

RICE UNIVERSITY

**Predicting wind induced damage to residential
structures: a machine learning approach**

by

Josue Salazar

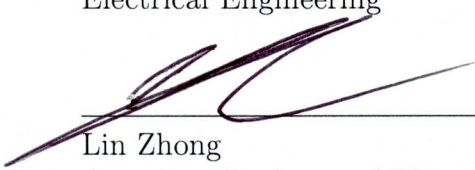
A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Science

APPROVED, THESIS COMMITTEE:



Devika Subramanian, Chair
Professor of Computer Science and
Electrical Engineering



Lin Zhong

Associate Professor of Electrical and
Computer Engineering and Computer
Science



Leonardo Dueñas-Osorio
Associate Professor of Civil and
Environmental Engineering

Houston, Texas

May, 2015

ABSTRACT

Predicting Wind Induced Damage to Residential Structures: A Machine Learning Approach

by

Josue Salazar

Hurricane winds can cause significant physical damage to residential properties. Pre-storm prediction of wind damage risk allows residents and city emergency officials to plan actions to reduce loss of life and property. In this thesis, I have developed a data-driven machine learning framework to estimate the probability of structural damage risk to a home subject to hurricane force winds. The modeling framework maps a set of predictor variables with the potential to explain structural damage to actual observations of homes damaged by hurricane winds. Widely used wind damage prediction models are parametric and are based on the physics of a structure responding to a wind load. Using a wind damage dataset gathered from about 700,000 residential buildings after Hurricane Ike in 2008 over Harris County, I have built a hybrid machine learning model that combines classification trees and logistic regression. My model is 47.5% more accurate than the physics-based approach at predicting expected damage at the one-kilometer square block level. I demonstrate the robustness of the model by using it to predict wind damage to homes in Harris County for simulated hurricanes of category 1 through 5 on the Saffir-Simpson scale.

Contents

Abstract	ii
List of Illustrations	vi
List of Tables	xi
1 Introduction	1
1.1 Importance of hurricane damage prediction	1
1.2 Formulation of the prediction problem	3
1.3 Contributions	8
1.4 Outline	9
2 Background	11
2.1 Literature review	11
2.1.1 Probabilistic component-based vulnerability modeling	11
2.1.2 Supervised machine learning modeling	16
2.1.3 Summary	18
2.2 Current state of the art wind damage prediction model: HAZUS	
Multi-Hazard	19
2.2.1 Methodology	19
2.2.2 Validation studies	20
3 Machine learning framework for wind damage modeling	24
3.1 Hybrid framework: LogRFT	24
3.2 Dealing with uncertainty in wind speeds in training process	26
3.2.1 Incorporation of virtual samples	27

3.2.2	Training LogRFT with static and dynamic variables	29
-------	---	----

4 Case study: Prediction of wind-induced damage caused by Hurricane Ike 31

4.1	Model evaluation	31
4.1.1	Learning objective: areal level accuracy	31
4.1.2	Training and validation protocol	32
4.2	Description, pre-processing, and analysis of data	33
4.2.1	Target variable: Harris County House of Authority survey data	33
4.2.2	Spatial analysis of target variable	35
4.2.3	Spatial analysis based on location and values: spatial autocorrelation of roof damage	37
4.2.4	Predictor variables	41
4.3	Feature selection	57
4.3.1	Correlation-based Feature Selection (CFS)	58
4.3.2	Selection of dynamic variables through LASSO analysis	70
4.3.3	Selection of static variables through random forest variable importance	73
4.4	Fitting hybrid model	75
4.4.1	Variable subset selection based on single hybrid tree: LogTree	78
4.4.2	Constructing LogRFT hybrid model	80

5 Analysis of results 84

5.1	Analysis of logistic regressions in the LogRFT hybrid model	84
5.1.1	Clustering and description of logistic regression coefficients	85
5.1.2	Geographical distribution of logistic models	87
5.1.3	Characterization of logistic models with respect to static variables	90

5.2	Assessing LogRFT generalization to unseen wind speeds	93
5.2.1	Comparing generalization among the LogRFT and HAZUS-MH models	93
5.2.2	Construction and analysis of LogRFT based on 3-second peak wind gusts	98
5.3	Summary of results	102
6	Conclusion and future work	104
	Bibliography	107

Illustrations

1.1	(i) Individual level representation of residential structures. (ii) Region Set A divides the area into three regions and the corresponding expected damage computed from individual homes within each region is shown. (iii) Region Set B shows the scale effect of the MAUP compared to Region Set A. (iv) Region Set C shows the effect of zoning compared to Region Set A.	4
2.1	Random forest variable importance score of explanatory variables used to discriminate between correctly and incorrectly predicted one-kilometer square blocks made by the HAZUS-MH wind damage methodology.	22
2.2	a) Distribution of errors made by the HAZUS-MH wind damage methodology at the one-kilometer square block level. b) Spatial distribution of individual level explanatory variable “quality of structure” considered as the most important variable that explains HAZUS-MH errors. HAZUS-MH errors appear to correlate with neighborhoods having low quality structures.	23
3.1	Logistic regression trained with synthetic data from Table 3.1.	28
3.2	Logistic regression trained with synthetic data in Table 3.1 and virtual samples generated according to Algorithm 3.1.	30

4.1	Left: Histogram showing the distribution of categories indicating roof damage to residential structures (0:No damaged, 1:Minor, 2:Moderate, 3:Severe, 4:Destruction). Right: Histogram showing the distribution of roof damage indicator (0:No damage, 1:Damaged) . . .	35
4.2	Spatial distribution of observed roof damage caused by Hurricane Ike in 2008 to residential structures over the Harris County, Texas. Source of data: HCHA	36
4.3	Moran's I spatial analysis of observed roof damage at different spatial radius distances r (330ft to 920ft).	40
4.4	Geographical distribution of land value and building value (in 2008 US dollars)	45
4.5	Geographical distribution of building age and remodeled age.	46
4.6	Distribution of wind variables over the Harris County, Texas observed in Hurricane Ike, 2008. a) Wind swath (WINDSWAMPH). b) Maximum sustained wind speeds (MAXWIND). c) Wind direction (WINDDIR). d) Wind steadiness (WINDSTEAD).	56
4.7	3-dimensional plot of Equation 4.4 for k values of 5 (left) and 15 (right).	59
4.8	Q-Q plots for variable BLDGVALUE (left) and its logarithmic transformation (right).	60
4.9	The graph MAXWIND vs. COAST_DIST in (a) visually exposes non-linearity among the variables. Joined together by multiplication, the distribution of combined variable MAXWIND_COAST_DIST is shown in (b).	62
4.10	Application of CFS with greedy backward elimination search on the expanded set of 94 predictor variables. The graph shows the iteration number, the variable eliminated at the given iteration, and the "merit" heuristic measure evaluating the worth of the remaining subset of variables.	65

4.11	Heatmap of Pearson's correlation among static variables. Variables are ordered from left-right and bottom-up in decreasing order of Pearson's correlation with roof damage. Three clusters of variables with positive correlation among them, as seen in this heatmap, correlate distinctively with roof damage: bottom-left cluster correlates positively, top-right cluster correlates negatively, and middle cluster correlates weakly.	68
4.12	Heatmap of Pearson's correlation among dynamic variables. Variables are ordered from left-right and bottom-up in decreasing order of Pearson's correlation with roof damage.	69
4.13	Effect of regularization parameter λ on LASSO Logistic Regression based on dynamic variables.	71
4.14	Trace plot of wind variable coefficients fit by LASSO as a function of λ (5-fold cross validated).	72
4.15	Random forest variable importance based 47 static variables and an ensemble of 50 trees. Top three most important variables include: distance to freeway, a combined terrain variable describing the percentage of open and wooded terrain in a radius of 362 ft, and distance to coast.	74
4.16	Exploring the effect of regularization parameter <i>MinLeaf</i> for a single CART based on AUC metric.	76
4.17	Tree depth and number of terminal leaves of a single CART based <i>MinLeaf</i> = 700.	77
4.18	Variable subset selection based on hybrid model with a single CART tree and logistic regressions at the leaves (LogTree) based on <i>MinLeaf</i> = 700 and a single dynamic variable: (MAXWIND_WINDSWAPMH).	79

4.19	Areal accuracy of LogRFT as the number of trees in the ensemble increases. The LogRFT is based on a random forest tree constructed with the top 15 most important static predictors and based on logistic regressions at the leaves fitted with the top most important dynamic wind variables.	82
4.20	Comparison of one-kilometer square areal accuracy between LogRFT and HAZUS-MH model based on expected roof damage from Hurricane Ike. LogRFT performs 45.5% better than HAZUS-MH. The performance of HAZUS-MH is based on the Hurricane Ike simulated 3-second wind gusts.	83
5.1	3-D plot of the 7,056 logistic regression models fitted in LogRFT#1 partitioned into 3 clusters by K-means algorithm.	86
5.2	Graphical representation of the three clusters of logistic models from the LogRFT#1. These models are plotted as a function of MAXWIND. The (MAXWIND_WINDSWAPMH) variable is set to MAXWIND ²	88
5.3	Geographical distributions of unique training samples used to fit logistic regression models in cluster 1 to 3. Samples from each cluster are aggregated to the one-kilometer square blocks by computing the count of residential properties within each block.	89
5.4	Graphs describing the distribution of values of each static variable in LogRFT#1 across the three logistic regression clusters. The distribution of values in each cluster is represented using box plots. Single variables log-transformed in Section 4.3.1 are transformed back to their original units of measure.	92

5.5	Distribution of expected damage predicted by the LogRFT model at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.	96
5.6	Distribution of expected damage predicted by HAZUS-MH fragility curves at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.	97
5.7	3-D plot of the 7,056 logistic regression models fitted in LogRFT_windgust_factored#1 partitioned into 3 clusters using K-means algorithm.	99
5.8	Graphical representations of the three clusters of logistic models obtained from model LogRFT_windgust_factored#1. Models plotted as a function of MAXWIND. The (MAXWIND_WINDSWAPMH) variable was set to MAXWIND ²	100
5.9	Distribution of expected damage predicted by LogRFT_windgust_factored at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.	101

Tables

3.1	Synthetic data representing a set of samples from a single node. The samples reflect variability in the target variable with respect to wind speed approximations. Target variable: “not damaged”=0, “damaged”=1.	28
4.1	Descriptive statistics for structure predictor variables.	44
4.2	Descriptive statistics for construction code enforcement variables. . .	47
4.3	Mapping of National Land Cover Database classes into roughness length values based on Table 3.9 (page 128) in the HAZUS-MH 2.1 Technical Manual.	48
4.4	Categorization of HGAC land cover classes into wooded, open, and developed terrain.	51
4.5	Descriptive statistics for terrain variables Part 1.	52
4.6	Descriptive statistics for terrain variables Part 2.	53
4.7	Descriptive statistics for Hurricane Ike wind hazard characteristics. .	55
4.8	Set of selected variables exhibiting highly skewed distributions which were transformed using Equation 4.5.	61
4.9	Descriptive statistics for static variables in decreasing order of absolute Pearson’s correlation with roof damage.	66
4.10	Descriptive statistics for dynamic variables in decreasing order of absolute Pearson’s correlation with roof damage.	67

4.11	Order of importance among dynamic variables as result of LASSO analysis. The descending order corresponds to the sequence in which the variable coefficients were set to zero by LASSO as λ increases. . .	72
4.12	One-kilometer square level prediction performance for the different wind damage models with respect to actual observed roof damage caused by Hurricane Ike over Harris County, Texas.	81
5.1	Descriptive statistics for the three clusters of logistic regression functions fitted in the LogRFT#1 model.	86
5.2	Saffir-Simpson hurricane hurricane categories.	93
5.3	One-kilometer square level prediction performance for the different wind damage models with respect to actual observed roof damage caused by Hurricane Ike over Harris County, Texas.	95

List of Algorithms

3.1	Algorithm for generating virtual samples based on random sampling values for w' from a uniform distribution.	29
-----	---	----

Chapter 1

Introduction

1.1 Importance of hurricane damage prediction

Hurricane hazards have caused significant loss of life and economic harm to the United States. Over the years 2003-2012, they induced property damage of up to \$122 billion dollars and more than 1,000 deaths all along the Gulf and East Coast [1]. According to a study of the economic impact of hurricanes in the United States, yearly ratios of damage to gross domestic product (GDP) for nine years between 1900 and 2008 have exceeded 20% [2]. Among these years, 2005 stands out with a damage to GDP ratio of about 90%, mainly due to damage caused by the landfall of five hurricanes in US territory (hurricanes Denis, Emily, Katrina, Rita, and Wilma). The source of such significant economic impact does not solely originate from the hurricane hazards but also from the level of economic development where the hurricane made landfall. As of 2005, New Orleans and three other coastal cities (Miami coast, Houston, and Tampa) were areas holding capital stock greater than \$100 billion [2]. In the United States, during 2003, 53% of the total population lived in coastal counties. Among these, 56% resided in coastal counties from the Northeast, Southeast, and Gulf of Mexico regions where hurricane activity is most experienced [3]. As long as economic development and population migration continue to increase in coastal regions, hurricanes will have a significant impact on the economy and population of the United States.

The harmful effects hurricanes cause have motivated the development of prediction tools with the goal to mitigate disaster and loss of life. In particular, this thesis focuses on the development of a data-driven machine learning framework which estimates the probability of structural damage risk to homes subject to hurricane wind forces. While most models in civil engineering for wind damage prediction are physics based, mine is based on the largest damage data set collected to date. The framework maps a set of predictor variables with the potential to explain structural damage to actual observations of homes damaged by hurricane winds. Using the wind damage dataset composed of about 700,000 residential buildings after Hurricane Ike in 2008 over Harris County, I build a hybrid machine learning model that combines classification trees and logistic regression. Since it is difficult to predict damage at the individual home level, I assess the model quality by aggregating damage predictions at the level of blocks which divide the Harris County into $1km^2$ areas. That is, I compare the model's prediction of damage at the one-kilometer square block level with the actual (or observed) damage percentage at the same level. My model correctly predicts the percentage of actual observed damage caused by Hurricane Ike for 75.2% of the one-kilometer square blocks across Harris County with a mean absolute error of 0.125. This outcome corresponds to a 47.5% increase in performance compared to the state-of-the-art wind damage model.

Improved prediction performance of wind-induced damage to residential structures has important applications for the federal government, local governments, and community residents. The federal government uses damage risk predictions to estimate economic losses of affected areas in order to decide the amount of economic aid needed by local governments for restoration. Local governments use damage risk predictions to identify neighborhoods most at risk, estimate the number of people likely

to evacuate, the number of temporary shelters needed, and the number of personnel and rescue teams needed to assist the displaced population. Community residents can use predictions prior to or during a hurricane threat to become informed about the risk of wind-induced damage specific to their neighborhood. Once informed, they can make decisions about improvements to their homes to increase resistance to wind damage, and better assess whether to shelter or evacuate in the event of a hurricane threat. These important applications accentuate the need for efficient, practical, and accurate wind-damage prediction models which enable communities to respond to hurricane threats more effectively.

1.2 Formulation of the prediction problem

The wind damage prediction problem can be formulated at two different geographical levels of aggregation: individual and areal. The individual level corresponds to having data values represent single family residences. For example, Figure 1.1(i) shows a set of five homes, two of which have been labeled as not damaged and three as damaged. The areal level corresponds to the case where a summary statistic is used to aggregate individual units within a geographical area. Figure 1.1(ii) shows the same set of houses, but now divided by three regions. Each region has been assigned the expected damage of homes within (100%, 50%, and 0%). In the following paragraphs, I expose the difficulties encountered when fitting wind damage models using data at the individual and areal levels of aggregation. I compare and contrast the advantages and disadvantages of using either of the two levels and use that as a guide to devise the best strategy for modeling the risk prediction problem.

Predicting at the individual level increases the difficulty of prediction due to two types of uncertainties: uncertainties due to naturally stochastic wind currents and

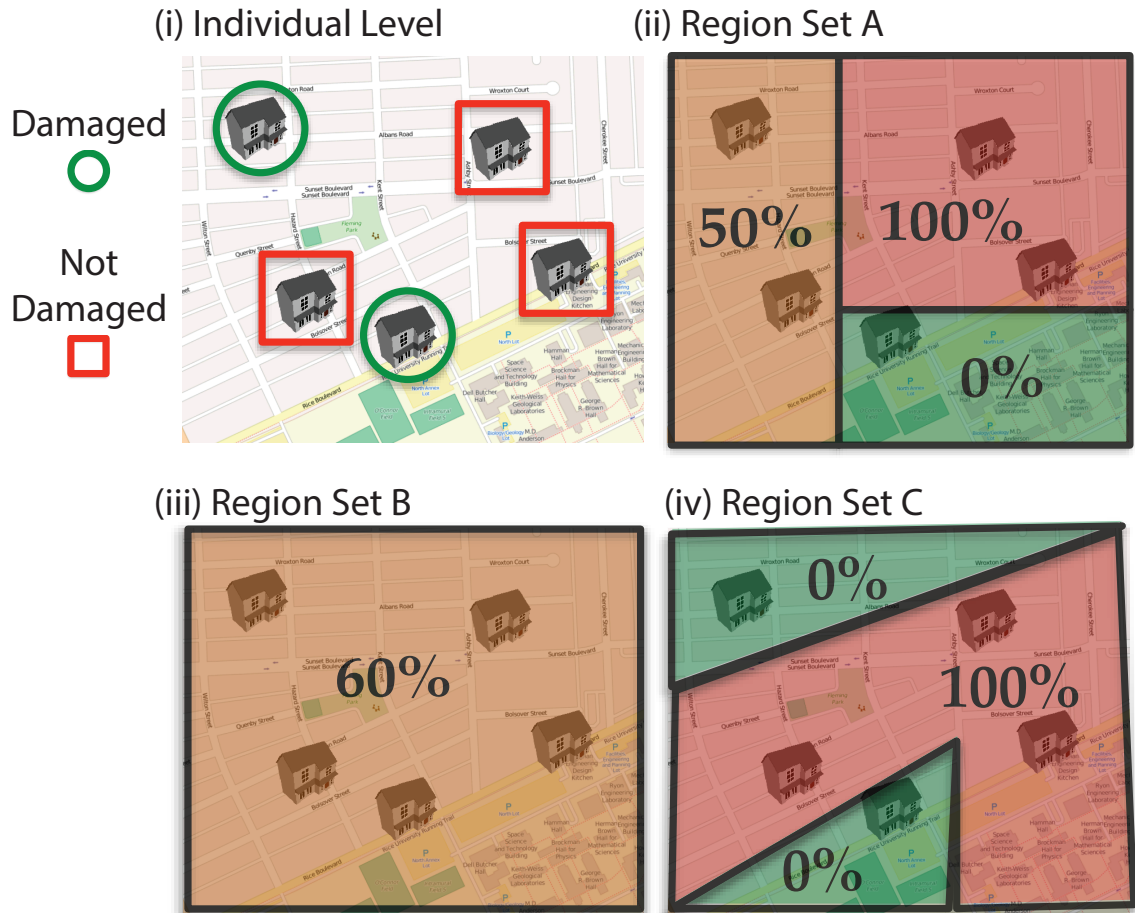


Figure 1.1 : (i) Individual level representation of residential structures. (ii) Region Set A divides the area into three regions and the corresponding expected damage computed from individual homes within each region is shown. (iii) Region Set B shows the scale effect of the MAUP compared to Region Set A. (iv) Region Set C shows the effect of zoning compared to Region Set A.

uncertainties due to unknown stochastic factors pertaining to wind resistive capacity of structures. The first type of uncertainty corresponds to the effects of turbulent wind flows interacting with the environment. Turbulent wind flows might create wind gusts defined as sudden, brief increases in wind speed followed by a sudden decrease

[4]. When severe, these can exert significant loads on buildings, causing structural failure. Formed due to friction (turbulence caused by wind blowing around obstacles), wind shear (changes in wind speed and direction over distance in both vertical and horizontal directions), and solar heating of the ground (which causes warm air to rise and cooler air to sink), wind gusts can also blow down trees over homes and pick up debris, transforming them into high velocity projectiles that collide with homes [5]. The second source of uncertainty corresponds to the unknown stochastic factors characterizing the resistance capacity of built structures. The structure's resistance as a whole is determined by the individual components' capacity to tolerate stress. Failure at any given component compromises the integrity of the whole structure. Even though probabilistic knowledge of the components' resistance capacity exists, stochastic factors such as quality degradation due to aging and unknown prior damage (e.g. prior hurricane, water infiltration, termite damage) increase uncertainty [6].

Predicting with data spatially aggregated to the areal level generates two types of problems: the Modifiable Areal Unit Problem (MAUP) and the ecological fallacy. The MAUP emerges from the fact that spatial aggregation can be performed at areal zones of different sizes having different boundary delineations [7]. Constructing statistical models based on aggregated summary statistics produce statistical analyses dependent on the specific areal aggregation. MAUP can be characterized by two effects: scale and zoning effects. The scale effect relates to the variation introduced due to the geographic level of aggregation (e.g. census blocks, census tracts, county). For example Figure 1.1(iii) shows a different outcome in expected damage after aggregating individual homes to a larger areal zone. The zoning effect relates to the variation introduced by the geographic delineation of zones (there are infinite ways to partition a study area into zones). Figure 1.1(iv) shows an example of how modifying

the zone boundaries generate different results. The second problem when predicting at the areal level corresponds to the ecological fallacy. Ecological fallacy occurs when making an inference about individuals based on spatially aggregated data [8]. Therefore, prediction estimates produced at the areal level can only be interpreted at that level of aggregation. For example, it is not correct to infer from Figure 1.1(iii) that given the probability damage in Region B is 60%, an individual home within Region B has a probability of damage of 60%. In fact, two homes within Region B have 0% probability of damage and three have 100% probability of damage.

Predicting at the individual level directly answers the question, “What is the probability of damage to a home?” This is actually the question home owners would like answered each time a hurricane threatens their community. Nonetheless, there are advantages as well as disadvantages when predicting at this level. The advantages of using individual level data are twofold. First, individual data allows for the construction of models based on residential structures that share similar components and characteristics by using information at the finest grain. Aggregating over a predefined zoning replaces the individuality of each structure by a summary statistic computed from different nearby structures. Second, estimating individual level damage risk allows for the possibility to aggregate damage probabilities to any areal level. Doing so relaxes the burden of predicting accurately at the individual level and allows us to evaluate prediction performance at the areal level of choice. The disadvantages of using individual level data are also twofold. First, as explained in previous paragraphs, difficulty exists in predicting whether a specific home will be damaged due to the stochastic processes taking place in a hurricane event. It is an open question whether it is even possible to predict wind damage at the individual level given the

available information. Second, if damage risk predictions were to be released to the public, the privacy of residents would be violated.

Predicting at the areal level answers the question, “What is the probability of expected damage in a neighborhood?” City emergency officials are interested in answering this question to identify neighborhoods with higher damage risk, so they can plan mitigation steps accordingly. The advantages of areal level prediction include analysis at larger scales over geography and protection of privacy of residents. The disadvantages of predicting at the areal level are the heavy dependence of results on the specific scale and zoning (MAUP). Also, predictions at the areal level limit the ability to infer damage at other aggregation levels (ecological fallacy).

The examination of the advantages and disadvantages of individual and areal level prediction motivates the combination of the best of both levels for learning data-driven wind damage models. First, model construction at the individual level avoids the MAUP problem. Predicting at this level allows the model construction to benefit from information at the finest grain. Second, once the model has been constructed, individual level wind damage predictions can be aggregated to the areal level of choice by computing the expected probability of damage. Aggregating individual level predictions and assessing the model accuracy at the areal level improves overall performance of the model. Note that as the scale of geographical aggregation increases (i.e. census blocks to census tracts), the impact of uncertainties and prediction errors is diminished as a consequence of summarizing predictions to a single statistic (i.e. expectation), therefore, improving prediction accuracy with respect to observed damage. However, predictions aggregated over a large area do not provide detailed information useful to residents who are trying to decide whether to evacuate to a safer place within the same locality. Additionally, irregular delineations of areal

units with different scales (as observed in census tracts) make geographical analysis difficult. For these reasons, I use one-kilometer square blocks to aggregate and evaluate predictions. This areal level configuration contains uniform areas of small enough size to make predictions at this level useful to residents.

1.3 Contributions

With a large data set of about 700,000+ observations of wind damage gathered after Hurricane Ike in 2008 over Harris County, I have made the following contributions:

- Developed a hybrid data-driven machine learning framework to construct damage functions for different types of residential structures that relate wind speed to the probability of minor wind damage. The hybrid framework is composed of a hierarchy of two well known algorithms fitted in sequence: random forest of classification trees with logistic regression models at the leaves. Although this combination technique has been previously researched [9][10][11], I contributed with the novel idea of training the models with different sets of features to accommodate the needs of the wind damage prediction problem. At the first level, a random forest separates the training data into similar building types, based on building, construction code, and terrain variables. The second level corresponds to fitting a logistic regression model with wind speed and observed damage at every terminal leaf in the random forest. Each logistic regression is fitted with wind speed variables and observed damage from samples at each terminal leaf. The hybrid model (referred to as LogRFT) is equivalent to having a function for each building type that estimates the probability of at least minor damage to an individual home given a wind speed.

- Developed a strategy to introduce prior knowledge in the learning process enabling generalization of models to wind speeds outside the scope of those present in the training data.
- Trained a random forest/logistic regression hybrid model (LogRFT) to optimize the number of correctly predicted one-kilometer square blocks with respect to the observed damage across Harris County, Texas. My model is 47.5% more accurate than the most widely used wind damage model at predicting expected damage at the one-kilometer square block level. I demonstrate the robustness of the hybrid model by using it to predict wind damage to homes in Harris County for simulated hurricanes of category 1 through 5 on the Saffir-Simpson scale. This analysis shows the hybrid model capable of making predictions with a smooth transition from low to high damage probabilities. Given that the existence of damage data is limited, it is not possible to evaluate the model rigorously, therefore, model overfitting might exist.

1.4 Outline

My thesis is organized as follows: In chapter 2, I assemble a literature review of studies addressing prediction of wind-induced damage to structures, dividing them into probabilistic component-based vulnerability models (physics based) and supervised machine learning models (data-driven). I focus attention on a component-based wind damage model called HAZUS-MH, widely used by local and federal governments to estimate wind damage risk. I motivate the machine learning approach by citing prior work which characterized prediction errors made by the HAZUS-MH methodology, by identifying parameters not modeled in the component-based approach. In chapter 3, I

describe the hybrid machine learning framework developed for modeling wind-induced damage, based on a combination of the random forest algorithm [12] and logistic regression models. In chapter 4, I present the model evaluation metrics, detailing data sources, analysis, and pre-processing of the target and predictor variables. I apply the hybrid machine learning framework to construct a wind-damage model based on the observed damage survey of residential structures from Hurricane Ike. Since a large number of predictor variables was obtained, I include feature selection techniques. In chapter 5, I analyze the logistic regression models fitted in the previous chapter by: identifying the different types of logistic functions, observing the geographic distribution of residences described by similar logistic models, and characterizing the types of residences corresponding to similar logistic functions. Additionally, I assess the generalization of the hybrid model over unseen hurricane category wind fields and analyze the observations. Lastly, in chapter 6, I conclude a summary of contributions of my thesis and suggest directions for future research.

Chapter 2

Background

2.1 Literature review

Divided into two categories, studies addressing the prediction of wind-induced damage to structures include: probabilistic component-based vulnerability models (based on the physics and interaction between built structures and the wind) and supervised machine learning models (derived from observed damage data, knowledge of structural properties of buildings, terrain characteristics, and the wind). The following subsections present a brief summary of models proposed by these studies, the variables used to infer damage risk to structures, and the validation of predictions produced by these models.

2.1.1 Probabilistic component-based vulnerability modeling

Probabilistic component-based vulnerability modeling consists of explicitly accounting for the physical resistance capacity of the structure's components subjected to increasing wind loads in order to estimate the probability of damage to the entire structure.

Unanwa et al. (2000) proposed the concept of wind damage bands which define the upper and lower thresholds of the relationship between the degree of damage and wind speeds for structure types with similar components (i.e. mid-rise, residential, commercial, and institutional structures) [13]. The upper and lower damage thresholds

are obtained by considering the highest and lowest probabilities of failure of the set of components and connections characterizing the structure type. Probabilities of failure for a structure, acquired from fragility curves, represent the probability of exceeding a level of damage as a function of wind speeds. These fragility curves are generated by analyzing a multiple fault tree scheme modeling predominant sequences of failure modes in hurricane events that significantly contribute to the amount of structural damage. The model considers components such as roof covering, roof structure, exterior doors and windows, exterior wall, interior, structural system, and foundation. A subsequent paper, Unanwa and McDonald (2000), details the procedure for the implementation of this model to predict wind-induced damage to individual structures and groups of structures [14]. This study validates model predictions by comparing them to observed damage to three buildings from Hurricane Fran. They report that “the actual damage degrees were found to be within 95% of the model prediction intervals” [14]. The size of the validation set limits the scope of assessment and makes it difficult to project its accuracy in new settings.

Ellingwood et al. (2004) presents a fragility-based methodology for assessing the resistance capacity of light-frame building constructions with various roof configurations subjected to hurricane winds [15]. Roof and construction parameters considered are roof type, slope, roof height, nailing pattern, connector type, and truss spacing. In contrast to Unanwa et al. (2000), Ellingwood et al. (2004) fails to consider individual component failures that propagate to complete collapse. Instead, Ellingwood et al. (2004) simply combine the conditional probability of damage of all individual component fragilities by marginalizing them over a given wind speed [15]. Predictions were compared to observed damage surveys performed by the National Association of Home Builders (NAHB) in the aftermath of Hurricanes Andrew (1992) and Iniki

(1992). Predictions made by the fragility models were found to over-predict the probability of roof panel damage for one-story residences during Hurricane Andrew. The survey indicated that $69 \pm 5\%$ of homes lost at least one roof panel while the fragility models predicted roof panel failure of 92%.

Pinelli et al. (2004) presents another probabilistic component-based model which relies on Monte Carlo simulations. The model determines the probabilities of occurrence of combinations of basic damage modes given 3-second average gust wind speeds while taking into consideration wind-driven debris [16]. These basic damage modes correspond to breakage of openings, loss of shingles and roof or gable end sheathing, and roof to wall connection and masonry wall damage. After selecting a building class (1-story gable roof masonry/wood frame, 1-story hip roof masonry/wood frame, etc.) for the Monte Carlo engine simulation, each component is assigned a resistance capacity based on probabilistic information acquired from previous tests and research [17]. Failure checks to individual components are performed after deterministically simulated wind loads to uncover component failures. The combination of failures determine the categorical damage state of the building. This process is repeated thousands of times for the same building class and the same wind conditions. The result of the Monte Carlo simulation is a matrix that contains the probability of occurrence of various damage states given 3-second wind gust speeds. Pinelli et al. (2004) presents a detailed implementation of the model for a hypothetical residential community containing different building types but does not provide any validation of predictions made by the model.

The HAZUS-MH physical model introduced by Vickery et al. (2006b) is the most widely used component-based model [18]. This model, currently used by local and federal governments to provide pre-storm prediction analyses of wind-induced damage

risk, was developed by the Federal Emergency Management Agency (FEMA). The methodology for developing structural damage functions relates to that of Pinelli et al. (2004). The use of simulated hurricane wind fields over 15 minutes intervals instead of using deterministic wind speeds, differentiates the Monte Carlo engine in the HAZUS-MH model. A detailed description of this model and studies performed to validate its predictions is provided in Section 2.2.

Another component-based vulnerability model, the Florida Public Hurricane Loss Project (FPHLP) model funded by the Florida Office of Insurance Regulation, assesses the risk of insured residential property associated with hurricane wind-induced damage in the State of Florida [19]. This model, based on Monte Carlo simulations, not only simulates wind speeds but also different wind directions for the variety of structure types common in the State of Florida. Components considered include the roof type, roof sheathing, roof-to-wall connections, walls, windows, doors, garage doors, region (north, central, south), and sub-region (high wind velocity zone, wind born debris region, other) [20]. Hamid et al. (2010) provides validation of the FPHLP model by comparing predicted economic losses to insurance data of historical losses at the county and state level from different hurricane events. Their validation shows a high correlation between the modeled and actual losses but at the very coarse county and state levels.

Chung et al. (2010) [21] present the application of an integrated model introduced by Lin et al. (2010) [22], combining the debris risk model developed by Lin and Vandemarcke (2008, 2010) [23] [24] and the FPHLP component-based vulnerability model. Chung et al. (2010) analyze the effect of wind-driven debris from neighboring structures by performing a Monte Carlo simulation with structural resistances randomly assigned to every building component in each run. For each wind

speed and direction, the wind-pressure damage model is applied to each building in the neighborhood. If any component failure causes an opening on the building or when debris damage occurs via windows, doors, or garage doors the internal pressure is adjusted. Unlike the FPHLP model, which only adjusts the internal pressure once, the model in Chung et al. (2010) keeps updating the internal pressure of the structure until no more openings are created and an equilibrium is reached. The simulations presented in the papers show that wind-borne debris from damaged buildings causes significant damage to neighboring buildings. Chung et al. (2010) do not provide any validation due to lack of observed damage data. In addition, wind damage computation time for a neighborhood of 358 structures becomes significant. This suggests that the Monte Carlo simulation becomes intractable over a significantly higher amount of structures.

Building upon previous research, the most recent component-based vulnerability model approach for wind damage prediction is presented by Grayson et al. (2013) [25] where they integrate the FPHLP model with a three-dimensional (3D) probabilistic wind-borne debris trajectory model developed by Grayson et al. (2012) [26]. This integrated model is broken into three phases allowing for an increase in computational efficiency. The first phase defines all parameters related to the area of interest. The second, composed of a Monte Carlo simulation, samples resistance capacities for each structure component. A separate module determines if wind loads generate wind-borne debris damage to individual structures. The final phase aggregates the damage to the individual structures for analysis. Grayson et al. (2012) provide an example of how the integrated model works. They acknowledge that one of the limitations of their model is that it cannot be verified due to the lack of data from real building

envelope failures during hurricanes. Neither Chung et al. (2010) nor Grayson et al. (2013) take windborne debris from vegetation into account in their models.

2.1.2 Supervised machine learning modeling

Previous studies have also been based on *supervised machine learning* to model wind-induced damage to residential structures. Supervised machine learning refers to the task of building models from supplied training data consisting of a target variable (which summarizes the outcome we would like to predict) and a collection of predictor variables (which potentially explain the target variable). Supervised machine learning explores, within a space of functions, solutions that map the predictor variables to the target variables. The solution which best explains the target variable given the predictor variables is used to make new predictions.

Huang et al. (2001) [27] and Sparks (2003) [28] present a very simple model for wind-induced damage learned from an insurance-claim training data. The training set is comprised of residential homes in South Carolina affected by Hurricanes Hugo and Andrew. Created using regression techniques to relate the mean surface wind speed to the damage ratio (amount paid by insurer divided by total insured value) at the zip code areal level, the model seems to fit the damage ratio well as wind speeds increased. Nonetheless, the papers fail to validate the model on a new damage dataset.

Dehring and Halek (2006) [29] investigate damage based on a sample of residential structures affected by Hurricane Charley in 2004. Although their goal is to investigate whether adopting the National Flood Insurance Program construction code affected damage outcomes in residential structures, they constructed a logistic regression model based on a binary damage variable. As predictor variables, they include

four indicator variables related to the construction code in place when the residential structure was built, as well as distance from the Gulf, area of building, area of parcel, real price per square foot, age relative to 2003, and base flood elevation. The logistic model's goodness of fit, measured by the Nagelkerke's R^2 at 22.03%, corresponds to the proportion of unexplained variance reduced by the predictor variables. The paper concluded that there was no evidence that increased construction code standards reduce property risk exposure.

Moreover, Highfield and Peacock (2010) [30] use a set of approximately 1,500 wind-damage assessments independently collected in the aftermath of Hurricane Ike from random sampled residential structures within the Galveston Island and the Bolivar Peninsula. In this paper, the authors learn Ordinary Least Squares (OLS) models using the collected information to describe and analyze the pattern of damage. Four observed assessments in a five-point scale including foundation and structural damage, roof damage, exterior damage, and overall damage, summarized into a damage index is used as target variable. The predictor variables used by the model include structure elevation, home age, distance to water, distance to seawall, proportion in FEMA A Zone, proportion in FEMA V Zone, maximum inundation, whether the structure is located in Galveston Island, improvement value, and proportions of hispanic and black population. Although their focus is to analyze the impact of each of these predictor variables, the best R^2 of 35.1% reflects the limitations of using a single linear model to explain individual residence level damage.

Another study made by Kim et al. (2013) [31] presents a supervised machine learning approach for predicting wind-induced damage based on insurance claims of commercial buildings from eight affected counties in the aftermath of Hurricane Ike. As target variable, they used the ratio of property damage loss paid by the insurer

divided by the structure’s appraised value. As predictor variables, they considered maximum sustained wind speeds, floor area, building age, FEMA Flood Zones, Hurricane Surge Zones, distance from water, and an indicator variable of whether the structure is located to the right side of the hurricane track. After selecting a random sample of 500 commercial structures and analyzing correlations between the damage ratio and variables, they applied a backward feature elimination method to find a best-fit multiple regression model. The resulting model has an adjusted R^2 of 33.7% which corresponds to the proportion of target variable variance explained by the set of significant predictor variables: right side of the hurricane track, building age, hurricane surge zones, and distance from property to shoreline. They attribute the low adjusted R^2 score to the lack of inclusion of unknown predictor variables, but do not consider wind damage prediction as a non-linear problem, as previous component-based engineering research suggests [32]. Surprisingly, the backward feature elimination procedure in Kim et al. (2013), drops maximum sustained wind speeds as a predictor variable contradicting the results derived in the probabilistic component-based approach [32].

2.1.3 Summary

Both data-driven and physics based research aimed at predicting wind-induced damage to residential structures exhibit limitations. Limitations in the machine learning literature include: failure to consider non-linear models (i.e. decision trees), failure to acquire important predictor variables for explaining wind-damage, and failure to apply techniques that enable models to generalize over unseen wind speeds. In general, work in component-based modeling, lacks model validations based on actual damage data. In prior work done in collaboration with other colleagues based on a

large damage data collected after Hurricane Ike, I performed a validation analysis on the most widely used component-based wind damage model: HAZUS-MH physical model. In the next section, I describe the HAZUS-MH methodology for wind-damage risk modeling, comment on prior validation studies, and report on the findings in the validation study we performed.

2.2 Current state of the art wind damage prediction model: HAZUS Multi-Hazard

FEMA developed a software based on Geographic Information Systems (GIS) called HAZUS Multi-Hazard (HAZUS-MH), for use by local city emergency management teams. HAZUS-MH contains prediction models for estimating potential physical, economic, and social impacts from earthquakes, hurricanes, and floods [33]. These estimations currently help emergency management teams develop damage mitigation strategies, discover susceptible communities at-risk, plan evacuation recommendations, and allocate emergency resources.

2.2.1 Methodology

The HAZUS-MH hurricane physical damage model, introduced by Vickery et al. 2006b [18] as previously mentioned, estimates potential damage to a structure by calculating the wind-induced loads acting on the building and then assessing the resistance of the building and its components under such wind-loads. Fragility curves, which relate wind gust speeds to the probability of exceeding a level of damage, estimate the resistance of buildings given different combinations of exterior components (frame composition, number of floors, roof shape, shutters, garage door type, roof-deck attachment, and sheeting nails) [34]. These curves are developed by iteratively

comparing each component’s resistance to pressure loads obtained from wind fields simulated at 15 minute intervals. Resistance for each component is sampled from its corresponding probabilistic distribution derived based on experimental data and engineering analysis [35]. Failure at each component is then assessed by comparing internal and external pressures. At the same time, the probability of damage caused by wind-borne debris originating from surrounding structures is simulated based on a missile impact model introduced in Twisdale et al. (1996) [36] and Vickery et al. (1999) [37]. After a number of simulations for a given building type, the resulting overall damage at each simulation is categorized into one of five damage levels (0-4) defined in Vickery et al. 2006b [18]: no damage, minor, moderate, severe, and destruction. The HAZUS-MH model outputs expected damage at the areal level of choice taking into consideration the building stock within each zone. Expected damage percentage for a given zone is computed by using the predictions of fragility models of all represented building types and their respective proportion relative to the total number of buildings within zone. Finally a count of damaged buildings is estimated by multiplying each zone’s probability of damage by the zone’s total building population.

2.2.2 Validation studies

In order to compare actual observed damage to predictions made by the HAZUS-MH model, the FEMA study [38] introduced a validation protocol for wind-damage models that accounts for the uncertainty in wind speed. Validation results based on this protocol at the county areal level indicate “good agreement” with observed damage caused by Hurricane Charley; however, results also show significant under-estimation of observed damage caused by Hurricane Ivan. The same study on Hurricanes Charley

and Ivan indicates significant under-estimation of observed damage at the individual home level. Another validation study performed by Vickery et al. (2006b) [18] for Hurricanes Andrew, Hugo, Erin, and Opal, finds that the HAZUS-MH fragility curves under-estimate the probability of potential damage at wind speeds less than 100mph.

Validation of HAZUS-MH using Hurricane Ike damage data

In collaboration with other colleagues, in Subramanian et al. (2013) [39], I performed a data intensive validation analysis of the HAZUS-MH model using damage data from Hurricane Ike. In this study, I reverse engineered and implemented the HAZUS-MH methodology for predicting wind damage as a function of wind speed. I evaluated the methodology's performance based on the same validation protocol introduced in previous validation studies [38]. The HAZUS-MH physical model uses deterministic simulation of Hurricane Ike 3-second peak gust speeds generated by the HAZUS-MH hazard model introduced in Vickery et al. (2006a) [40]. Results show that the HAZUS-MH physical model correctly predicts observed damage for 51% of the one-kilometer square blocks, under-estimates 25.6% of the blocks, and over-estimates damage in 23.4% of the blocks.

Based on a random forest [12] classifier built to discriminate correctly (in good agreement with observed damage) predicted blocks from incorrectly predicted blocks in Harris County, characterization of prediction errors made by HAZUS-MH is possible through twenty-one potential explanatory variables. The classification algorithm computes estimates of explanatory variable importance. We used these variable importance scores as a tool to reveal important variables that explain prediction errors made by HAZUS-MH. Figure 2.1 shows the variable importance score for each of the explanatory variables ordered in decreasing order from left to right. The top

three variables include quality of the structure, years passed since last remodeling (remodeled age), and building value. Figure 2.2 shows the distribution of errors at the one-kilometer square blocks along with a map that shows the spatial distribution of the quality of structure. Neighborhoods with homes having low quality structures appear to correlate spatially with under-estimated one-kilometer square blocks. For example, under-estimated blocks east of US-59 inside the I-610 loop tend to have low-quality structures. Additionally, over-estimated blocks in the northeast of Beltway 8 seem to include homes with structural quality.

The results of our validation study motivated me to ask whether it would be possible to construct a supervised machine learning model based on important exploratory variables in Figure 2.1.

In the following chapter, I layout a machine learning framework I developed to construct wind-damage models from data with the potential to explain damage at the one-kilometer square level.

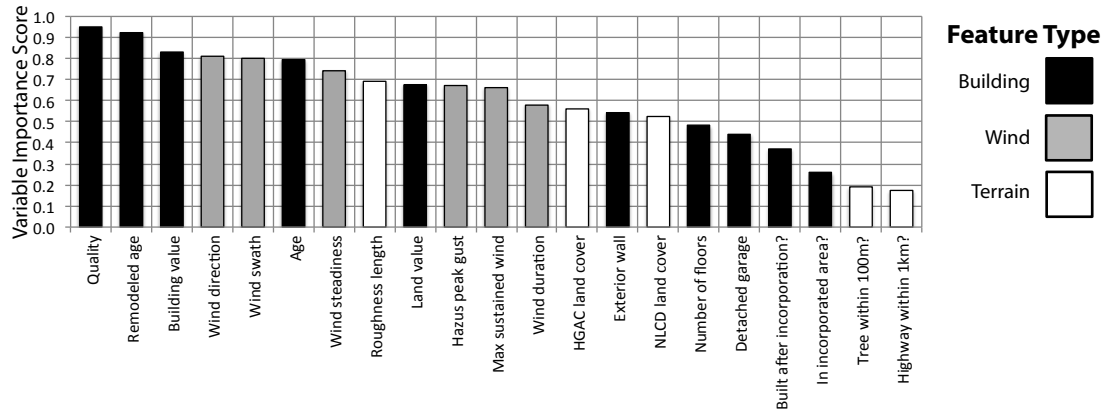


Figure 2.1 : Random forest variable importance score of explanatory variables used to discriminate between correctly and incorrectly predicted one-kilometer square blocks made by the HAZUS-MH wind damage methodology.

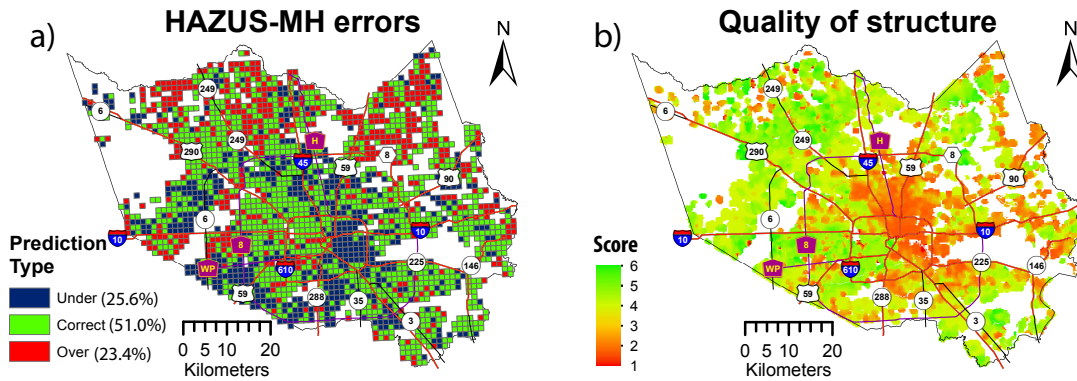


Figure 2.2 : a) Distribution of errors made by the HAZUS-MH wind damage methodology at the one-kilometer square block level. b) Spatial distribution of individual level explanatory variable “quality of structure” considered as the most important variable that explains HAZUS-MH errors. HAZUS-MH errors appear to correlate with neighborhoods having low quality structures.

Chapter 3

Machine learning framework for wind damage modeling

One of the most difficult aspects in modeling wind-induced damage to residential structures is dealing with the uncertainty introduced by stochastic factors governing the process. Probabilistic component-based models developed directly from probability functions describe the resistance capacity to wind loads of different building types. Some component-based models in the literature review Section 2.1.1 handle uncertainty by modeling the effects of debris generated from damage to nearby structures. In contrast, when modeling wind-induced damage directly from observed data, I have to deal with the variability of observed damage outcomes from similar building types experiencing similar wind fields. Such variability results from un-modeled and un-observed factors such as collisions from flying debris or trees falling over the structure. Therefore, supervised machine learning models have to deal with the challenge of estimating damage with an incomplete set of predictors.

3.1 Hybrid framework: LogRFT

Recognizing the difficulties in modeling wind-induced damage, I devised a machine learning framework (named LogRFT), composed of a hybrid combination of two well-established machine learning algorithms: random forest [12] (an ensemble of classification trees [41]) and logistic regression [42]. These two models are combined in a sequential hierarchical fashion. The first is composed of a random forest trained on

the target variable (observed damage to a building) and a set of predictor variables excluding wind variables. The second level is composed of logistic regression models at each terminal leaf in the random forest, based on wind variables and the target variable. This learning configuration enables us to model the variation of damage probabilities with wind speeds as a smooth logistic function, where a small change in the input wind speed results in a small change in the resulting damage probability.

The idea of learning hybrid decision trees with classifiers constructed at each terminal leaf is not new [9] [10] [11]. Seewald et al. (2001) [10] explored the construction of pruned decision trees with leaf classifiers based on linear regression, the nearest neighbor algorithm, and the Naive Bayes classifier. The experimental results in Seewald et al. (2001) show that hybrid decision trees with leaf classifiers perform slightly better than the original decision tree. Abu-Hanna and Keizer (2003) [11] explored the construction of classification trees with logistic regressions at the leaves to predict the probability of intensive care patient survivals, dividing the training set into patient sub-groups, to learn specialized models. These sub-groups of patients were created by a classification tree based on predictor variables such as temperature and heart rate, to discriminate between surviving and the nonsurvivor patients. They constructed logistic regression models at the terminal leaves based on a single variable that describes the severity of illness.

The hybrid combination of a single CART (Classification and Regression Trees) [41] tree and logistic regressions at the leaves (referred to as LogTree in this thesis) takes advantage of the strengths from both classifiers. On the one hand, CART trees, are capable of capturing non-linear relationships between the predictor variables and the target variable. A small change in the predictor variables could result in a large change in prediction. On the other hand, logistic regression models, which

capture linear relationships between predictors and damage probability (or non-linear relationships with appropriate choice of basis functions), exhibit a smooth continuous response for predicting the probability damage. Small changes in the predictor variables of a logistic model result in a small change in the predicted probability.

The LogRFT model previously introduced, composed of an ensemble of LogTree models put together via the Random Forest algorithm, benefits from the ability of random forests to generalize by bagging and by feature selection. One of the disadvantages of using a single classification tree is that it tends to overfit the data. The Random Forest algorithm tries to control the loss of generalization due to overfitting by introducing randomness into the construction of each classification tree in the ensemble [12]. By learning different hybrid trees that make uncorrelated errors, the generalization and accuracy of the overall is improved.

3.2 Dealing with uncertainty in wind speeds in training process

The hope is that the classification trees partition the training samples into groups with respect to the predictor variables (excluding wind variables) to enable effective discrimination between “damaged” and “not damaged” homes. Any variation left in the target variable within these groups, found at the terminal leaves, is assumed to be the result of stochastic wind currents experienced by the residential structures. Approximations of maximum wind speeds, necessary to explain the variation in the target variable, introduce two problems. First, while a home experiencing a given wind speed is labeled as “not damaged”, another home from the same homogeneous group at a terminal leaf experiencing a lower wind speed is labeled as “damaged”. This

occurs because of deterministic wind speed approximations utilized to describe the stochastic wind patterns that a home experiences throughout the hurricane. Second, using wind data specific to a hurricane, limits the range of speeds used in the training process. The absence of lower and higher wind speeds in the training data creates overfitted models.

To observe the effect of using wind speed approximations, consider the synthetic data presented in Table 3.1. These data represents a set of samples corresponding to a terminal leaf in the random forest. Observe that, while some “non-damaged” samples experience higher wind speeds, some reveal damage at lower speeds. Figure 3.1 shows a logistic regression trained using the synthetic data in Table 3.1. The logistic model represents the cumulative density of the probability of damage as a function of wind speed. Instead of obtaining a function that rises from low to high damage probabilities as wind speed increases, the opposite behavior is observed. Since the training data does not contain samples outside the range of 60mph-90 mph, logistic regression overfits the training data resulting in lack of generalization over higher and lower wind speeds. The model over-predicts damage at low wind speeds $P(\text{Damage}|\text{windspeed} = 0\text{mph}) = 0.62$ and under-predicts damage at high wind speeds $P(\text{Damage}|\text{windspeed} = 180\text{mph}) = 0.34$. This shows the need for a strategy to overcome the effects of overfitting on observed wind speeds.

3.2.1 Incorporation of virtual samples

I have designed a procedure that enables logistic regression to generalize over wind speed ranges absent in the training data. This procedure is based on virtual samples created from existing ones by incorporating prior knowledge [43]. In the context of wind damage prediction, a virtual sample is added according to the following policy:

Synthetic Data	
Wind speed (Mph)	Target
60	0
65	1
70	1
72	0
78	1
80	0
84	1
86	0
88	1
90	0

Table 3.1 : Synthetic data representing a set of samples from a single node. The samples reflect variability in the target variable with respect to wind speed approximations. Target variable: “not damaged”=0, “damaged”=1.

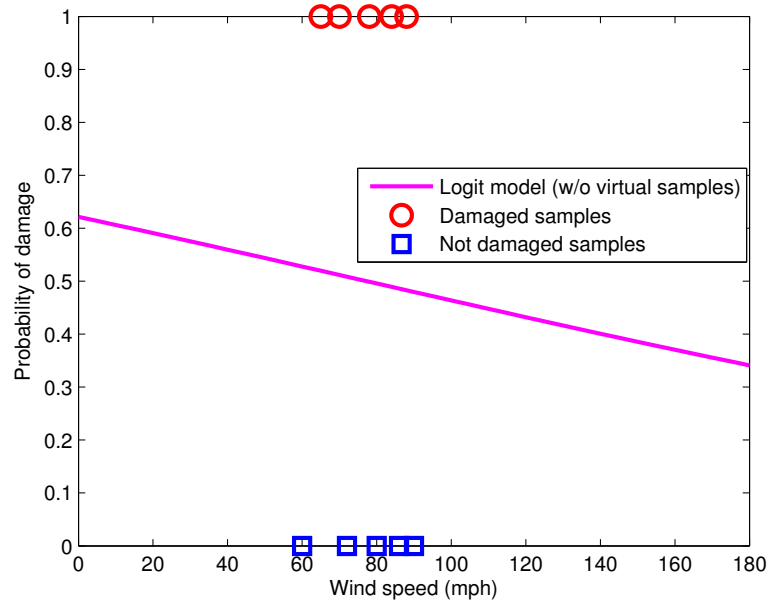


Figure 3.1 : Logistic regression trained with synthetic data from Table 3.1.

a residential structure “damaged” at wind speed w remains “damaged” at $w' > w$; if a structure is “not damaged” at w , then it remains “not damaged” at $w' < w$. In Algorithm 3.1 for adding virtual samples, a uniform probability distribution randomly generates values for w' . I use an upper bound of 177mph for the top wind speed based on largest recorded wind speeds on Earth.

Algorithm 3.1: Algorithm for generating virtual samples based on random sampling values for w' from a uniform distribution.

```

 $U(a, b) \leftarrow$  uniform distribution on the interval (a,b);
 $b \leftarrow$  original sample;  $b' \leftarrow$  virtual sample;
 $w \leftarrow$  windspeed(b);  $w' \leftarrow$  windspeed(b');
if  $class(b) = damaged$  then
    |  $w' = U(w, 177 \text{ mph});$ 
else
    |  $w' = U(0 \text{ mph}, w);$ 

```

Creating virtual samples for the synthetic data in Table 3.1 based on Algorithm 3.1 and adding them to the training set to fit a logistic regression model, enables generalization over unseen wind speeds. Figure 3.2 shows how virtual samples help logistic regression avoid over-estimating damage for low speeds and under-estimating damage for high speeds.

3.2.2 Training LogRFT with static and dynamic variables

Trained in sequence, the two-level LogRFT machine learning model begins by first fitting the random forest and then fitting logistic regressions at each terminal leaf. Each

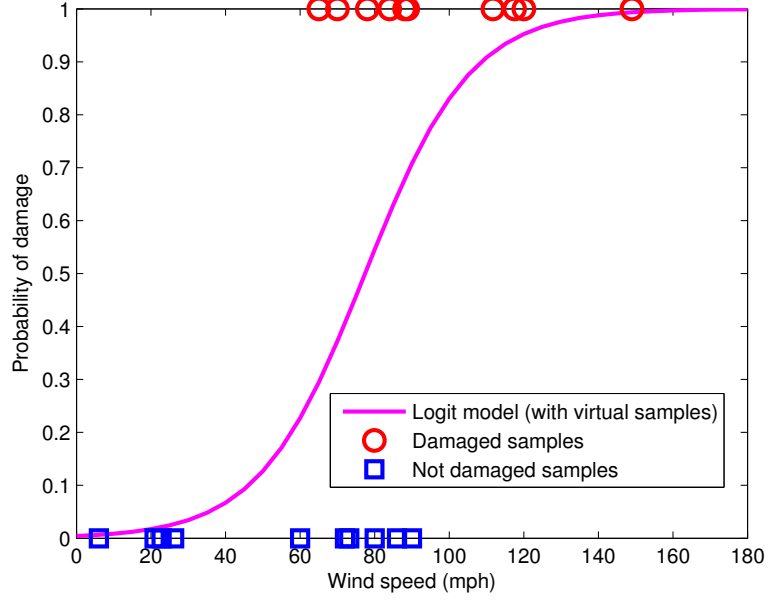


Figure 3.2 : Logistic regression trained with synthetic data in Table 3.1 and virtual samples generated according to Algorithm 3.1.

level utilizes a different set of predictor variables which divide into static and dynamic variables. Static variables, used to construct Random Forest, do not change during the hurricane event (e.g. distance to coast, distance to highway, etc.). Dynamic variables, that change during the hurricane event (e.g. wind speed, wind direction, etc.), are used to fit the logistic regression models. In order to enable logistic regressions to generalize over unseen wind speeds, the training set of dynamic variables is expanded to contain virtual samples with w' values sampled from a uniform distribution, as described by Algorithm 3.1.

Chapter 4

Case study: Prediction of wind-induced damage caused by Hurricane Ike

I applied the hybrid machine learning framework designed in chapter 3 to a large damage data set collected after Hurricane Ike in Harris County, Texas. This chapter is organized as follows: (1) I state the model evaluation metrics, learning objectives, and model validation protocol. (2) I describe the training data and results from a spatial autocorrelation analysis of the target variable. (3) I perform feature selection based on a correlation filter and embedded methods. (4) I select the subset of best predictor variables using the hybrid framework and construct a LogRFT model.

4.1 Model evaluation

4.1.1 Learning objective: areal level accuracy

Due to the difficulty of accurately predicting the probability of damage for individual residential structures, the learning objective optimizes the number of correctly predicted one-kilometer square blocks. The definition of correctly predicted blocks (in good agreement with actual observed damage) is introduced in a validation protocol in a study conducted by FEMA related to HAZUS-MH [38]. The validation protocol directly considers the uncertainty in wind speed by defining an interval of predictions made evaluating the wind-damage model at $\pm 10\%$ wind speed. I define expected damage for areal unit u evaluated at $\pm 10\%$ wind speed as the interval

$[p_{0.9w}(u), p_{1.1w}(u)]$, and $o(u)$ as the actual percentage of observed damage at the areal level. The decision rule in Equation 4.1 classifies a unit u as under predicted, over predicted, or correctly predicted. Equation 4.2 indicates the overall accuracy of the wind-damage model specified by the percentage of correctly predicted areas. Note that the averaged individual probability estimates of all residential structures inside the geographic extent of u , determines the expected probability of damage for areal unit u .

$$class(u) = \begin{cases} correct & \text{if } p_{0.9w}(u) \leq o(u) \leq p_{1.1w}(u) \\ underpredicted & \text{if } o(u) > p_{1.1w}(u) \\ overpredicted & \text{if } o(u) < p_{0.9w}(u) \end{cases} \quad (4.1)$$

$$areal_accuracy = \frac{1}{n} \sum_{i=1}^n [class(u_i) = correct] \quad (4.2)$$

Since areal level accuracy is defined using sensitivity to wind speed, models constructed without wind variables are assessed at the level of individual homes using the area under the ROC curve (AUC)[44] - a standard measure for classification accuracy.

4.1.2 Training and validation protocol

The dataset of 578,666 residences used in the construction and validation of the LogRFT hybrid model, is split randomly without replacement into two sets: true test set (30%) and training set(70%). The true test set is composed of 173,600 samples, and is used to evaluate generalization and performance of the fitted model. The training data of 405,066 samples is further decomposed into five separate equally

sized sets of approximately 81,014 samples each. These five sets are used as folds to compute cross validation estimates. Consequently, training and test sets for each cross validated fold contain approximately 324,052 and 81,014 samples respectively. The curse of dimensionality problem limits the number of samples of the final model to approximately no more than 18 predictor variables ($\log_2(324,052) = 18.3$) [45].

Three types of accuracy estimates are reported in generated graphs. True Test estimates correspond to those resulting by evaluating the trained model on the set aside true test dataset. Test estimates correspond to the performance of the trained model based on the fold test sets. Train estimates correspond to the performance of the trained model evaluated on the fold training set. Analyzing these accuracy estimates together helps identifying under- and over-fitting effects.

4.2 Description, pre-processing, and analysis of data

Supervised machine learning requires a set of predictor variables and a target variable for each residential structure. The target variable summarizes the observed outcome for each residential structure, corresponding to whether or not it was damaged by the hurricane, while predictor variables potentially explain damage incurred to a residence. In the following sections, I describe the acquisition, pre-processing, and analysis of each of these variables.

4.2.1 Target variable: Harris County House of Authority survey data

In the aftermath of Hurricane Ike's passing through Harris County, Texas in September 13, 2008, the Harris County House of Authority (HCHA) utilized 200 inspectors to conduct a survey of approximately 774,000 physical residential damage assessments [46]. Planned assessment over the entire Harris County used spatial sampling tech-

niques, ensuring surveys were performed evenly reflecting the residential density of the local communities. This residence by residence survey, between September 23, 2008 and November 12, 2008, recorded the following information about each residential structure: damage to the overall building, damage to its components (roof, wall, foundation, garage), and damage to facing and landscaping. Mostly caused by wind, roof damage is chosen as the target variable for the supervised machine learning model developed in this thesis. Inspectors categorized roof damage for each residential structure in the same discrete scale of 0 (no damage) to 4 (destruction) used by FEMA [18]. Among the total number of surveyed residential properties, only single-family residential properties were considered to construct the wind-damage model. There are 578,666 single-family residential homes in the dataset. Figure 4.1 (left) shows a histogram of the distribution of the categories of roof damage among single-family residences. Given that roof damage categories moderate (2) through total destruction (4) compose only 5.06% of the total damage, I focus this study on the prediction of whether residential structures experience any level of roof damage or none at all. Therefore, I map FEMA’s the discrete scale of 0-4 to 0 (“not damaged”) and 1 (“damaged”), where label 1 contains the merged categories from minor damage to destruction. The histogram Figure 4.1 (right) shows the distribution of “damaged” (25%) and “not damaged” (75%) residential structures.

Each residential structure is mapped to its latitude and longitude coordinates by merging the dataset with the centroids of the residential parcels provided by the Harris County Appraisal District (HCAD) from 2008 [47]. This processing is done using the spatial analysis software called ArcGIS, and it creates and saves data in an ArcGIS geodatabase of point geometry data type. Once in this format, ArcGIS offers a range of spatial analysis tools to help understand the distribution of damage to the

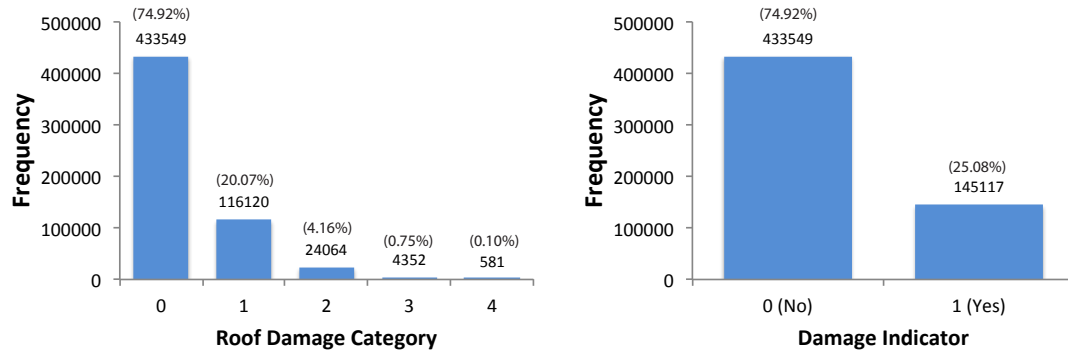


Figure 4.1 : Left: Histogram showing the distribution of categories indicating roof damage to residential structures (0:No damaged, 1:Minor, 2:Moderate, 3:Severe, 4:De-struction). Right: Histogram showing the distribution of roof damage indicator (0:No damage, 1:Damaged)

residential structures surveyed by the HCHA. Figure 4.2 shows the distribution of roof damage across the Harris County. Damage outcome is mixed in areas south of I-610 and in between SH-288 and I-35. The figure reveals of predicting wind-induced damage at the individual residence level.

4.2.2 Spatial analysis of target variable

Spatial analysis techniques applied to the damage survey data help us understand geographical patterns present in the data. Spatial analysis can be performed on location alone as well as on values and location (spatial autocorrelation). Location based analysis classifies the geographical distribution of the dataset as clustered, dispersed, or randomly distributed, whereas analysis based on location and values identifies patterns on the spatial distribution of values.

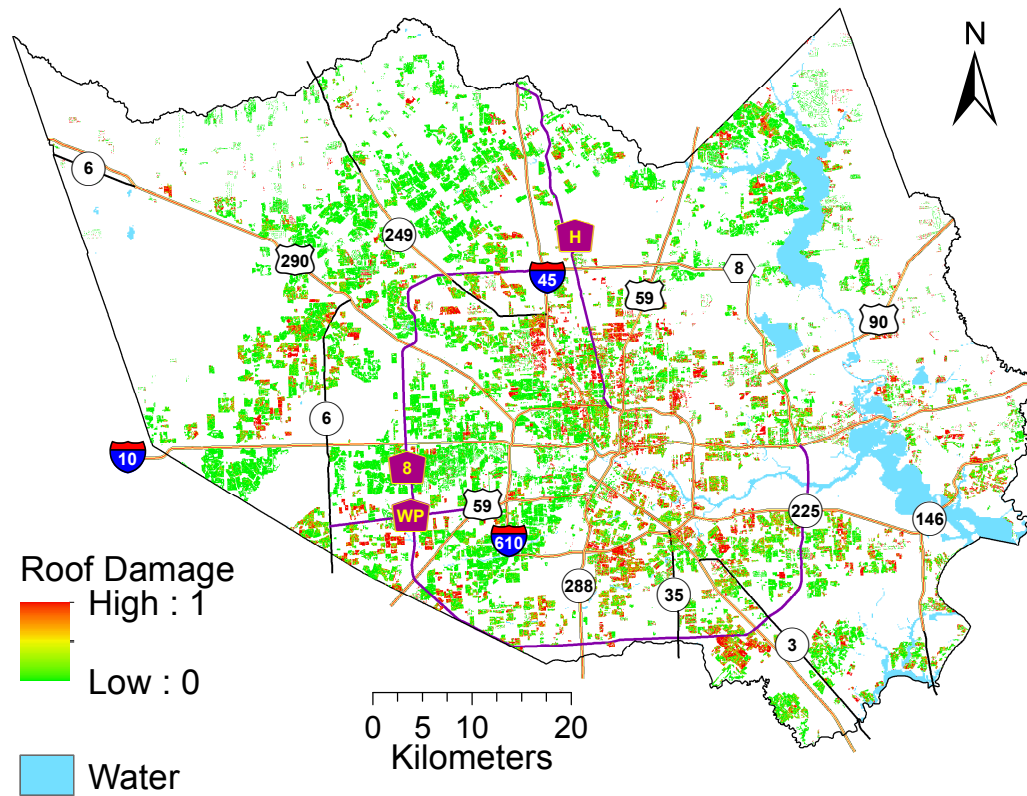


Figure 4.2 : Spatial distribution of observed roof damage caused by Hurricane Ike in 2008 to residential structures over the Harris County, Texas. Source of data: HCHA

Spatial analysis based on location itself: average nearest neighbor

The spatial distribution of damaged homes helps us understand global patterns and the closeness among them. Although residential homes in urban areas tend to lie near each other, analyzing their geographical locations gives an empirical evaluation of whether they are actually clustered. Additionally, knowing homes' proximity to each other provides an initial understanding of the spatial extent of neighborhoods. In order to analyze the geographical distribution of surveyed residential structures, I use the Average Nearest Neighbor tool from ArcGIS [48]. This tool measures the distance

between each geospatial point (representing a residential structure) and its nearest neighbor point location. The tool computes the observed average of all the nearest neighbor distances in the survey database, as well as the expected average distance based on a hypothetical spatially random distribution of points. The expected average helps analyze clustering or dispersement. The hypothetical random distribution of points contains the same number of points as the survey data and covers the same total area. If the observed average is less than the expected average from the hypothetical random distribution of points, the distribution pattern reveals clustering. However, dispersed distribution exists if the observed average is greater than the expected average. An average nearest neighbor ratio is computed by dividing the observed average distance by the expected average distance. When the ratio is less than 1, the spatial pattern exhibits clustering, and when the ratio is greater than 1, the pattern leans toward a dispersed spatial distribution [48]. The Average Nearest Neighbor tool computed an observed mean distance of 71.8ft, an expected mean distance of 145.5ft, and a nearest neighbor ratio of $71.8/145.5 = 0.493$ with a z-score of -852.56 and significance level of less than 1%. Based on these measures, the geographical distribution of residence locations reveals a clustering pattern with a less than 1% likelihood resulting from random chance.

4.2.3 Spatial analysis based on location and values: spatial autocorrelation of roof damage

Analyzing observed damage in the spatial dimension assists in understanding the effect of stochastic processes within neighborhoods in two ways: first, spatial analysis uncovers whether damage to residential structures is clustered, dispersed, or randomly distributed over the geographical space; and secondly, if damage is found to be

clustered, spatial analysis uncovers the more prominent neighborhood size at which such clustering occurs.

Nearby built and natural environments experience similar high-speed wind currents, wind direction dynamics, wind-generated debris, and exposure to wind stress for similar periods of time. Therefore, damage that nearby residential structures experience possibly originates directly from wind-loads or indirectly through debris released by neighboring damage to residential structures, trees, or other objects.

The observation that close residential structures experience more similar damage than distant structures is called *spatial dependence*. Spatial dependence, described by Tobler (1970) through his "First Law of Geography," states that everything relates to everything else, but near things relate more than distant things [49]. Spatial dependence quantified through *spatial autocorrelation*, refers to the correlation of a variable with itself over space. When the values of neighboring homes are similar, the variable exhibits positive spatial autocorrelation indicating a tendency toward clustering. In dissimilar neighboring values, negative spatial autocorrelation exists, indicating a dispersed pattern. If the neighboring values are neither similar nor dissimilar, spatial independence and values exhibit a random distribution [50].

Widely measured through Moran's I [51], spatial autocorrelation of a variable is computed with:

$$I = \frac{N \sum_{i=1}^N \sum_{j=1}^N w_{i,j} (x_i - \bar{x}) (x_j - \bar{x})}{\left(\sum_{i=1}^N \sum_{j=1}^N w_{i,j} \right) \left(\sum_{i=1}^N (x_i - \bar{x})^2 \right)} \quad (4.3)$$

where, in relationship to the the survey damage data: N corresponds to the total number of surveyed residential structures, $w_{i,j}$ to a matrix with binary values indicating whether structure i and j are within spatial distance r , and x_i to the

binary value indicating whether structure i experienced roof damage. The Moran's I equation computes a deviation from the mean for each structure's damage value x , multiplying the deviation values from each structure i by the deviation values from structures j within spatial distance r identified in matrix $w_{i,j}$. Results remain positive if both structures are damaged or both are not damaged, indicating spatial clustering. Negative product of deviations indicates spatial dispersion. When the multiplication of deviances balances to a sum of zero, a spatial random distribution of damage values exists. The Moran's I equation is divided by the variance of damage values of the surveyed structures and by $\sum_{i=1}^N \sum_{j=1}^N w_{i,j}$ in order to normalize the value to the interval $[-1.0, 1.0]$.

In order to study the spatial distribution of roof damage among residential structures, I measure spatial autocorrelation of roof damage (0/1 scale) using Moran's I at different spatial radius distances r using ArcGIS [52]. Figure 4.3 shows the Moran's I z-scores for each distance radius r in the range from 330ft to 920ft in increments of 10 ft. This figure depicts spatial autocorrelation remains positive for each spatial distance r , suggesting that roof damage is clustered spatially at multiple levels. Therefore, processes explaining roof damage are in action at many spatial levels and not confined to a single level. Therefore, choosing any distance r within the range 330ft to 920ft is justified for performing spatial analysis of roof damage.

Assuming that the probability of roof damage for a given home is higher when surrounded by damaged homes, I analyze spatial autocorrelation at the extent of census blocks (the smallest geographical unit used by the United States Census Bureau). The U.S. Census Bureau describes census blocks within a city as "generally bounded on all sides by streets" [53]. By computing the count of surveyed homes within each census block, I compute the average count of homes k within all census

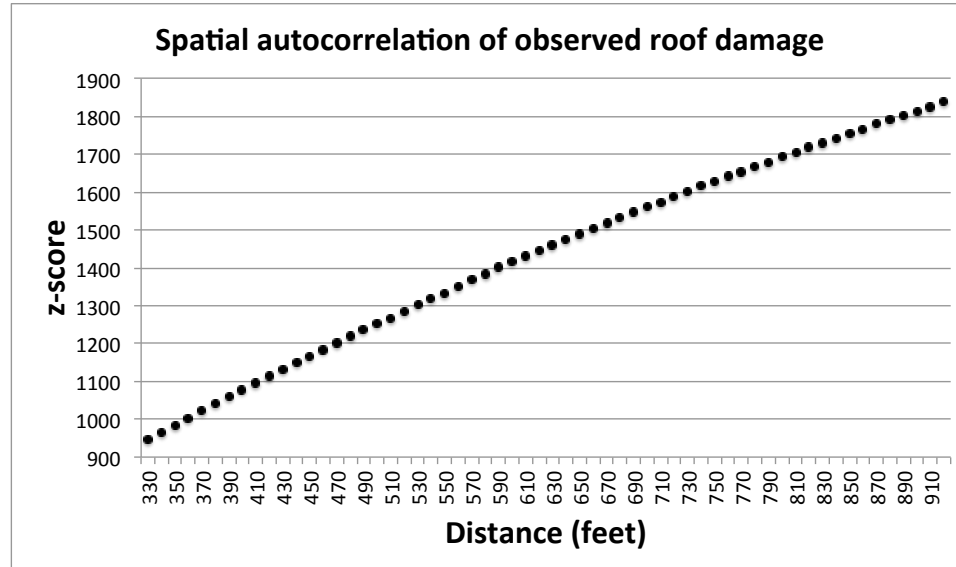


Figure 4.3 : Moran's I spatial analysis of observed roof damage at different spatial radius distances r (330ft to 920ft).

blocks across the Harris County, Texas. Interpreting this average count as the average number of neighbors around surveyed homes, I select a distance band radius with the same average number of neighbors for use in spatial analysis [54]. To do so, I first obtain the census block geospatial dataset from 2008 for the Harris County; second, surveyed homes within each census block aggregate by counting; third, I compute the average number of homes within the census blocks. The mean number of surveyed homes within the census blocks is 19. Finally, to obtain the average distance from each surveyed home to its nineteenth nearest neighbor, I use the ArcGIS tool called Calculate Distance Band from Neighbor Count [55]. This tool calculates the average distance to the 19th nearest neighbor to be $r = 362\text{ft}$. This neighborhood level of analysis at which roof damage clusters, helps me create predictor variables that

reflect structural and terrain characteristics of the neighborhood surrounding each residential structure.

4.2.4 Predictor variables

The selection of predictor variables used in the development of the hybrid machine learning model include those deemed influential toward damage outcomes. These variables obtained from different sources, fall into the following categories: structure-related variables, construction code enforcement variables, terrain configuration variables, and wind hazard variables.

Structure-related variables

I acquired information about residential structures as of 2008 through the public records maintained by the Harris County Appraisal District (HCAD) [47]. As proxy for the structural quality of the residential structure and its capacity to withstand damage, I selected the following variables from these public records:

- *building value* (in 2008 dollars). The value of the building is a reasonable proxy for its structural quality. This feature was acquired from table “real_acct” column “impr_value”.
- *land value* (in 2008 dollars). Highly assessed residential property land correlates with the economic solvency of the owners and the quality of their homes. Acquired from table “real_acct” column “land_value”, the spatial distribution of land and building value observed in Figure 4.4, demonstrates that expensive land to the west of downtown and inside I-610 correlates with expensive residences around the same geographical area.

- *extra features value* (in 2008 dollars). Acquired from table “real_acct” column “extra_features_value” this value includes additions to the building and land.
- *number of stories*. Acquired from table “fixtures” column “units” with column “type”={’STY’, ’STC’}. Buildings with a higher number of stories are more exposed to wind currents.
- *age* in years (2008 - the year of home construction). The vulnerability of a residential structure to strong winds relates directly to its age. In other words, the older the structure, the more vulnerable it becomes. By subtracting the year in which a home was built from the year 2008, I obtained the age of the structure.
- *remodeled age* in years (2008 - the year of a home’s last remodeling). Acquired from table “building_res” column “yr_remodel”, to account for structures previously remodeled and improved. I calculated the age of the structure relative to its last remodeled date. If no remodeling is recorded, then age is calculated relative to the date of construction. Figure 4.5 shows the spatial distribution of the residential structures’ age and remodeled age. Observe that residential structures within Beltway 8, west of Houston’s downtown, and south of I-10, remodeled age is lower than the construction age, indicating recent remodeling to these residential structures.
- *building quality*. Acquired from table “building_res” column “quality”, this feature maps the quality of a building on a discrete scale of 1 to 6, 6 being the best and 1 being the worst.

- *detached garage* (0 for no, 1 for yes). Acquired from table “extra_features” column “cd”={’CRG4’, ’CRG5’, ’RRG1’, ’RRG2’, ’RRG3’, ’RRG4’}, the presence of detached garages increase the probability of damage to a residential structure because they are likely to be damaged by wind-borne debris.
- *exterior wall* (0 for no, 1 for yes). Acquired from table “structural_elem” column “type”=’XWR’, the exterior wall greatly influences the resistive capacity of a residential structure. I identified residential structures with strong exterior wall materials such as brick/masonry, frame/concrete, stucco, brick/veneer, and stone.

Given that observed roof damage significantly clusters at multiple spatial extents, I only considered spatial autocorrelation in neighborhoods delineated by a radius of 362 ft surrounding each residential structure. To do so, I averaged the previously mentioned structure variables based on a 362 ft radius and added them to the set of predictor variables (refer to Section 4.2.3 for explanation on this radius selection).

Furthermore, I acquired 2012 building footprint database corresponding to buildings in Harris County from Rice University’s GIS/Data Center [56]. From this database I obtained the *building perimeter length* (in feet) and *building area* (in squared feet) of just the residential structures found in the survey dataset from 2008. These two variables capture the surface area exposed to high wind currents.

The descriptive statistics for all structure variables are shown in Table 4.1. The table includes the Pearson’s correlation with roof damage and the Moran’s I spatial autocorrelation measure with respect to a 362 ft radius.

Descriptive Statistics of Structure Variables							
ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
1	Land value in 2008 dollars (LANDVALUE)	54637.357	121281.693	100.0	6845400	-0.0939	-0.0942
2	Mean land value within 362 feet (LANDVALMN)	54430.860	116408.375	900.0	6775130	-0.0982	-0.0979
3	Building value in 2008 dollars (BLDGVALUE)	112559.282	111959.806	5000.0	6073131	-0.1527	-0.1393
4	Mean building value within 362 feet (BLDGVALMN)	112396.509	89649.437	5937.7	3078260	-0.1751	-0.1741
5	Extra features value in 2008 dollars (EXTFEATVAL)	1416.356	3531.653	0.0	180612	-0.0564	-0.0540
6	Mean extra features value within 362 feet (EXTFMN)	1413.026	1642.699	0.0	36904	-0.1168	-0.1166
7	Quality of structure (QUALITYCD)	4.192	0.740	1.0	6	-0.1888	-0.1801
8	Mean quality of structure within 362 feet (QUALMN)	4.192	0.634	1.0	6	-0.2107	-0.2099
9	Age of structure with respect to 2008 (AGE)	33.624	20.434	2.0	210	0.1535	0.1347
10	Mean age within 362 feet (AGEMN)	33.642	18.466	2.0	120	0.1504	0.1497
11	Remodeled age of structure with respect to 2008 (REMOAGE)	30.727	19.407	2.0	210	0.1647	0.1453
12	Mean remodeled age within 362 feet (REMAGEMN)	30.739	16.038	2.0	99	0.1773	0.1767
13	Number of stories (NUMSTORIES)	1.323	0.488	1.0	4	-0.1157	-0.1227
14	Mean number of stories within 362 feet (NUMSTMN)	1.323	0.349	1.0	4	-0.1712	-0.1706
15	Building perimeter in feet (BLDNG_LEN)	243.908	102.229	50.121	4992.41	-0.1042	-0.0979
16	Building area in squared feet (BLDG_AREA)	2704.786	1890.554	156.214	215126.97	-0.0843	-0.0790
17	Mean of exterior brick wall indicator within 362 feet (EXTWALLMN)	0.750	0.367	0.000	1	-0.1339	-0.1338
18	Mean of detached garage indicator within 362 feet (DETGARMN)	0.248	0.287	0.000	1	-0.1063	-0.1063
ID	Indicator Variables	Mean	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)			
19	Brick wall (EXTWALL)	0.750	-0.1181	-0.1133			
20	Detached garage (DETCGAR)	0.248	-0.0712	-0.0704			

Table 4.1 : Descriptive statistics for structure predictor variables.

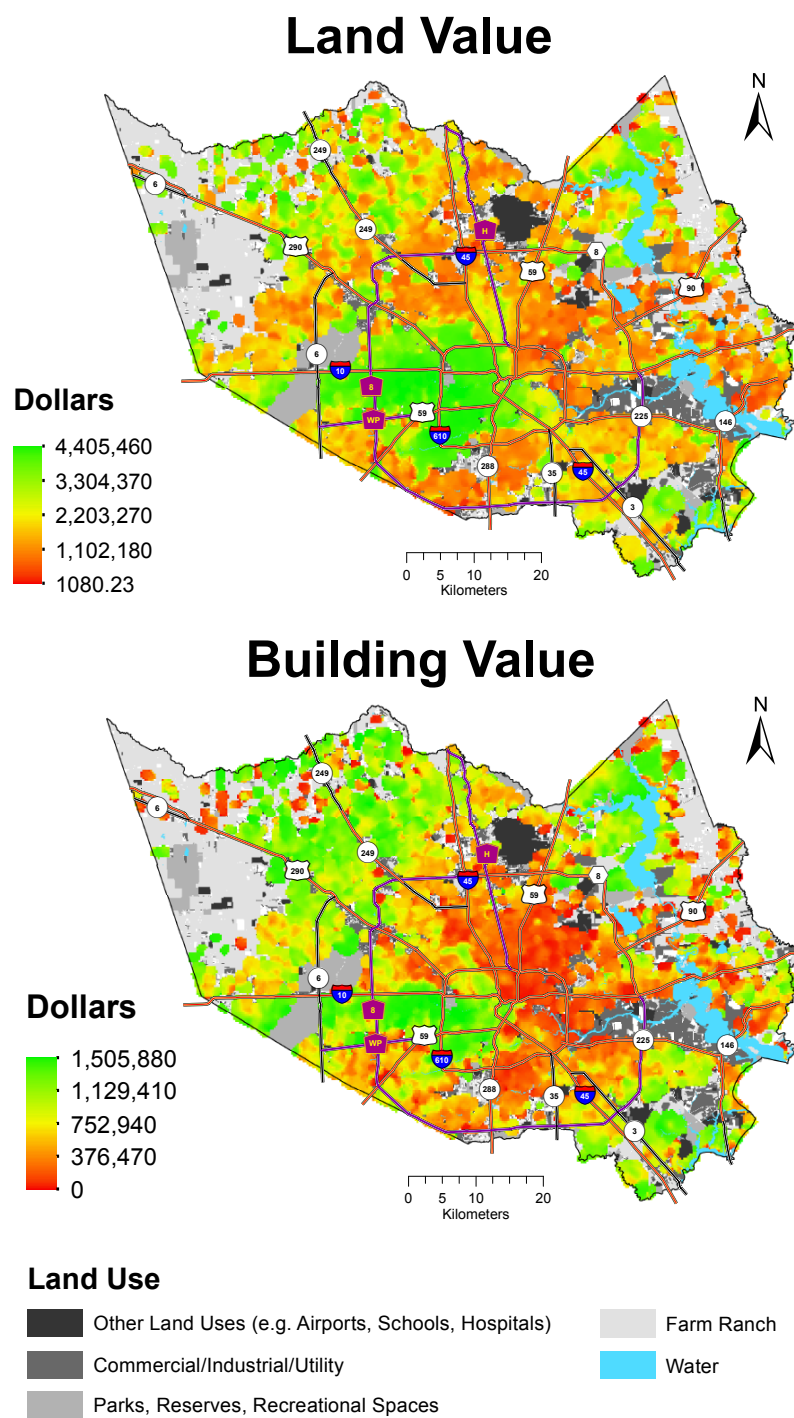


Figure 4.4 : Geographical distribution of land value and building value (in 2008 US dollars)

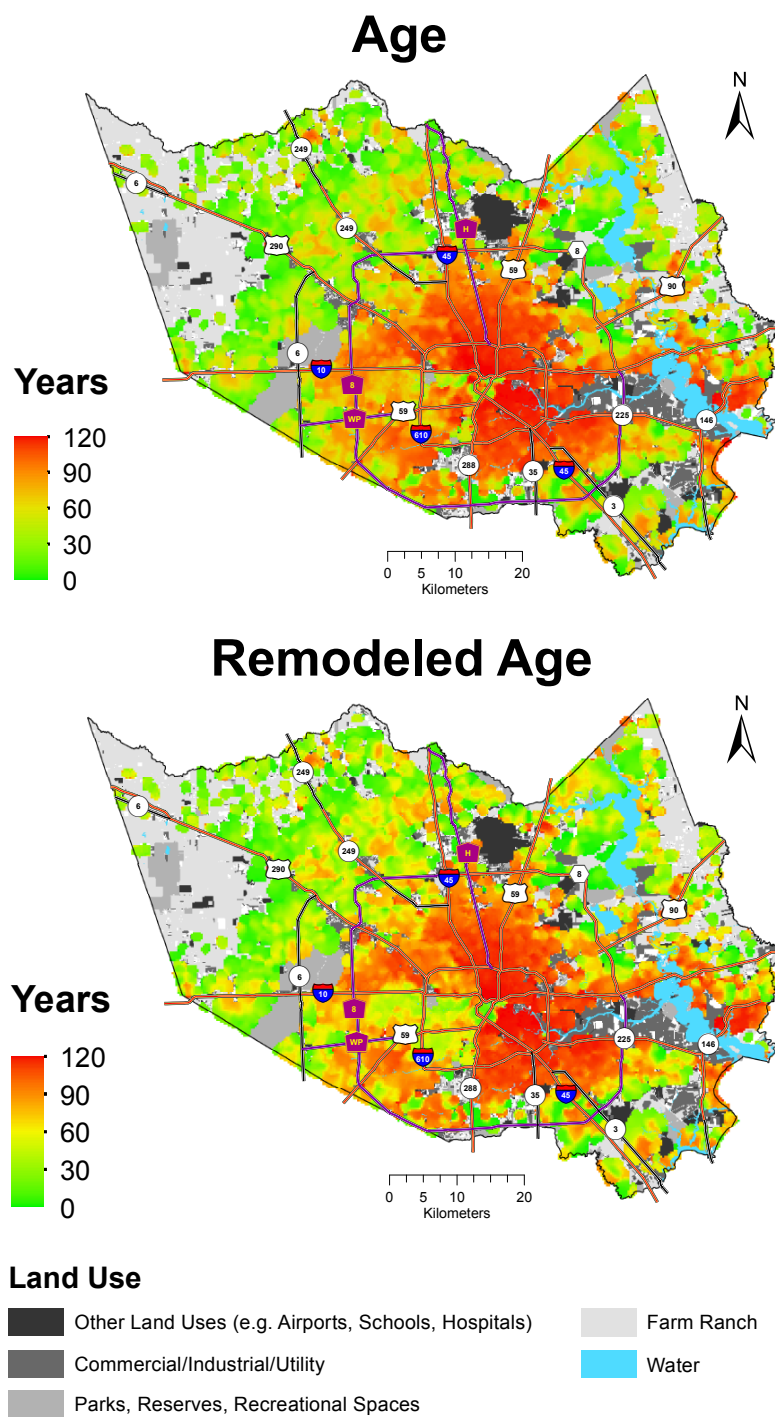


Figure 4.5 : Geographical distribution of building age and remodeled age.

Construction code enforcement variables

Construction codes for residential structures within city limits are enforced in order to regulate the quality of new residences. Residential structures built outside or before its inclusion into a city, are likely to be less rigid and have inferior quality. To account for differences in construction quality, I included the following two variables: *incorporation year* and *within incorporated area* (0 for no, 1 for yes). The *incorporation year* from each of the thirty-four cities that intersect Harris County, Texas was collected from public information and recorded in each city polygon acquired from GIS HCAD [57]. I spatially joined the city polygons with the GIS survey dataset. Incorporation years range between 1837 and 1970. After the spatial join, I calculated the variable *within incorporated area* to indicate whether or not the building is inside an incorporated zone. Table 4.2 shows the descriptive statistics for these variables.

Descriptive Statistics of Construction Code Enforcement Variables							
ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
21	Incorporation year (INCRP_YR)	1044.328	923.052	0	1970	0.0702	0.0702
Indicator Variables		Mean	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)			
22	Within incorporated area (INCAREAQ)	0.562	0.0707	0.0707			

Table 4.2 : Descriptive statistics for construction code enforcement variables.

Terrain configuration variables

I acquired and computed a number of variables from geographical datasets as proxies for the exposure of residential structures to wind loads and flying debris. The first variable, *roughness length*, is equivalent to the height at which the wind speed theoretically becomes zero depending on the height of terrain elements in the area. Acquired from the NLCD 1992 land cover data set [58] featuring a 30 by 30 meter resolution and classified according to the National Land Cover Database (NLCD) classification system [59], the *roughness length* variable, identifies exposure to higher wind loads. With the help of aerial photographs and NLCD classification descriptions, roughness lengths empirically estimated for individual areas of Texas, were published in the HAZUS-MH 2.1 Technical Manual Table 3.9 (page 128) [60]. Using this table, I mapped the ten NLCD land cover categories to roughness length values as shown in Table 4.3.

Land Cover Class	NLCD 1992 Class	HAZUS-MH
		Technical Manual Table 3.9 Roughness Length
Open Water	11	0.01
Cultivated	81-85	0.07
Herbaceous Wetland	92	0.1
Grassland/Shrub	51,71	0.1
Bare	31-33	0.15
Developed open space	21	0.35
Developed lower intensity	23	0.44
Woody wetlands	91	0.5
Developed higher intensity	22	0.55
Forest	41-43	0.56

Table 4.3 : Mapping of National Land Cover Database classes into roughness length values based on Table 3.9 (page 128) in the HAZUS-MH 2.1 Technical Manual.

The *distance to coast* in feet from each residential structure was computed in order to account for the decreasing wind force of the hurricane, as it makes its way through land. In order to compute this feature, I created a line in ArcGIS which delineated the gulf coast area closer to Harris County, Texas. Given the geospatial survey dataset represented as points, I calculated the closest distance to such line delineating the coast with the aid of the Near (Analyst) ArcGIS tool [61].

Roads and highways provide an unobstructed channel for storm winds that directly hit nearby exposed residential structures. In order to identify the structures probably damaged by this exposure, I merge major Houston highways with other major routes, and used them to compute the variable *highway distance*. This variable is equal to the Euclidean distance in feet from each residential structure to the closest highway. I acquired the two road datasets from the GIS system of the city of Houston’s public works department [62].

According to Han et al. (2009) [63], soil moisture impacts the stability of electricity poles and trees. When the soil is highly saturated with water, the probability of poles and trees falling down increases, and it affects the surrounding built environment. To account for this type of hazard, I acquired soil variables from the State Soil Geographic (STATSGO) database available through the United States Geological Survey (USGS) [64]. These variables include: *available water capacity* (in inches per inch); *percentage of soil consisting of clay* (in percent of material less than 2mm in size); and *permeability of soil* (in inches per inch). Nateghi et al. (2011) [65] considered features measuring the soils’s moisture and clay content of soil as important for predicting duration of power outage.

I also computed the feature “*tree within 100 meters (328ft)*” from a dataset found in the GIS system of the Houston’s Department of Public works containing 193,000

geo-referenced trees located mostly in Houston’s downtown [62]. I base the radius of analysis chosen on the spatial autocorrelation analysis in Section 4.2.3. Although this dataset does not report tree data for the entire Harris County, it aids assessment of damage due to tree blown debris caused to residential structures within Houston’s downtown.

From the Houston-Galveston Area Council [66], I acquired the 2008 HGAC Land Cover raster dataset categorizing the entire Harris County in the NLCD classification system of 10 categories: developed, higher intensity (1); developed, lower intensity (2); developed, open space (3); cultivated (4); grassland/shrub (5); forest (6); woody wetland (7); herbaceous wetland (8); bare (9); and open water (10). For each of the 10 categories, I computed a variable characterizing the Euclidean distance in meters from each residential structure to the closest area of each category. This set of variables captures the surrounding terrain around each residential structure. Additionally, I mapped the 10 classes into three terrain types: wood, open, and developed terrain. Table 4.4 shows the mapping of HGAC land cover classes into the tree terrain types. For each residential structure, I computed the percentage area covered by each of these three terrain types with respect to a circle of radius 362 feet around each home. I called these features: *wood terrain percent*, *open terrain percent*, and *developed terrain percent*. The collection of these three predictors capture the surrounding terrain of each residential structure. Wooded terrain is source of debris and trees that causes increase in probability of damage. High wind currents flowing over open terrain collide freely against exposed residential structures. Obstructed wind currents flowing through developed terrain cause turbulence and creation of wind gusts, which are considered a source of damage to homes.

Terrain Type	Land Cover Class Description	HGAC 2008 Class Code
Developed	Developed, higher intensity	1
	Developed, lower intensity	2
Wooded	Forest	6
	Woody wetland	7
Open	Developed, open space	3
	Cultivated	4
	Grassland/shrub	5
	Herbaceous wetland	8
	Bare	9
	Open water	10

Table 4.4 : Categorization of HGAC land cover classes into wooded, open, and developed terrain.

To further capture the effects of hurricane wind currents given the built and natural environment, I obtained from a remote sensing technology called Light Detection and Ranging (LIDAR), raster datasets describing the continuous terrain height measures (in feet) over Harris County, Texas from the year 2008 [67]. Using the 2012 building footprint database for Harris County from HCAD [47], I computed the *maximum height of building area* and the *mean height of building area*. These variables are vital for capturing the discrepancy in height within the building’s footprint due to external objects going over the home (i.e. a tree extending its branches above the home). Additionally, I computed the *maximum height*, *mean height*, and *standard deviation height* from each residential structure to a radius of 25 meters (82 ft) and from 25 meters to 50 meters (164 ft). Used as proxy, these variables, explain damage caused by sources of debris and trees falling over the structure.

The descriptive statistics for all the terrain configuration variables are shown in Tables 4.5 and 4.6.

Descriptive Statistics of Terrain Variables (Part 1)

ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
24	Roughness length (ROUGHLEN)	0.365	0.182	0.010	0.560	0.0093	0.0071
25	Distance to coast in feet (COAST_DIST)	260011.629	60710.024	111691.000	430150.000	-0.1525	-0.1525
26	Euclidean distance in feet to closest freeway (FREEWAYED)	4424.843	3650.064	0.000	20810.200	-0.0629	-0.0629
27	Soil available water capacity in inches per inch (AWC)	0.162	0.014	0.170	0.076		0.0763
28	Percent of soil consisting of clay (CLAY)	31.052	11.845	-0.100	49.800	0.1116	0.1118
29	Permeability of the soil in inches per hour (PERM)	0.808	1.030	-0.100	7.000	-0.0874	-0.0874
30	Euclidean distance in meters to HGAC LC Category 1: Developed, Higher Intensity (ED1)	31.105	37.015	0.000	1020.000	-0.0435	-0.0411
31	Euclidean distance in meters to HGAC LC Category 2: Developed, Lower Intensity (ED2)	9.582	25.844	0.000	2188.360	-0.0113	-0.0134
32	Euclidean distance in meters to HGAC LC Category 3: Developed, Open Space (ED3)	342.400	247.751	0.000	3905.300	-0.0827	-0.0822
33	Euclidean distance in meters to HGAC LC Category 4: Cultivated (ED4)	9025.717	5956.358	0.000	23700.600	0.0571	0.0571
34	Euclidean distance in meters to HGAC LC Category 5: Grassland/Shrub (ED5)	694.110	710.094	0.000	5208.000	-0.0152	-0.0152
35	Euclidean distance in meters to HGAC LC Category 6: Forest (ED6)	790.678	676.411	0.000	4150.860	-0.0018	-0.0019
36	Euclidean distance in meters to HGAC LC Category 7: Woody Wetland (ED7)	957.289	863.617	0.000	5677.530	-0.0301	-0.0301
37	Euclidean distance in meters to HGAC LC Category 8: Herbaceous Wetland (ED8)	1007.877	812.856	0.000	6179.850	-0.0282	-0.0283
38	Euclidean distance in meters to HGAC LC Category 9: Bare (ED9)	1701.120	1244.665	0.000	7097.900	0.0097	0.0097
39	Euclidean distance in meters to HGAC LC Category 10: Open Water (ED10)	1094.562	782.387	0.000	5587.580	0.0150	0.0150
40	Percentage of open terrain within a radius of 362 feet based on HGAC LC (OPEN_PER36)	0.072	0.101	0.000	1.000	-0.0221	-0.0225

Table 4.5 : Descriptive statistics for terrain variables Part 1.

Descriptive Statistics of Terrain Variables (Part 2)

ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
41	Percentage of developed terrain within a radius of 362 feet based on HGAC LC (DEVT_PER36)	0.895	0.138	0.000	1.000	0.0370	0.0371
42	Percentage of wood terrain within a radius of 362 feet based on HGAC LC (WOOD_PER36)	0.033	0.078	0.000	0.914	-0.0369	-0.0366
43	Maximum height of building area in feet based on LIDAR 2008 (BLDGHTMAX)	31.327	15.048	8.013	253.461	-0.0904	-0.0990
44	Mean height of building area in feet based on LIDAR 2008 (BLDGHTMEAN)	15.743	5.684	0.113	104.696	-0.1113	-0.1219
45	Maximum terrain height within a radius of 25 meters based on LIDAR 2008 (MAX0_25M)	43.900	19.265	-0.331	253.849	-0.0709	-0.0718
46	Mean terrain height within a radius of 25 meters based on LIDAR 2008 (MEA0_25M)	12.414	8.506	-0.331	109.193	-0.0850	-0.0863
47	Standard deviation of terrain height within a radius of 25 meters based on LIDAR 2008 (STD0_25M)	11.227	5.372	0.000	51.465	-0.0751	-0.0770
48	Maximum terrain height within a radius of 25 to 50 meters based on LIDAR 2008 (MAX25_50M)	43.782	21.227	-0.470	498.658	-0.0701	-0.0685
49	Mean terrain height within a radius of 25 to 50 meters based on LIDAR 2008 (MEA25_50M)	12.194	8.975	-1.714	362.346	-0.0829	-0.0805
50	Standard deviation of terrain height within a radius of 25 to 50 meters based on LIDAR 2008 (STD25_50M)	10.900	5.867	0.000	179.803	-0.0742	-0.0718
ID	Indicator Variables	Mean	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)			
51	Within 100 meters of tree (TREE100M)	0.076	0.0031	0.0034			

Table 4.6 : Descriptive statistics for terrain variables Part 2.

Wind hazard variables

The National Oceanic and Atmospheric Administration (NOAA) released real time snapshots of wind fields in 6 hour intervals before, during, and after hurricane Ike made landfall near Galveston Texas. Available in ArcGIS shapefile format, these data releases contain grids of points uniformly separated by 5 km with measurements of 1-minute maximum sustained wind speeds [68]. Taking into consideration all grid datasets released before and after Hurricane Ike that intersect Harris County, I converted wind speed data to raster form based on spatial kriging interpolation, and merged them together by computing the maximum speed among all the rasters. Kriging interpolation estimates the geographical surface of values from a scattered set of points. Note that kriging does not take into consideration the terrain configuration and its effect on earth surface wind speeds. Then, the resulting single raster contains the *maximum wind speed* used to assign each of the buildings its corresponding speed. Additionally, NOAA computed the H*wind analysis [69] over Harris County after Hurricane Ike. This analysis provides wind characteristics, after analyzing and adjusting for the height and exposure of anemometers, used to capture the storm. These wind hazard variables obtained from the H*wind analysis include: 1-minute maximum sustained *wind swath*, *wind direction*, *wind duration*, and *wind steadiness*. *Wind duration* is the duration of sustained winds over 34 meters/second, and it is a measure of the cycles of gusts and lulls in a turbulent wind field [70]. *Wind steadiness* is the ratio of the vector mean of wind velocity to its scalar mean over the time period required for a storm to traverse the region [70]. Values range from 0% to 100%. Areas receiving large wind direction shifts due to the passage of the eye experience low values of steadiness, on the order of 10%-20%. Strong winds combined with low steadiness account for significant damage to structures. Descriptive statistics of these

wind hazard variables are shown in Table 4.7. Shown in Figure 4.6, the spatial distributions of variables *maximum wind speed*, *wind swath*, *wind direction*, and *wind steadiness* are included.

Descriptive Statistics of Wind Hazard Variables							
ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
52	Hurricane Ike maximum wind speed in MPH (MAXWIND)	69.488	4.478	51.775	79.992	0.02065	0.0206
53	Hurricane Ike wind swath in MPH (WINDSWAMPH)	76.638	6.092	50.696	85.154	0.07971	0.0797
54	Hurricane Ike wind direction in geographic degrees (WINDDIR)	250.162	19.223	187.139	287.430	-0.09782	-0.0978
55	Hurricane Ike wind duration in secs (WINDDUR)	0.920	0.657	-0.052	2.351	0.13293	0.1329
56	Hurricane Ike wind steadiness in secs (WINDSTEAD)	0.344	0.093	0.120	0.489	-0.09303	-0.0930

Table 4.7 : Descriptive statistics for Hurricane Ike wind hazard characteristics.

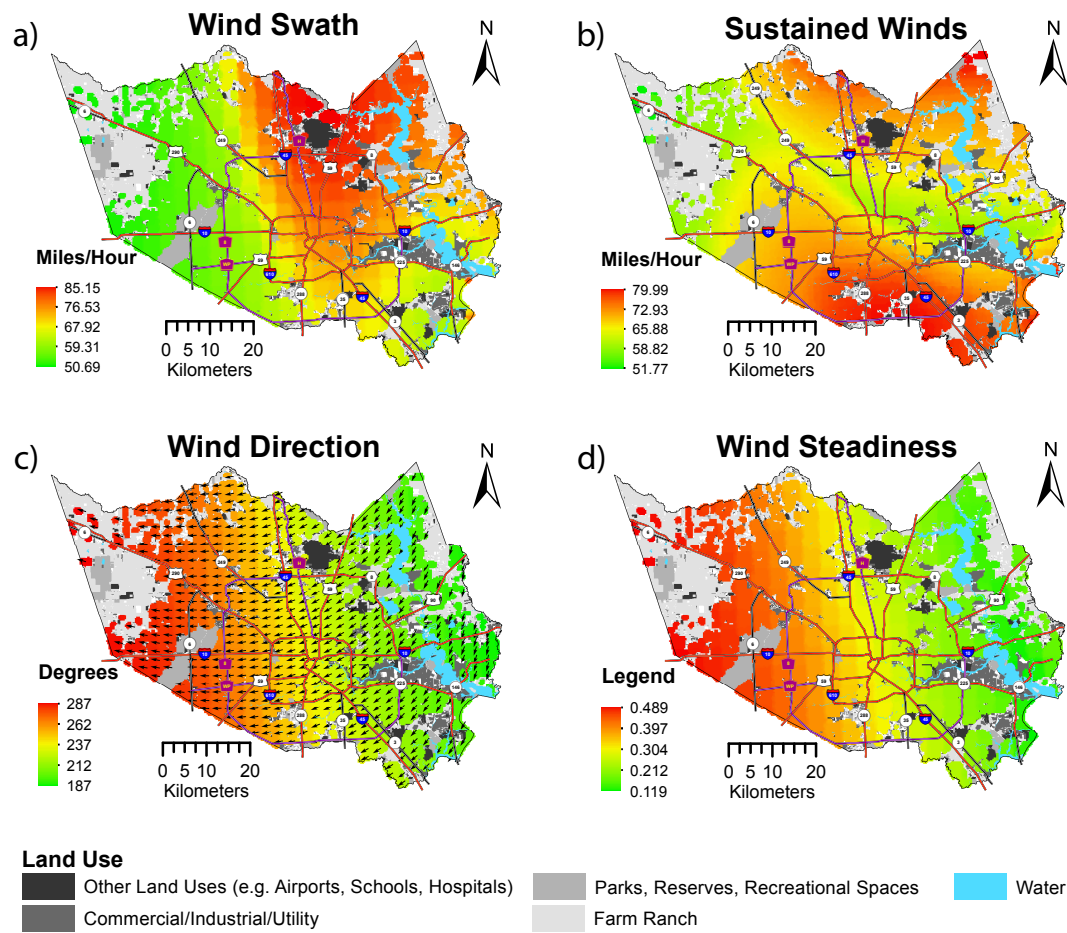


Figure 4.6 : Distribution of wind variables over the Harris County, Texas observed in Hurricane Ike, 2008. a) Wind swath (WINDSWAMPH). b) Maximum sustained wind speeds (MAXWIND). c) Wind direction (WINDDIR). d) Wind steadiness (WINDSTEAD).

4.3 Feature selection

By fitting parameters based on available data, supervised machine learning models recognize patterns in the data that describe the underlying process generating observed outcomes. Irrelevant or redundant predictor variables degrade the model's performance. Irrelevant variables add noise to the learning process. Redundant variables do not add new knowledge relative to that already covered by other variables. Identifying these types of variables and selecting the subset of variables that best captures the patterns of the underlying process is called *feature selection* [71]. The benefits of feature selection include: increase in predictive performance, reduction in the time and space of the model construction process, and reduction of the effects of the curse of dimensionality [72] [73].

Feature selection methods in machine learning research fall into three categories: *wrappers*, *filters*, and *embedded methods* [72]. The methods differ in the techniques used to evaluate the value of subsets of variables. Wrappers are those that apply a specific learner to a subset of variables to assess their value. Filter methods are independent of a learning method. Instead, filters evaluate the value of a subset of variables by analyzing the interactions between the variables through information theoretic or correlation measures [72]. Embedded methods use the set of all variables to fit a model, and then analyze the model to infer the importance of the variables. Finding the optimal subset of predictor variables is a problem known to be NP-hard [74]. Therefore, feature selection methods find an approximately optimal subset of predictor variables based heuristic search techniques.

Despite the fact that wrappers explore the space of features optimal for the intended learning algorithm, due to the repeated evaluation each time a new feature subset is considered, this method becomes computationally intensive. Much faster

than wrappers, filter and embedded methods combined, reduce the number of variables [75].

First, I choose to apply a correlation-based filter to eliminate irrelevant and redundant variables. I proceed by applying embedded feature selection methods for static and dynamic variables separately. Using LASSO logistic regression, I find the most important dynamic (wind) variables. Advancing, I apply feature selection based on the variable importance measure embedded in the Random Forest to static (structure and terrain related) variables.

4.3.1 Correlation-based Feature Selection (CFS)

One of the most popular feature selection filters is Correlation-based Feature Selection (CFS) developed by Hall (1999) [75]. This filter defines irrelevant and redundant variables in terms of correlation evaluated among subsets with combinations of predictor variables and the target variable. While a non-correlated variable with respect to the target variable is considered irrelevant, a highly correlated variable with predictor variables already in the subset, is considered redundant. Based on the assumption that the optimal variable subset includes highly correlated variables with the target variable, with minimal mutual correlation, CFS measures the quality of a subset of variables with the following equation:

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (4.4)$$

where M_S is the “merit” of subset S containing k variables, $\overline{r_{cf}}$ is the average correlation between the target variable c and every predictor variable $f \in S$, and $\overline{r_{ff}}$ is the average correlation among all pair combinations of predictor variables S . Figure

4.7 shows a 3-dimensional plot of Equation 4.4 for k values of 5 and 15. As the number of variables in S increases, the plane stretches towards higher values of M_S for high values of \bar{r}_{cf} and low values of \bar{r}_{ff} . Furthermore, adding a new predictor variable to S highly correlated with the target variable and not correlated to other features, results in a higher merit measure. While CFS assumes conditionally independent predictor variables given the target variable, it works well when this assumption is moderately violated [75]. Non-linear interactions between predictor variables can be considered by adding product terms to the set.

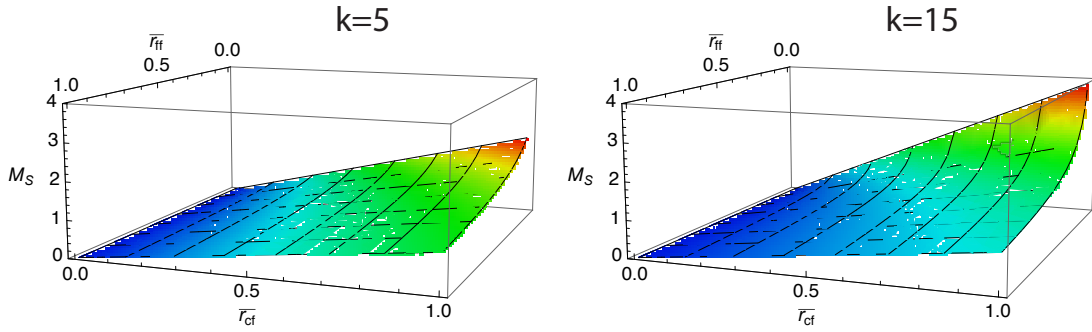


Figure 4.7 : 3-dimensional plot of Equation 4.4 for k values of 5 (left) and 15 (right).

Selecting relevant wind damage predictor variables

Given the large number of acquired predictor variables described in Section 4.3.1, I consider CFS to filter out redundant and irrelevant variables. Before applying CFS, I perform two pre-processing steps: *variable transformation* and *joining of variables*.

By transforming high skewed variables with a logarithmic function, their distribution becomes more normally distributed. The specific transformation is:

$$\hat{X} = \ln(X + 2) \quad (4.5)$$

where \hat{X} is the transformed variable X . Adding two to X avoids infinity values in the computation of \hat{X} where X is zero. I chose the addition of constant two instead of one due to the minimum value of variable MEA25_50 = -1.717 as observed in Table 4.6. Negative values of MEA25_50 occur when estimations of bare earth elevation in the LIDAR dataset are higher than the surface elevation [76]. Table 5.2 includes the complete set of variables chosen for transformation. Although transformed, the name of these variables remains the same in future references. For illustration, Figure 4.8 shows the effect of the logarithmic transformation on variable BLDGVALUE. On the left of this figure, the Q-Q plot of the original values for BLDGVALUE shows a high positive skew and non-linear relation with respect to the normal distribution. After applying the logarithmic transformation in Equation 4.5, the transformed values of BLDGVALUE becomes statistically closer to a normal distribution as shown in the Q-Q plot in the right of Figure 4.8.

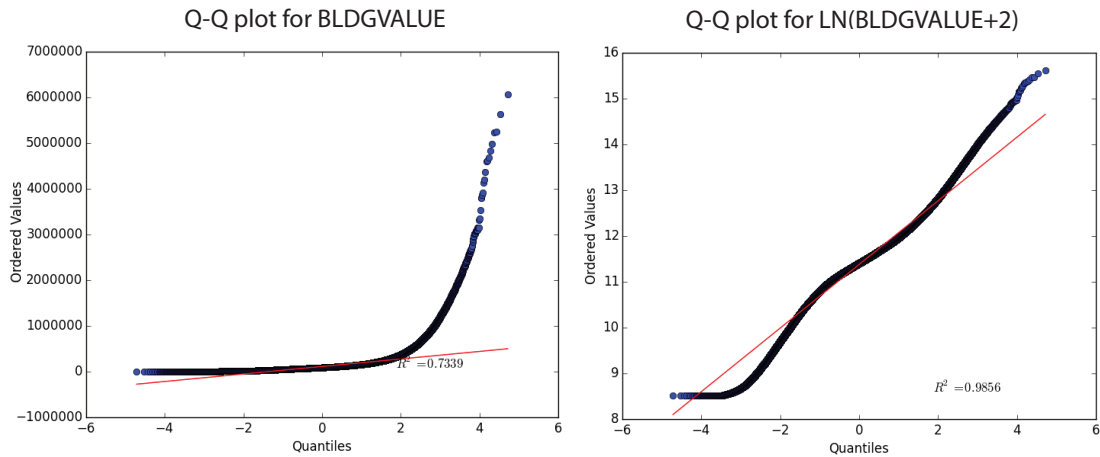


Figure 4.8 : Q-Q plots for variable BLDGVALUE (left) and its logarithmic transformation (right).

Transformed Variables	
Terrain Variables	Structure Variables
COAST_DIST	LANDVALUE
FREEWAYED	BLDGVALUE
ED1	EXTFEATVAL
ED2	AGE
ED3	REMOAGE
ED4	LANDVALMN
ED5	BLDGVALMN
ED6	EXTFMN
ED7	AGEMN
ED8	REMAGEMN
ED9	BLDNG_LEN
ED10	BLDG_AREA
OPEN_PER36	
DEVT_PER36	
WOOD_PER36	
BLDGHTMAX	
BLDGHTMEAN	
MAX0_25M	
MEA0_25M	
STD0_25M	
MAX25_50M	
MEA25_50M	
STD25_50M	

Table 4.8 : Set of selected variables exhibiting highly skewed distributions which were transformed using Equation 4.5.

To enable CFS to consider non-linear interactions among variables, the second pre-processing step consisted of joining pairs of variables by multiplication. After graphing their values against each other, I visually selected the complete set of join combinations of variable pairs. I chose a pair of variables to form a new joined variable only if they exhibited some form of non-linear relation. Figure 4.9(a) shows an example of two variables joined together: MAXWIND and COAST_DIST. Non-linearity is observed in Figure 4.9(a) as the values of COAST_DIST increase. For low values of COAST_DIST, MAXWIND remains high for a number of residential properties up to a certain COAST_DIST value (at about 12.1) and after that, MAXWIND decreases.

Combining the values of these two variables occurred by multiplication in order to form a new variable named MAXWIND_COAST_DIST. Figure 4.9(b) depicts the distribution of values for MAXWIND_COAST_DIST. Created from selected variable combinations, together, original and 38 new variables, add up to 94. In Figure 4.10, the complete set of variables appear in the order in which CFS eliminates them (implemented using a greedy search backward elimination).

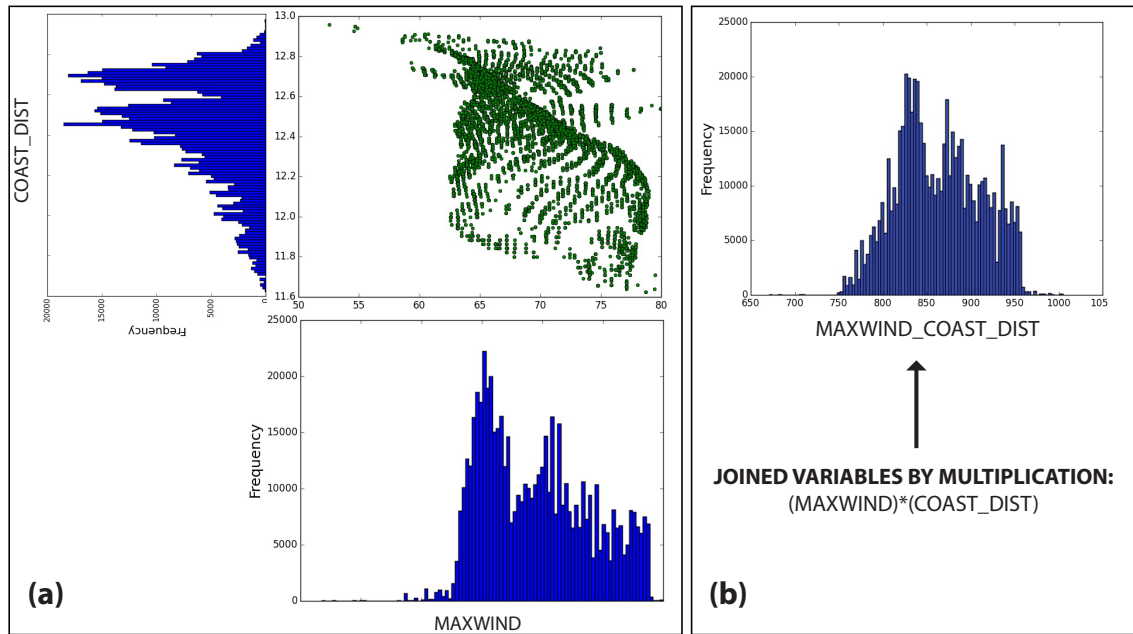


Figure 4.9 : The graph MAXWIND vs. COAST_DIST in (a) visually exposes non-linearity among the variables. Joined together by multiplication, the distribution of combined variable MAXWIND_COAST_DIST is shown in (b).

As the space of solutions is explored, the CFS algorithm uses the heuristic measure called “merit” to assess the value of subsets of variables using Equation 4.4. Greedy feature selection methods can either do: *forward selection* or *backward elimination*. Forward selection begins with an empty set of variables and incorporates variables

progressively based on the heuristic measure. Backward elimination begins with the full set of variables and progressively eliminates those to improve merit. Based on previous experiments, backward elimination produces better subsets of variables, reasoning that, as variables are dropped, it assesses the interactions in the context of the entire set [75]. In contrast, forward elimination assesses the relevance of variables in the context of the ones that have already applied.

Utilizing the set of 94 predictor variables found in the pre-processing stage, the CFS algorithm executed using backward elimination. Figure 4.10 shows a graph containing the iteration number, the variable eliminated in the iteration, and the “merit” measured after eliminating the indicated variable. As the algorithm proceeds with variable elimination, the heuristic reaches its peak at iteration 81. The set of variables listed from iteration 81 to 94 correspond to the sub-optimal solution found by the backward elimination algorithm. The CFS algorithm discovered 14 variables relevant to explaining the target variable.

I selected the 57 variables in iterations within one standard deviation from the maximum “merit” to accommodate for the possibility of non-linear interactions not captured by the CFS. This translates to retaining variables from iterations 38 to 94. Other selection techniques described in sections that follow are used to further reduce the number of predictor variables. Meanwhile, I divide this new set of predictors into two groups as explained in Section 3.2.2: static and dynamic variables. Table 4.9 and 4.10 show descriptive statistics for these variables respectively sorted in decreasing order of Pearson’s correlation absolute value with the target variable. Shown in decreasing order of correlation with the target variable from left-right and bottom-up, Figures 4.11 and 4.12 display a heatmap of correlations among the static and dynamic variables respectively. Shown in Figure 4.11, three clusters of variables

with positive correlation among them also correlate distinctively with roof damage: bottom-left cluster correlates positively, top-right cluster correlates negatively, and middle cluster correlates weakly.

Small measures of “merit”, as observed over all iterations in Figure 4.10, result from holding minimal mean correlation between target and predictor variables (see 3-dimensional plot of “merit” in Figure 4.7). This suggests that the target variable cannot be explained using linear interactions among predictor variables, and that more sophisticated techniques are needed to capture non-linearities. In the next section, I use machine learning algorithms to further select variables and ultimately create a model to predict wind-induced damage to residential structures.

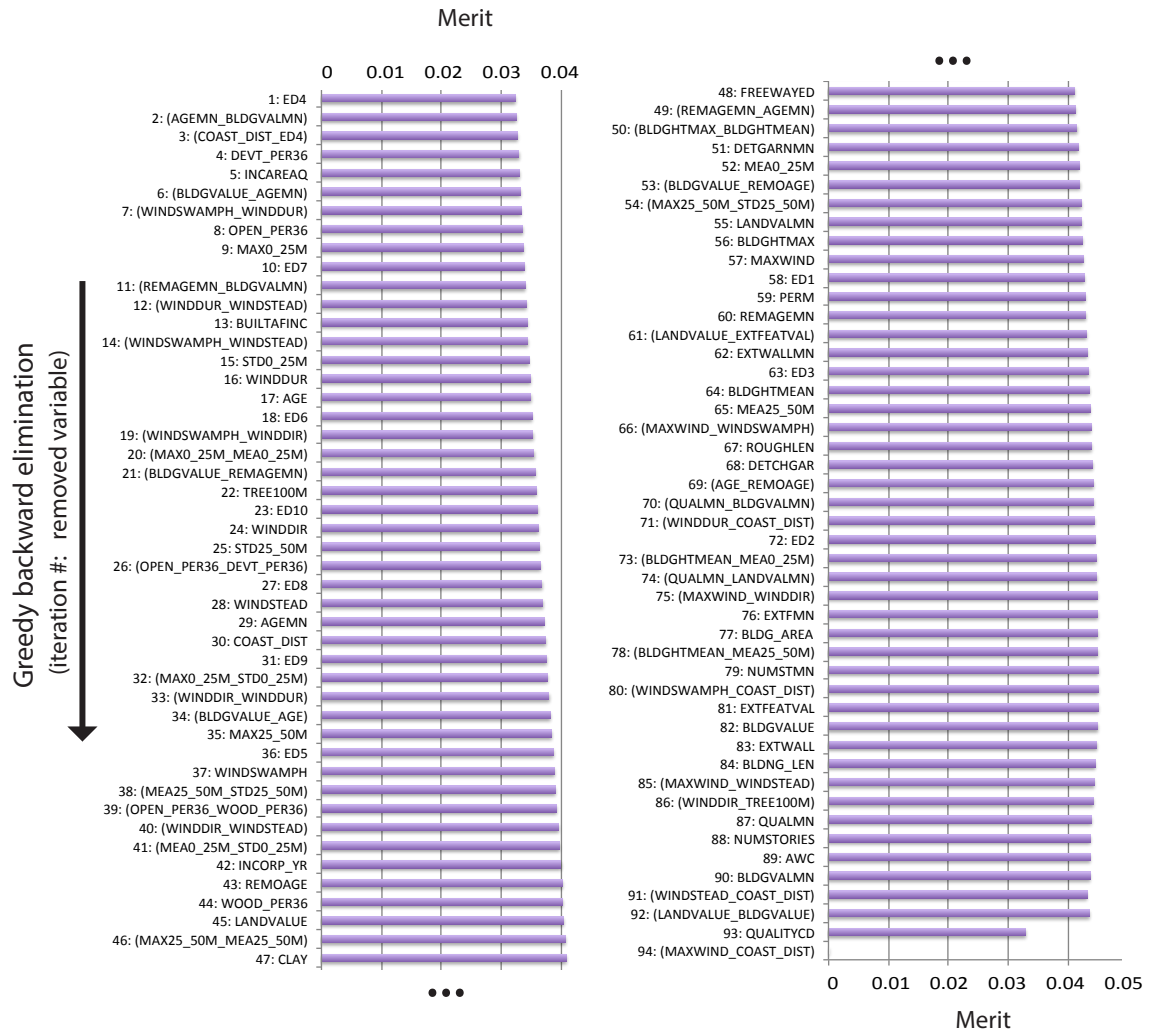


Figure 4.10 : Application of CFS with greedy backward elimination search on the expanded set of 94 predictor variables. The graph shows the iteration number, the variable eliminated at the given iteration, and the “merit” heuristic measure evaluating the worth of the remaining subset of variables.

Static Variables							
ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
1	BLDGVALMN	11.439	0.598	8.689	14.940	-0.2365	-0.2353
2	(QUALMN_LANDVALMN)	43.195	9.103	10.733	86.998	-0.2269	-0.2262
3	(QUALMN_BLDGVALMN)	48.221	9.319	12.376	81.754	-0.2266	-0.2257
4	(LANDVALUE_BLDGVALUE)	116.733	15.410	55.138	238.327	-0.2244	-0.2142
5	BLDGVALUE	11.379	0.705	8.518	15.619	-0.2195	-0.1992
6	QUALMN	4.192	0.634	1.000	6.000	-0.2107	-0.2099
7	QUALITYCD	4.192	0.740	1.000	6.000	-0.1888	-0.1801
8	LANDVALMN	10.248	0.925	6.805	15.729	-0.1761	-0.1757
9	LANDVALUE	10.236	0.935	4.625	15.739	-0.1744	-0.1744
10	NUMSTMN	1.323	0.349	1.000	4.000	-0.1712	-0.1706
11	(AGE_REMOAGE)	11.439	4.307	1.922	28.693	0.1631	0.1420
12	REMAGEMN	3.334	0.607	1.386	4.611	0.1607	0.1598
13	(REMAGEMN_AGEMN)	11.715	3.894	1.922	21.260	0.1581	0.1574
14	REMOAGE	3.271	0.711	1.386	5.357	0.1579	0.1367
15	BLDG_AREA	7.804	0.403	5.064	12.279	-0.1502	-0.1382
16	BLDNG_LEN	5.456	0.289	3.954	8.516	-0.1475	-0.1367
17	COAST_DIST	12.44	0.26	11.62	12.97	-0.1456	-0.1456
18	EXTWALLMN	0.750	0.367	0.000	1.000	-0.1339	-0.1338
19	BLDGHTMEAN	2.833	0.282	0.748	4.670	-0.1215	-0.1343
20	EXTWALL	0.750	0.433	0.000	1.000	-0.1181	-0.1133
21	NUMSTORIES	1.323	0.488	1.000	4.000	-0.1157	-0.1227
22	(BLDGHTMAX_BLDGHTMEAN)	9.771	2.052	1.997	25.886	-0.1146	-0.1260
23	CLAY	31.052	11.845	-0.100	49.800	0.1116	0.1118
24	DETGARNMN	0.248	0.287	0.000	1.000	-0.1063	-0.1063
25	(BLDGVALUE_REMOAGE)	36.956	7.160	12.174	63.955	0.1060	0.0890
26	(BLDGHTMEAN_MEA25_50M)	7.027	2.172	-3.684	21.122	-0.1035	-0.1046
27	(BLDGHTMEAN_MEA0_25M)	7.163	2.163	0.970	20.073	-0.1022	-0.1058
28	BLDGHTMAX	3.417	0.415	2.304	5.543	-0.0994	-0.1091
29	PERM	0.808	1.030	-0.100	7.000	-0.0874	-0.0874
30	(MEA0_25M_STD0_25M)	6.482	2.400	0.355	18.444	-0.0803	-0.0809
31	(MAX25_50M_MEA25_50M)	9.361	3.352	-0.531	35.014	-0.0774	-0.0744
32	MEA0_25M	2.503	0.595	0.512	4.711	-0.0769	-0.0761
33	AWC	0.162	0.014	-0.100	0.170	0.0757	0.0763
34	MEA25_50M	2.463	0.641	-1.250	5.898	-0.0740	-0.0703
35	DETCGAR	0.248	0.432	0.000	1.000	-0.0712	-0.0704
36	INCRP_YR	1044.328	923.052	0.000	1970.000	0.0702	0.0702
37	FREEWAYED	8.017	0.981	0.693	9.943	-0.0672	-0.0662
38	(MAX25_50M_STD25_50M)	9.284	2.974	0.409	32.341	-0.0659	-0.0635
39	ED3	5.568	0.846	0.693	8.271	-0.0656	-0.0654
40	EXTFMN	6.075	2.269	0.693	10.516	-0.0442	-0.0443
41	WOOD_PER36	0.709	0.036	0.693	1.069	-0.0375	-0.0372
42	(OPEN_PER36_WOOD_PER36)	0.516	0.046	0.480	0.822	-0.0372	-0.0373
43	ED1	2.935	1.296	0.693	6.930	-0.0170	-0.0113
44	(LANDVALUE_EXTFEATVAL)	29.580	35.022	3.206	162.888	-0.0166	-0.0172
45	ED2	1.465	1.277	0.693	7.692	-0.0113	-0.0157
46	ROUGHLEN	0.365	0.182	0.010	0.560	0.0093	0.0071
47	EXTFEATVAL	2.878	3.358	0.693	12.104	-0.0039	-0.0043

Table 4.9 : Descriptive statistics for static variables in decreasing order of absolute Pearson's correlation with roof damage.

Dynamic Variables							
ID	Continuous Variables	Mean	Standard Deviation	Min	Max	Pearson's Correlation with Damage	Spatial Autocorrelation with Damage (Moran's I)
1	(WINDDUR_COAST_DIST)	11.352	8.073	-0.654	27.720	0.1300	0.1300
2	(WINDSTEAD_COAST_DIST)	4.292	1.206	1.394	6.345	-0.0990	-0.0990
3	(WINDDIR_WINDSTEAD)	87.764	28.588	25.285	140.692	-0.0979	-0.0979
4	(MAXWIND_WINDSTEAD)	23.779	6.205	7.856	31.402	-0.0875	-0.0875
5	(MAXWIND_WINDDIR)	17361.401	1490.183	12153.513	20214.600	-0.0684	-0.0684
6	(MAXWIND_WINDSWAMPH)	5334.112	612.517	2624.800	6665.670	0.0631	0.0631
7	(WINDSWAMPH_COAST_DIST)	952.657	72.030	657.197	1077.936	0.0443	0.0443
8	(MAXWIND_COAST_DIST)	863.658	47.893	671.185	1004.523	-0.0305	-0.0306
9	MAXWIND	69.488	4.478	51.775	79.992	0.0207	0.0206
10	(WINDDIR_TREE100M)	19.200	66.787	0.000	264.227	0.0010	0.0012

Table 4.10 : Descriptive statistics for dynamic variables in decreasing order of absolute Pearson's correlation with roof damage.

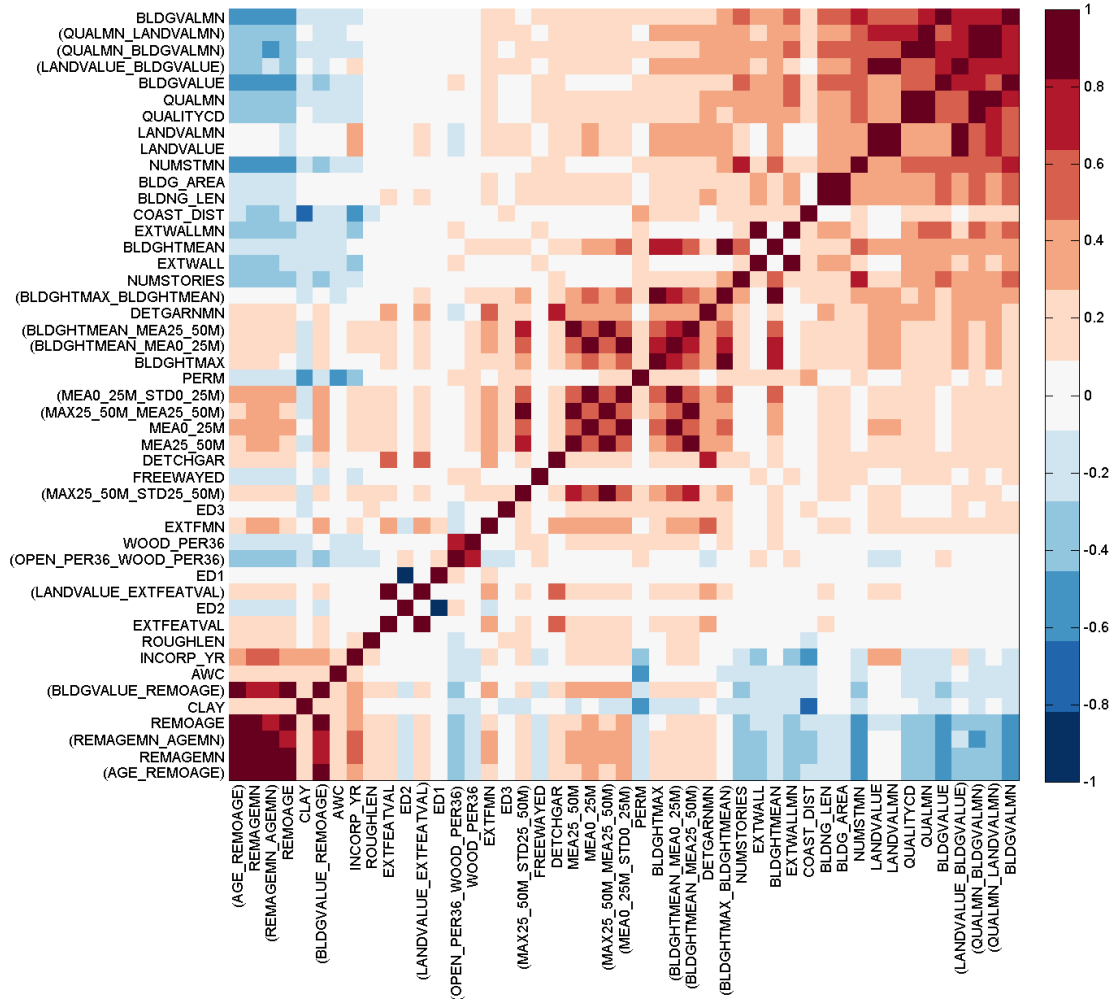


Figure 4.11 : Heatmap of Pearson's correlation among static variables. Variables are ordered from left-right and bottom-up in decreasing order of Pearson's correlation with roof damage. Three clusters of variables with positive correlation among them, as seen in this heatmap, correlate distinctively with roof damage: bottom-left cluster correlates positively, top-right cluster correlates negatively, and middle cluster correlates weakly.

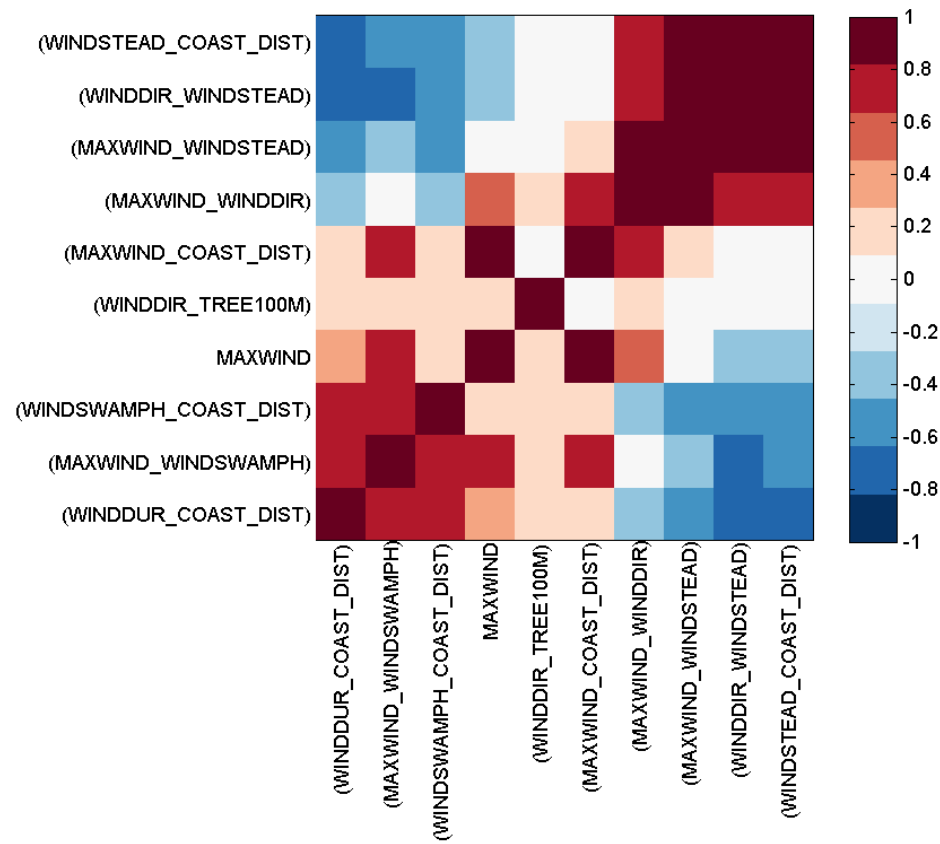


Figure 4.12 : Heatmap of Pearson's correlation among dynamic variables. Variables are ordered from left-right and bottom-up in decreasing order of Pearson's correlation with roof damage.

4.3.2 Selection of dynamic variables through LASSO analysis

Ten relevant dynamic variables, listed in Table 4.10, emerge after pre-processing and selection via CFS. Constructing logistic regression models at the terminal leaves of the hybrid model requires identification of the most important among these wind related variables.

The Least Absolute Shrinkage and Selection Operator (LASSO), a well-established technique for feature subset selection in linear models, solves the L1-penalized regression problem for finding the set of linear coefficients that minimize the sum of squares subject to the regularization parameter λ [77]. For the case of logistic regression, LASSO minimizes the deviance between the observed values and the expected values subject to the same regularization parameter. The parameter $\lambda \geq 0$ regularizes the amount of shrinkage applied to the coefficient estimates. As the λ parameter increases, it causes some coefficients to decrease and others to be set at exactly zero. At the same time, the impact of predictor variables with coefficients set to zero reflects in the increase of deviance. Therefore, as the regularization parameter increases, LASSO tries to retain only the features which best explain the target variable.

Using the Matlab implementation of LASSO, I executed a 5-fold cross validated analysis to measure the impact of the regularization parameter λ . Virtual samples are added to the training sets of each cross-validated fold based on Algorithm 3.1 in Section 3.2.2. Figure 4.13 shows the deviance as the λ parameter increases. The deviance mostly remains constant for values $\lambda < 10^{-2}$. For values $\lambda \geq 10^{-2}$, the impact of important coefficients set to zero becomes more evident by observing the larger increments in deviance. Figure 4.14 shows a trace plot of the wind variable coefficients fit by LASSO. On observation, λ has a rapid shrinking effect up to about $\lambda = 10^{-3}$. When $\lambda > 10^{-3}$, most variable coefficients reduce to zero. For example, at

$\lambda = 10^{-1.56}$ only the coefficients for MAXWIND, (WINDSWAMPH.COAST_DIST), and (MAXWIND_WINDSWAMPH) remain non-zero. Furthermore, a zoomed in plot in Figure 4.14 shows (MAXWIND_WINDSWAMPH) as the only variable maintaining a non-zero coefficient at $\lambda = 10^{-1.5592}$. Table 4.11 shows dynamic variables in the order of importance relative to the sequence in which variable coefficients were set to zero as λ increases. The two most important variables, (MAXWIND_WINDSWAMPH) and MAXWIND, are later used to fit logistic regressions at the random forest leaves in the hybrid model.

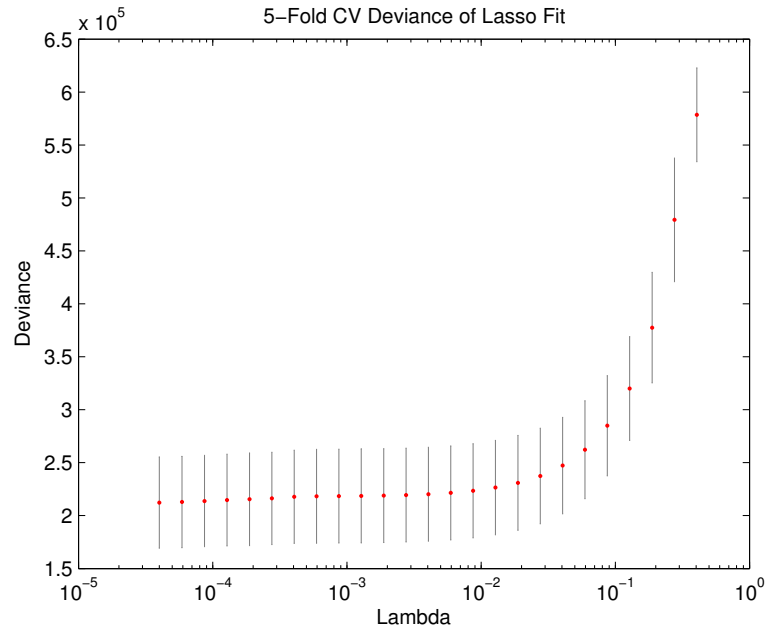


Figure 4.13 : Effect of regularization parameter λ on LASSO Logistic Regression based on dynamic variables.

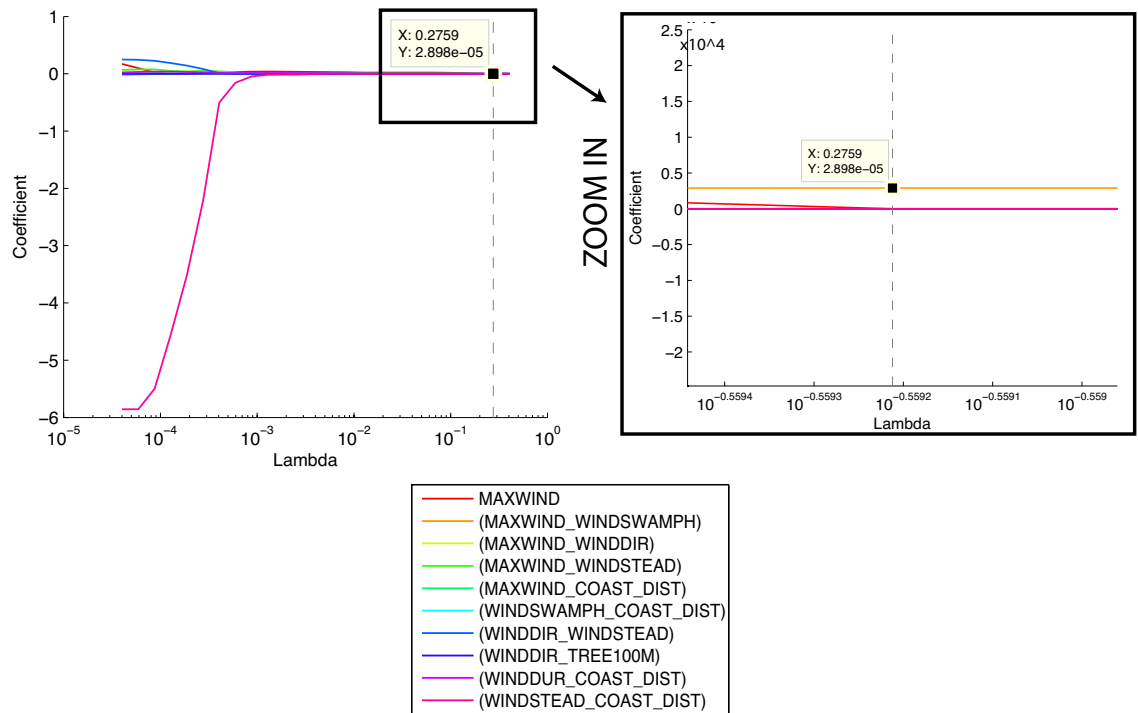


Figure 4.14 : Trace plot of wind variable coefficients fit by LASSO as a function of λ (5-fold cross validated).

Order of Importance	Variable Name
1	(MAXWIND_WINDSWAMPH)
2	MAXWIND
3	(WINDSWAMPH_COAST_DIST)
4	(WINDDUR_COAST_DIST)
5	(WINDSTEAD_COAST_DIST)
6	(WINDDIR_TREE100M)
7	(MAXWIND_WINDSTEAD)
8	(WINDDIR_WINDSTEAD)
9	(MAXWIND_COAST_DIST)
10	(MAXWIND_WINDDIR)

Table 4.11 : Order of importance among dynamic variables as result of LASSO analysis. The descending order corresponds to the sequence in which the variable coefficients were set to zero by LASSO as λ increases.

4.3.3 Selection of static variables through random forest variable importance

As a result of the correlation-based feature selection performed in Section 4.3.1, I considered the forty-seven static variables describing the individual buildings, construction codes, and surrounding terrain listed in Table 4.9. I used the importance measure embedded in the Random Forest algorithm to find the most important static variables [12] [71]. Variable importance measures in the random forest become stable as the number of trees in the ensemble increases. Random Forest supplies a measure of importance based on the amount of error increase in the tree by tree evaluation of the out-of-bag samples after randomly permuting its values [78]. The more the error increases, the more important the variable is considered.

I constructed a 5-fold cross validated random forest model with 50 trees based on the Matlab function `TreeBagger`. This model only utilizes the set of static variables as predictors. The `TreeBagger` function computes the out-of-bag variable importance score for each feature in each tree by permuting the out-of-bag values and then computing the error increase [79]. Figure 4.15 shows the 5-fold variable importance scores for the 47 static variables in decreasing order from left to right. The importance scores are stable across the 5-fold evaluations. The top three important variables include: distance to closest freeway, a combined terrain variable describing the percentage of open and wooded terrain in a radius of 362ft, and the distance to coast. These variables explain damage: residential structures closer to freeways and the Gulf Coast, and located mostly in open terrain, are more exposed to higher turbulence and wind gusts; and structures surrounded mostly by wooded terrain are more susceptible to damage caused by tree blow downs and wind-borne debris. In Section 4.4, I proceed to

select the optimal subset of these variables based on the areal prediction performance specific to the machine learning hybrid framework.

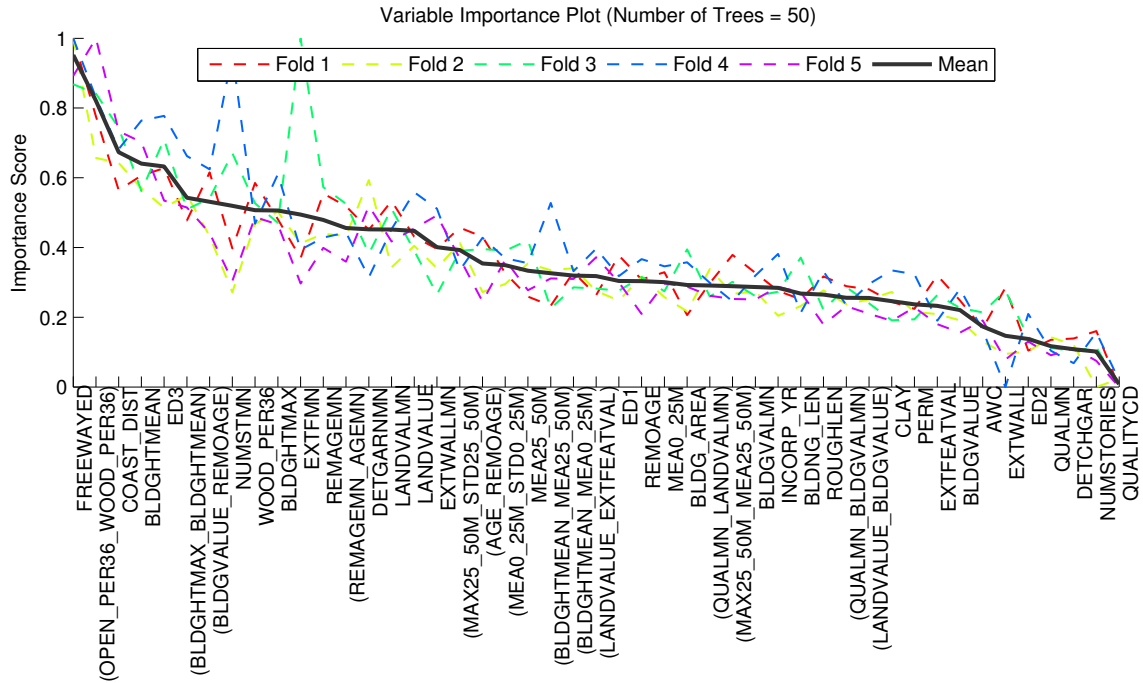


Figure 4.15 : Random forest variable importance based 47 static variables and an ensemble of 50 trees. Top three most important variables include: distance to freeway, a combined terrain variable describing the percentage of open and wooded terrain in a radius of 362 ft, and distance to coast.

4.4 Fitting hybrid model

By discovering an ordering of variable importance among both static and dynamic variables, I am able to perform a search for the subset of variables that maximizes the areal accuracy of the hybrid model. In this section, I present the need to prune the decision trees in the hybrid model in order to allow enough training samples at the terminal leaves to fit logistic regressions, I explore the optimal subset of static variables based on a single hybrid classification tree, and I finally construct the random forest hybrid model which performs significantly better than the HAZUS-MH wind damage model.

Choosing regularization parameter for minimum samples at tree terminal leaves

Growing the random forest without any pruning produces overfitted trees with a single sample at each terminal leaf, leaving an insufficient number of samples for logistic regression models to be fitted as required in the hybrid model. Therefore, I explore the optimal number of samples to allow at the terminal leaves.

Since random forests are based on a single classification trees (CART), I analyzed the response of a single tree to the regularization parameter *MinLeaf* based on the Matlab implementation of CART (fitctree). The *MinLeaf* parameter enables the user to select the minimum number of samples allowed at the terminal leaves of each tree in the ensemble. As the value for *MinLeaf* increases, the trees become smaller in size. I explored the effect *MinLeaf* on predicting damage to individual homes as quantified by the AUC metric. Figure 4.16 shows the AUC values of a 5-fold cross validated analysis for 100 CART with *MinLeaf* values ranging from 1 up to 10^5 constructed with the 47 selected static variables. Individual level metrics for each

CART are plotted based on the True Test set, fold Test set, and fold Train set. As *MinLeaf* increases, the True Test AUC reaches a maximum at $MinLeaf \approx 25$. At this point, evaluation on the fold Train set suggest overfitting as indicated in the figure. The overfitting effect becomes minimal up to $MinLeaf = 700$ where the True Test set AUC is 73%. Choosing $MinLeaf = 700$ enables the logistic regression models at the terminal leaves to be fitted with at most 9 wind variables ($\log_2(700) = 9.45$). A CART fitted with $MinLeaf = 700$ contains approximately 100 terminal leaves and a depth of approximately 28 levels as Figure 4.17 shows.

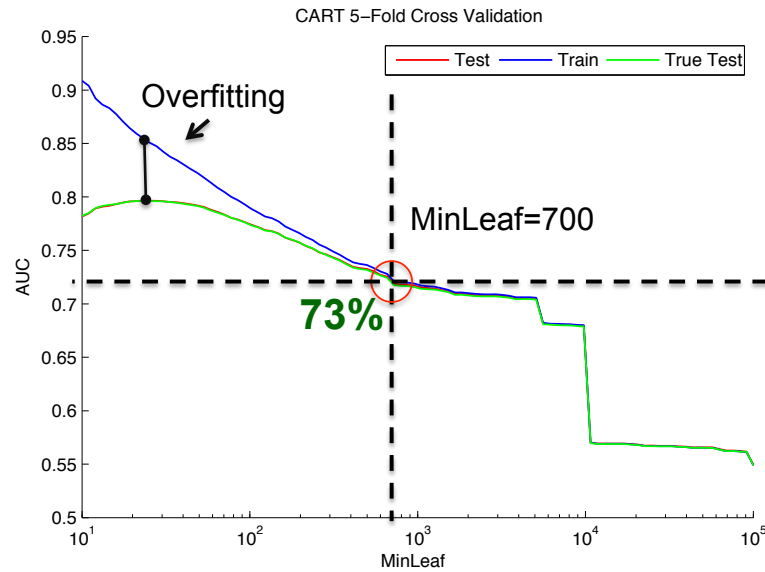


Figure 4.16 : Exploring the effect of regularization parameter *MinLeaf* for a single CART based on AUC metric.

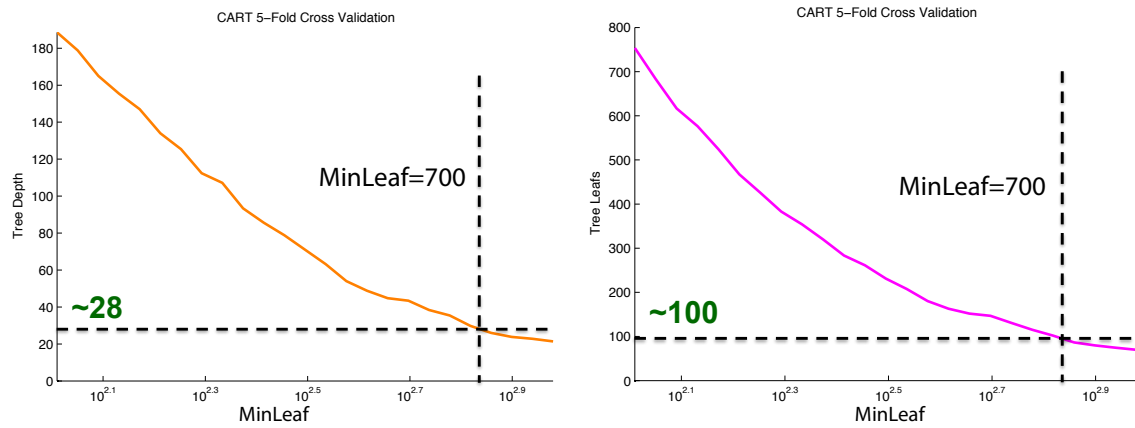


Figure 4.17 : Tree depth and number of terminal leaves of a single CART based $MinLeaf = 700$.

4.4.1 Variable subset selection based on single hybrid tree: LogTree

In order to select the best subset of static variables, I evaluate the performance of multiple subsets based on a single hybrid tree (named LogTree) and in terms of the accuracy of that tree at the one-kilometer square block level. Subsets with combinations of variables are created in a greedy fashion based on their importance shown in Figure 4.15. The first subset contains the most important variable, the second subset contains the top two most important variables, and so on until a subset containing the total amount of static variables is assembled. The LogTree model, constructed with these variable subsets, is composed of a single CART with $MinLeaf = 700$ and logistic regressions fitted at the leaves with the top important wind variable (MAXWIND_WINDSWAPMH). In order to train the hybrid model, I followed the training protocol introduced in Section 3.2.2 adding virtual samples based on Algorithm 3.1 for fitting the logistic regressions. Figure 4.18 shows the cross validated accuracy at the areal level of one-kilometer square blocks based on the LogTree hybrid model. The last significant increase in areal True Test set accuracy occurs with subset 15 where 62.19% one-kilometer square blocks are predicted correctly. For subsequent subset evaluations, the areal accuracy stabilizes at about 62.3%. The low areal accuracy is due to high variance in prediction errors made by a single tree. True Test set accuracy can be improved by building an ensemble of trees that make uncorrelated prediction errors with new data. In the next section I proceed to construct the hybrid LogRFT which is based on an ensemble of trees.

Subset 15 is composed of fifteen variables with the highest variable importance score in Figure 4.15. In decreasing order of importance, these variables include: distance to closest freeway, combined percentage of open and wooded terrain within 362 ft, mean height within building's footprint, closest distance to open terrain, com-

binned maximum height and mean height within building's footprint, combined building value and years since building remodeling, mean number of stories within 362ft, percentage of wooded terrain within 362 ft, maximum altitude within building's footprint, mean extra features value of homes within 362 ft, mean of years since building remodeling of homes within 362 ft, combined mean of years since building remodeling and mean building age of homes within 362 ft, mean detached garage presence in homes within 362 ft, and mean land value of homes within 362 ft. Eight of these variables describe surrounding terrain and seven describe structural characteristics of the residential buildings.

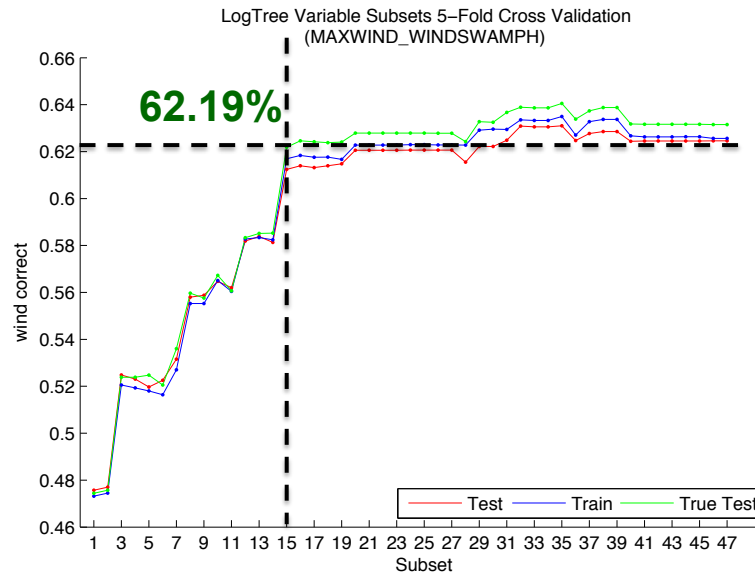


Figure 4.18 : Variable subset selection based on hybrid model with a single CART tree and logistic regressions at the leaves (LogTree) based on $MinLeaf = 700$ and a single dynamic variable: (MAXWIND_WINDSWAMPH).

4.4.2 Constructing LogRFT hybrid model

As a result of the analysis done with LogTree, I selected variables in subset 15 to fit a LogRFT hybrid model. The random forest tree includes up to 50 individual trees based on $MinLeaf = 700$. After observing the variation of individual level accuracy with the number of trees increase, I choose to use 20 trees since cross-validated True Test AUC stabilizes after 20 trees to about 76%. At 20 trees, the average total number of terminal leaves per cross-validated fold is 7,036 and the mean number of leaves per tree is 352.

Given these parameters and the two most important wind variables (MAXWIND _WINDSWAPMH) and MAXWIND, the LogRFT model is constructed using the training protocol described in Section 3.2.2. Figure 4.19 shows that as the number of trees in the ensemble increases, the one-kilometer square block accuracy varies up and down, finally stabilizing at about 75.2% after 15 trees. Compared to the performance of the HAZUS-MH physical model evaluated with the Ike simulated 3-second wind gusts introduced in Section 2.2.2, LogRFT performs 45.5% better at predicting expected wind damage risk at one-kilometer square blocks. The performance of the HAZUS-MH physical model drops significantly when 1-minute maximum sustained wind speeds (MAXWIND) is used, instead of the 3-second peak wind gusts. performance significantly drops. With respect to MAXWIND, the HAZUS-MH model accuracy is 3.2%, under-prediction is 89.6%, and over-prediction 7.2%. For ease of comparison, Table 4.12 shows the areal performance of these different models. The LogRFT model is compared against best performing HAZUS-MH model.

Figure 4.20 shows the spatial distribution of errors across the Harris County made by LogRFT (left) and HAZUS-MH (right) based on observed damage from Hurricane Ike. In contrast to the balanced errors in the HAZUS-MH model, errors in the

Wind damage predictive model	Accuracy	Under-prediction	Over-prediction
LogRFT			
- Trained and evaluated with MAXWIND and (MAXWIND_WINDSWAPMH).	75.2%	7.1%	17.7%
HAZUS-MH physical model evaluated with Hurricane Ike simulated 3-second wind gusts.	51.0%	25.6%	23.4%
HAZUS-MH physical model evaluated with 1-minute maximum sustained wind speeds (MAXWIND).	3.2%	89.6%	7.2%

Table 4.12 : One-kilometer square level prediction performance for the different wind damage models with respect to actual observed roof damage caused by Hurricane Ike over Harris County, Texas.

LogRFT model skew towards over-prediction (7.1% under and 17.7% over). Under-predicted block regions appear mainly clustered in three areas: first, in the north west between US-290 and SH-249; second, in the west of SH-6 between US-290 and I-10; and third, west of Beltway 8 in the area south of I-10 and east of SH-6. Over-predicted block areas mainly clustered in three areas: first, north of Beltway 8 from east to west beginning at US-90 and ending at US-290; second, between SH-6 and Beltway 8 south of US-290; and third, in the west of Harris County around I-10. Further analysis of these results and the LogRFT hybrid model itself is documented in the next chapter.

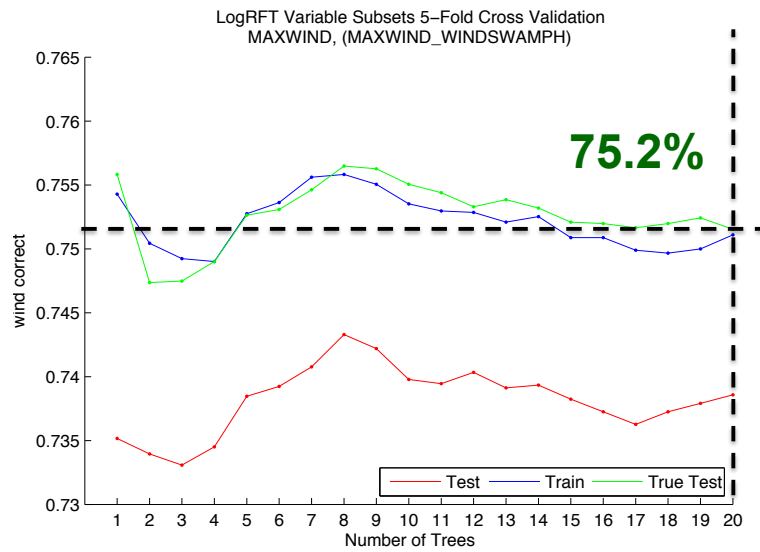


Figure 4.19 : Areal accuracy of LogRFT as the number of trees in the ensemble increases. The LogRFT is based on a random forest tree constructed with the top 15 most important static predictors and based on logistic regressions at the leaves fitted with the top most important dynamic wind variables.

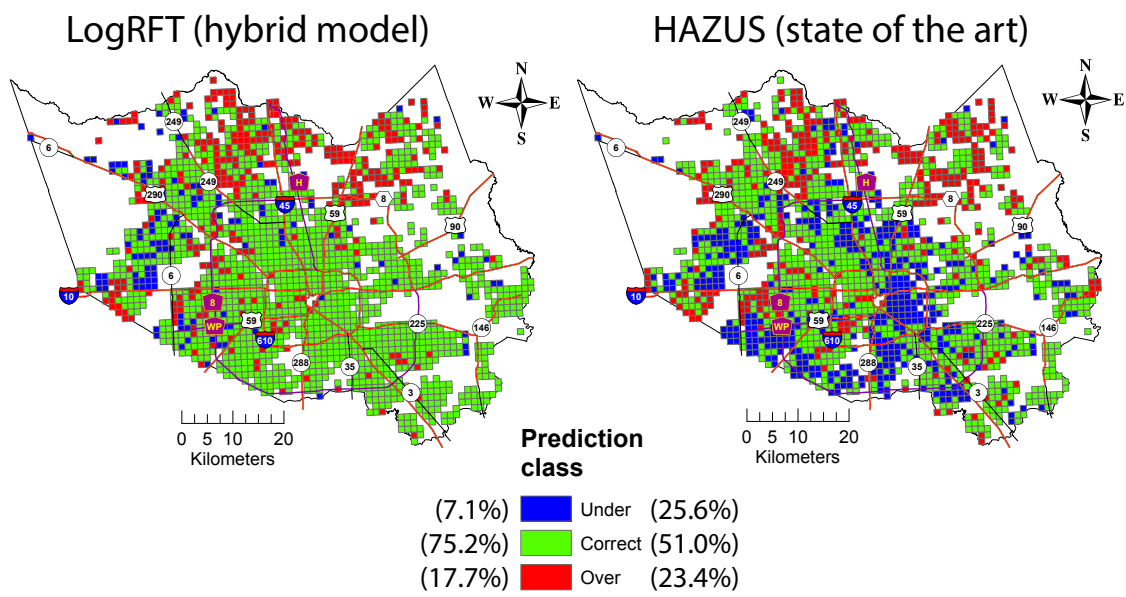


Figure 4.20 : Comparison of one-kilometer square areal accuracy between LogRFT and HAZUS-MH model based on expected roof damage from Hurricane Ike. LogRFT performs 45.5% better than HAZUS-MH. The performance of HAZUS-MH is based on the Hurricane Ike simulated 3-second wind gusts.

Chapter 5

Analysis of results

By constructing a LogRFT hybrid model that exhibits significant improvement over the HAZUS-MH model, I have demonstrated it is possible to construct probabilistic wind damage risk models from data. Nonetheless, the LogRFT model makes incorrect predictions for 24.8% of the one-kilometer square blocks. Analysis of the spatial distribution of errors over the Harris County suggests that under- and over-prediction errors are clustered in certain geographical areas. Analyzing the logistic regression models fitted by the hybrid framework helps us in understanding their source. In this chapter, I examine the logistic regression models fitted at the random forest leaves, I explore the geographical location of homes used in logistic models, and I describe the difference between groups of similar logistic models characterized by the static variables. Furthermore, in a subsequent section, I evaluate the generalization of the LogRFT hybrid model to unseen hurricane category wind fields and report the results.

5.1 Analysis of logistic regressions in the LogRFT hybrid model

In order to understand the wind-damage prediction estimates produced by the LogRFT model, I examine the logistic regressions fitted at each leaf in the random forest. These logistic functions describe the rate at which the probability of wind damage increases/decreases for a group of similar residential structures as wind vari-

ables change. Functions exhibiting sharp increments in damage probability within a small wind speed range reflect model overfitting and need further analysis. This behavior triggers under- and over-prediction errors depending on where the sharp increase in damage probability happens. For example, logistic models with sharp probability increments at low wind speeds, predict high damage probabilities for low speeds. This behavior consequently results in over-prediction errors. Furthermore, functions with sharp damage probability increments at high wind speeds produce under-prediction errors. Analyzing the different logistic functions fitted by the hybrid model helps diagnose prediction errors and plan future techniques to improve model performance.

5.1.1 Clustering and description of logistic regression coefficients

To explore the logistic regression models fitted in the LogRFT methodology, I select the LogRFT model from cross-validation fold #1 (from now on referred to as LogRFT#1). Each leaf in LogRFT#1 is associated with a logistic regression model described by one constant and two coefficients. The first coefficient corresponds to MAXWIND and the second coefficient corresponds to (MAXWIND_WINDSWAPMH). Using the constant and the two coefficients from each logistic model, I applied the K-means algorithm (Matlab function kmeans) to cluster the logistic models into similar groups. Before clustering, the variables in the logistic models were min-max normalized to make all equally important. Executing the K-means algorithm multiple times with 10 replicates finds logistic regressions mainly clustered into three distinctive groups. Figure 5.1 shows a 3D-plot of the 7,056 logistic models in LogRFT#1 and the partitioning among the 3 clusters found by K-means. Observe that 20.8% of the logistic models belong to clusters 1, 74.2% belong to cluster 2,

and 5.0% belong to cluster 3. Table 5.1 shows descriptive statistics of the logistic regression coefficients for each cluster. The majority of logistic functions (cluster 2) concentrate in the intercept interval $[-17.47, -3.261]$, MAXWIND coefficient interval $[-0.212, 0.229]$, and (MAXWIND_WINDSWAPMH) coefficient interval $[-4E-05, 0.005]$.

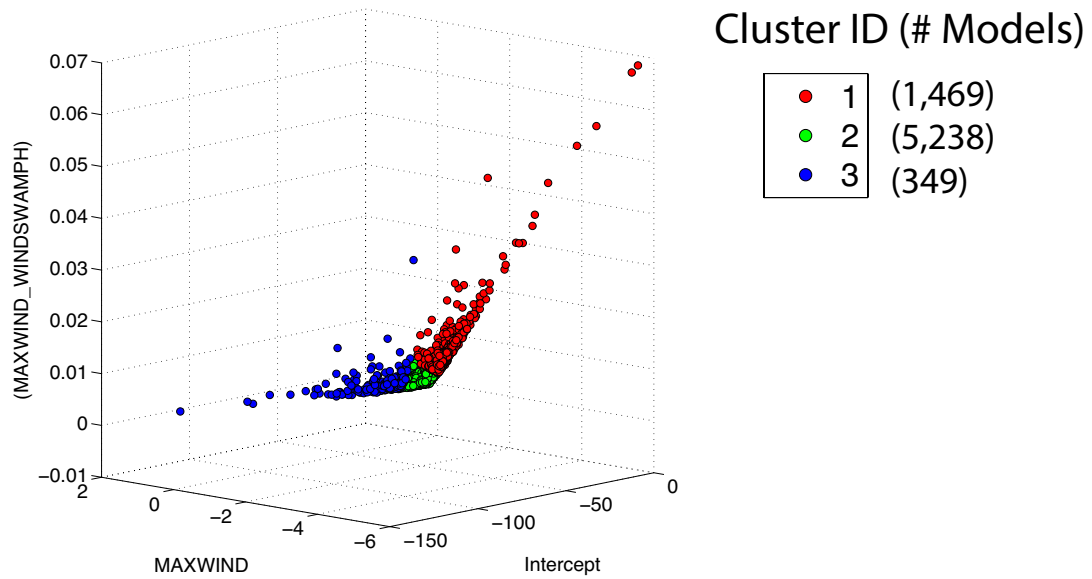


Figure 5.1 : 3-D plot of the 7,056 logistic regression models fitted in LogRFT#1 partitioned into 3 clusters by K-means algorithm.

Cluster ID	Number of Models	Intercept				MAXWIND				(MAXWIND_WINDSWAPMH)			
		MEAN	STD	MIN	MAX	MEAN	STD	MIN	MAX	MEAN	STD	MIN	MAX
1	1469	-5.572	2.596	-37.5	-1.187	-0.393	0.37	-5.695	-0.168	0.006	0.005	0.003	0.068
2	5238	-8.88	2.668	-17.47	-3.261	0.014	0.083	-0.212	0.229	0.001	8E-04	-4E-05	0.005
3	349	-25.36	11.26	-115.1	-16.89	0.228	0.207	-1.733	1.514	0.001	0.002	-2E-04	0.029

Table 5.1 : Descriptive statistics for the three clusters of logistic regression functions fitted in the LogRFT#1 model.

To interpret the logistic regressions describing each cluster, I graph the logistic functions and observe their behavior as wind variables change. To simplify the graphs, I plotted the probability of wind damage as a function of MAXWIND with values ranging from 0mph-180mph. The values used for variable (MAXWIND _WINDSWAPMH) are set to MAXWIND². Figure 5.2 shows the logistic regression functions for clusters 1 through 3. By plotting all logistic functions within each cluster on the same graph, it is possible to observe the overall behavior describing the cluster. Logistic functions from clusters 1 and 3 exhibit a sharp increase mostly within the interval of 60mph-90mph. This behavior reflects model overfitting to wind speeds due to the lack of variation in the observed wind estimates. Figure 4.10 shows that MAXWIND range from 52mph-80mph for Hurricane Ike. Some functions in cluster 2 exhibit transitioning of probabilities beginning from 0mph and reaching 100% at about 80mph. This type of logistic model describes groups of residential structures more vulnerable to wind speed loads. Within the same cluster, other logistic models describe groups of residential structures capable of withstanding stronger wind loads. For example, the probability of damage of the right most logistic model in cluster 2, begins to rise from 0% at about 70mph and reaches 100% damage probability at 160mph.

5.1.2 Geographical distribution of logistic models

Every logistic regression model describes the probability of damage for a group of similar residential structures. By plotting the location of such structures, I analyze the geographical impact of the logistic models in each cluster. To do so, I extract the training samples used to construct each logistic function in the LogRFTR#1 model and group them according to their cluster membership. Since the classification trees in the random forest are constructed from random samples with replacement, an

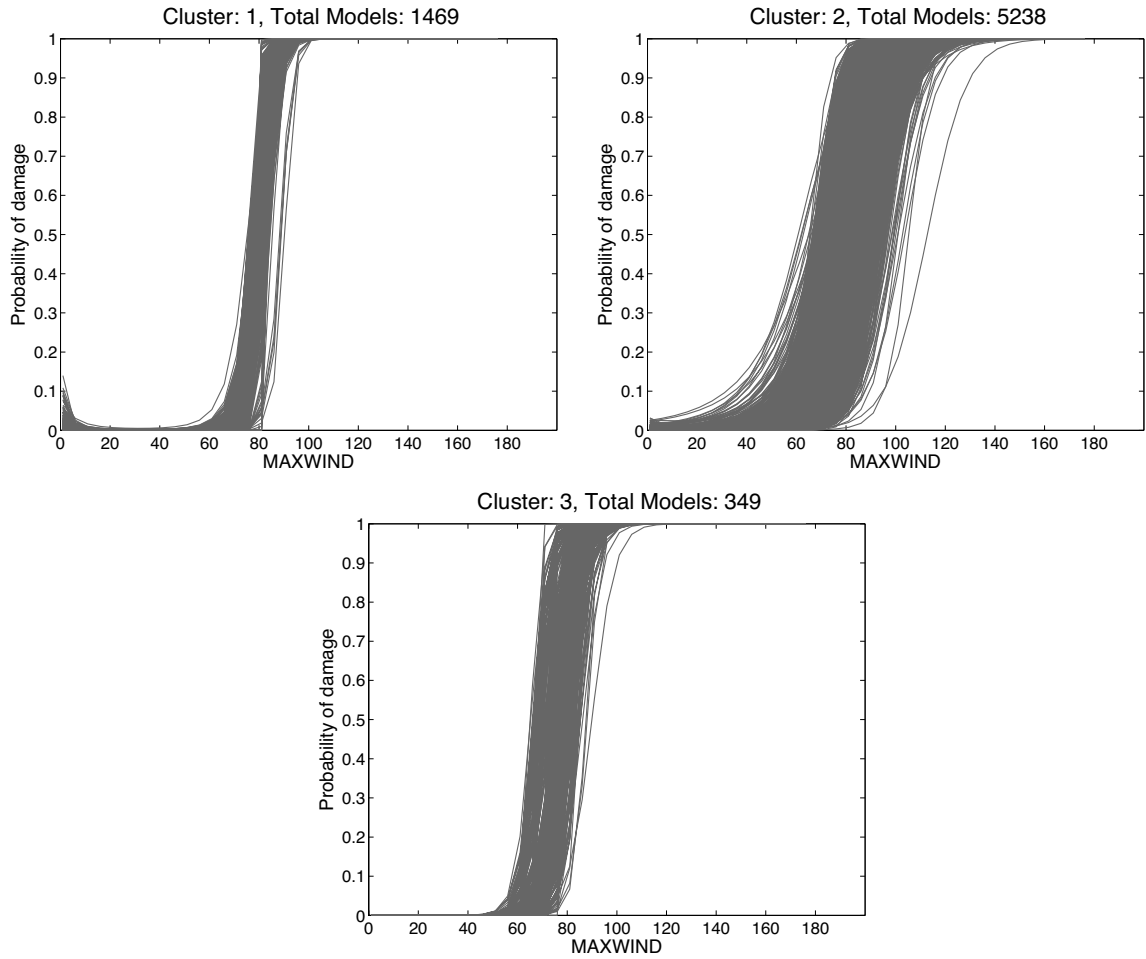


Figure 5.2 : Graphical representation of the three clusters of logistic models from the LogRFT#1. These models are plotted as a function of MAXWIND. The (MAXWIND_WINDSWAPMH) variable is set to MAXWIND^2 .

individual sample can appear multiple times in the training set of a given logistic model. Therefore, I only keep the set of unique samples that describe each cluster of logistic functions. The unique samples from each cluster are gathered and aggregated to the one-kilometer square blocks by computing the count of samples within each

block. Figure 5.3 depicts the geographical distribution of training samples in clusters 1 to 3.

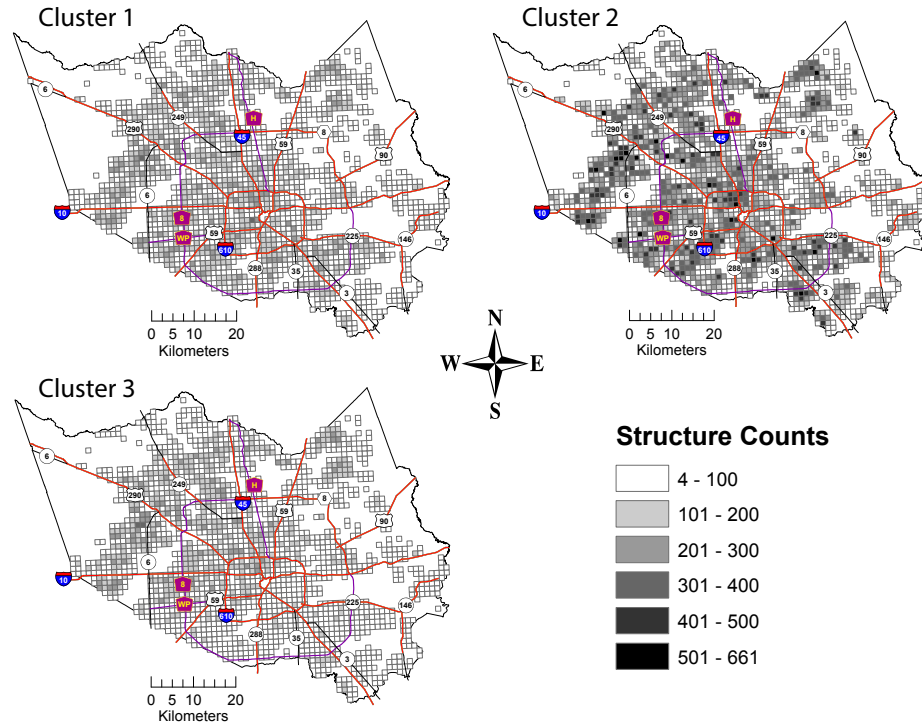


Figure 5.3 : Geographical distributions of unique training samples used to fit logistic regression models in cluster 1 to 3. Samples from each cluster are aggregated to the one-kilometer square blocks by computing the count of residential properties within each block.

The one-kilometer square blocks in the geographical distribution map into six categories depending on the number of structures present within each block. Most of the structures in cluster 3 are located toward the northwest of Harris County. According to Section 5.1.1, the logistic regression models in this cluster exhibit sharp increments in damage probability within a small wind speed range. The geographic location of this cluster coincides with most under- and over-prediction errors made

by the LogRFT model north and west of the Harris County in Figure 4.20 (left). Since this cluster contains blocks holding residential property counts mostly in the 1 to 200 range, overfitted logistic models explain only part of the errors made by LogRFT. Given the large density of residential properties explained by cluster 2, logistic functions within this cluster have a higher potential to affect the areal level prediction. Therefore, a second source of over-prediction errors is introduced by slightly shifted functions to lower wind speeds such as those present in cluster 2 Figure 5.2. Along the same lines, logistic functions in this cluster exhibiting low probabilities of damage up to 90 mph, possess a high potential to be the source of under-prediction errors made by the LogRFT model.

5.1.3 Characterization of logistic models with respect to static variables

After analyzing the spatial distribution of training samples, I proceed to characterize the differences among them with respect to the top fifteen predictor static variables selected in Section 4.4.1. Knowledge of distinctive building types across all clusters is obtained by contrasting the values of these variables. In Figure 5.4, I create a graph for each static variable containing a box plot based on the unique training samples at each cluster. Note that Section 4.3.1 performed log-transformation of all static variables, with the exception of NUMSTMN and DETGARNMN. In order to facilitate the interpretation of values, only single variables are transformed to their original scale for plotting in this figure. Joint variables such as (OPEN_PER36_WOOD_PER36) are not transformed back to original values.

As explained in Subsections 5.1.1 and 5.1.2, under- and over-prediction errors appear more prominent in locations where cluster 1 and 3 samples are geographically located. According to Figure 5.4, residential properties in cluster 1 are constructed

on less expensive land, are mostly older/unremodeled one-story homes surrounded by the smallest percentage of wooded terrain, are closest to the freeway system and to the Gulf coast. Under-estimation errors in cluster 1 can be potentially explained by fact that the residences in that cluster are more vulnerable to wind damage. Properties in cluster 3 are farthest from the Gulf coast (also observed in Figure 5.3), have the highest mean/max building heights, are the newest/remodeled buildings, have the highest number of stories (a mixture of 1, 2, and 3 story building), have the highest cost of exterior features, the highest percentage of detached garages in neighborhood, and are built on more expensive land. Residences in cluster 3 are thus more resistant toward damage, therefore, over-estimation errors could potentially occur.

In summary, further investigation of under- and over-prediction errors needs to be performed. In particular, understanding the effects of overfitted functions rising within a small range of wind speeds help resolve errors made by clusters 1 and 3. Additionally, investigating the details of how logistic functions in the extremes of cluster 2 are fitted could help devise strategies to enable better generalization.

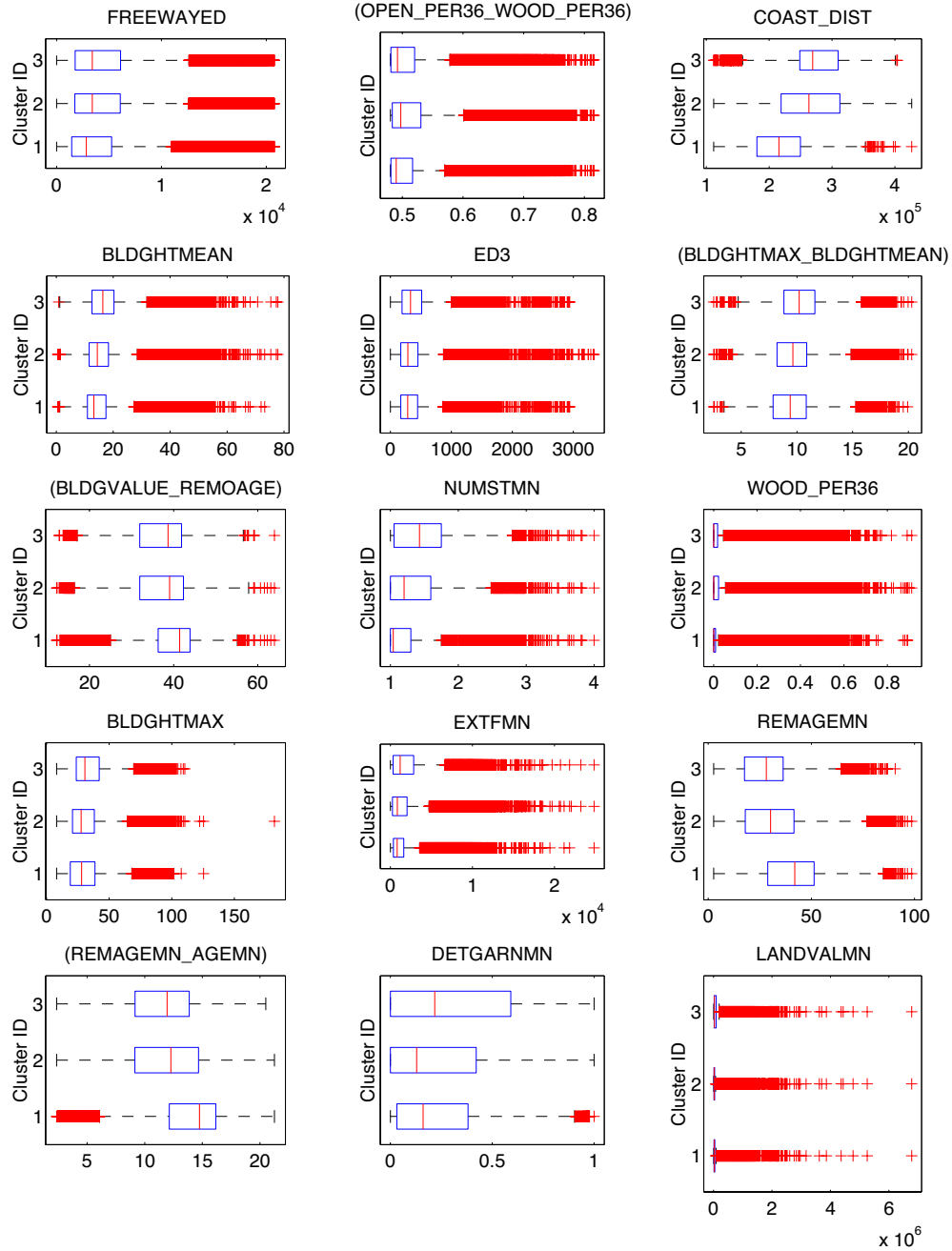


Figure 5.4 : Graphs describing the distribution of values of each static variable in LogRFT#1 across the three logistic regression clusters. The distribution of values in each cluster is represented using box plots. Single variables log-transformed in Section 4.3.1 are transformed back to their original units of measure.

5.2 Assessing LogRFT generalization to unseen wind speeds

Since the purpose of the LogRFT model is to predict wind-induced damage to residential structures for future hurricane events, generalization is judged based on wind fields outside Hurricane Ike’s wind speed range. A deterministic wind field for each of the five Saffir-Simpson hurricane categories is used to simulate wind speeds over the Harris County. For reference, Table 5.2 shows the range of 1-minute speeds defining the Saffir-Simpson categories. The simulated wind fields, computed by the wind hazard model integrated in the HAZUS-MH software [40], are used to evaluate the LogRFT and HAZUS-MH wind damage models. Figures 5.5 and 5.6 show the spatial distribution of predicted expected damage for hurricane categories one through five respectively for both models at the one-kilometer square block level.

Category	Wind speed range (mph)
1	74-95
2	96-110
3	111-130
4	131-155
5	>155

Table 5.2 : Saffir-Simpson hurricane hurricane categories.

5.2.1 Comparing generalization among the LogRFT and HAZUS-MH models

By observing the response of both the LogRFT and HAZUS-MH models to lower and higher wind speeds, we can evaluate the generalization differences between them. Figure 5.5 shows that the LogRFT model predicts a sharp increase in damage proba-

bilities from hurricane category 1 to category 2. This this is due to the use of logistic functions in which the transition from low to high damage probabilities occurs within the wind speed interval of 60mph-100mph (see Figure 5.2). For hurricane category wind fields 3 to 5, the LogRFT model predicts expected damage probabilities of 80% to 100% for all one-kilometer square blocks. The HAZUS-MH model also predicts a sharp increase in damage probabilities from hurricane categories 2 to 3 and from categories 3 to 4 as shown in Figure 5.6. This suggests that the model's fragility curves transition from low to high damage probabilities mostly in the wind speed range of about 96mph-130mph. In contrast, the logistic functions fitted by the LogRFT exhibit low to high probability transitions at a much lower wind speed range.

Both the logistic functions and fragility curves describe the cumulative probability of wind damage as a function of wind speed. The difference in their structure stems from the fact that they have been developed based on different wind speed ranges. While the LogRFT model is trained using 1-minute wind speeds (MAXWIND and MAXWIND_WINDSWAMPH), the HAZUS-MH model is developed based on 3-second peak wind gust speeds [18]. According to a FEMA mitigation assessment report, 3-second peak wind gusts are 1.3 times (or 30% higher than) the 1-minute maximum sustained wind speed [80].

To compare the performance of the LogRFT methodology based on 3-second peak wind gusts, I convert the 1-minute wind speed variables to 3-second peak wind gusts to construct a LogRFT model (from now on referred to as LogRFT_windgust_factored). This model exhibits a one-kilometer square level accuracy of 43.7%, an under-prediction of 43.9%, and an over-prediction of 12.4%. To facilitate comparing the performance of the different wind models discussed, Table 5.3 shows their one-kilometer square block accuracies and prediction errors. Contrasting the performance of mod-

els LogRFT_windgust_factored and LogRFT, areal accuracy decreased 41.9%, under-prediction increased 518.3%, and over-prediction decreased 29.9%. These numbers reflect a large percentage of one-kilometer square blocks previously classified as correctly predicted now classified as under-predicted. Furthermore, evaluating the HAZUS-MH model with variable MAXWIND converted to 3-second wind gusts yields to an areal accuracy of 41%, under-prediction rate of 43.1%, and over-prediction rate of 15.9%. This results based on the HAZUS-MH model are very similar to those obtained with the LogRFT_windgust_factored model.

Wind damage predictive model	Accuracy	Under-prediction	Over-prediction
LogRFT			
1) - Trained and evaluated with MAXWIND and (MAXWIND_WINDSWAPMH).	75.2%	7.1%	17.7%
LogRFT_windgust_factored			
2) - Trained and evaluated with MAXWIND and (MAXWIND_WINDSWAPMH) converted to 3-second wind gusts.	43.7%	43.9%	12.4%
HAZUS-MH physical model evaluated			
3) with Hurricane Ike simulated 3-second wind gusts.	51.0%	25.6%	23.4%
HAZUS-MH physical model evaluated			
4) with MAXWIND converted to 3-second wind gusts.	41.0%	43.1%	15.9%
HAZUS-MH physical model evaluated			
5) with 1-minute maximum sustained wind speeds (MAXWIND).	3.2%	89.6%	7.2%

Table 5.3 : One-kilometer square level prediction performance for the different wind damage models with respect to actual observed roof damage caused by Hurricane Ike over Harris County, Texas.

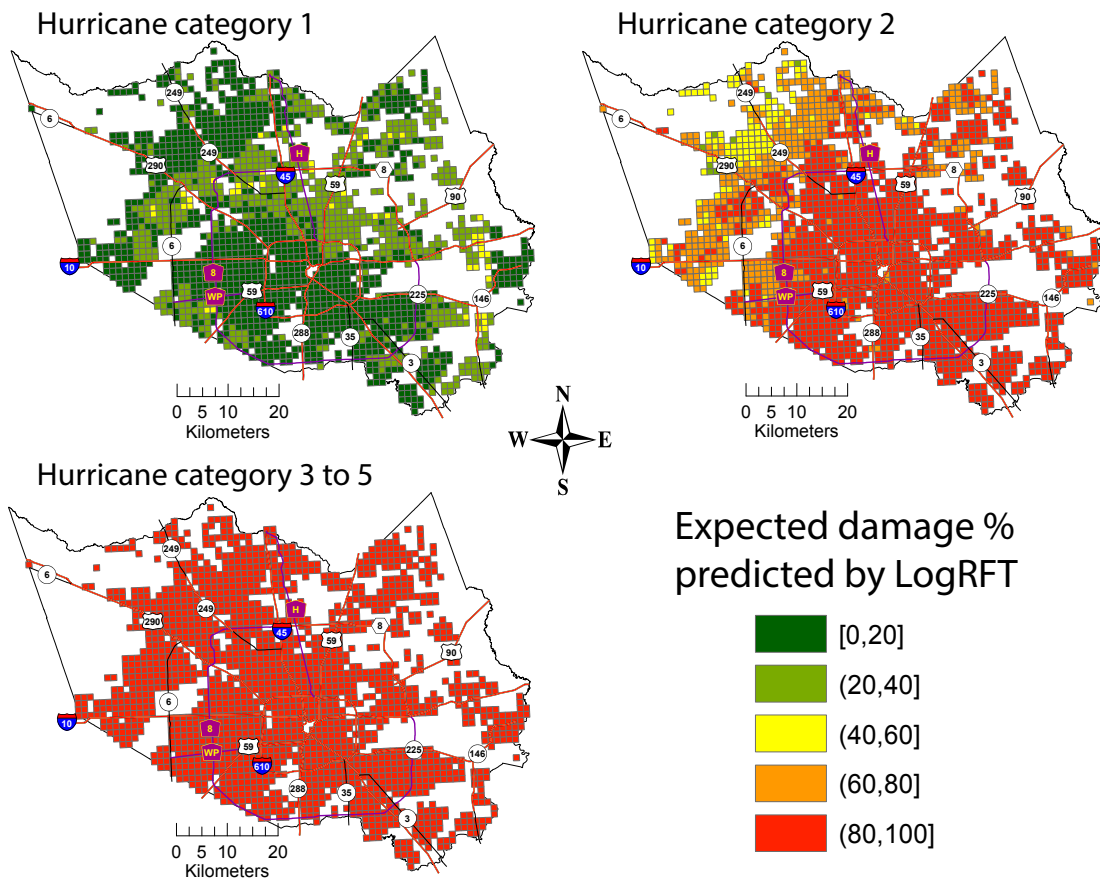


Figure 5.5 : Distribution of expected damage predicted by the LogRFT model at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.

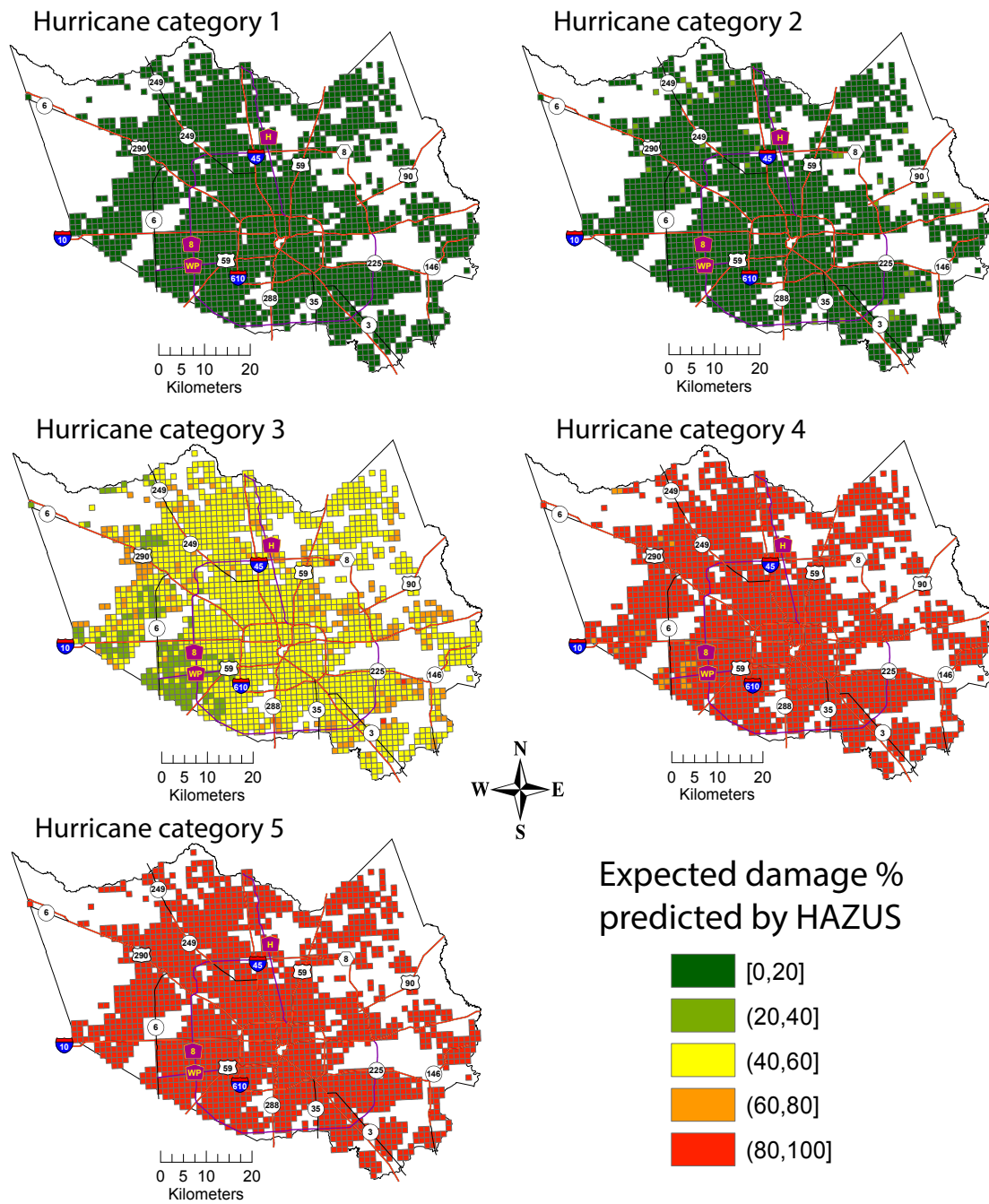


Figure 5.6 : Distribution of expected damage predicted by HAZUS-MH fragility curves at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.

5.2.2 Construction and analysis of LogRFT based on 3-second peak wind gusts

Logistic coefficients fitted with 3-second wind gust speeds are analyzed following the same strategy as in Section 5.1. Obtained from the LogRFT_windgust_factored#1 model, these coefficients cluster into three groups based on the K-means algorithm as depicted in Figure 5.7. Comparison to coefficients in the LogRFG#1 model indicates that the new intercepts and (MAXWIND_WINDSWAMPH) coefficients distribute over a smaller range, while MAXWIND coefficients appear in a similar range. Furthermore, the configuration of the new logistic models as a function of wind speed compared those fitted with 1-minute maximum sustained winds, exhibit a transition from low to high damage probabilities at the higher wind speed range of about 70 mph-130 mph according to Figure Figure 5.8.

In order to assess the generalization of the LogRFT_windgust_factored model to unseen hurricane category wind fields, I evaluate the model with the same simulated wind fields used to assess the generalization of models LogRFT and HAZUS-MH. Figure 5.9 shows the distribution of expected predicted damage at the one-kilometer square blocks for these wind fields. In contrast to the generalization observed in LogRFT, LogRFT_windgust_factored predicts less damage for category 1, 2, and 3. Prediction estimates for category 1 and 2 exhibit a smoother transition to higher damage probabilities. Additionally, comparing the model's generalization to that of the HAZUS-MH model, predictions are identical for both models with the exception of category 1 and 2. While HAZUS-MH uniformly predicts 0%-20% expected damage for category 2, LogRFT_windgust_factored predicts 20%-60% expected damage for one-kilometer square blocks that spatially correlate with roof damage in Figure 4.2. These

observations suggest that `LogRFT_windgust_factored` identifies regions of vulnerable residential structures across the Harris County.

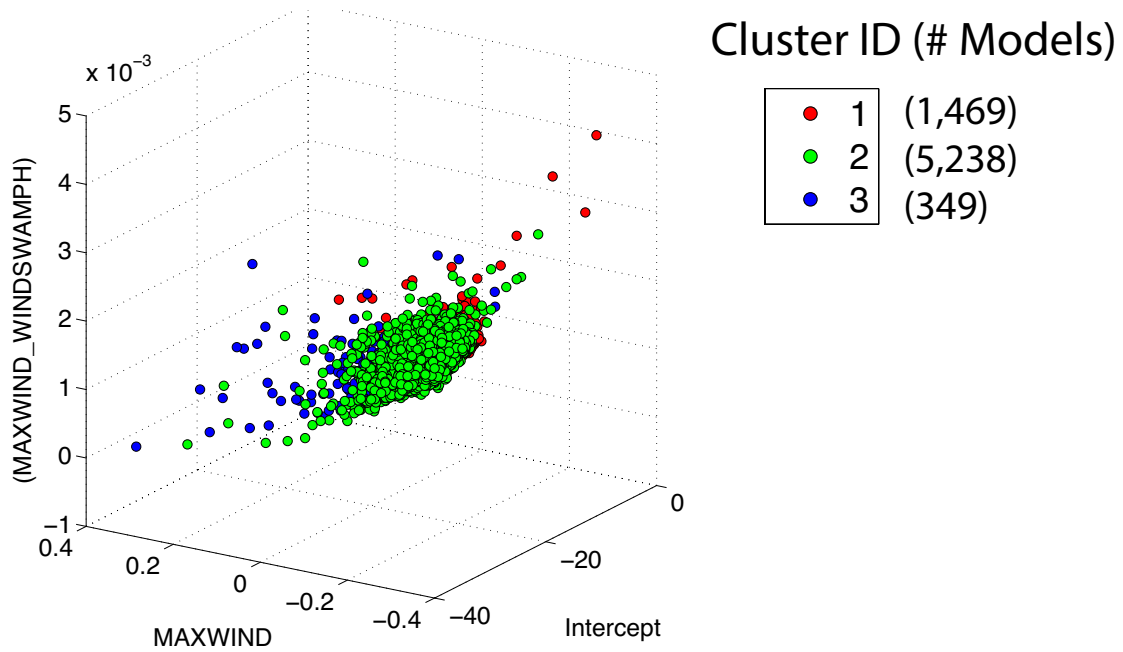


Figure 5.7 : 3-D plot of the 7,056 logistic regression models fitted in `LogRFT_windgust_factored#1` partitioned into 3 clusters using K-means algorithm.

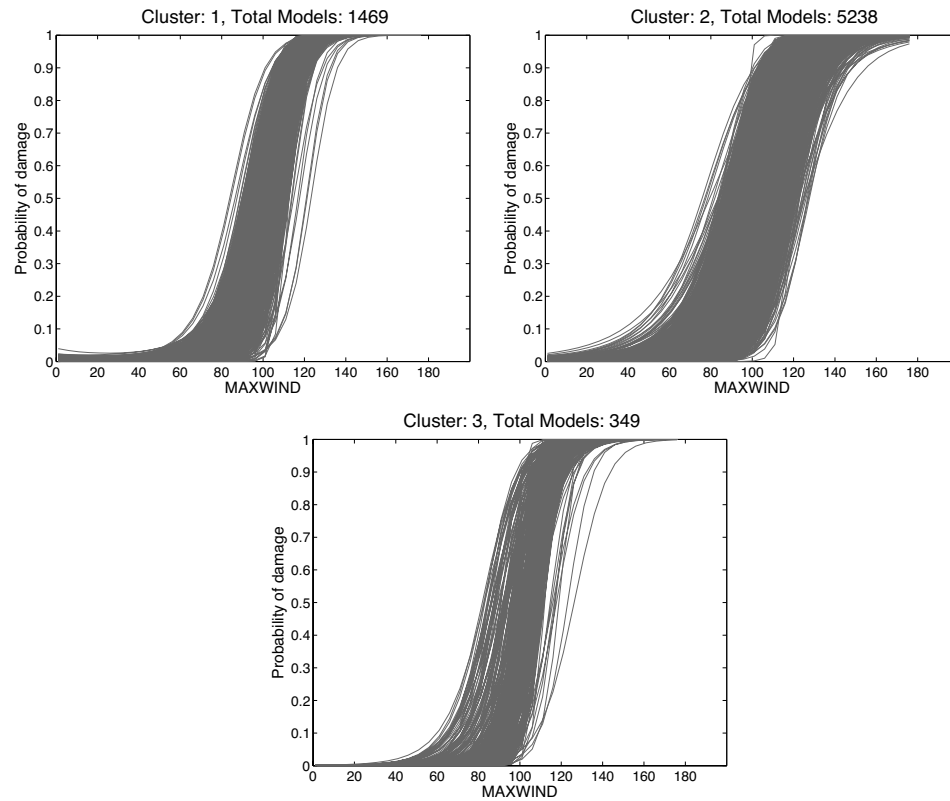


Figure 5.8 : Graphical representations of the three clusters of logistic models obtained from model LogRFT_windgust_factored#1. Models plotted as a function of MAXWIND. The (MAXWIND.WINDSWAPMH) variable was set to MAXWIND^2 .

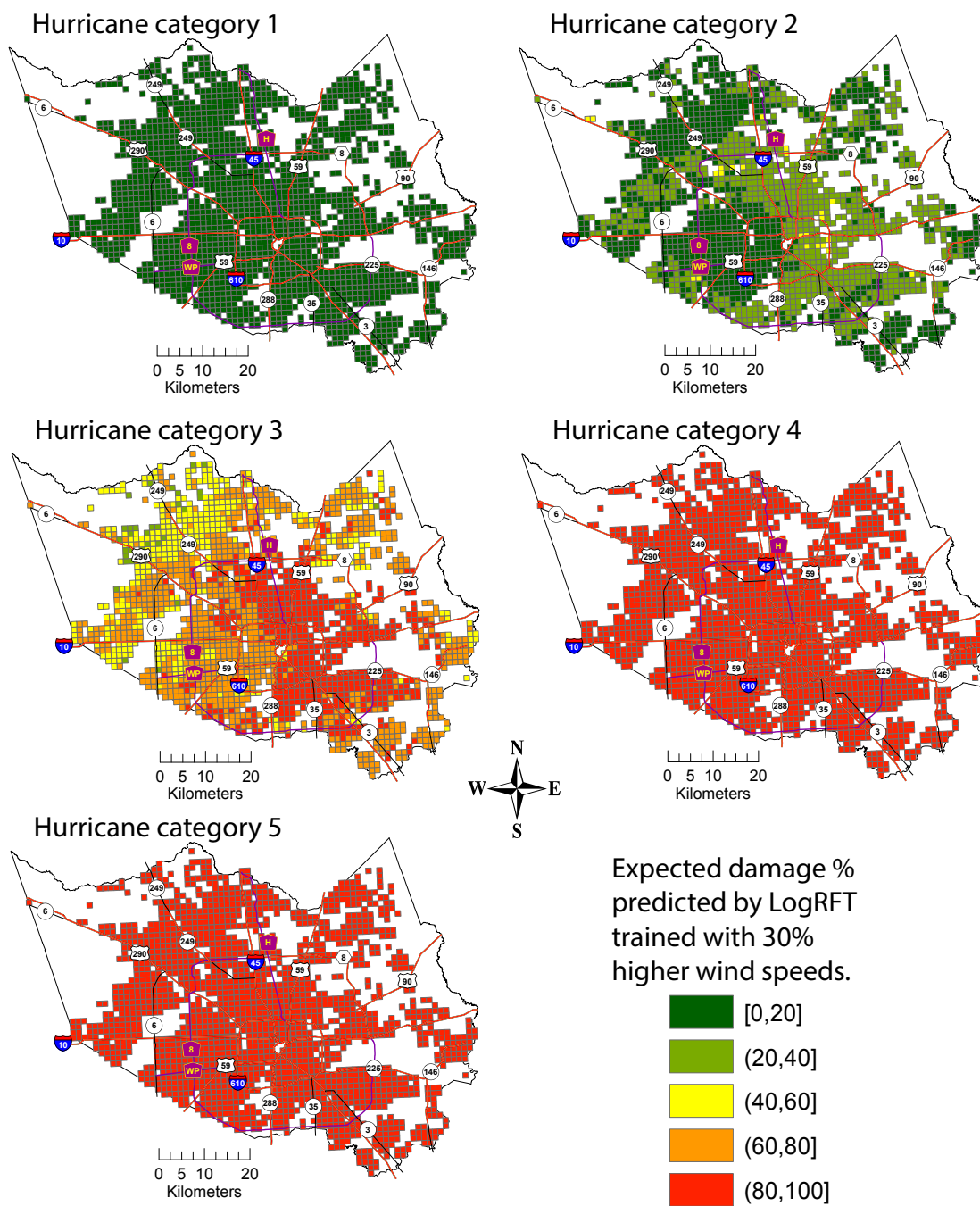


Figure 5.9 : Distribution of expected damage predicted by LogRFT_windgust_factored at the one-kilometer square block level for Saffir-Simpson hurricane categories 1 to 5.

5.3 Summary of results

The logistic regressions fitted by the LogRFT methodology for multiple building types confirm that these cumulative density functions relate the probability of roof damage to wind speeds. Because of the form of these functions, the constructed LogRFT model can predict damage probabilities from as a function of increasing wind speeds. This generalization ability is observed over LogRFT#1 predictions for hurricane categories 1-5, although they appear to over-estimate damage relative to the HAZUS-MH model predictions. Nonetheless, by converting the 1-minute wind speeds to 3-second wind gusts and using them to train the LogRFT_windgust_factored model, I demonstrate the capability of the LogRFT methodology to resemble the generalization observed in the HAZUS-MH model. In addition, the LogRFT methodology identifies regions vulnerable to wind damage. Constructed based on individual level data describing Harris County residences, LogRFT models surpass fragility curves in the HAZUS-MH model which were developed for a limited number of building types described by a few parameters as presented in Section 2.2.

Models developed from 3-second wind gusts predict observed damage with significantly lower areal accuracy compared to those developed from 1-minute sustained wind speeds. The one-kilometer square accuracies and errors from models developed and evaluated based on 3-second wind gusts are very similar, as observed in Table 5.3. Furthermore, logistic functions fitted with 3-second wind gusts shift to higher wind speeds in comparison to those fitted with 1-minute wind speeds. This change produces a drop in areal accuracy from 75.2% to 41.0%. Therefore, it appears that in order to obtain higher prediction accuracies, wind damage models based on cumulative density functions need augmentation with 1-minute wind speeds. Additional independent observed damage information caused by a different hurricane will allow

the validation of these models. Meanwhile, a more detailed study of prediction errors made by models developed with 3-second wind speeds is necessary.

Chapter 6

Conclusion and future work

In the work presented in this thesis I developed the first purely data-driven machine learning framework to predict the probability of wind damage risk to individual residences and to areal units, in particular, one-kilometer square blocks. While most models in civil engineering for damage prediction are physics based, the modeling framework presented in this thesis is based entirely on the largest damage data set collected to date. The model constructed by the machine learning framework (LogRFT) performs 47.5% better than the state of the art HAZUS-MH physical model at predicting wind damage from Hurricane Ike at the one-kilometer square level. To achieve this performance improvement, I made two technical contributions. First, since observed wind estimates recorded over the Harris County, Texas for Hurricane Ike were limited to the range from 50mph-80mph, I added virtual samples to the training set to improve generalizability over unseen wind speeds. The virtual samples are added based on the rationale that buildings damaged at a given wind speed remain damaged at higher wind speeds and that buildings undamaged at a given wind speed, remain undamaged at lower wind speeds. The addition of virtual samples enables models to predict damage smoothly as wind speed increases. In particular, logistic regression models provide the best smooth response for increasing wind speeds. This finding led me to my second technical contribution: to divide the predictor variables into two groups; and to adapt a two-level hybrid model (called LogRFT) using a random forest tree trained on static variables and logistic regression models

at the leaves trained with dynamic variables. By implementing the LogRFT methodology, I demonstrate the possibility of building good models of wind damage risk at the kilometer square level in an entirely data-driven fashion.

I analyze logistic regression models fitted at the leaves of the LogRFT model by: exploring the behavior of the logistic functions as wind speed increases; plotting the geographical location of residential structures described by the logistic functions; and characterizing clusters of logistic models based on the predictor static variables used to construct the random forest model. This analysis helped identify groups of logistic functions which explain under- and over-prediction errors. Characterization of the errors made by the LogRFT model needs to be investigated in future work in order to improve the performance of the hybrid framework.

In addition, analysis of model generalization over unseen wind speeds uncovered significant over-estimation of wind damage risk made by the LogRFT model compared to the HAZUS-MH model. Nonetheless, training the hybrid model with 3-second wind gust speeds enables the construction of logistic regressions with prediction errors similar to those made by the HAZUS-MH model. Through this observation, the model trained with 3-second wind gust speeds and the HAZUS-MH model produce very similar areal performance when evaluated on the same wind field. Most of the errors persist as under-prediction errors. Future work is needed to investigate why cumulative probability densities of wind damage based on 3-second wind speeds result in a decrease in areal accuracy and significantly increase under-prediction errors when compared to models constructed with 1-minute wind speeds.

The hybrid machine learning methodology presented in this thesis can be used to construct wind damage models that provide real-time predictions for any coastal region. In predicting wind damage accurately, I recommend constructing the model

based on 1-minute maximum sustained wind speeds. However, constructing the model based on 3-second peak wind gusts, produces damage estimates that do not seem to over-estimate at lower wind speeds.

My thesis helps provide more accurate estimations of wind damage risk to residents and emergency agencies for future hurricanes events. The hybrid model I constructed can be used to provide real-time wind damage estimates for Harris County, Texas. With better damage risk estimates, emergency agencies can guarantee the best support to affected communities, and residents can make decisions that potentially lead to the reduction of loss of life and property due to hurricane hazards.

Bibliography

- [1] National Weather Service, “Weather Fatality, Injury and Damage Statistics.” <http://www.nws.noaa.gov/om/hazstats.shtml>, 2015.
- [2] W. D. Nordhaus, “The economics of hurricanes and implications of global warming,” *Climate Change Economics*, vol. 1, pp. 1–20, May 2010.
- [3] K. M. Crosset, *Population trends along the coastal United States: 1980-2008*. Government Printing Office, 2005.
- [4] National Weather Service, “Wind gust.” <http://graphical.weather.gov/definitions/defineWindGust.html>.
- [5] Wisconsin state journal, “Ask the Weather Guys: What causes wind gusts?.” http://host.madison.com/news/local/ask-weather-guys/ask-the-weather-guys-what-causes-wind-gusts/article_3c4ccc46-37db-11e1-ae38-001871e3ce6c.html.
- [6] C. Gutierrez, R. Cresanti, and W. Jeffrey, “Performance of physical structures in hurricane katrina and hurricane rita: A reconnaissance report,” *NIST Technical Note*, vol. 1476, 2006.
- [7] D. E. Jelinski and J. Wu, “The modifiable areal unit problem and implications for landscape ecology,” *Landscape ecology*, vol. 11, no. 3, pp. 129–140, 1996.

- [8] N. Pearce, “The ecological fallacy strikes back,” *Journal of epidemiology and community health*, vol. 54, no. 5, pp. 326–327, 2000.
- [9] P. E. Utgoff, “Perceptron trees: A case study in hybrid concept representations,” *Connection Science*, vol. 1, no. 4, pp. 377–391, 1989.
- [10] A. K. Seewald, J. Petrak, and G. Widmer, “Hybrid decision tree learners with alternative leaf classifiers: An empirical study,” in *FLAIRS Conference*, pp. 407–411, 2001.
- [11] A. Abu-Hanna and N. de Keizer, “Integrating classification trees with local logistic regression in intensive care prognosis,” *Artificial Intelligence in Medicine*, vol. 29, no. 1, pp. 5–23, 2003.
- [12] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [13] C. Unanwa, J. McDonald, K. Mehta, and D. Smith, “The development of wind damage bands for buildings,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 84, no. 1, pp. 119–149, 2000.
- [14] C. O. Unanwa and J. R. McDonald, “Building wind damage prediction and mitigation using damage bands,” *Natural Hazards Review*, vol. 1, no. 4, pp. 197–203, 2000.
- [15] B. R. Ellingwood, D. V. Rosowsky, Y. Li, and J. H. Kim, “Fragility assessment of light-frame wood construction subjected to wind and earthquake hazards,” *Journal of Structural Engineering*, vol. 130, no. 12, pp. 1921–1930, 2004.

- [16] J.-P. Pinelli, E. Simiu, K. Gurley, C. Subramanian, L. Zhang, A. Cope, and J. J. Filliben, “Hurricane damage prediction model for residential structures,” *Journal of Structural Engineering (ASCE)*, vol. 130, no. 11, pp. 1685–1691, 2004.
- [17] A. Cope, K. Gurley, J.-P. Pinelli, and S. Hamid, “A simulation model for wind damage predictions in florida,” in *Proc., 11th Int. Conf. on Wind Engineering*, 2003.
- [18] P. J. Vickery, P. F. Skerlj, J. Lin, L. A. Twisdale Jr, M. A. Young, and F. M. Lavelle, “Hazus-mh hurricane model methodology. ii: Damage and loss estimation,” *Natural Hazards Review*, vol. 7, no. 2, pp. 94–103, 2006.
- [19] K. Gurley, J. Pinelli, C. Subramanian, A. Cope, L. Zhang, J. Murphree, A. Artiles, P. Misra, S. Culati, and E. Simiu, “Florida public hurricane loss projection model engineering team. final report,” *International Hurricane Research Center, Florida Internat’l University*, 2005.
- [20] S. Hamid, B. G. Kibria, S. Gulati, M. Powell, B. Annane, S. Cocke, J.-P. Pinelli, K. Gurley, and S. Chen, “Predicting losses of residential structures in the state of Florida by the Public Hurricane Loss Evaluation Model,” *Statistical methodology*, vol. 7, no. 5, pp. 552–573, 2010.
- [21] S. Chung Yau, N. Lin, and E. Vanmarcke, “Hurricane damage and loss estimation using an integrated vulnerability model,” *Natural Hazards Review*, vol. 12, no. 4, pp. 184–189, 2010.
- [22] N. Lin, E. Vanmarcke, and S.-C. Yau, “Windborne debris risk analysis-part ii. application to structural vulnerability modeling,” *Wind and Structures*, vol. 13, no. 2, pp. 207–220, 2010.

- [23] N. Lin and E. Vanmarcke, “Windborne debris risk assessment,” *Probabilistic Engineering Mechanics*, vol. 23, no. 4, pp. 523–530, 2008.
- [24] N. Lin and E. Vanmarcke, “Windborne debris risk analysis-part i. introduction and methodology,” *Wind and Structures*, vol. 13, no. 2, pp. 191–206, 2010.
- [25] J. Michael Grayson, W. Pang, and S. Schiff, “Building envelope failure assessment framework for residential communities subjected to hurricanes,” *Engineering Structures*, vol. 51, pp. 245–258, 2013.
- [26] M. Grayson, W. Pang, and S. Schiff, “Three-dimensional probabilistic windborne debris trajectory model for building envelope impact risk assessment,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 102, pp. 22–35, 2012.
- [27] Z. Huang, D. Rosowsky, and P. Sparks, “Hurricane simulation techniques for the evaluation of wind-speeds and expected insurance losses,” *Journal of wind engineering and industrial aerodynamics*, vol. 89, no. 7, pp. 605–617, 2001.
- [28] P. Sparks, “Wind speeds in tropical cyclones and associated insurance losses,” *Journal of wind engineering and industrial aerodynamics*, vol. 91, no. 12, pp. 1731–1751, 2003.
- [29] C. A. Dehring and M. Halek, “Do coastal building codes mitigate hurricane damage to residential property?,” *Available at SSRN 928009*, 2006.
- [30] W. E. Highfield, W. G. Peacock, and S. Van Zandt, “Determinants & characteristics of damage in single-family island households from hurricane ike1,” in *The Association of colle-25 giate schools of planning conference, Minneapolis, Minnesota*, pp. 7–10, 2010.

- [31] J. Kim, P. Woods, Y. Park, T. Kim, J. Choi, and K. Son, “Predicting the hurricane damage ratio of commercial buildings by claim payout from hurricane ike,” *Natural Hazards and Earth System Sciences Discussions*, vol. 1, no. 4, pp. 3449–3483, 2013.
- [32] Y. Li and B. R. Ellingwood, “Hurricane damage to residential construction in the us: Importance of uncertainty modeling in risk assessment,” *Engineering structures*, vol. 28, no. 7, pp. 1009–1018, 2006.
- [33] FEMA, “HAZUS Multi-Hazard Software.” www.fema.gov/hazus.
- [34] D. V. Rosowsky and B. R. Ellingwood, “Performance-based engineering of wood frame housing: Fragility analysis methodology,” *Journal of Structural Engineering*, vol. 128, no. 1, pp. 32–38, 2002.
- [35] P. J. Vickery, J. X. Lin, and L. A. Twisdale Jr, “Analysis of hurricane pressure cycling following missile impact for residential structures,” *Journal of wind engineering and industrial aerodynamics*, vol. 91, no. 12, pp. 1703–1730, 2003.
- [36] L. Twisdale, P. Vickery, and A. Steckley, “Analysis of hurricane windborne debris impact risk for residential structures,” *Applied Research Associates Inc., Raleigh, North Carolina, USA*, 1996.
- [37] P. Vickery, J. Lin, and L. Twisdale, “Analysis of hurricane windborne debris impact risk for residential structures. ii, applied research associates,” *Inc., Raleigh, NC*, 1999.
- [38] FEMA, “HAZUS-MH Hurricane Wind Model Validation Study – Florida.” <http://www.fema.gov/library/viewRecord.do?id=2757>, April 2007.

- [39] D. Subramanian, J. Salazar, L. Duenas-Osorio, and R. Stein, “Building and validating geographically refined hurricane wind risk models for residential structures,” *Natural Hazards Review*, vol. 10, 2013.
- [40] P. Vickery, J. Lin, P. Skerlj, L. A. Twisdale, and K. Huang, “HAZUS-MH hurricane model methodology I: hurricane hazard, terrain, and wind load modeling,” *Natural Hazards Review*, vol. 7, no. 2, pp. 82–93, 2006.
- [41] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [42] C. M. Bishop *et al.*, *Pattern recognition and machine learning*, vol. 1. springer New York, 2006.
- [43] P. Niyogi, F. Girosi, and T. Poggio, “Incorporating prior information in machine learning by creating virtual examples,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2196–2209, 1998.
- [44] A. H. Fielding and J. F. Bell, “A review of methods for the assessment of prediction errors in conservation presence/absence models,” *Environmental conservation*, vol. 24, no. 01, pp. 38–49, 1997.
- [45] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.
- [46] Harris County Housing Authority, “Harris County Damage Assessment: Helping Harris County Communities recover from Hurricane Ike.” http://www.harriscountytexas.gov/cmpdocuments/103/ike/harris_county_damage_assessment_ike.pdf, 2009.

- [47] Harris County Appraisal District, “Harris County Appraisal District Property records.” <http://pdata.hcad.org/download/2008.html>, 2008.
- [48] ArcGIS Help, “Average Nearest Neighbor Tool.” <http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000008000000>.
- [49] W. R. Tobler, “A computer movie simulating urban growth in the detroit region,” *Economic geography*, pp. 234–240, 1970.
- [50] D. W.-S. Wong and J. Lee, *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. John Wiley & Sons Hoboken, NJ, USA, 2005.
- [51] P. A. Moran, “Notes on continuous stochastic phenomena,” *Biometrika*, pp. 17–23, 1950.
- [52] ArcGIS Help, “Spatial autocorrelation (Global Moran’s I).” http://resources.arcgis.com/en/help/main/10.1/index.html#/Spatial_Autocorrelation_Global_Moran_s_I/005p0000000n000000/.
- [53] United States Census Bureau, “Census Blocks.” https://www.census.gov/geo/reference/gtc/gtc_block.html.
- [54] United States Census Bureau, “Census Blocks 2008 TIGER/Line Shapefile for Harris County.” <https://www.census.gov/cgi-bin/geo/shapefiles2008/county-files?county=48201>.
- [55] ArcGIS Help, “Calculate Distance Band from Neighbor Count (Spatial Statistics).” <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//005p00000003r0000000>.

- [56] Rice University GIS/Data Center, “Harris County building footprint database from 2012.” <http://library.rice.edu/services/gdc>.
- [57] Harris County Appraisal District GIS Maps, “GIS HCAD Public Data.” <http://pdata.hcad.org/GIS/index.html>, 2008.
- [58] United States Geological Survey, “USGS Seamless Data Warehouse.” <http://seamless.usgs.gov>, 2011.
- [59] National Land Cover Database, “NLCD land cover class definitions.” <http://landcover.usgs.gov/classes.php>, 2011.
- [60] Federal Emergency Management Agency, “HAZUS-MH 2.1 user manual.” <http://www.fema.gov/library/viewRecord.do?id=5120>, 2009.
- [61] ArcGIS Help, “Near (Analyst) tool.” <http://resources.arcgis.com/en/help/main/10.2/index.html#/00080000001q000000>.
- [62] City of Houston Department of Public Works, “Geographic Information Management System.” <http://www.gims.houstontx.gov/PortalWS/MainPortal.aspx>, 2008.
- [63] S.-R. Han, S. D. Guikema, S. M. Quiring, K.-H. Lee, D. Rosowsky, and R. A. Davidson, “Estimating the spatial distribution of power outages during hurricanes in the gulf coast region,” *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 199–210, 2009.
- [64] United States Geological Survey, “State soil geographic (statsgo).” <http://water.usgs.gov/GIS/metadata/usgswrd/XML/ussoils.xml>, 1994.

- [65] R. Nateghi, S. Guikema, and S. Quiring, “Comparison and validation of statistical methods for predicting power outage durations during hurricanes,” *Risk Analysis*, vol. 31, no. 12, pp. 1897–1906, 2011.
- [66] Houston-Galveston Area Council, “15 county land use and land cover data.” <http://www.h-gac.com/community/socioeconomic/land-use-data.aspx>, 2008.
- [67] Houston-Galveston Area Council, “H-GAC LiDAR 2008 database.” <http://www.h-gac.com/rds/imagery/lidar-imagery.aspx>, 2008.
- [68] National Atmospheric and Oceanic Administration, “Ike wind analyses.” http://www.hwind.co/legacy_data/Storms/ike2008.html, 2008.
- [69] M. D. Powell, S. H. Houston, L. R. Amat, and N. Morisseau-Leroy, “The HRD real-time hurricane wind analysis system,” *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 77-78, pp. 53–64, 1998.
- [70] M. D. Powell, S. H. Houston, and I. Ares, “Real-time Damage Assessment in Hurricanes,” *21st American Meteorological Society Conference on Hurricanes and Tropical Meteorology*, 1995.
- [71] E. Tuv, A. Borisov, G. Runger, and K. Torkkola, “Feature selection with ensembles, artificial variables, and redundancy elimination,” *The Journal of Machine Learning Research*, vol. 10, pp. 1341–1366, 2009.
- [72] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

- [73] M. Köppen, “The curse of dimensionality,” in *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, pp. 4–8, 2000.
- [74] M. Binshtok, R. I. Brafman, S. E. Shimony, A. Martin, and C. Boutilier, “Computing optimal subsets,” in *Proceedings of the National Conference on Artificial Intelligence*, vol. 22, p. 1231, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2007.
- [75] M. A. Hall, *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.
- [76] ArcGIS Help, “Estimating forest canopy density and height.” <http://resources.arcgis.com/en/help/main/10.1/index.html#//015w00000056000000>.
- [77] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.
- [78] M. Sandri and P. Zuccolotto, “Variable selection using random forests,” in *Data analysis, classification and the forward search*, pp. 263–270, Springer, 2006.
- [79] MathWorks Documentation, “Treebagger class.” <http://www.mathworks.com/help/stats/treebagger-class.html>.
- [80] FEMA, “FEMA P-757, Hurricane Ike in Texas and Louisiana: Mitigation Assessment Team Report, Building Performance Observations, Recommendations, and Technical Guidance.” <https://www.fema.gov/media-library/assets/documents/15498>, 2009.