How important are socioeconomic factors for hurricane performance of power systems? An analysis of disparities through machine learning

Alexys H. Rodríguez A.*, Abdollah Shafieezadeh† and Alper Yilmaz‡
Department of Civil, Environmental and Geodetic Engineering, The Ohio State University
Columbus, Ohio 43210
Email: *rodriguezavellaneda.1@osu.edu, †shafieezadeh.1@osu.edu, ‡yilmaz.15@osu.edu

Abstract—This paper investigates whether socioeconomic factors are important for the hurricane performance of the electric power system in Florida. The investigation is performed using the Random Forest classifier with Mean Decrease of Accuracy (MDA) for measuring the importance of a set of factors that include hurricane intensity variables, a group of independent variables (see (1)) including hurricane performance model. This indicates that social disparities may exist in the occurrence of power outages, which directly impact the resilience of communities and thus require immediate attention.

Index Terms—extreme winds, electric power system, performance, socioeconomic, random forest

I. INTRODUCTION

Hurricane winds induce physical damage to power grid infrastructure and consequently outages that are often widespread. The outages may subsequently disrupt the services of other infrastructure systems, thus impacting communications, emergency services, and humanitarian aid in the critical times after the hazard. Aware of this complex environment, researchers have focused their efforts on understanding the physical vulnerabilities and assessing risks facing the power system, primarily from engineering perspective, to mitigate risks through optimal pre-allocation of resources [1]–[3] or post-disaster recovery plans [4]–[6].

A. Social impacts of infrastructure disruptions

Infrastructure systems have a strategic priority for the society [7]–[9]. However, the main requirements and guidelines that direct infrastructure design and engineering practices lack consideration of the socioeconomic impacts of infrastructure service disruptions [10], which is deemed as a critical gap [10], [11]. Chang [12] argues the importance of engineering, social science, interdisciplinary research, planning, and a growing amount of information sources for reducing infrastructure systems disruptions caused by natural hazards and for improving the resilience of certain socioeconomic vulnerable groups to disasters.

The socioeconomic impacts caused or exacerbated by infrastructure disruptions and their cascading effects during extreme events have been investigated by several multidisciplinary teams. For example, Mitsova et al. [13] studied the electric power recovery times (dependent variable) in Florida caused by Hurricane Irma (2017). Analysis of data using spatial lag regression models at the county scale, and later on at the household scale [14], showed that racial minority groups with poor socioeconomic status are more impacted by power service disruptions.

Environmental justice is a framework that focuses on socioeconomic differences between spatial units that are impacted and not impacted by environmental hazards. The analysis considers demographic (e.g. race, ethnicity, income or poverty, education) and/or environmental (distance to environmental hazards, others) indicators to analyze differential recovery rates in advance to natural disasters, therefore, facilitating identification of communities that are more vulnerable. Using this framework and quantitative regression analyses with three environmental justice vulnerability indexes (demographic, environmental hazard, environmental justice), and other indicator variables (terrain and distance to electric transmission lines), [15] found that specific vulnerable community groups in Puerto Rico had lower rates of recovery (restoration times) for their electric power system in the aftermath of Hurricane Maria (2017).

Contributing to understanding the relationship between electricity outages and socioeconomic vulnerability, this paper presents a comprehensive study with four important qualities: (a) it analyzes the significance of socioeconomic variables using the Random Forest machine learning technique to classify outage performance, (b) it integrates data from multiple historical hurricanes in a single analysis, (c) it exclusively uses public data, and (d) it uses wind gust data to characterize the extreme wind hazard (velocities and times), see II-C ERA5 reanalysis data set. Studying the outage distribution at the county level in Florida in the aftermath of hurricane hazards, this study seeks to understand the impact of socioeconomic variables on system performance that is represented by binary states of damaged system (1) or undamaged system (0), when comparing the “maximum power outages” with the outage threshold of 2% of the customers (see in the Figure 1 the performance threshold of 98%). In addition to socioeconomic variables, the model considers two groups of independent variables (see (1)) including hurricane wind-related variables (current and past maximum wind speed and recovery times for each county, section II-C) and other variables including the number of customers and county

978-1-6654-1775-4/22/$31.00 ©2022 IEEE
urban/rural classification (see section II-D).

\[
\text{system performance} \in \{\text{damaged, undamaged}\} = \text{function} (\text{socioeconomic variables, wind hazard, number of customers, urban/rural county})
\]

Given the system quality threshold of 98%, the performance of the system can be measured in terms of maximum percentage of outages. Figure 1 shows the system performance for the Gulf County during Hurricane Alberto (2018). In this case, the x-axis shows the dates and times of 13 different damage reports, and the y-axis shows the percentage of the system running. The maximum percentage of outages is reached when the system quality is at the lowest. For classification purposes using machine learning, the system performance can be considered to have two possible states, undamaged system (value of 0) for outages above the threshold, and damaged system (value of 1) for outages below the threshold.

II. DATA SET

We studied the effect of five hurricanes including Alberto (2018), Michael (2018), Dorian (2019), Eta (2020), and Isaias (2020) in the State of Florida by considering the outages and socioeconomic parameters at county level.

A. Outage Data

The Florida Public Service Commission - PSC provides public power outage reports in PDF format for important wind events related to Florida, since 2018 [16]. An automated data post-processing code is developed to generate a spatial database and extract the maximum number of outages for the hurricanes in the study. The original data source is summarized in Table I. The main characteristics of the reports, i.e. the number of reports, and the first and last time reported, vary notably across the hurricanes (see section II-E for the explanation of outage data integration of five hurricane outage in one database).

B. Socioeconomic Variables

Socioeconomic county indicators that are traditionally relevant in environmental justice frameworks and infrastructure disruption studies are considered here. These indicators relate to race/ethnicity, general population profile, population dependence, housing and households, educational attainment, income, and poverty. Table II shows detailed information about the socioeconomic variables used in this study, coming from the 5-year American Community Survey [17] for the previous year to each hurricane landfall. Therefore, ACS-5 2017, 2018, and 2019 are considered in this study.

C. Wind Hazard

The European Centre for Medium-Range Weather Forecasts (ECMWF) provides the data set ERA5, which among many climate variables (land, atmospheric, and oceanic), supplies wind gust \((fg10)\) velocities. ERA5 reanalysis data set reconstruct past weather and climate based on model information (past short-range weather forecast) combined with observations, i.e. data assimilation. The observations includes field measurements and satellite information from multiple sources and scales. The spatial resolution of ERA5 is a grid of 0.25 decimal degrees covering the entire world surface at different pressure levels, with hourly information from 1979 to the present. The variable \(fg10\) is available in the data set “ERA5 hourly data on single levels” and represents 3-seconds wind gusts \((V_g)\) in m/s, at 10 meters anemometer height, and under open space conditions [18]. For Hurricane Alberto, in Figure 2, the maximum \(V_g\) from ERA5 is compared against the best track information provided by the National Hurricane Center (NHC) - National Oceanic and Atmospheric Administration (NOAA)\(^1\). In order to compare, the intensity \(V_{60}\) in knots provided by NHC-NOAA was transformed to

---

**Table I**

| Hurricane | Reports | First Report | Last Report |
|----------|---------|--------------|-------------|
| Alberto  | 13      | 5/27/2018 9:01 | 5/29/2018 9:20 |
| Dorian   | 12      | 9/2/2019 18:51 | 9/4/2019 18:00 |
| Eta      | 15      | 11/8/2020 21:00 | 11/12/2020 18:00 |
| Isaias   | 6       | 8/17/2020 18:00 | 8/2/2020 18:00 |
| Michael  | 131     | 11/9/2020 19:00 | 11/8/2018 11:18 |

**Table II**

| Type            | Variable                  | Id. |
|-----------------|---------------------------|-----|
| Race/Ethnicity  | % population White        | 2   |
|                 | % Latino                  | 3   |
|                 | % African                 | 4   |
|                 | % Asian                   | 5   |
|                 | % Indian                  | 6   |
|                 | % Other                   | 7   |
| General Population Profile | Population          | 8   |
|                 | Population Density        | 9   |
| Dependence      | % population < 5 years    | 10  |
|                 | % population > 65 years   | 11  |
|                 | % population no vehicle   | 12  |
|                 | % population public assistance | 13  |
|                 | % population limited English | 14  |
|                 | % population disability   | 15  |
|                 | % population health insurance | 16  |
| Housing & Households | % renter occupied housing | 17  |
|                 | % income for renting > 30% | 18  |
|                 | % single > 65 years (rented) | 19  |
|                 | % single > 65 years (owned) | 20  |
| Educational Attainment | % < High School | 21  |
| Income & Poverty | Unemployment rate         | 22  |
|                 | % poverty level           | -23 |
|                 | % income for renting > 30% | 18  |
V₃ in m/s using [19]. From the hourly time history in the time frame of each hurricane outage report (see Table I), a spatiotemporal analysis was performed to extract wind hazard relate features including (a) the maximum V₃ velocity reported for the ERA5 pixels intersecting each county area (maximum in space) at the time of each outage report and (b) the past maximum V₃ velocity reported for the pixels time history intersecting each county area (maximum in space and in time). The maximum velocity in (a) represents the wind speed at the moment of each outage report, and the one in (b) represents the maximum wind speed faced by each county during the hurricane time frame previous to each outage report. The time in days from the event (b) to each outage report provides an additional feature that represents (c) the amount of time that has been available to each county to recover. For this analysis, these three features corresponding to the “maximum outage report” are considered.

D. Other Variables

Two additional independent variables are included in the model: (a) urban (1) / rural (0) county classification and (b) the number of power utility customers per county.

E. Database Integration

An integrated spatial database is created with all the feature information related to the five hurricanes’ “maximum outage report”, all the socioeconomic variables (22), the hazard features (3), urban/rural classification, and the number of customers, at the county scale. Table III shows a summary of the variables. The total number of columns is 28 (1 dependent - 27 independent). The Id column is used later in this document to identify (label) variables in the results graphs.

1NHC-NOAA maintains the Hurricane Databases (HURDAT) and provides best track files for multiple historical storms, including 6-hours interval estimations of 60-seconds (V₆₀) maximum sustained surface winds in knots. ([https://www.nhc.noaa.gov/gis/](https://www.nhc.noaa.gov/gis/))

### III. RANDOM FOREST

Random Forest, developed by Breiman [20], is an ensemble of many decision trees from different bootstrap samples (bagging), but with de-correlated trees. It can capture the general (linear or non-linear) structure of a data set (low bias), and it can be robust to outliers and noise. Handling binary, categorical, and numerical features without re-scaling or transforming, the implementation of Random Forest is simple, flexible, without exigent parameter tuning, and with low computational cost. It can also be used for regression, classification, or survival analysis in high dimensional data. For more details about the method with code examples, see Chapter 8 of [21].

#### A. Variable Importance

Node impurity is a measure of the uncertainty or diversity (non-homogeneity) of the labels at each tree node. When the labels inside a node belong to the same class the impurity is zero, and on the contrary, high impurity is related to mixed labels in a node. The most important feature with higher impurity reduction when splitting will be located as the root node (top node of the inverted tree) resulting in more pure children nodes (groups of majority class are isolated) compared with parent nodes. Gini index and Entropy are similar concepts directly related to the degree of inhomogeneity of the samples inside a node. The Gini index measures the probability of picking two distinct classes inside the node, being directly proportional to node impurity, and similarly, the Entropy is the expected self-information, a measure of uncertainty, or disorder of a sample set contained in a node. In Random Forest, the most important attributes can be selected, considering maximum information gain when splitting, a technique called Mean Decrease Impurity (MDI), but this approach also considers noisy features as important [22] and it is biased towards (a) categorical variables with multiple classes, (b) continuous variables, or (c) variables with high category frequencies [23]–[25]. A preferred method that overcomes [26] the deficiencies of MDI is permutation importance, also known as Mean Decrease of Accuracy (MDA), but it requires extra computing cost. With MDA an important variable has a considerable decrease (big difference) in the model accuracy comparing its Out-of-Bag (OOB) prediction error (model including original variable) with the new OOB average prediction error obtained from performing multiple random permutations (shuffles) of the variable content inside the model [20], and the process is repeated independently for each variable in the model. The mentioned difference is the variable permutation importance, therefore, MDA measures how the score decreases when the variable is not available. Neither MDI nor MDA reflect the predictive significance of the model.
a feature, but its relative importance for the specific model [27].

IV. EXPERIMENTS

From the 335 county records in the integrated database (see II-E), only 11% (37 samples) of the reported outage percentages correspond to “damaged” system performance status (using an outage threshold of 2%), and the remaining 89% (298 samples) correspond to “undamaged”. The implementation of Random Forest was performed 1000 times, each one with a different testing-training (30%-70%) random splitting (see section IV-A below), variable importance analysis, and optimized search for hyper-parameters through k-fold cross-validation, using HalvingRandomSearchCV Scikit-Learn (Machine Learning in Python) [27].

A. Training and testing data sets

In machine learning supervised classification, the training and testing sets are assumed to have the same distribution [28]; hence, the random partitioning in training (70%) and testing (30%), and the whole Random Forest analysis was performed 1000 thousand times to verify similar model results across different runs.

B. Tuning Hyper-parameters

The selection of an optimum set of hyper-parameters is based on increasing the performance of random search using successive halving [29], [30], based on cyclic evaluation of parameters using few system resources at the first try, then filtering top-scoring candidate across iterations, and finally, in the last iteration, using an exhaustive search in few optimal parameters.

C. Model performance

The relevant elements are those counties reporting a “damaged” (positive class) system, and after running Random Forest, the selected elements are counties classified in the same positive class.

Class-specific evaluation of the model can be analyzed using sensitivity, specificity, precision, recall, and f1-score. The trade-off between sensitivity and specificity, defined by a probability threshold of belonging to the positive or negative class, can be seen in The Receiver Operating Characteristics (ROC) and the corresponding Area Under the Curve (AUC). A good classification model for all the classes has an AUC close to one (1), the maximum value. For a detailed compendium of this evaluation metrics in Machine Learning, see Chapter 2 of [31].

V. RESULTS AND DISCUSSION

The optimum sets of hyper-parameters using random search successive halving for the mean model are (a) criterion: entropy, (b) maximum number of tree nodes: 10, (c) maximum number of features per tree: 4 out of 27, (d) number of trees in the forest: 10, and (e) the minimum number of samples to split a node: 7.

The minimum (worst) and the maximum (best) model accuracies across 1000 iterations are 78% and 98%, respectively. These different results (maximum difference of 20%), indicate a non similar probability distribution function (PDF) for all the random training and testing partitions, more so considering that the average model accuracy is 91%. Figure 3 shows that all the 1000 implemented models are different ranging from really bad performance, the min model with AUC of 0.594, to excellent performance with the best model having AUC of 0.995. Despite the average model having a good AUC value of 0.949, Figure 4 shows that this model cannot detect the “damaged” class (1) very well due to the low f1-score of 31% (high precision of 100% and low recall of 18%), but is highly trustable when it does. In Figure 4, the results confirm really good performance of the mean model for the “undamaged” class (0).

The results for variable importance are analyzed for the three representative models, namely the best (maximum), the worst (minimum), and the mean model, in addition to the average importance across the 1000 models. It should be noted that MDA addresses the MDI bias (see section III-A) by removing wrongly selected variables or reducing the importance of others. Therefore, the analyses here are focused on MDA results, whose importance in percentage are ordered in Table IV. Each variable in each model has importance in percentage (from 0% to 100%) represented by the bars height in the Figure 5, where the standard deviation is depicted as a thin vertical line, or by a number in Table IV, which in addition has the variable importance average (not only the representative models, but the 1000 repetitions). Note that the worst model (min) can have higher importance for the same variable, in comparison with the best model (max).

The analysis using MDA reveals that the wind hazard variables represent the 49% of the importance, including the past maximum velocity with an average score of 39.81%, the current maximum velocity with 5.42%, and the recovery time (days) each county had from the time of highest impact (stronger wind) to the time of maximum outages report, with
3.5%. The top-five socioeconomic variables, which represent the 48% of the model importance, are (a) percentage of Other population (Id = 7) with 6.31%, (b) percentage of the population with Limited English speaking ability (Id = 14) with 4.99%, (c) percentage of the Latino population (Id = 3) with 3.35%, (d) Unemployment Rate (Id = 22) with 2.91%, and (e) percentage of the Indian-American population (Id = 6) with 2.77%.

**CONCLUSION**

Socioeconomic variables are found to be considerably important for the outage performance of electrical power system in Florida against hurricanes. Results indicate that the social disparities are represented mainly by race/ethnicity (percentage of the Other, Latino, and Indian-American population), income and poverty (unemployment rate), and dependence (percentage of the population with limited English speaking ability). These relevant socioeconomic conditions should be considered when making resilience decisions for future wind hazards in the study area. Nonetheless, more studies need to be conducted to reach a sound conclusive remarks due to the fact that (a) the issue of unbalanced data set may cause variation between training and testing probability distributions across the repetitive experiments, and (b) additional data of hurricane events may need to be included in the database to increase the fidelity of the models.

**REFERENCES**

[1] N. L. Dehghani, A. B. Jeddidi, and A. Shafieezadeh, “Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning,” *Applied Energy*, vol. 285, p. 116355, 2021.

[2] Y. M. Darestani, K. Sanny, A. Shafieezadeh, and E. Fereshtehnejad, “Life cycle resilience quantification and enhancement of power distribution systems: A risk-based approach,” *Structural Safety*, vol. 90, p. 102075, 2021.

[3] A. M. Salman and Y. Li, “Multi-hazard risk assessment of electric power systems,” *Journal of Structural Engineering*, vol. 143, no. 3, p. 04016198, 2017.

[4] N. L. Dehghani and A. Shafieezadeh, “Multi-stage resilience management of smart power distribution systems: A stochastic robust optimization model,” *IEEE Transactions on Smart Grid*, 2022.

[5] M. O. Roman, E. C. Stokes, R. Shrestha, Z. Wang, L. Schultz, E. A. S. Carlo, Q. Sun, J. Bell, A. Molthan, V. Kalb, et al., “Satellite-based assessment of electricity restoration efforts in puerto rico after hurricane maria,” *PloS one*, vol. 14, no. 6, p. e0218883, 2019.

[6] E. Kabir, S. D. Guikema, and S. M. Quiring, “Predicting thunderstorm-induced power outages to support utility restoration,” *IEEE Transactions on Power Systems*, vol. 34, pp. 4370–4381, nov 2019.

[7] NEHRP, “Strategic plan. fiscal years 2009–2013,” tech. rep., National Earthquake Hazards Reduction Program (NEHRP), Oct. 2008.

[8] N. R. Council, *Disaster Resilience: A national imperative*, National Academies Press, nov 2012.

[9] N. I. for Standards and Technology, “Community resilience planning guide for buildings and infrastructure systems: A playbook,” tech. rep., NIST, oct 2020.

[10] A. T. Council, “Critical assessment of lifetime system performance: Understanding societal needs in disaster recovery,” tech. rep., U.S. Department of Commerce. National Institute of Standards and Technology, Engineering Laboratory, Gaithersburg, MD, apr 2016.

[11] A. R. Berkeley, M. Wallace, and C. Coo, “A framework for establishing critical infrastructure resilience goals,” *Final Report and Recommendations by the Council, National Infrastructure Advisory Council*, pp. 18–21, 2010.

[12] S. E. Chang, “Socioeconomic impacts of infrastructure disruptions,” oct 2016.

[13] D. Mitsova, A.-M. Esnard, A. Sapat, and B. S. Lai, “Socioeconomic vulnerability and electric power restoration timelines in florida: the case of hurricane irma,” *Natural Hazards*, vol. 94, pp. 689–709, aug 2018.

[14] D. Mitsova, M. Escaleras, A. Sapat, A.-M. Esnard, and A. Lamadrid, “The effects of infrastructure service disruptions and socio-economic vulnerability on hurricane recovery,” *Sustainability*, vol. 11, p. 516, jan 2019.

[15] M. Sotolongo, L. Kuhl, and S. H. Baker, “Using environmental justice to inform disaster recovery: Vulnerability and electricity restoration in puerto rico,” *Environmental Science & Policy*, vol. 122, pp. 59–71, aug 2021.

[16] F. P. S. C. (PSC), “Hurricane season power outage reports,” Web Page, May 2018–2022.

[17] U. S. C. Bureau, “American community survey (acs).” Online, 2017, 2018, 2019.

[18] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, et al., “The era'5 global reanalysis,” *Quarterly Journal of the Royal Meteorological Society*, vol. 146, no. 730, pp. 1999–2049, 2020.

[19] B. C. S. Durst, O.B.E. “Wind speeds over short periods of time,” *The Meteorological Magazine*, vol. 89, p. 181–187, July 1960.

[20] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[21] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[22] H. Le, Y. Wang, S. Basu, K. Kumbier, and B. Yu, “A debiased md1 feature importance measure for random forests,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, “Bias in statistical learning,” *Advances in Neural Information Processing Systems*, vol. 112, pp. 1–21, 2007.

[24] S. Nembrini, I. R. König, and M. N. Wright, “The revival of the gini importance,” *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[25] K. K. Niedermus, “On the stability and ranking of predictors from random forest variable importance measures,” *Briefings in bioinformatics*, vol. 12, no. 4, pp. 369–373, 2011.
[26] A. Ziegler and I. R. König, “Mining data with random forests: current options for real-world applications,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 4, no. 1, pp. 55–63, 2014.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., “Scikit-learn: Machine learning in python,” the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[28] T. Coleman, M. F. Dorn, K. Kaufeld, and L. Mentch, “Forecasting hurricane-related power outages via locally optimized random forests,” Stat, vol. 10, feb 2021.

[29] K. Jamieson and A. Talwalkar, “Non-stochastic best arm identification and hyperparameter optimization,” in Artificial intelligence and statistics, pp. 240–248, PMLR, 2016.

[30] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” The Journal of Machine Learning Research, vol. 18, no. 1, pp. 6765–6816, 2017.

[31] A. Zheng, Evaluating machine learning models: a beginner’s guide to key concepts and pitfalls. O’Reilly Media, 2015.