land
$lSHRUB$: Ratio of Shrub land
$lWATER$: Ratio of water body
$lWOODY_{WETLANDS}$: Ratio of woody wetlands
$tALDER$: Proportion of the tree species ALDER
$tASH$: Proportion of the tree species ASH
$tBIRCH$: Proportion of the tree species BIRCH
$tBUCKEYE$: Proportion of the tree species BUCKEYE
$tCHERRY$: Proportion of the tree species CHERRY
$tCOTTONWOOD$: Proportion of the tree species COTTONWOOD
tCYPRESS : Proportion of the tree species CYPRUS
tDOGWOOD : Proportion of the tree species DOGWOOD
tDOUGLAS_FIR : Proportion of the tree species DOUGLAS_FIR
tELM : Proportion of the tree species ELM
tFIR : Proportion of the tree species FIR
tGUM : Proportion of the tree species GUM
tHACKBERRY : Proportion of the tree species HACKBERRY
tHEMLOCK : Proportion of the tree species HEMLOCK
tHICKORY : Proportion of the tree species HICKORY
tJUNIPER : Proportion of the tree species JUNIPER
tLARCH : Proportion of the tree species LARCH
tLAUREL : Proportion of the tree species LAUREL
tMAGNOLIA : Proportion of the tree species MAGNOLIA
tMANGROVE : Proportion of the tree species MANGROVE
tMAPLE : Proportion of the tree species MAPLE
tOAK : Proportion of the tree species OAK
tOTHER : Proportion of OTHER type of tree species
tPEA : Proportion of the tree species PEA
tPERSIMMON : Proportion of the tree species PERSIMMON
tPINE : Proportion of the tree species PINE
tSPRUCE : Proportion of the tree species SPRUCE
tSYCAMORE : Proportion of the tree species SYCAMORE
tWALNUT : Proportion of the tree species WALNUT
tWILLOW : Proportion of the tree species WILLOW
logOutDailyAve : log average daily outages
logOutDailyMax : log maximum daily outages
outDailyAveNorm : Average daily outages normalized by the population density
outDailyMaxNorm : Maximum daily outages normalized by the population density
logOutDailyAveNorm : log daily average outages normalized by the population density
logOutDailyMaxNorm : log daily maximum outages normalized by the population density
logPrcp : log precipitation
logPopDensity : log population density
References


Figure 1: Elbow plot to determine the number of clusters for tree types and land usages. We look at this plot to find $k$ values that minimizes the intra-cluster variation, commonly known as k-means total within sum of squares. We can see that the total within sum of squares is sufficiently minimized at $k = 6$. 
Figure 2: Map of fastest two-minute wind speed (m/s) along with the weather stations. The circle dots represent the weather stations where data is collected. The increasing gradient in red shade shows the faster wind speed, with the numbers of the scale being quantile levels of such wind speeds.
Figure 3: Number of daily mean outages per person per squared mile for hurricane Irma (2017). This is the map that shows the numbers of daily mean outages adjusted for population density, given in number of outages per person per squared mile for the Atlantic basin states. The light grey area denotes the not reported area where no outages data is available, where the darker the purple shade, the more outages reported, where the numbers of the scale are quantile levels of the outages to account for the great differences among areas in respect with outages.
Figure 4: Plot of semivariogram of the outages ratio. From this plot we can get information of nugget, sill, and range, showing that the Matérn is a reasonable fit.
Figure 5: Distribution of storms into training, validation, and test set for the classification model and the regression model.

<table>
<thead>
<tr>
<th>Storm Name</th>
<th>Year</th>
<th>Classification</th>
<th>Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ana</td>
<td>2015</td>
<td>Train</td>
<td>Train</td>
</tr>
<tr>
<td>Bill</td>
<td>2015</td>
<td>valid</td>
<td>Valid</td>
</tr>
<tr>
<td>Bonnie</td>
<td>2016</td>
<td>Train</td>
<td>Train</td>
</tr>
<tr>
<td>Colin</td>
<td>2016</td>
<td>Train</td>
<td>Test</td>
</tr>
<tr>
<td>Eight</td>
<td>2016</td>
<td>Train</td>
<td>Train</td>
</tr>
<tr>
<td>Hermine</td>
<td>2016</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td>Julia</td>
<td>2016</td>
<td>Train</td>
<td>Train</td>
</tr>
<tr>
<td>Matthew</td>
<td>2016</td>
<td>valid</td>
<td>Train</td>
</tr>
<tr>
<td>Cindy</td>
<td>2017</td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td>Emily</td>
<td>2017</td>
<td>Test</td>
<td>Test</td>
</tr>
<tr>
<td>Harvey</td>
<td>2017</td>
<td>Train</td>
<td>Valid</td>
</tr>
<tr>
<td>Irma</td>
<td>2017</td>
<td>Train</td>
<td>Valid</td>
</tr>
<tr>
<td>Nate</td>
<td>2017</td>
<td>valid</td>
<td>Train</td>
</tr>
</tbody>
</table>
Figure 6: Performance of classifiers against the test set. Performance of a classifier for a category is defined as the ratio of the number of observations in that category that are correctly classified to the total number of observations in that category. Overall performance of a classifier is defined as the ratio of the number of observations that are correctly classified to the total number of observations.
(a) Confusion matrix for random forest classifier.  
(b) Confusion matrix for $k$–NN classifier where $k = 10$.

Figure 7: Confusion matrices for classifiers against the test set. Percentages are the ratio of observations that the random forest classifier assigned to the predicted class but actually fall in the actual class to the number of observations that in the actual class. Numbers in parentheses are the number of observations that the random forest classifier assigns to the predicted class but fall in the actual class. Color scale is based on percentages.
Figure 8: The partial effect and dependence of log ratio with respect to different individual predictors. Here the predictors include principal components of temperature, principal component of wind, precipitation, longitude, latitude, storm distance, maximum gust speed, gust duration, sustained wind duration, all the tree and land clusters.