Electric Power Systems Research 146 (2017) 236–245

Contents lists available at ScienceDirect

Electric Power Systems Research

journal homepage: www.elsevier.com/locate/epsr

Using vegetation management and LiDAR-derived tree height data to improve outage predictions for electric utilities

D.W. Wanik[a,*], J.R. Parent[b], E.N. Anagnostou[a], B.M. Hartman[c]

[a] Department of Civil and Environmental Engineering, University of Connecticut, United States
[b] Department of Natural Resources and the Environment, University of Connecticut, United States
[c] Department of Statistics, Brigham Young University, United States

ARTICLE INFO

Article history:
Received 23 October 2016
Received in revised form 27 January 2017
Accepted 31 January 2017

Keywords:
Electric distribution network
Severe weather
Outage prediction model
Vegetation management
Resilience
LiDAR

ABSTRACT

The interaction of severe weather, overhead electric infrastructure and surrounding vegetation contributes to power outages. Given that 90% of storm outages in Connecticut are tree-related, accurate modeling of power outages before a storm arrives could result in improved pre-staging of crews and equipment resulting in improved electric reliability. The authors have generated a light detection and ranging (LiDAR) data product that provides a 1-m resolution measurement of vegetation that is tall enough to strike overhead distribution powerlines, called "ProxPix". These data, along with other vegetation management (e.g. tree trimming) and infrastructure data were evaluated for their improvement an outage prediction model over eastern Connecticut during Hurricane Sandy. A random forest model utilizing a repeated balanced sampling (RBS) approach with 10,000 iterations was used to evaluate which model forcing data accurately predicted the occurrence of a power outage in a 0.5 km grid cell. The authors found that models inputted with infrastructure, vegetation management, ProxPix, performed up to 5–13% better than simpler models depending on model evaluation criteria and input data; and that the model forced with utility infrastructure and ProxPix had the best overall performance. The ProxPix data created for this study have application to other research topics such as prioritizing areas for vegetation management near utilities and providing data on potential tree threats to roads, railways, or other infrastructure networks.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Motivation

Connecticut has been subjected to prolonged power outages due to damage caused primarily by the interaction of trees and overhead power lines during extraordinary storms (i.e. Storm Irene and the October nor'easter in 2011, and Hurricane Sandy in 2012). Hurricane Sandy was especially impactful to Eversource Energy (formerly Connecticut Light & Power), with >500,000 customers without power and >15,000 outages (defined as individual locations requiring the manual intervention of a utility restoration crew for repair) caused mostly by branches or entire trees falling onto the overhead lines [1]. Contributing to these high storm-related outages, Connecticut has the highest wildland-urban interface in the United States [2] with a majority of residents living under a rural or urban tree canopy.

Trees provide a host of benefits including: habitat for wildlife [3], shade that moderates temperatures [4], and aesthetic benefits [5]. Electric utility companies are tasked [6] with managing these trees surrounding overhead lines to maintain acceptable reliability for customers (e.g. limiting the number of interruptions and duration of outages). The management of trees and other flora around overhead lines is known as vegetation management (VM); a multifaceted program of managing trees by trimming above, below and on the side of overhead lines; and the management of vines and shrubs.

1.2. Literature survey

There are many research papers that try to predict the impact of weather on overhead distribution lines using a variety of methods, including multiple linear regression [7], artificial neural networks [8], tree-based models [9,10], and hybrid models where combinations of binary and count data models are used in combination

* Corresponding author.
E-mail address: dave.wanik@uconn.edu (D.W. Wanik).

http://dx.doi.org/10.1016/j.epsr.2017.01.039
0378-7796/© 2017 Elsevier B.V. All rights reserved.

[11]. Among these studies, [12] showed that increasing trimming on distribution circuits could lead to decreased outages on the Duke Power system in the Carolinas. Other researchers [13] have investigated including VM data and other weather and geographic data into outage prediction models and found that the data could help the accuracy of these models. Further, although multiple VM trimming strategies exist (e.g. trim overhead lines every 2–7 years), [14] showed that an optimized vegetation management program as a function of utility cost and customer cost could yield an improvement in reliability (4–6%) and a reduction in total cost (9%).

Understanding the vegetation risk of an infrastructure network can lead to better management of the vegetation. Utilities typically track where VM has occurred on the overhead lines in a geographic information system. Airborne Light Detection And Ranging (LiDAR) data complement VM data by making it possible to develop accurate models of tree heights and locations over large areas. Airborne LiDAR data have been used for more than a decade to model the heights of forest canopies [15]. Canopy height models (CHM) estimate forest canopy height at any given location and make it possible to identify trees that are within striking distance of power lines. The identification of these risk trees provides a direct physical basis for outage prediction models to better incorporate the environmental conditions surrounding overhead power lines.

1.3. Contributions

The objective of this paper is to evaluate the available data on trees and infrastructure for their effect on improving hurricane outage model predictions. Specifically, the authors compare models incorporating these datasets to more traditional models that incorporate only limited environmental data. The authors introduce a newly created 1 m proximal tree pixel dataset that encapsulates the risk of an overhead line to potential vegetation failures, provide detailed information on physical architecture of the distribution network (covered vs. bare wire, backbone and lateral circuit) and evaluate the benefit to a model predicting the occurrence of an outage in a 0.5 km grid cell. The gridded spatial resolution is believed to be the highest of other outage modeling papers in the literature, which the authors hope will provide more detailed results on the processes that contribute to power outages. In addition, the authors present partial dependence and variable importance plots to demonstrate how each variable influences the model predictions. Given the temporal and spatial constraints of the data, the authors focus the paper exclusively on outages occurring during Hurricane Sandy [2012] in eastern Connecticut.

1.4. Paper structure

This paper is divided into six additional sections: Section 2 describes the study area; Section 3 describes the data used in more detail; Section 4 describes the methodology and error metrics; Section 5 presents the results; Section 6 discusses the results; and Section 7 presents the conclusion and future areas of research.

2. Study area

The study area is focused on eastern Connecticut (Fig. 1) due to the availability of LiDAR data over the region. Eastern Connecticut has a diverse landscape with a lowland coastal southern region and a hilly northern region encompassing the Thames River valley. Population is most heavily concentrated along the shoreline and along the Thames River valley. Eversource Energy-Connecticut delivers power to nearly every town in eastern Connecticut except for three which are served by municipal utilities.

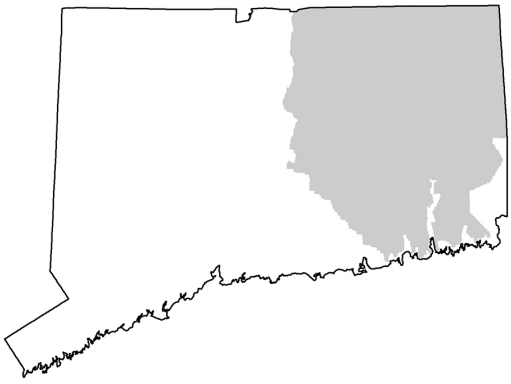

Fig. 1. Distribution of the 4600 km^2 LiDAR data in eastern CT over the study area.

3. Data

The data used in this study (described below) were aggregated using a grid with 0.5 × 0.5 km cell sizes; representing a total of 9000 grid cells covering the study area. All datasets were averaged within each grid cell. Weather simulation data were processed at 2 km spatial resolution using the Weather Research and Forecasting (WRF) model [16], while all other explanatory data were processed at the 0.5 km grid resolution. For this study, the authors used the weather simulation and land cover data [17] for Hurricane Sandy as described in Ref. [9]—see this article for a full description of the weather simulation methodology and validation of winds. Examples of variables from the weather simulation include the maximum gust and wind at 10 m, the total accumulated precipitation, and the duration of wind at 10 m above specific thresholds (i.e. 9, 13, and 18 m/s). To join the 2 km weather data to the 0.5 km aggregated data, the centroid of each 0.5 km grid cell was joined to the nearest 2 km centroid and assigned the corresponding data. See Table 1 for a description of all variables included in the model.

3.1. Utility infrastructure

The authors considered attributes of the conductors related to their circuit material (e.g. bare or covered) and designation (e.g. backbone or lateral) to make the model more physically-meaningful. Conductor material was deemed important because the overhead lines suffer from different types of outages (i.e. incidental touching of trees for bare conductors, destruction of conductors for bare and covered conductors). Circuit designation was included because backbone circuits typically serve many more businesses and essential town functions (e.g. police, fire, ambulance) than lateral circuits, and are expected to be more resilient due to enhanced VM activities from 1994 to 2007 (Personal Communication, Sean Redding, Eversource Energy).

3.2. Vegetation management

Using vegetation management annual planning data for years 2009 through 2012, the authors calculated the percentages of overhead lines that received Standard Maintenance Trimming (SMT) and Enhanced Tree Trimming (ETT) treatment as a function of conductor material and circuit designation for a given year in each 0.5 × 0.5 km grid cell. A linear decay function (Eq. (1)) was applied to the SMT to express the diminishing benefit of such treatment as time passes due to regrowth. A cumulative function (Eq. (2)) was applied to ETT because the benefit of such trimming activities are

Table 1
List of variables included in models.

Variable	Abbreviation	Description	Type	Units	Model 1	Model 2	Model 3	Model 4	Model 5
Duration of wind at 10 m above 9 m/s	wgt9	Weather	Continuous	h	X	X	X	X	X
Duration of wind at 10 m above 13 m/s	wgt13	Weather	Continuous	h	X	X	X	X	X
Duration of wind at 10 m above 18 m/s	wgt18	Weather	Continuous	h	X	X	X	X	X
Continuous duration of wind at 10 m above 9 m/s	cowgt9	Weather	Continuous	h	X	X	X	X	X
Continuous duration of wind at 10 m above 13 m/s	cowgt13	Weather	Continuous	h	X	X	X	X	X
Continuous duration of wind at 10 m above 18 m/s	cowgt18	Weather	Continuous	h	X	X	X	X	X
Duration of wind gusts above 18 m/s	ggt18	Weather	Continuous	h	X	X	X	X	X
Duration of wind gusts above 27 m/s	ggt27	Weather	Continuous	h	X	X	X	X	X
Duration of wind gusts above 36 m/s	ggt36	Weather	Continuous	h	X	X	X	X	X
Duration of wind gusts above 45 m/s	ggt45	Weather	Continuous	h	X	X	X	X	X
Total accumulated precipitation	TotPrec	Weather	Continuous	mm	X	X	X	X	X
Wind stress[a]	WStress	Weather	Continuous	unitless	X	X	X	X	X
Wind gust[a]	Gust	Weather	Continuous	m/s	X	X	X	X	X
Wind at 10 m height[a]	Wind10m	Weather	Continuous	m/s	X	X	X	X	X
Precipitation rate[a]	PreRate	Weather	Continuous	mm/h	X	X	X	X	X
Sum of assets	SumAsset	Infrastructure	Continuous	count	X	X	X	X	X
Total length of overhead lines	TotOHLength	Infrastructure	Continuous	m	X	X	X	X	X
Percentage backbone bare lines	PercBBBare	Infrastructure	Continuous	m		X	X	X	X
Percentage backbone covered lines	PercBBCov	Infrastructure	Continuous	m		X	X	X	X
Percentage of lateral bare lines	PercLATBare	Infrastructure	Continuous	m		X	X	X	X
Percentage of lateral covered lines	PercLATCov	Infrastructure	Continuous	m		X	X	X	X
Percent forested	PERCForest	Land cover	Continuous	%	X	X	X		
Percentage of backbone bare—ETT	ETTValBBBare	Vegetation management	Continuous	%			X		X
Percentage of lateral bare—ETT	ETTValLatBare	Vegetation management	Continuous	%			X		X
Percentage of Lateral Covered—ETT	ETTValBBCov	Vegetation management	Continuous	%			X		X
Percentage of Lateral Covered—ETT	ETTValLatCov	Vegetation management	Continuous	%			X		X
Percentage of backbone bare—SMT	SMTValBBBare	Vegetation management	Continuous	%			X		X
Percentage of lateral bare—SMT	SMTValLatBare	Vegetation management	Continuous	%			X		X
Percentage of lateral covered—SMT	SMTValBBCov	Vegetation management	Continuous	%			X		X
Percentage of lateral covered—SMT	SMTValLatCov	Vegetation management	Continuous	%			X		X
ProxPix per kilometer	ProxPix_km	Hazardous tree pixels	Continuous	ProxPix/km				X	X

[a] Denotes maximum and mean variables calculated.

thought to be longer lasting and more effective than SMT (Personal Communication, Sean Redding, Eversource Energy).

In summary, for each conductor material and circuit designation, the authors calculate the SMT or ETT value per grid cell as follows:

$$SMT\ \ Value\ =\ \sum_{j=2009}^{2012} SMT_j \cdot r \cdot (2012\ -\ j) \tag{1}$$

$$ETT\ \ Value\ =\ \sum_{j=2009}^{2012} ETT_j \tag{2}$$

In the above two equations; SMT and ETT refer to the percentage of lines that were trimmed during the year, respectively; the year is j; the decay rate (25% per year) is *r*. Note that Sandy occurred in 2012 and the earliest date for which trim data were available is 2009. Fig. 2 shows the distribution of trimming by type (ETT vs. SMT) across the entire study area from 2009–2012. Refer to Table 1 for a summary of the different variables related to vegetation management.

Note that the Enhanced Tree Trimming (ETT) specification during 2009–2012 along the roadside differed for backbone and lateral circuits. Regarding ETT roadside clearance for lateral circuits, the clearance zone from the conductor was 8 ft to the side, 20 ft overhead and cut brush flat to the ground; for backbone circuits, the clearance zone from the conductor was clear overhead, 8 ft to the side, and brush was cut to the ground. Exceptions to these specifications could be granted in some cases at the request of a tree owner or town (Personal Communication, Sean Redding, Eversource Energy). Comparatively, the Standard Maintenance Trimming (SMT) specification during 2009–2012 was less intensive. The roadside SMT clearance for lateral circuits was a minimum of 8 ft to the side, 15 ft overhead, 10 ft below clearance within reach of a 55 ft lift unit; for backbone circuits, trimming must re-clear to previous overhead clearances within reach of a 70 ft lift unit.

3.3. LiDAR-derived tree height information

Airborne LiDAR data were acquired for nearly 4600 km^2 in eastern Connecticut from November 3 to December 11, 2010 [18] (Fig. 1). The data were collected with a Leica ALS60 Airborne Laser Scanner at an altitude of approximately 2000 m above ground level. At this altitude, the ALS60's beam divergence of 0.22 mrad creates a footprint of roughly 44 cm on the ground. The scanner's pulse rate was 117.9 kHz and the pulse wavelength was 1064 nm. The flight line overlap was 50% and the data provider eliminated data between the geometrically usable portions of the swaths. The maximum scan angle of the sensor was 16.5° from nadir and it recorded up to 4 returns per laser pulse. The dataset has an overall density of 1.56 returns/m^2 with a maximum point spacing of 0.7 m, excluding water bodies. The horizontal accuracy of the dataset is equal to or better than 1 m RMSE. The project's principle contractor processed the LiDAR data to create a bare-earth digital elevation model (DEM), at a 1 m resolution, with building features removed. Dewberry [18] evaluated the accuracy of the DEM using 62 surveyed ground control points distributed through non-vegetated, grass, and forested terrains. The vertical RMSE for the DEM, based on ground control points, was estimated at 5 cm in non-vegetated terrain, 17 cm for grassy terrain, and 21 cm in forest terrain. The primary purpose of the LiDAR dataset was to develop the bare-earth DEM for use in conservation planning, floodplain mapping, dam safety assessments, and hydrological modeling [18].

3.3.1. Canopy height model

A canopy height model, based on airborne LiDAR, was created following the methods described in Ref. [19]. The CHM was created by subtracting the bare-earth DEM, created by Dewberry [18],

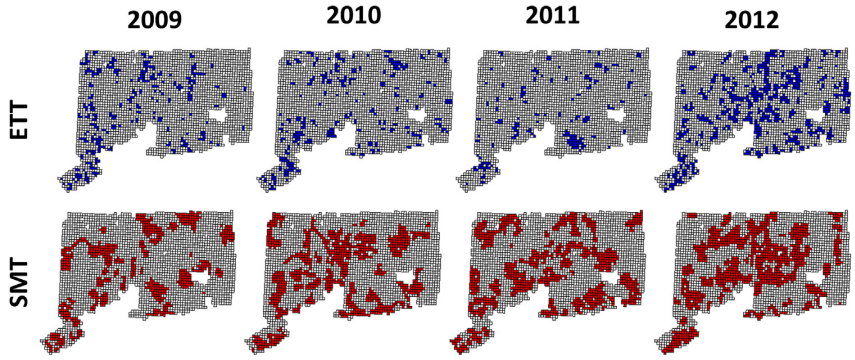

Fig. 2. Comparison of grid cells across Eversource CT service territory. Grid cells with color represent that trimming occurred (does not reflect how intense the trimming was, just that some form of intervention took place).

from a digital surface model (DSM) that the authors created from the LiDAR data. The DSM corresponded to the maximum elevations in the tree canopy and was aligned to the DEM grid and had the same 1 m resolution. The cell values for the DSM were determined by taking the maximum of all non-ground first-return points within a given cell. Because the overall density of the LiDAR dataset was 1.56 returns/m^2, the majority of non-water pixels in the DSM grid contained at least one first-return point. The first-returns were filtered to remove points that obviously did not correspond to features on the earth's surface (e.g. large birds in flight). These anomalous points were identified by comparing each first-return to all points within a 2.5 m radius. Points were discarded if they were more than 30 m taller than any other points within the neighborhood. The authors selected the 30 m threshold because it approximates the upper limit of canopy heights in northeastern forests and thus it represented a reasonable maximum elevation difference for points along forest gaps and edges. Considering the point spacing of our airborne LiDAR data, the authors assumed that continuous data gaps larger than 3 m in radius were likely to correspond to water, which tends to absorb LiDAR energy [20]. The bare-earth DEM values were assigned to the cells in these larger data gaps that were presumed to correspond to water bodies. A test of 52 1 × 1 km sample areas showed that only 12.9%, on average (std. dev. = 3.4), of the areas consisted of data gaps for which there was no first-return data. More than 91% of these gaps were less than 1 m in radius; approximately 7% of the gaps were 1–2 m in radius; and approximately 1% of the gaps were 2–3 m in radius. Thus, the authors interpolated the values for cells in the DSM data gaps, smaller than 3 m in radius, by taking the median of the known values in the cells' eight nearest neighbors. Cells with fewer than three known nearest neighbors were filled using a 2nd or 3rd interpolation pass.

3.3.2. Proximal tree pixels

Previous studies have suggested that mapping of individual tree crowns requires LiDAR data with a spatial resolution of at least 4–8 pts/m^2 [21,22]. Because the resolution of our data (<2 pts/m^2) was insufficient for mapping crowns, the authors used "proximity tree pixels" as a surrogate for risk trees. The authors define proximity tree pixels (ProxPix) as 1 m pixels in a canopy height model (CHM) that are tall enough and close enough to contact a power line in the event of a whole or partial tree failure. Multiple ProxPix can correspond to a single tree; however, intuitively, the authors expected a high correlation between the number of ProxPix and number of trees corresponding to those ProxPix. ProxPix were identified from the canopy height model as pixels with a height larger than the distance from the pixel at ground-level to a point 10 m above ground at the location of the nearest power line (Fig. 3,Fig. 4). The ground-level for a pixel was determined from the bare-earth DEM; 10 m was assumed to be the height of the primary lines above ground level.

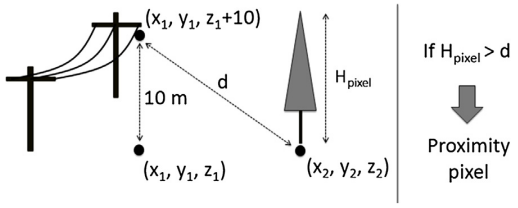

Fig. 3. Calculation of the minimum hazard pixel height for a given location.

The power line dataset had a median horizontal position error of 11.8 m (std. dev. = 9.9 m). Therefore, the authors created a 15 m buffer around the reported power line locations and mapped ProxPix treating the buffer zone as the power line. ProxPix were extracted throughout the study area (Fig. 1) and aggregated into counts for the 0.5 km grid. The counts were then normalized by total overhead line length (resulting in ProxPix/km) for use in the models.

4. Methodology

4.1. Model forcing complexities

To ensure that predictors were not contributing unnecessary complexity in the model structure, five model forcing complexities (Table 1) were evaluated to investigate the added value to the outage prediction model of incorporating: (a) vegetation management, (b) the LiDAR-derived proximal tree pixel dataset and (c) detailed infrastructure. The baseline model (Model 1) was used for comparison to the more complex models and consisted of a full set of weather variables, land cover, and the count of isolating devices per grid cell to represent the overhead infrastructure. This model is most similar to those in Refs. [9] and [10]. Model 2 builds upon the baseline model with the addition of the circuit material and designation data, as well as land cover data. Model 3 builds upon Model 2 with the addition of vegetation management data. Model 4 also builds upon Model 2 with the addition of ProxPix data. Model 5 is the most comprehensive, including detailed infrastructure, vegetation management, and ProxPix data. To compare the benefit of ProxPix and land cover, Models 1–3 used land cover data to represent the local tree conditions while Models 4 and 5 used ProxPix as an alternative to land cover.

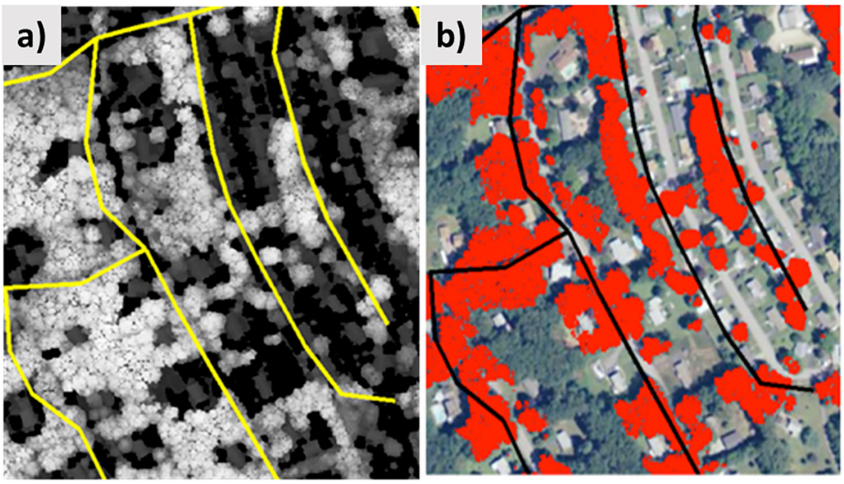

Fig. 4. (a) Canopy height model with 1 m spatial resolution based on LiDAR data, lighter colors indicate taller features; (b) model of ProxPix (red) near power lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Description of random forest algorithm

The statistical software program R [23] was used to complete all modeling and analyses. The R package "randomForest" [24] was used to predict the binary response variable. The authors selected the random forest model due to its efficiency and satisfactory performance in previous literature that predicted hurricane damages [9] and outages [13]. Random forest [25] is an extension of the classification and regression tree ("decision tree") model, and a key benefit of the random forest algorithm is that predictions tend to be more robust compared to simpler models [26]. The decision tree algorithm recursively partitions the training data set into groups of records with similar values for the target variable (Eq. (3)). The tree is grown by conducting for each decision node, an exhaustive search of all variables and all possible splitting values, selecting the optimal split according to the following criteria [27,28].

$$\varphi(s|t) = 2P_L P_R \sum_{j=1}^{\#of\ classes} |P(j|t_L) - P(j|t_R)| \quad (3)$$

where:

$t_L =$ the left child node of node t

$t_R =$ the right child node of node t

$$P_L = \frac{number\ of\ records\ at\ t_L}{number\ of\ records\ in\ training\ data}$$

$$P_R = \frac{number\ of\ records\ at\ t_R}{number\ of\ records\ in\ training\ data}$$

$$P\left(j/t_l\right) = \frac{number\ of\ class\ j\ records\ at\ t_L}{number\ of\ records\ at\ t}$$

$$P\left(j/t_R\right) = \frac{number\ of\ class\ j\ records\ at\ t_R}{number\ of\ records\ at\ t}$$

The optimal split is whichever split maximizes this measure $\varphi(s|t)$ over all possible splits at node t.

Whereas a decision tree makes a series of logical "if-then" statements from a single pass through the training partition to make a prediction, the random forest creates several decision trees using random subsets of the training data records and explanatory variables to fit multiple decision trees [25]. The predictions from all of the decision trees are referred to as the "forest", and the average of the individual tree predictions in the forest are used as the final prediction of the model.

Table 2
Counts of which model was the winner using repeated balanced sampling.

Model	AUC	FOR	TPR	TNR	PC	K
Model 1	658	477	477	2475	658	658
Model 2	1530	1486	1486	1833	1530	1530
Model 3	1646	1754	1754	1700	1646	1646
Model 4	3706	3118	3118	2486	3706	3706
Model 5	2460	3165	3165	1506	2460	2460

4.3. Binary response variable and model evaluation

Outages are defined as locations that require a restoration crew to manually intervene to restore power. Of the 9000 grid cell in the 0.5 km spatial domain, 1320 had one outage, 440 had more than 1 outage, and no grid cell had more than eight outages. Because of this distribution, the authors used binary response models with a balanced sampling approach to investigate the accuracy of outage predictions. An indicator of "1" was assigned if an outage occurred in the grid cell, otherwise a "0" was assigned. Grid cells without an outage (n = 7240) are referred to as the majority class, grid cells with an outage (n = 1760) are referred to as the minority class. The balanced random forest (BRF) algorithm proposed by Ref. [29] consists of down-sampling the majority class to learn from imbalanced by fitting a single model. In order to maximize the information of the majority class, the authors repeated the BRF algorithm 10,000 times and compared the error metrics for each iteration. This approach resulted in a distribution of the error metrics for each of the five model complexities the authors investigated. For each of the 10,000 iterations, the model complexity with the most improved error metrics was selected as the "winner". The frequency of how many times each model was selected the "winner", based on the error metrics, is summarized in Table 2. A flowchart of the modeling framework is provided in Fig. 5.

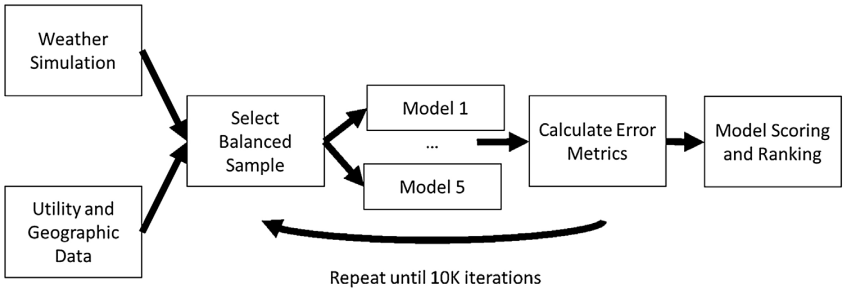

Fig. 5. Flowchart of the algorithm evaluated featuring data preparation, model fit and model scoring.

4.4. Error metrics

The R package "SDMTools" [30] was used to calculate various contingency table metrics to describe the model performance more completely. Specifically, the authors used the following metrics for model comparison: area under the curve (AUC), false omission rate (FOR), true positive rate (TPR), true negative rate (TNR), proportion correct (PC), and Cohen's kappa (K). For the following equations (Eq. (4) through 7); TP (true positive) refers to the counts of true positives; TN (true negative) refers to the counts of true negatives; FP (false positives) refers to the counts of false positives; and FN (false negatives) refers to the counts of false negatives.

$$FOR = \frac{FN}{FN + TN} \tag{4}$$

$$TPR = \frac{TP}{TP + FN} \tag{5}$$

$$TNR = \frac{TN}{TN + FP} \tag{6}$$

$$PC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$K = \frac{2((TP * TN) - (FP * FN))}{(TP + FN)(FN + TN) + (TP + FP)(FP + TN)} \tag{8}$$

Each of the metrics the authors evaluated has desirable properties that describe the model performance. The area under the Receiver Operating Characteristic curve (AUC) is an accuracy metric that shows the model discrimination (the ability to correctly identify true negatives positives); an AUC of 1 represents a perfect prediction whereas an AUC of 0.5 or less represents a test that is no better than chance. The false omission rate (Eq. (4)) is the proportion of false negatives given the test outcome was negative; a FOR of 1 indicates that all predictions for negatives were false whereas a FOR of 0 indicates that there were no false negative. The true positive rate (TPR), also known as sensitivity, is how many times TP was correctly predicted given a positive reading; a TPR of 1 indicates a model is perfect at predicting true positives, a TPR of 0 indicates a model that is incapable of predicting true positives. The true negative rate (TNR), also known as specificity, is a measure of how many times true negatives are actually predicted; a TNR of 1 indicates a model is perfect at predicting where damage will not occur, a TNR of 0 indicates a model that fails to predict where outages will not occur. The proportion correct (PC) is a measure of how many observations are correctly identified; a PC of 1 indicates a perfect predictor, a PC of 0 indicates a predictor that fails to predict where actual outages occur. Cohen's kappa statistic (K), also known as Heidke Skill Score, is a measure of agreement between categorical variables based on the proportion correct. A K of 1 indicates a perfect prediction, and zero or negative values can happen when forecasts are equal to or worse than the reference.

Table 3
Relative improvement in the error metrics compared to baseline model (Model 1).

Model	AUC	FOR	TPR	TNR	PC	K
Model 1	–	–	–	–	–	–
Model 2	132.5%	211.5%	211.5%	−25.9%	132.5%	132.5%
Model 3	150.2%	267.7%	267.7%	−31.3%	150.2%	150.2%
Model 4	463.2%	553.7%	553.7%	0.4%	463.2%	463.2%
Model 5	273.9%	563.5%	563.5%	−39.2%	273.9%	273.9%

5. Results

5.1. Comparison with baseline model performance

The error metrics of Model 1 were used as baseline to compare subsequent models with increasingly complex model forcings (Models 2 through 5). As previously mentioned, the validation strategy used repeated balanced sampling (RBS) with 10,000 iterations to select which model performed the best for each of the five models. The frequency of how many times each model was selected the "winner", based on the error metrics, is summarized in Table 2. Models 2–5 were selected more frequently for most error metrics than the baseline model (Model 1). Model 4 was the most frequently selected for AUC, TNR, PC, and K error metrics, while Model 5 was the most frequently selected for TPR and FOR. Although Model 5 was selected most frequently for these two metrics, Model 4 was the second most selected model. Under the validation strategy the authors used to compare the five model forcing complexities, Model 4 proved to be the best model for the following error metrics: AUC, TNR, PC and K (Tables 2 and 3). Although Model 4 was "better" (more frequently selected) than Model 1 for TNR, it was only marginally better, which may indicate that a simpler model forcing complexity may yield better predictions of the true negatives. The next best model was Model 5, the most complex model, and was most frequently selected for two error metrics (FOR and TPR). To compare model performance, each model was ranked [1–5] across all error metrics based on frequency (where 1 is considered the most frequent and 5 is least frequent). The average of these individual error metric rankings was computed to select the best overall model. Overall, Model 4 was the best model followed by Models 5, 3 and 2 (Table 4). From a physical perspective, the superiority of Model 4 is attributed to the model forcing complexity that included ProxPix, which in turn gave the best representation of local tree risk and more accurate outage predictions (this will be discussed further in Section 7).

5.2. Comparison of error metrics

The kernel density plots in Fig. 6 show the improved performance of Model 4 compared to the Model 1 (baseline model) for all error metrics. The value of AUC, TPR, TNR and PC ranged between

Table 4
Selecting the best model overall using the ranking scheme.

Model	Rank (AUC)	Rank (FOR)	Rank (TPR)	Rank (TNR)	Rank (PC)	Rank (K)	Average rank	Final rank
Model 1	5	5	5	2	5	5	4.5	5
Model 2	4	4	4	3	4	4	3.8	4
Model 3	3	3	3	4	3	3	3.2	3
Model 4	1	2	2	1	1	1	1.3	**1**
Model 5	2	1	1	5	2	2	2.2	2

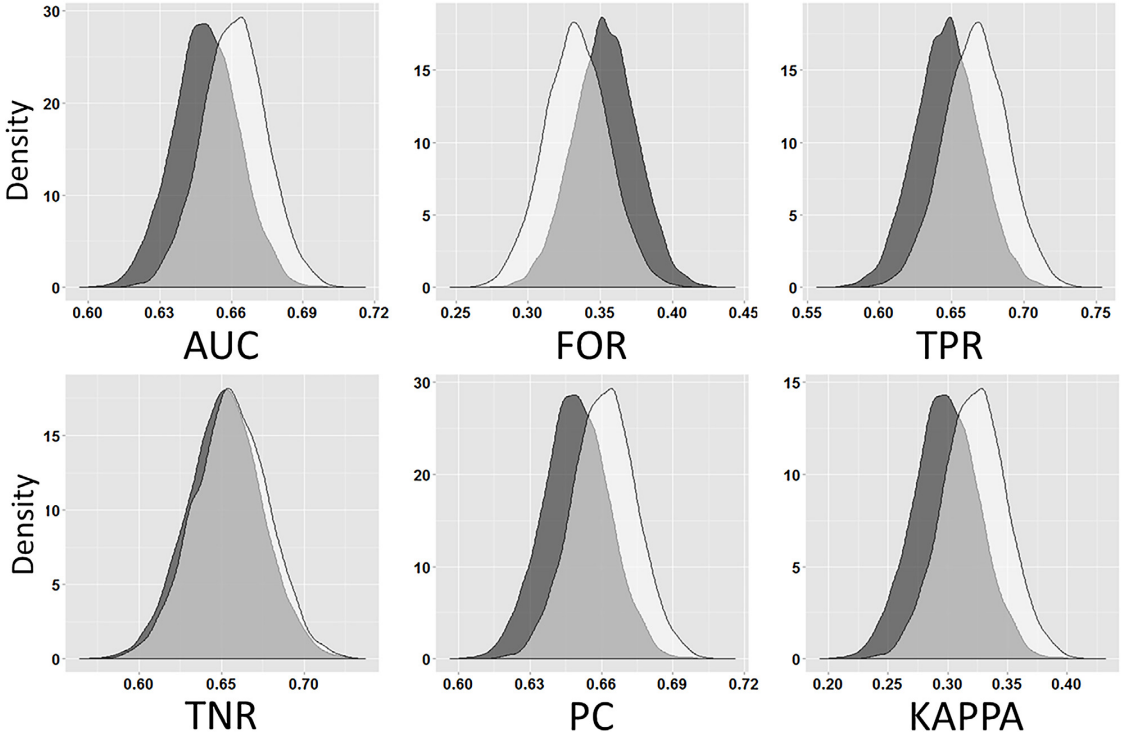

Fig. 6. Panel plot of all contingency metrics evaluated (dark gray = Model 1, white = Model 4, light gray = overlap of Model 1 and Model 4).

0.58–0.72; while K values varied between 0.19–0.43. The maximum improvement for each metric between Models 1 and 4 was 6% for AUC, 5% for FOR, 9% for TPR, 7% for TNR, 6% for PC, and 13% for K. However, the average improvement of Model 4 compared to Model 1 was 1% for AUC, 2% for FOR, 2% for TPR, 1% for PC, 2% for K and no improvement on average for TNR. Fig. 6 shows a clear separation between the kernel density plots for each category except for TNR where the distribution of values nearly overlapped.

5.3. Variable importance and partial dependence plots

The authors now focus on the results of Model 5, as it includes the interaction of all variables of interest related to weather, infrastructure and tree conditions. The variable importance plot in Fig. 7 details the degree to which a specific covariate contributed to Model 5 as measured by the mean decrease in the Gini coefficient. The three most important variables were the total length of overhead lines, the sum of the assets, and the ProxPix normalized by overhead line length. Also among the top ten most important variables were the storm accumulated precipitation, the mean and maximum gust, and the precipitation rate. Figs. 8 and 9 shows partial dependence plots, which details the dependence between the response variable and a specific covariate, marginalizing over the values of all other variables. The covariate of interest was plotted on the x axis, while the y axis provided the change in response variable (in this case, it is the logit of the predicted outages). Variables that contributed to decreased predicted outages included the total overhead length (Fig. 8) and the count of isolating devices (not shown). As expected, increased mean gust and wind at 10 m contributed to increased predicted outages. Interestingly, the ProxPix and percentage of a grid cell with forested land use had nonlinear patterns. The "U-shaped" patterns exhibited in these plots may be influenced by stand-alone risk trees in urban areas that increase the probability of an outage. The plateau that is seen in the normalized ProxPix panel may be attributed to tree sheltering processes in which a neighboring tree can dampen another trees wind exposure, while the increase in forested land use may increase outages by virtue of there being more trees to interact with the overhead power lines.

Although not the most important predictors in the model, the type of circuit, wire and vegetation management treatment applied can influence predicted outage probability. On backbone lines, ETT and SMT on backbone bare and covered lines consistently showed a decrease in predicted outages, while lateral lines had a mixed response depending on conductor material and circuit type.

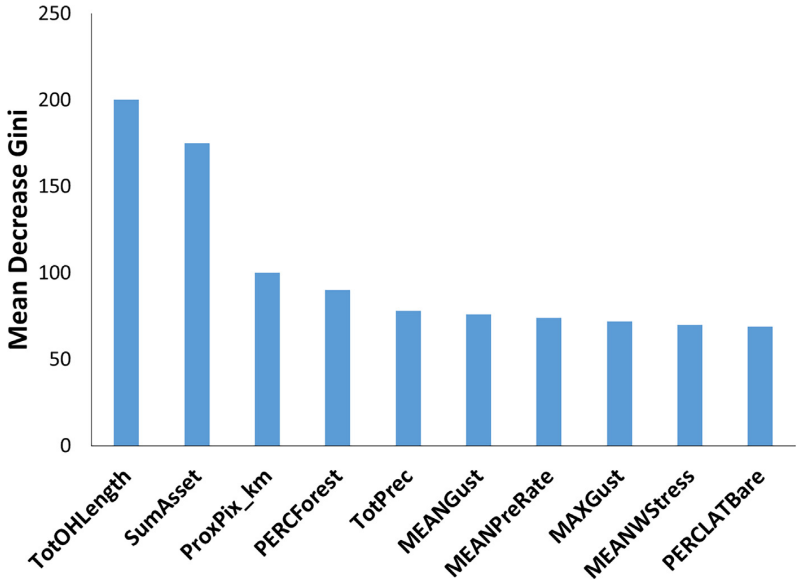

Fig. 7. Variable importance plot from Model 5.

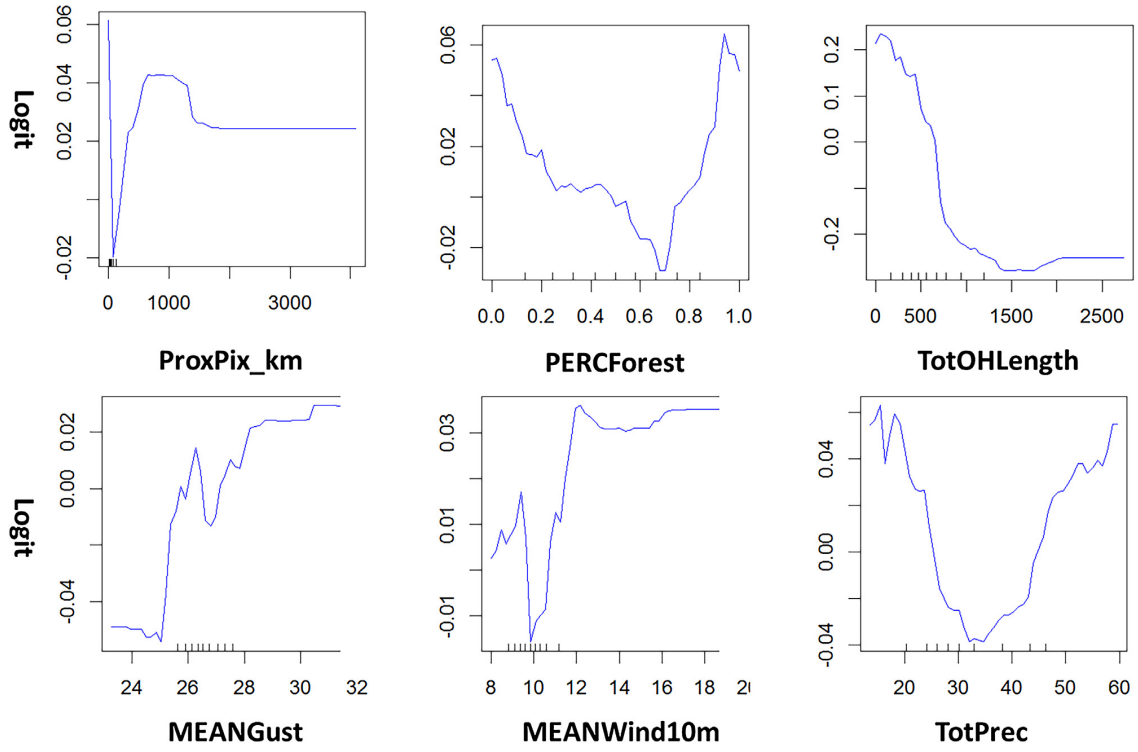

Fig. 8. Partial dependence plots from Model 5 related to infrastructure, tree condition and weather. See Table 1 for a list of abbreviations.

6. Discussion

6.1. Comparison of models evaluated

The results have demonstrated the improvement of including additional tree height, vegetation management and infrastructure data into an outage prediction model with a binary target variable. The use of the repeated balanced sample approach prevented the models from overfitting on a single dataset, and the scoring of the models according to multiple error metrics allowed for a comparison of model performance. As shown by Table 3, all models with more data performed better than the baseline model (Model

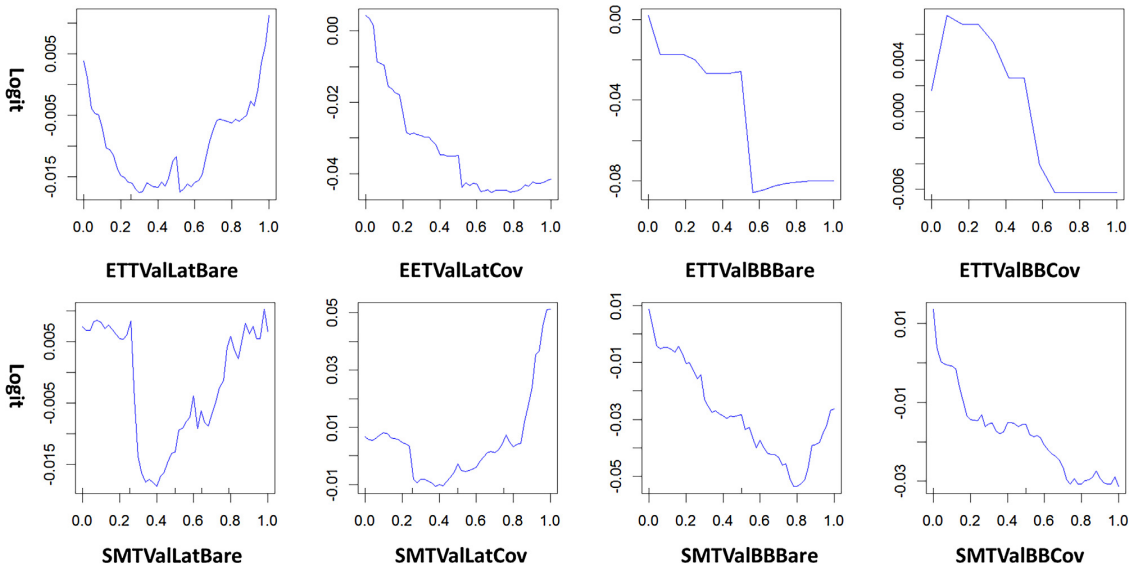

Fig. 9. Partial dependence plots from Model 5 related to vegetation management, circuit and wire type. See Table 1 for a list of abbreviations.

1). A comparison of Model 1 and Model 2 shows that the model with additional infrastructure data on circuit material and designation performed better than the reduced infrastructure model, with nearly a threefold increase in the number of times Model 2 was selected the winner. A comparison of Model 2 and Model 3 suggests that VM data and infrastructure data may be a substitute for tree canopy height data given that Model 3 was selected more frequently than it. However, Tables 2 and 4 show that Models 4 and 5, which included tree height and detailed infrastructure data, were selected most frequently across the majority of error metrics evaluated. These were selected as the up to 61% of the time depending on which error metric was evaluated.

6.2. Comparison to results from the literature

Our results are consistent with other papers in the literature that show that incorporating additional data on infrastructure and environmental conditions can yield at improved outage predictions (read below for further explanation). In particular, this current study supports our previous study [9], which showed that including additional information on land cover and infrastructure can be used to improve the spatial accuracy of outage predictions during hurricanes.

Many recent outage modeling papers that use a grid-averaged approach predict the count of outages per grid cell rather than modeling the probability of an outage occurring [13,31–33]. Despite difference in response variable type, grid cell resolution, geographic region, storms and model forcing, the authors can still draw comparisons to other papers on the influence of certain variables contributing to the predicted outages. Nateghi et al. [13] presented partial dependence plots that demonstrate how VM contributed to increased predicted hurricane outages in two Gulf region states, and also showed that the mean absolute error per grid cell increased 43–53% when outage models were fit without VM data. The authors build on these results and show that VM can contribute to either an increase or decrease in predicted outages depending on the infrastructure attributes. Specifically, the increase of ETT and SMT on backbone lines contributes to lower predicted outages regardless of conductor material; while the pattern varied by treatment, conductor material and circuit designation for lateral lines (Fig. 9).

The benefits of VM during non-storm conditions in the Duke Power System (Carolinas) was previously modeled by [12], who showed that going from a four year to a three year trimming cycle would eliminate 0.9 outages per circuit over a 43 month period. The authors are unable to make the conclusion that VM reduces hurricane outages because our paper is limited by a single region and storm. However, seeing change in predicted outages as a function of conductor material and circuit type is promising (Fig. 9), and provides motivation to incorporate additional data on the overhead infrastructure and local tree conditions in future studies. In addition, building on the previous work by Ref. [34] who showed how storm attributes would contribute to longer restoration times, the authors believe ProxPix may be able to enhance outage duration predictions by providing information on local tree conditions which may be correlated with damage severity (i.e. broken wire vs. broken pole). As time progresses, the authors hope to extend our methodology to analyze how VM, ProxPix and infrastructure contribute to outage predictions from more frequently occurring storms (e.g. thunderstorms and nor'easters)

7. Conclusion

The authors have presented a study that shows that the inclusion of data related to localized tree conditions (ProxPix and vegetation management) and utility infrastructure can improve outage prediction models with a binary response variable. The authors found that more complex models with vegetation management, ProxPix, and infrastructure performed better than the simple baseline model; and that the model forced with utility infrastructure and ProxPix had the best overall performance.

The analysis of partial dependence plots and variable importance plots have shown how various forcing data were interpreted by the models. Similar to other studies [9,13,35], variables related to the infrastructure (total line length and the sum of isolating devices) were among the most important predictors in the model. A key finding of this study was that ProxPix was also shown to be more important than weather-related variables such as wind and rain. Further, vegetation management was shown to decrease the predicted outages on backbone lines across all wire types, but

not so on lateral lines. Although many factors have been considered, the authors have not accounted for all possible factors that contribute to outage predictions, including voltage and the structural integrity of the overhead distribution network, which may enhance the accuracy of outage prediction models. The methodology evaluated in this study may be useful to regions with both similar and different climates given that environmental variables are allowed to be interpreted by the machine learning model, and can learn the unique interactions between weather, environment and infrastructure.

Guikema et al. [36] previously showed how an outage model calibrated with publicly available data in the Gulf region could be applied to show the impact of historic storms in different geographic regions. This approach could over- or underestimate outages as each electric utility has different utility infrastructure and vegetation conditions, which are not well represented by land cover data alone. With the use of LiDAR, the methodology in Ref. [36] could potentially benefit from the quantitative measurement of tree conditions along roadsides across all utilities and be used to enhance these models. In addition, given that the majority of overhead lines follow roadways, LiDAR-derived tree height data near utility lines could also be used for determining which roads might be most vulnerable to downed trees.

Acknowledgements

The authors greatly acknowledge the funding from the Electric Power Research Institute and the data from Eversource Energy. Many thanks to three reviewers who provided helpful comments to improve this manuscript.

References

[1] Connecticut Light & Power Transmission and Distribution Reliability Performance Report, Connecticut Light & Power, Berlin, CT, 2014 http://www.dpuc.state.ct.us/DOCKCURR.NSF/60903cc7b9de44728525746b006e8ffb/027ee9c77ff179fb85257cac006627be?OpenDocument&scrollTop=0&scrollTop=1418
[2] S.I. Stewart, V.C. Radeloff, R.B. Hammer, The Public and Wildland Fire Management: Social Science Findings for Managers, United States Department of Agriculture, Forest Service, Newtown Square, PA, 2006, Report No.: NRS-1 Available from: http://www.fs.fed.us/nrs/pubs/gtr/gtr_nrs1.pdf.
[3] M.L. Cadenasso, S.T.A. Pickett, Linking forest edge structure to edge function: Mediation of herbivore damage, J. Ecol. [Internet] 88 (1) (2000) 31–44 [cited 16 May, 2016].
[4] D. Armson, P. Stringer, A.R. Ennos, The effect of tree shade and grass on surface and globe temperatures in an urban area, Urban For. Urban Green. [Internet] 11 (3) (2012) 245–255 [cited 16 May 2016].
[5] T. Gerstenberg, M. Hofmann, Perception and preference of trees: a psychological contribution to tree species selection in urban areas, Urban For. Urban Green. [Internet]. 15 (2016) 103–111 [cited 16 May, 2016].
[6] IEEE, C2-2007 National Electric Safety Code—PDF Single User Professional Edition: VuSpec (NESC). [Internet], 2007 https://doi.org/10.1109/ieeestd.2006.322219.
[7] B.J. Cerruti, S.G. Decker, A statistical forecast model of weather-related damage to a major electric utility, J. Appl. Meteorol. Climatol. 51 (2) (2012) 191–204.
[8] A.I. Sarwat, M. Amini, A. Domijan Jr., A. Damnjanovic, F. Kaleem, Weather-based interruption prediction in the smart grid utilizing chronological data, J. Mod. Power Syst. Clean Energy [Internet] 4 (2) (2016) 308–315 [cited 26 January, 2017].
[9] D.W. Wanik, E.N. Anagnostou, B.M. Hartman, M.E.B. Frediani, M. Astitha, Storm outage modeling for an electric distribution network in northeastern USA, Nat. Hazards [Internet] 79 (2) (2015) 1359–1384 [cited 13 June, 2016].
[10] J. He, D.W. Wanik, B.M. Hartman, E.N. Anagnostou, M. Astitha, M.E.B. Frediani, Nonparametric tree-based predictive modeling of storm outages on an electric distribution network, Risk Anal. [Internet] (2016) [cited 20 September, 2016].
[11] S.D. Guikema, S.M. Quiring, Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data, Reliab. Eng. Syst. Saf. [Internet] 99 (2012) 178–182 [cited 22 July, 2014].
[12] S.D. Guikema, R.A. Davidson, H. Liu, Statistical models of the effects of tree trimming on power system outages, IEEE Trans. Power Deliv. [Internet] 21 (3) (2006) 1549–1557 [cited 2 September, 2014].
[13] R. Nateghi, S. Guikema, S.M. Quiring, Power outage estimation for tropical cyclones: Improved accuracy with simpler models, Risk Anal. [Internet] 34 (6) (2014) 1069–10788 [cited 22 July, 2014].
[14] P.A. Kuntz, R.D. Christie, S.S. Venkata, Optimal vegetation maintenance scheduling of overhead electric power distribution systems, IEEE Trans. Power Deliv. [Internet] 17 (4) (2002) 1164–1169 [cited 21 December, 2014].
[15] K. Lim, P. Treitz, M. Wulder, B. St-Ongé, M. Flood, LiDAR remote sensing of forest structure, Prog. Phys. Geogr. [Internet] 27 (1) (2003) 88–106 [cited 18 January, 2015].
[16] W.C. Skamarock, et al., A Description of the Advanced Research WRF Version 3, NCAR, Boulder, Colorado, 2008, Report No.: TN–475+STR Available from: http://www2. mmm.ucar.edu/wrf/users/docs/arw_v3. pdf.
[17] University of Connecticut [Internet], Center for Land Use Education and Research, University of Connecticut, Storrs, CT, 2006 [cited 22 July, 2014]. Available from: http://clear.uconn.edu/projects/landscape/download.asp.
[18] Dewberry Project Report for the U.S, Corps of Engineers High Resolution LiDAR Acquisition & Processing for Portions of Connecticut. Dewberry, 2011, 77 p. Report No.: USACE Contract W912P9-10-D-053.
[19] J.R. Parent, J.C. Volin, Modeling forest canopy height using moderate resolution leaf-off airborne light detection and ranging (LiDAR), Int. J. Remote Sens. (2015), Revisions pending: December 2014.
[20] J.B. Campbell, Introduction to remote sensing, Introd. Remote Sens. [Internet] (1987) [cited 18 January. 2015].
[21] J.S. Evans, A.T. Hudak, R. Faux, A.M.S. Smith, Discrete return lidar in natural resources: recommendations for project planning, data processing, and deliverables, Remote Sens. [Internet] 1 (4) (2009) 776–794 [cited 19 September, 2016].
[22] D. Laes, S. Reutebuch, B. McGaughey, P. Maus, T. Mellin, C. Wilcox, J. Anhold, M. Finco, K. Brewer, Practical Lidar Acquisition Considerations for Forestry Applications, 2008, 7 p. Report No.: RSAC-0111-BRIEF1.
[23] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[24] A. Liaw, M. Wiener, Classification and regression by randomForest, R. News 2 (3) (2002) 18–22.
[25] L. Breiman, Random forests, Random For. [Internet] (2001) [cited 18 January, 2015].
[26] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees [Internet], 1984 [cited 22 July, 2014].
[27] Rea Kennedy, Solving Data Mining Problems Through Pattern Recognition, Pearson, Upper River Saddle, New Jersey, 1995.
[28] D. LaRose, C. LaRose, Data Mining and Predictive Analytics, Wiley, Hoboken, New Jersey, 2015.
[29] C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Department of Statistics, UC Berkeley, 2004, Report No.: 666.
[30] J. VanDerWal, L. Falconi, L.S. Januchowski, L. Storlie, SDMTools: Species Distribution Modelling Tools: Tools for Processing Data Associated with Species Distribution Modelling Exercises, 2014.
[31] S.- Han, S.D. Guikema, S.M. Quiring, K.- Lee, D. Rosowsky, R.A. Davidson, Estimating the spatial distribution of power outages during hurricanes in the gulf coast region, Reliab. Eng. Syst. Saf. [Internet] 94 (2) (2009) 199–210 [cited 22 July, 2014].
[32] S.- Han, S.D. Guikema, S.M. Quiring, Improving the predictive accuracy of hurricane power outage forecasts using generalized additive models, Risk Anal. [Internet] 29 (10) (2009) 1443–1453 [cited 15 August, 2014].
[33] H. Liu, R.A. Davidson, T.V. Apanasovich, Spatial generalized linear mixed models of electric power outages due to hurricanes and ice storms, Reliab. Eng. Syst. Saf. [Internet] 93 (6) (2008) 897–912 [cited 22 July, 2014].
[34] R. Nateghi, S.D. Guikema, S.M. Quiring, Forecasting hurricane-induced power outage durations, Nat. Hazards [Internet] 74 (3) (2014) 1795–1811 [cited 8 February, 2015].
[35] R. Nateghi, S.D. Guikema, S.M. Quiring, Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes, Risk Anal. [Internet] 31 (12) (2011) 1897–1906 [cited 22 July, 2014].
[36] S.D. Guikema, R. Nateghi, S.M. Quiring, A. Staid, A.C. Reilly, M. Gao, Predicting hurricane power outages to support storm response planning, Access IEEE [Internet] 2 (2014) 1364–1373.