A Machine Learning Method to Quantify the Role of Vulnerability in Hurricane Damage

Laura Szczyrba
Virginia Tech
lszczyrba@vt.edu

Yang Zhang
Virginia Tech
yang08@vt.edu

Duygu Pamukcu
Virginia Tech
duygu@vt.edu

Derya Ipek Eroglu
Virginia Tech
deryaipek@vt.edu

ABSTRACT

Accurate pre-disaster damage predictions and post-disaster damage assessments are challenging because of the complicated interrelationships between multiple damage drivers, including various natural hazards, as well as antecedent infrastructure quality and demographic characteristics. Ensemble decision trees, a family of machine learning algorithms, are well suited to quantify the role of social vulnerability in disaster impacts because they provide interpretable measures of variable importance for predictions. Our research explores the utility of an ensemble decision tree algorithm, Random Forest Regression, for quantifying the role of vulnerability with a case study of Hurricane María. The contributing predictive power of eight drivers of structural damage was calculated as the decrease in model mean squared error. A measure of social vulnerability was found to be the model's leading predictor of damage patterns. An additional algorithm, other methods of quantifying variable importance, and future work are discussed.

Keywords

Vulnerability, Impact, Damage, Machine Learning, Hurricane María.

INTRODUCTION

Factors that contribute to the impact of a disaster may exhibit complex, nonlinear relationships with structural damage and with each other. Because social vulnerability measures are especially challenging to measure and deconvolve, many hurricane impact assessment models rely on relationships between hazardous forcings and structural materials alone, while ignoring the role of social variables, e.g., socioeconomic status. However, studies have shown that socioeconomic status consistently emerges as a contributing factor to housing damage in a variety of disasters and that vulnerable populations are more likely to reside in homes that receive higher levels of damage during a hurricane (Fothergill and Peek 2004). To address this disconnect, machine learning methods can be applied to quantify the complicated role that social vulnerability plays in structural damage.

Because of the quantity and complexity of variables that influence structural damage, a large amount of data is required to build a representative model. Many statistical models require informed assumptions about the input data, i.e., if the relationships are expected to be linear or nonlinear. Ensemble decision trees, a family of machine learning algorithms, are useful for analyzing large, multivariate datasets in which the relationships between the variables and the target are unknown. These highly flexible algorithms are non-parametric, non-linear, and can accommodate highly dimensional datasets.

Random Forest is a common ensemble decision tree algorithm that constructs a group of independent classification or regression trees and leverages the majority vote of trees to determine the resultant prediction for classification or the average prediction per data instance for regression models (Breiman 1996; Breiman 2001). Each decision tree is constructed from a bootstrap subset of the given training dataset and a random subset of the predictive features (Ho 1995) which ensures that the decision trees are, ideally, uncorrelated. Each tree begins with a parent node and then the algorithm recursively splits the data with Boolean Logic towards increasingly homogeneous groups of child nodes by optimizing a splitting criterion (e.g., mean squared error) until the tree reaches its stopping criterion (e.g., maximum depth) at the terminal nodes, where the predictions are produced. As the number of trees increases, by the Strong Law of Large Numbers, the resultant predictions converge.

By leveraging learned, as opposed to assumed, statistical relationships in large datasets, ensemble decision tree algorithms can provide otherwise ignored approaches to disaster data analytics, yet few disaster management studies have taken advantage of this suite of tools. A 2017 study of wildfire damage in Portugal by Oliveira et al. used ensemble decision trees to explore the correlation between damage and demographic factors. They concluded that purchasing power and housing quality were significantly correlated with the extent of wildfire damage. In addition, they found that certain demographic groups, such as the elderly and households with lower education levels, were relatively more vulnerable to wildfire impacts. A study of flood damage across German households found ensemble decision trees were more accurate than traditional impact models, while precautionary measures were used as a social indicator and were negatively correlated with structural losses (Merz et al. 2013). By leveraging household-level damage assessments in Bangladesh, another flood study used machine learning algorithms, including ensemble decision trees, to conclude that larger households and higher education levels were associated with lower flood damage (Ganguly et al. 2019). The field of disaster management is just beginning to utilize machine learning for explanatory purposes. This case study applies the Random Forest ensemble decision tree algorithm to quantitatively explore the relative role that social factors played in the structural damage caused by Hurricane María in Puerto Rico.

HURRICANE MARIA DATASET

Hurricane María made landfall in Puerto Rico on September 20th, 2017 as an intense Category 4 storm that left the island in devastation for months to follow. It became the third costliest storm in U.S. history, after Hurricanes Katrina (2005) and Harvey (2017), with total approximate damages of $90 billion in the U.S. Virgin Islands and Puerto Rico (Pasch et al. 2018). Furthermore, an independent study commissioned by Puerto Rico's government estimated a death toll of 2,975 individuals, although the number of fatalities remains a contentious and politicized topic (Federal Emergency Management Agency 2018b). In addition to the storm's intensity, the severity of María's impact also resulted from pre-existing social disparities that have developed throughout the history of Puerto Rico. Included in these disparities, a shortage of adequate low-income housing has been exacerbated in the past two decades by the economic downturn in Puerto Rico (Santiago-Bartolomei 2018).

A diverse set of publicly available data was collected and processed in ArcGIS software to explore the role of vulnerability in the impact of Hurricane María (Table 1). Because census tracts are widely used for public policy and planning that specifically promote socioeconomic well-being (Krieger 2006), this study scaled to census tract spatial units. Each dataset contained full coverage of the island of Puerto Rico.

Table 1. Description of all variables included in this analysis.

	Measure	*Variable Type*	*Source*	*Abbreviation*
Wind	Distance from hurricane center	Predictive	NHC	HurTrack
	Peak gust	Predictive	ARA	PeakGust
Flood	Proportion of flooded area	Predictive	FEMA	PropFA
	Average depth of flooding	Predictive	FEMA	AveDepth
	Proportion of SFHA	Predictive	FEMA	PropSFHA
Landslide	Average landslide density	Predictive	USGS	AveLS
Vulnerability	Proportion of special communities	Predictive	PR OFSA	PropSC
	Social vulnerability	Predictive	CDC	CDCVuln
Damage	Damage ratio index	Target	FEMA	DamInd

Hurricane María threatened Puerto Rico with wind, flood, and landslide hazards (Figure 1). Data collected to represent these hazards required manipulation in ArcGIS to represent a census tract-level hazard summaries. The Feature to Point tool extracted a within-polygon center point (i.e., centroid) from each census tract. Then, the proximity tool Near calculated the nearest distance (in degrees) between each center point and the National Hurricane Center (NHC) Hurricane María best track line (National Hurricane Center 2017). Applied Research Associates (ARA) windfield data are partitioned to the census tract scale and did not require further processing (Applied

Research Associates 2017). Federal Emergency Management (FEMA) flood event depth grids provided mean depth of flooding (m) measurements and were also used to calculate flooded area (Federal Emergency Management Agency 2017). Zonal Statistics calculated the mean raster values per census tract. The Raster to Polygon conversion tool recast the raster grids to polygons, allowing measurement of flooded area per census tract. Depth grid polygons were intersected with the census tracts to calculate overlapping area (m^2) divided by the total area of the census tract to determine proportion of flooded area. The National Flood Insurance Program (NFIP) Special Flood Hazard Area (SFHA) data were processed similarly and provided a measure of general flood risk, quantified as the proportion of SFHA-covered area (Federal Emergency Management Agency 2018c). A categorical dataset from the United States Geological Survey (USGS) measured landslide density, i.e., number of landslides per four square kilometers (United States Geological Survey 2019). These data were converted from a 2 km x 2 km grid to census tract measures by intersecting the grid with the tracts and Zonal Statistics summarized the average landslide code value within each tract.

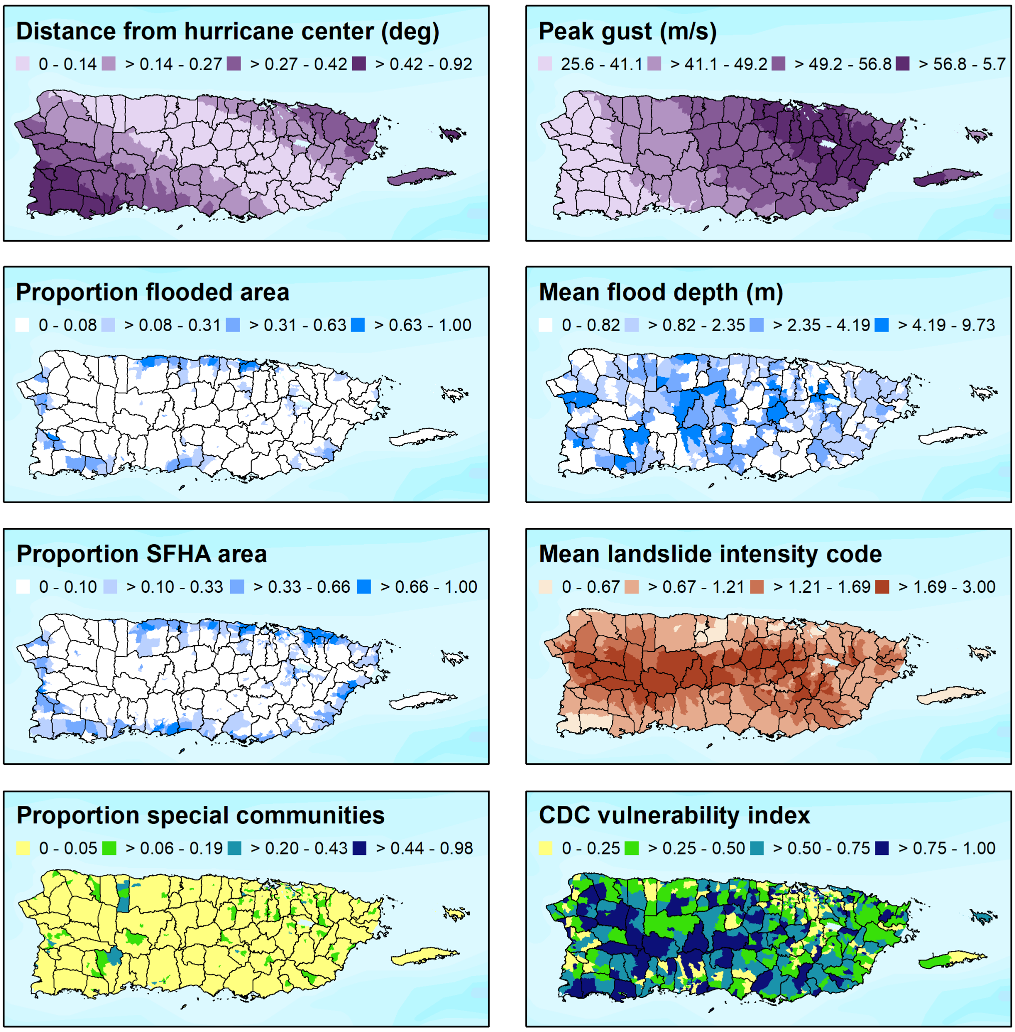

Figure 1. Predictive features used in the machine learning analysis.

Two pre-existing datasets represented vulnerability in this analysis (Figure 1). The Puerto Rican Geographic Information Systems Catalogue provided a dataset of "special communities" compiled by the Puerto Rican Office of the General Coordinator for Social Financing and Self-Management (PR-OFSA) (Oficina del Coordinador General para el Financiamiento Social y la Autogestión 2008). It contained a polygon shapefile delineating 713 identified disadvantaged communities throughout the island. The same methods used to calculate the proportional flooded area and proportional area of SFHA were also applied: the area of each census tract was intersected with the special communities data and then divided by the total area. The Center for Disease Control (CDC) created a social vulnerability index for Puerto Rico in 2017 with methods from Flanagan et al. (2011). These data leverage 15 vulnerability indicators from the American Community Survey (ACS) 5-year dataset from 2012-2016 including socioeconomic status, household composition and disability, minority status, language, housing, transportation, among other characteristics (Center for Disease Control 2017).

A preexisting dataset of Puerto Rican structural vulnerability does not exist and, furthermore, the island lacks accurate permitting and construction data. The few available datasets are incomplete and fail to capture the widespread abundance informal homes. Informal housing refers to structures that are self-built without proper permitting or licensed contractors, and they are often out of compliance with zoning and building regulations. A 2018 report by the Puerto Rico Home Builders Association estimated that 45% of structures on the island are informal (Asociación de Constructores de Puerto Rico 2018). Given the prevalence of these structures, the existing databases are unreliable and were not used in this analysis. However, future work will include the creation of a structural vulnerability index for Puerto Rico, utilizing Census and American Community Survey data (see Conclusions and Future Work section).

A structural damage index, the target variable in the Random Forest analysis, depicted Hurricane María's structural impact normalized by exposure. After the hurricane, FEMA created an aerial damage assessment database by categorically identifying structure-by-structure damage based on geospatial assessments of pre- and post-event imagery (Federal Emergency Management Agency 2018a). This dataset was augmented to include "Not Affected" points with centroids extracted from building footprint data from OpenStreetMap (OSM) (OpenStreetMap 2019) (Figure 2). The damage index measures the ratio of damaged structures to the total number of structures in each census tract (Burton 2010) and is a continuous numerical target, ranging from 0 to 0.36.

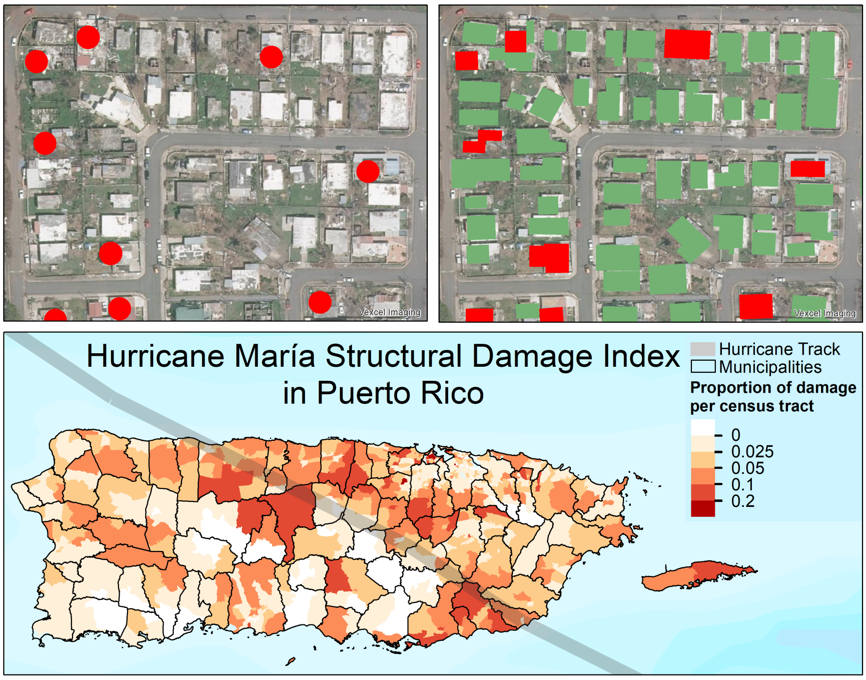
Vexcel Imaging
Vexcel Imaging
Hurricane María Structural Damage Index in Puerto Rico
Hurricane Track
Municipalities
Proportion of damage per census tract
0
0.025
0.05
0.1
0.2

Figure 2. Hurricane María damage index development. FEMA damage assessment points were combined with OSM building footprints and then aggregated to the census tract scale as the ratio of damaged to total structures.

ANALYSIS

After constructing a geodatabase aggregated to the census tract scale, the Random Forest regression algorithm, sourced from Python's SciKit Learn machine learning package (Pedregosa et al. 2011), was applied to extract the relative importance of predictive features. A total of 905 land-based census tracts are within Puerto Rico, 882 of which have complete Census information collected. The dataset was then split randomly into a training set containing 80% of the data and an evaluation set containing the remaining 20% of data. Automated optimization techniques, including grid search cross validation, tuned the model to optimal hyperparameters using variance (r^2), mean absolute error (MAE), and mean error (ME) measures to assess model performance.

The automated tuning process updated three of the Random Forest default parameters. It limited each tree's depth to 250 levels. A total of 1,000 predictive trees comprised the ensemble and the maximum number of random features analyzed at each decision tree split was constrained to two features because of the relatively low number of predictive features. The training model captured 72% of the variance (r^2) and generalized to represent 29% of the variance (r^2) on the separate, unseen evaluation dataset. The MAE measured 0.012 and 0.020 while the model's ME measured -0.000038 and -0.0017 for the training and evaluation dataset, respectively. Table 2 compares the performance of the training and evaluation data from the Random Forest regression model and Figure 3 plots the model predictions against the true data values.

Table 2. Summary of Random Forest model performance.

	Training Score	*Evaluation Score*
Variance (r^2)	0.72	0.29
Mean Absolute Error	0.012	0.020
Mean Error	-0.000038	-0.0017

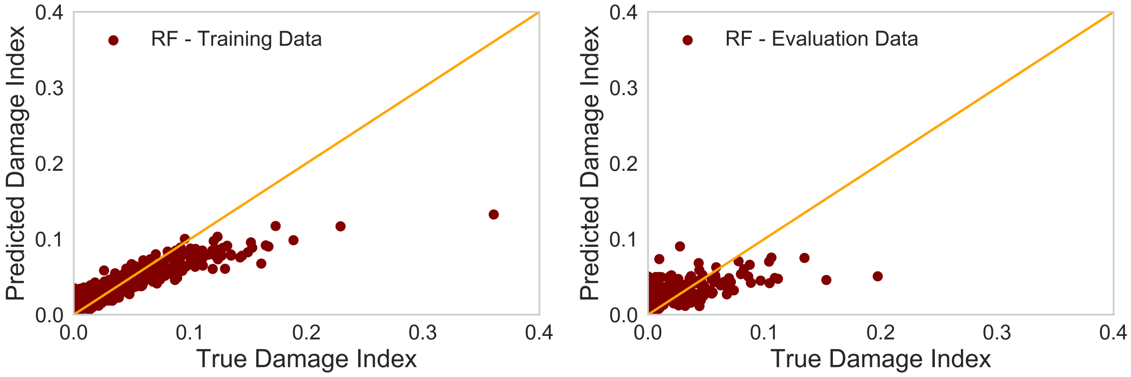

Figure 3. Random Forest regression model predictions compared to true target variable values.

If the models heavily relied upon vulnerability measures to determine damage patterns, it would indicate that vulnerability played an important role in this case study. Ensemble decision tree algorithms provide straightforward means to interpret the importance of predictive features (Breiman et al. 1984). The trees built in this study relied upon mean squared error (MSE) to split each node, thus they simultaneously quantified the quality of each split as the effectiveness of each predictive feature in reducing predicted target variable error. Therefore, the default importance calculation provided by SciKit Learn - mean decrease in MSE - measured the importance of each predictive feature (Pedregosa et al. 2011). Results from the importance calculation indicate that the CDC vulnerability index (CDCVuln) was the most important predictive feature of Hurricane María damage (Figure 4). The following three influential features include distance from hurricane track (HurTrack), peak gust (PeakGust), and then the proportion of special communities per census tract (PropSC).

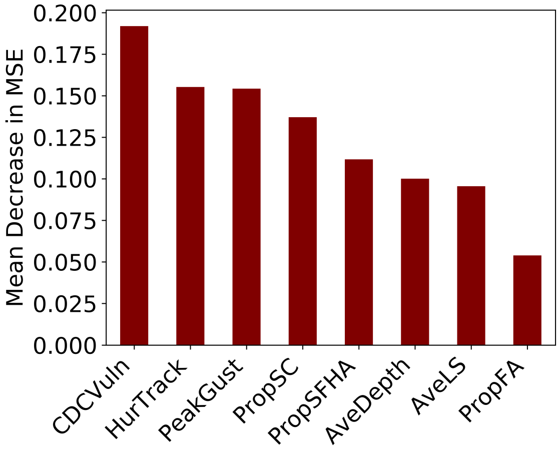

Figure 4. Feature importances measured as mean decrease in model mean squared error.

CONCLUSIONS AND FUTURE WORK

Ensemble decision tree algorithms can inform damage prediction, disaster response, and impact assessment efforts because they are readily interpretable, can accommodate large datasets from a variety of different sources and formats, and provide quantifiable measures of feature importance. Preliminary results show that social vulnerability is correlated with Hurricane María's damage patterns, with one measure as the leading variable informing the model above all other wind, flood, and landslide variables. Proportion of identified disadvantaged or "special" communities per census tract, the secondary measure of vulnerability, was the fourth most important predictive feature and was more informative than the flood and landslide variables. Overall, vulnerability played a critical role in the analysis of damage patterns.

The initial findings of this study indicate that hazardous forces alone do not sufficiently explain damage patterns. Hurricane impact assessment models should include social factors as input variables to accurately depict areas of priority for decision makers, improve resource allocation, and, ultimately, ensure a more efficient and equitable response effort. These findings also contribute to the conclusions of existing literature which encourage deliberate pre-disaster mitigation investment in vulnerable communities (Fothergill and Peek 2004). In order to prevent the exacerbation of social disparities in the wake of disasters, it is vital that the prevailing structural and institutional policies which lead to society-wide inequalities are reflected upon.

The reported model variance (r^2), MAE, and ME indicate that the model is overfitting the training dataset. This may be due to the imbalanced distribution in target variable values. The data contain a prominence of census tracts at low damage levels and areas of higher damage are considered outliers (Figure 5). Since machine learning algorithms optimize to reflect average conditions, this makes generalizability challenging on disproportionately distributed data. The paucity of data points also impedes generalizability. Research on the performance of machine learning algorithms on small, imbalanced datasets indicate that this is a common challenge (He and Garcia 2009) and previous studies leveraging similarly distributed targets reported comparable overfitting behaviors (Sadler et al. 2018).

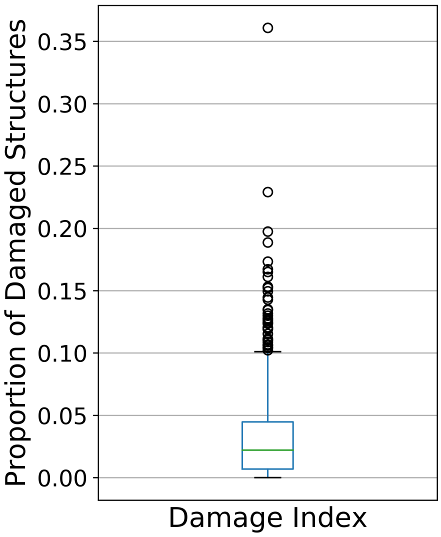

Figure 5. Distribution of damage index values.

The input data limited the predictive performance of the Random Forest model. On average, the model could be trained to account for approximately 29% of the variance in the evaluation dataset with an 72% fit in the training dataset. The dataset contains a high amount of variance due to a number of factors in addition to the distribution of the target variable, including the passage of Hurricane Irma to the north of the island just weeks before Hurricane María, the absence of structural and building material data, and the oversimplified wind data which did not incorporate topographic effects (such as wind speed-up in mountainous terrains). Given these limitations, a predictive performance of 29% on a testing set is satisfactory, however improvement is expected as future work progresses.

The original damage index target variable is imbalanced towards low damage values (Figure 5). Machine learning algorithms optimize on average performance criteria, therefore it is not surprising that the preliminary Random Forest model underpredicted damages due to the prevalence of low damage points in the target dataset. To improve upon this, two additional indices will be run through the analysis, one excluding census tracts with 0 damage, and another with a log transformation of the index without 0 damage records. Further manipulation of the target variable will be avoided in order to maintain the interpretability of model results.

An additional ensemble decision tree machine learning algorithm, Stochastic Gradient Boosting Trees (Friedman 2002), will also be implemented. This algorithm may perform better than the Random Forest algorithm because it builds stagewise decision trees that iteratively improve the performance of the previous estimator, known as model boosting. It is expected that after adjusting the target variable distribution and utilizing a boosting algorithm, model performance on the evaluation dataset will significantly improve.

To augment the preexisting vulnerability data, new vulnerability indices will be calculated based on the method proposed by Cutter et al. in 2003. The contributions of socioeconomic and structural vulnerabilities will be examined separately by categorizing variables into two groups where the socioeconomic vulnerability index will be measured by the variables representing the living conditions and population characteristics, and the structural vulnerability index will be measured based on the variables representing housing characteristics and the structural quality (Holand et al. 2011). These two original vulnerability measures will then be incorporated into the ensemble decision tree models to examine the relative influence of structural, socioeconomic, or comprehensive vulnerability on damage.

Additional measures of predictive feature importance and dependencies will also be calculated. Because the default impurity importance measure can potentially be biased towards favoring features with high-cardinality (Strobl et al. 2007), the importance of each predictive feature will also be calculated by permuting, or randomly shuffling, each feature's values while measuring changes in model variance (Breiman 2001) and by permuting related groups of features (e.g. wind, flood, landslide, vulnerability) while measuring changes in variance (Koch et al. 2019). While feature importance measures indicate which features are most-valuable to model performance, they provide little information in terms of how or why features are important. Learned partial dependencies demonstrate the expected damage response as a function of each feature and can be extracted from trained decision tree models (Friedman 2002). The average dependence of each feature with the target variable will be plotted as well as all individual instances (Goldstein et al. 2015) which can reveal hidden heterogeneities or interactions within the dataset.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number 1735139. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Applied Research Associates (2017). *Hurricane María Wind Data*. https://disasters.geoplatform.gov/publicdata/NationalDisasters/2017/HurricaneMaria/Data/Wind/ARA/.

Asociación de Constructores de Puerto Rico (2018). "Situación de la Industria de la Vivienda en Puerto Rico: Recomendaciones de Política Pública". In: *Informe Final*, pp. 1–63.

Breiman, L. (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140.

Breiman, L. (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). "Classification and regression trees. Wadsworth Int". In: *Group* 37.15, pp. 237–251.

Burton, C. G. (2010). "Social vulnerability and hurricane impact modeling". In: *Natural Hazards Review* 11.2, pp. 58–68.

Center for Disease Control (2017). *CDC Social Vulnerability Index Puerto Rico*. https://respond-irma-geoplatform.opendata.arcgis.com/datasets/39490368e512402ba5d7635458b18f30.

Cutter, S. L., Boruff, B. J., and Shirley, W. L. (2003). "Social vulnerability to environmental hazards". In: *Social science quarterly* 84.2, pp. 242–261.

Federal Emergency Management Agency (2017). *Modeled Preliminary Observations*. https://disasters.geoplatform.gov/publicdata/NationalDisasters/2017/HurricaneMaria/Data/DepthGrid/.

Federal Emergency Management Agency (2018a). *Historical Damage Assessment Database, Public Release*. https://communities.geoplatform.gov/disasters/historical-damage-assessment-database/.

Federal Emergency Management Agency (2018b). “Mitigation Assessment Team Report: Hurricanes Irma and Maria in Puerto Rico”. In: *FEMA P-2020 Mitigation Assessment Team Report*, pp. 1–296.

Federal Emergency Management Agency (2018c). *Puerto Rico, Commonwealth of: Effective Products*. https://msc.fema.gov/portal/advanceSearch#searchresultsanchor.

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., and Lewis, B. (2011). “A social vulnerability index for disaster management”. In: *Journal of homeland security and emergency management* 8.1.

Fothergill, A. and Peek, L. A. (2004). “Poverty and disasters in the United States: A review of recent sociological findings”. In: *Natural hazards* 32.1, pp. 89–110.

Friedman, J. H. (2002). “Stochastic gradient boosting”. In: *Computational statistics & data analysis* 38.4, pp. 367–378.

Ganguly, K. K., Nahar, N., and Hossain, B. M. (2019). “A machine learning-based prediction and analysis of flood affected households: A case study of floods in Bangladesh”. In: *International journal of disaster risk reduction* 34, pp. 283–294.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation”. In: *Journal of Computational and Graphical Statistics* 24.1, pp. 44–65.

He, H. and Garcia, E. A. (2009). “Learning from imbalanced data”. In: *IEEE Transactions on knowledge and data engineering* 21.9, pp. 1263–1284.

Ho, T. K. (1995). “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition*. Vol. 1. IEEE, pp. 278–282.

Holand, I. S., Lujala, P., and Rød, J. K. (2011). “Social vulnerability assessment for Norway: A quantitative approach”. In: *Norsk Geografisk Tidsskrift-Norwegian Journal of Geography* 65.1, pp. 1–17.

Koch, J., Stisen, S., Refsgaard, J. C., Ernstsen, V., Jakobsen, P. R., and Højberg, A. L. (2019). “Modeling Depth of the Redox Interface at High Resolution at National Scale Using Random Forest and Residual Gaussian Simulation”. In: *Water Resources Research* 55.2, pp. 1451–1469.

Krieger, N. (2006). “A century of census tracts: health & the body politic (1906–2006)”. In: *Journal of urban health* 83.3, pp. 355–361.

Merz, B., Kreibich, H., and Lall, U. (2013). “Multi-variate flood damage assessment: a tree-based data-mining approach”. In: *Natural Hazards and Earth System Sciences* 13.1, pp. 53–64.

National Hurricane Center (2017). *NHC GIS Archive - Tropical Cyclone Best Track - Hurricane María*. https://www.nhc.noaa.gov/gis/archive_besttrack.php?year=2017.

Oficina del Coordinador General para el Financiamiento Social y la Autogestión (2008). *Comunidades especiales*. http://www.gis.pr.gov/descargaGeodatos/Delimitaciones/Pages/Comunidades.aspx.

Oliveira, S., Zêzere, J. L., Queirós, M., and Pereira, J. M. (2017). “Assessing the social context of wildfire-affected areas. The case of mainland Portugal”. In: *Applied geography* 88, pp. 104–117.

OpenStreetMap (2019). *HOTOSM Puerto Rico Buildings*. https://data.humdata.org/dataset/hotosm_pri_buildings#.

Pasch, R. J., Penny, A. B., and Berg, R. (2018). “National hurricane center tropical cyclone report: Hurricane Maria”. In: *TROPICAL CYCLONE REPORT AL152017, National Oceanic And Atmospheric Administration and the National Weather Service*, pp. 1–48.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). “Scikit-learn: Machine learning in Python”. In: *Journal of machine learning research* 12.Oct, pp. 2825–2830.

Sadler, J., Goodall, J., Morsy, M., and Spencer, K. (2018). “Modeling urban coastal flood severity from crowd-sourced flood reports using Poisson regression and Random Forest”. In: *Journal of hydrology* 559, pp. 43–55.

Santiago-Bartolomei, R. (2018). “Notes for a Planning and Public Policy Framework for Housing in Puerto Rico”. In: *Center for a New Economy Housing and Land Initiative*, pp. 1–6.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). “Bias in random forest variable importance measures: Illustrations, sources and a solution”. In: *BMC bioinformatics* 8.1, p. 25.

United States Geological Survey (2019). *Map data from landslides triggered by Hurricane María in four study areas of Puerto Rico*. https://www.sciencebase.gov/catalog/item/5ca3c65fe4b0b8a7f6334309.