Environmental Research Communications

OPEN ACCESS

RECEIVED
2 March 2023

REVISED
31 May 2023

ACCEPTED FOR PUBLICATION
13 June 2023

PUBLISHED
23 June 2023

PAPER

Social media and volunteer rescue requests prediction with random forest and algorithm bias detection: a case of Hurricane Harvey

Volodymyr V Mihunov[1,*], **Kejin Wang**[1], **Zheye Wang**[2], **Nina S N Lam**[1] **and Mingxuan Sun**[3]

[1] Department of Environmental Sciences, Louisiana State University, 93 S Quad Dr Ste 1002, Baton Rouge, LA 70803, United States of America
[2] Rice University, Kinder Institute for Urban Research, Houston, TX, United States of America
[3] Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: vmihun1@lsu.edu

Keywords: AI fairness, random forest, regression fairness, explainable AI, disaster rescue

Abstract

AI fairness is tasked with evaluating and mitigating bias in algorithms that may discriminate towards protected groups. This paper examines if bias exists in AI algorithms used in disaster management and in what manner. We consider the 2017 Hurricane Harvey when flood victims in Houston resorted to social media to request for rescue. We evaluate a Random Forest regression model trained to predict Twitter rescue request rates from social-environmental data using three fairness criteria (independence, separation, and sufficiency). The Social Vulnerability Index (SVI), its four sub-indices, and four variables representing digital divide were considered sensitive attributes. The Random Forest regression model extracted seven significant predictors of rescue request rates, and from high to low importance they were percent of renter occupied housing units, percent of roads in flood zone, percent of flood zone area, percent of wetland cover, percent of herbaceous, forested and shrub cover, mean elevation, and percent of households with no computer or device. Partial Dependence plots of rescue request rates against each of the seven predictors show the non-linear nature of their relationships. Results of the fairness evaluation of the Random Forest model using the three criteria show no obvious biases for the nine sensitive attributes, except that a minor imperfect sufficiency was found with the SVI Housing and Transportation sub-index. Future AI modeling in disaster research could apply the same methodology used in this paper to evaluate fairness and help reduce unfair resource allocation and other social and geographical disparities.

1. Introduction

Machine learning (ML) as a subset of artificial intelligence (AI) is increasingly employed for decision making in many aspects of society. More ML algorithms have been developed and are made available for use. In disaster risk management, applications of ML are still limited. Existing applications in this field range from assessing hazard exposure, social vulnerability and risk assessment, to post-disaster damage estimation, prioritization of disaster aid distribution, and prediction modeling (Global Facility for Disaster Reduction and Recovery (GFDRR) 2018, Gevaert *et al* 2021). Given the complexity of disaster risk management and its urgent nature during disastrous events, more and better applications of ML algorithms in this field would be useful to emergency management and decision making.

However, despite the ability of AI to assist in decision making, there are concerns related to AI modeling, and among them are issues related to fairness. Fairness is best described as impartial treatment of data subjects by the AI (Fjeld *et al* 2020). Without considering fairness, AI may become biased at different stages of its life cycle. For example, historical, representation, and measurement biases occur during data generation, whereas learning, aggregation, evaluation, and deployment biases occur during model building and implementation (Suresh and Guttag 2021). Most existing fairness criteria fall under three general approaches: making the average predicted outcome similar for each sensitive group (*independence*), equalizing predicted outcomes given the ground truth outcomes (*separation*), and equalizing ground truth outcomes given the predicted outcome (*sufficiency*) (Barocas *et al* 2019).

This study investigates the fairness in the use of ML in predicting rescue requests from social media during emergencies. The increasing risk of floods prompts us to consider the case when stranded people request rescue through social media, like it happened during Hurricane Harvey (Mihunov *et al* 2020). Hurricane Harvey hit the Texas coast near Corpus Christi with category 4 wind speeds on 26 August 2017. It quickly dissipated and changed direction, unexpectedly delivering over 1.5 m (5 ft) of rainfall to the Houston metropolitan region (Watson *et al* 2018), causing unprecedented flooding and leaving many people stranded. Calls to the 911 were overloaded and could not be connected. As a result, many residents resorted to social media to seek rescue from flooded waters (Wang *et al* 2022, Zhou *et al* 2022).

Hurricane Harvey events left a digital footprint that prompted extensive research by disaster management scholars (Zou *et al* 2018, Mihunov *et al* 2022). Availability of social media data and growing accessibility of ML methods make it likely that black-box models predicting locations of rescue request will emerge and be adopted for real-world applications. If fairness of those models is not considered, potential biases may cause real-world harm, such as unfair allocation of rescue teams' resources and inadequate assistance to vulnerable communities.

The objective of this study is to examine if bias exists in AI algorithms used in disaster management and in what manner. More specifically, we use rescue request data derived from both Twitter and volunteer-collected data to examine how AI fairness could impact the prediction of rescue requests. We focus on two sources of potential biases—social vulnerability and digital divide. Our research questions are:

- RQ1: what are the significant predictors of the rescue request pattern using ML algorithms, and how do they influence model predictions?
- RQ2: does the model learnt from the social media rescue data exhibit unfairness based on social vulnerability and digital access?

We hypothesize that high social vulnerability and limited access to digital technology are associated with low predicted social media rescue rates, and fairness criteria of *independence*, *separation*, or *sufficiency* are violated based on the two characteristics.

Specifically, we selected the Social Vulnerability Index (SVI) and indices of its four themes (socioeconomic status, household composition and disability, minority and language, transportation and housing) as sensitive attributes because they encompass many characteristics of the communities that may have higher disaster assistance needs and may experience discrimination (Flanagan *et al* 2011). In addition, four digital access variables were assigned as sensitive attributes to evaluate potential algorithm bias introduced due to the digital divide. By addressing the two research questions, this study could offer new insights into the use of AI modeling with social media data in future rescue operations, as well as provide baseline information for the development and applications of fair AI methods for disaster resilience.

2. Background

Many decisions impacting individuals and communities are assisted or made by algorithms, including hiring, lending, policing, criminal justice, and stock trading, among others (Lepri *et al* 2018, Shang *et al* 2020). Despite the promise of eliminating the limitations and biases of human decision making, data-driven algorithmic systems retain some of the same ethical concerns, among which are fairness, transparency, and explainability (Fjeld *et al* 2020, Gevaert *et al* 2021, Lepri *et al* 2018). Transparency refers to the communication of the internal functioning of the model, whereas explainability is the ability to interpret how the model makes its predictions (Lepri *et al* 2018, Mittelstadt *et al* 2019). Explainability and transparency are both ways of validating ML models while attaining new domain knowledge about the modeled system (Mittelstadt *et al* 2019).

Depending on the application, an AI algorithm's impact on its subjects may be different, including potential harms. Among the most common types of harms are *allocation harms*, which happen when the opportunities, resources, or information are withheld through the AI decisions, and *quality-of-service harms,* which happen when a system does not work equally well for different users (Bird *et al* 2020). For instance, unfair hiring, lending, or underwriting decisions are examples of *allocation harms*, whereas poor quality of facial or speech recognition or incorrect health care decisions due to someone's race or gender are examples of *quality-of-service harms*.

The ML literature has proposed and refined dozens of fairness criteria (Barocas *et al* 2019, Chouldechova 2017, Steinberg *et al* 2020). These criteria aim at identifying discrimination of groups of individuals based on a set of defining characteristics, called sensitive features or protected attributes, for which some groups are considered privileged, and some are not. Consider a simple case with a binary classifier where A is the sensitive attribute, Y is the target variable (i.e., ground truth or dependent variable), $\hat{Y}$ is the predicted value, and score R is the probability of the predicted value. The three commonly used fairness criteria are *independence*, $R \perp A$; *separation*, $R \perp A|Y$; and *sufficiency* $Y \perp A|R$ (Barocas *et al* 2019).

The first criterion, *independence*, requires that R is independent of A. It has many equivalent and related definitions, such as *demographic parity*, *statistical parity*, *group fairness,* and others (Barocas *et al* 2019). It is satisfied when $\mathbb{P}\{\hat{Y} = 1|A = a\} = \mathbb{P}\{\hat{Y} = 1|A = b\}$. Consider $\hat{Y} = 1$ as 'acceptance', *independence* thus requires equal acceptance rate in both groups a, b of the sensitive attribute A.

The second criterion, *separation*, is used when the protected groups are not equally 'qualified' for the same outcome (Dwork *et al* 2012). A certain demographic group $A = a$ may have a higher rate of true event $Y = 1$, which may justify higher acceptance rate from the group a. This condition is called *separation.* It is expressed as $R \perp A|Y$, where R is independent of A given Y. In binary classification, *separation* requires $\mathbb{P}\{\hat{Y} = 1|Y = 1, A = a\}$ $=\mathbb{P}\{\hat{Y} = 1|Y = 1, A = b\}$ and $\mathbb{P}\{\hat{Y} = 1|Y = 0, A = a\}$ $=\mathbb{P}\{\hat{Y} = 1|Y = 0, A = b\}$. Since $1 - \mathbb{P}\{\hat{Y} = 1|Y = 1\}$ is the *false negative rate* and $\mathbb{P}\{\hat{Y} = 1|Y = 0\}$ is the *false positive rate* of a classifier, the first statement stipulates that groups a, b have the same *false negative rate*, whereas the second one equalizes their *false positive rates.* Combining the two constraints, *separation* calls for *error rate parity*.

The third criterion, *sufficiency*, is $Y \perp A|R$, or Y is independent of A given R. This corresponds to $\mathbb{P}\{Y = 1|R = r, A = a\}$ $=\mathbb{P}\{Y = 1|R = r, A = b\}$, which means that each group of the sensitive attribute A has the same accuracy of the predicted probabilities (Barocas *et al* 2019). A related model property, *calibration* is used to assess how close modeled probabilities are to the probabilities of true events in Y (Zadrozny and Elkan 2002, Niculescu-Mizil and Caruana 2005). Predicted probability scores R are *calibrated* with respect to the ground truth Y, when for all predicted probability values $r \in [0, 1]$, $\mathbb{P}\{Y = 1|R = r\} = r$. Accordingly, for any value r, a prediction of a class with probability r is correct in r fraction of cases. Calibrated probabilities can be directly interpreted as the model's confidence in each prediction, which is considered critical information, for example, in deciding a treatment plan or crime sentencing. Unfairness arises when different confidence thresholds are applied to different demographic groups for assigning them to a prediction class (Chouldechova 2017).

Popular metrics such as *independence*, *separation*, and *sufficiency* have been developed for evaluating classification fairness. However, few studies have used these metrics for evaluating regression fairness (Fitzsimons *et al* 2019, Steinberg *et al* 2020). It has been suggested that once model fairness is evaluated, biases can be mitigated by targeting the data with adjustment or weighting, applying regularization or penalty to the model, or post-processing of the predictions (Suresh and Guttag 2021).

Research into AI fairness is limited in disaster risk management and geospatial sciences (Gevaert *et al* 2021). One of the key fairness problems in disaster risk management is the uneven access to digital technology, resulting in disparate representation of vulnerable groups in the emerging geospatial big data (such as mobility and social media data). For example, timely estimates of population change before and after a disaster can be derived using mobile phone call records, but they are biased towards those without phone access (Yu *et al* 2018, Pestre *et al* 2020). Similarly, models and predictions that use social media data may be biased, due to their inability to control the representativeness of such data (Zou *et al* 2018, Wang *et al* 2019).

Furthermore, AI predictions are rarely explained or tested for bias. For example, the study by Behl *et al* (2021) is among the few explainable AI applications in disaster literature. Their study obtains Local Interpretable Model-Agnostic Explanations (LIME) from the black box algorithms trained to process Twitter data to identify people's needs after a disaster, and to show which words contained in the tweets contributed to tweet classifications. This study will address the lack of research into AI fairness and explainability in disaster risk management and geospatial sciences and will be among the first to apply explainable and fair AI principles to predict social-media derived rescue requests using social-environmental data.

3. Data collection and feature extraction

Procedures of data collection and analyses used in this study are depicted in figure 1.

3.1. Social media and volunteer-collected rescue requests

Hurricane Harvey tweets posted between August 17 and September 7, 2017 were purchased from Twitter using keywords 'hurricane, harvey, disaster, cajun navy, hurricaneharvey, txdps, txtf1, redcross, coastguard, coastguard, houstonpolice, houstonoem, salvationarmy, flood, sos, flooding, storm, rescue, sendhelp, cajunnavy, fema, salvation army'. This led to a dataset of close to 45 million tweets. We searched for rescue request tweets posted between August 27 and August 31 which yielded 4.1 million tweets (retweets excluded). Rescue requests are defined as rescue seeking messages with geographic addresses (Wang *et al* 2022). We selected

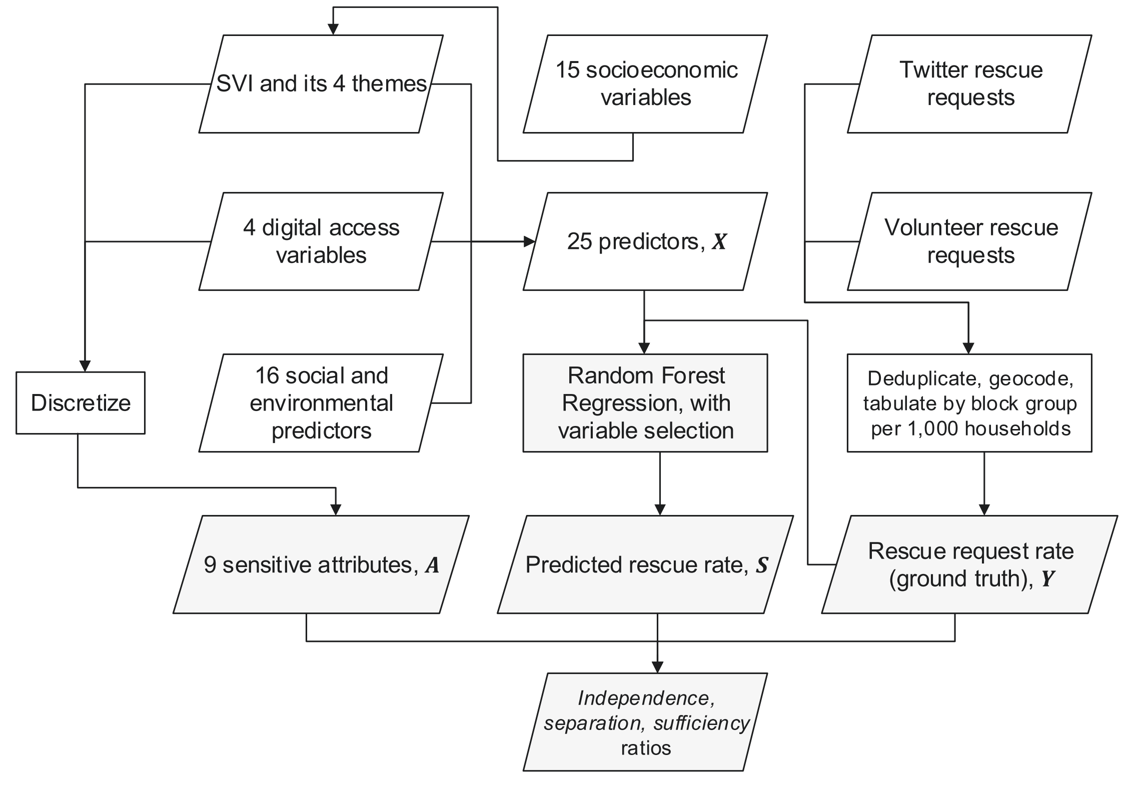
SVI and its 4 themes
15 socioeconomic variables
Twitter rescue requests
4 digital access variables
25 predictors, $\boldsymbol{X}$
Volunteer rescue requests
Discretize
16 social and environmental predictors
Random Forest Regression, with variable selection
Deduplicate, geocode, tabulate by block group per 1,000 households
9 sensitive attributes, $\boldsymbol{A}$
Predicted rescue rate, $\boldsymbol{S}$
Rescue request rate (ground truth), $\boldsymbol{Y}$
Independence, separation, sufficiency ratios

Figure 1. Flowchart of data and procedures used in modeling and fairness measurement of social media rescue requests.

Table 1. Social media rescue requests: verified and deduplicated by address from 2 datasets.

Twitter	Collected by volunteers	Matched between the two datasets	Total unique requests
962	1,229	383	1,808

tweets containing any ZIP code in the Houston Metropolitan Statistical Area (MSA) and then manually removed the tweets that did not contain rescue requests. Through this step, 2,105 tweets (rescue requests) were obtained. The addresses in the tweets were then geoparsed using Google Geocoding API, such that the coordinates were assigned based on the addresses mentioned in the body of the tweets. We then manually corrected spelling errors in the few unassigned tweets and matched them with the correct coordinates, and this resulted in 962 tweets (table 1).

Additionally, we acquired a dataset with rescue requests collected by volunteers from social media or through several online forms distributed at the time. The document includes a Google Spreadsheet containing the addresses and related information of the rescue requests, as well as information on the prioritization and fulfillment of the requests (Arrazolo 2017). All addresses in the document were identified and de-duplicated. They were then geocoded, and matching addresses were identified between the Twitter and the volunteer datasets (table 1). Combining the two datasets led to 1,808 unique rescue request tweets. The geocoded rescue requests were then tabulated by block group and normalized per 1,000 households, resulting in 535 out of 3,017 block groups having rescue requests. One of the block groups that had one Twitter rescue request was removed due to a missing value of the number of households, thus 534 block groups and 1,807 rescue requests were considered in subsequent analysis.

3.2. Social vulnerability index (SVI) data collection and calculation

We used the Social Vulnerability Index (SVI), which was designed and implemented by the Centers for Disease Control and Prevention (CDC), to characterize communities that are at a greater risk during disasters (Flanagan *et al* 2011). The index is a composite of a total of 15 variables and include four sub-indices: (1) socioeconomic status, (2) household composition and disability, (3) minority status and language, and (4) housing and

Table 2. Social vulnerability index variables ($n = 3,017$, all block groups in Houston MSA)[a]

SVI Theme	Variable	Min	Max	Mean	St. D.
1. Socioeconomic status	% below poverty	0	0.80	0.17	0.15
	Unemployment rate	0	0.22	0.04	0.03
	Per capita income (U.S. Dollars)	1,038	207,879	30,364	21,570
	% no high school diploma	0	0.38	0.02	0.03
2. Household composition and disability	% over 65 y.o.	0	0.65	0.11	0.07
	% under 17 y.o.	0	0.58	0.25	0.09
	% with disability	0	0.53	0.10	0.06
	% single parent households	0	1	0.32	0.25
3. Minority status & language	% with minority status	0	1	0.33	0.24
	% speaking English 'less than well'	0	1	0.10	0.12
4. Housing type & transportation	% housing with more than 10 units	0	1	0.18	0.26
	% mobile home units	0	0.89	0.05	0.12
	% households w/more people than rooms	0	1	0.06	0.08
	% households with no vehicle	0	0.68	0.05	0.08
	% living in group quarters	0	0.99	0.01	0.06

[a] Percentage values are listed as proportions.

transportation (table 2). Its early applications included hazard mitigation and planning research (Flanagan *et al* 2018). More recent studies evaluated its associations with cardiovascular disease and surgical outcomes, house fire events, health outcomes related to heat events, Hurricanes Katrina, Sandy, and Harvey, as well as health behaviors and outcomes during the COVID-19 pandemic (Flanagan *et al* 2018, Li *et al* 2022, Ramesh *et al* 2022). Moreover, Wang *et al* (2019) used SVI to evaluate whether social disparities existed in social media use leading up to Hurricane Sandy.

Justifications for SVI variables' inclusion in each vulnerability theme are similar to our rationale of using them as characteristics that may make vulnerable communities subject to algorithm bias in a rescue allocation model (Morrow 1999, Cutter *et al* 2003, Flanagan *et al* 2011). For example, income, poverty, employment, and education variables are related to presumed lesser resources for evacuation, preparation, and recovery, and being served by less robust public infrastructure. Second, age, single parenting, and disability variables capture the presence of people that may require special planning and accommodations and additional resources. Third, race, ethnicity, and English proficiency variables describe groups that are often underserved due to racial discrimination and language barrier. Fourth, housing type, crowding, and vehicle availability variables represent lack of access to secure infrastructure. Finally, the composite SVI represents broad vulnerability of the community. In this study, we assume that vulnerable communities will have more people who need rescue (Wang *et al* 2019). Testing algorithm fairness using SVI and its four sub-indices allows us to identify potential bias towards protected groups, who are considered vulnerable in disasters.

We calculated the SVI for the 3,017 Houston MSA block groups by summing the percentile rank of each of the 15 variables from highest to lowest (except per capita income, which is ranked from lowest to highest). A percentile rank is a proportion of scores in a distribution that a specific score is greater than or equal to:

$$Percentile\ Rank = (Rank - 1)/(N - 1) \tag{1}$$

where N is the total number of data points. A percentile rank of the sum is then estimated. The result is a score with a range $[0, 1]$, and its higher values correspond to higher vulnerability. The 15 variables are further grouped by the four themes, for which separate scores are calculated.

Most variables used in the SVI calculation were accessed from the American Community Survey (ACS) 2015 5-year estimates, except the data on institutionalized and non-institutionalized group quarters' residents, which were from the 2010 Decennial Census. The overall SVI and its four sub-indices are included in a training dataset and used as sensitive attributes (table 3).

3.3. Access to digital devices and internet services

The four digital access variables (percent of households with no computer or device, percent of households with no Internet, percent of households with cellular data plan, and percent of households with smartphone) were acquired from the ACS 5-year 2017 estimates (table 3). We employ these data to test whether they are significant predictors of social media rescue requests and whether the predictive model may exhibit bias due to digital divide. The groups with low digital adoption rates, such as older adults, are at a disadvantage during a major disaster, when critical information is disseminated, and relief efforts are coordinated on social media (Dargin *et al* 2021, Choi *et al* 2022). Thus, we assess AI fairness based on these characteristics.

Table 3. The 25 social-environmental predictors ($n = 534$, block groups with rescue requests)[a]

Category	Variable	Min	Max	Mean	St D
Target	Rescue requests per 1000 households	0.17	250.00	5.57	13.83
Land cover (percent of block group area)	% Agricultural land area	0	0.88	0.06	0.14
	% Wetland area	0	0.50	0.04	0.08
	% Water area	0	0.68	0.01	0.04
	% Herbaceous, forested, and shrub area	0	0.57	0.07	0.12
	% Developed land area	0.02	1	0.82	0.25
Environment	Road length per block group area (m/m^2)	0.001	0.09	0.02	0.01
	% Road length in flood zone	0	1	0.24	0.29
	Mean elevation (10 meters)	1.03	65.07	20.06	10.67
	% Area in Flood zone	0	1	0.32	0.32
Other socioeconomic characteristics	% Renter occupied housing	0	1	0.41	0.30
	% Civilian labor force	0.26	0.95	0.67	0.10
	% Households with income < $40,000	0	1	0.36	0.22
	% With no college degree	0.05	1	0.71	0.24
	Total rooms in housing units per household	3.52	10.59	6.34	1.30
	% Over 16, no vehicle to commute to work	0	0.84	0.21	0.12
	% With no health insurance	0	0.67	0.22	0.14
Access to digital devices and Internet services[b]	% Households with no computer or device	0	0.74	0.13	0.13
	% Households with no Internet	0	0.77	0.19	0.16
	% Households with cellular data plan	0	1	0.53	0.16
	% Households with smartphone	0.18	1	0.74	0.15
CDC Social Vulnerability Index[b]	SVI Theme 1 Socioeconomic Status	0	1	0.51	0.29
	SVI Theme 2 Household comp. & Disability	0.005	0.999	0.49	0.29
	SVI Theme 3 Minority Status & Language	0	0.999	0.55	0.27
	SVI Theme 4 Housing and Transportation	0	1	0.50	0.30
	SVI	0	0.999	0.51	0.29

[a] Percentage values are listed as proportions.

[b] Also used as sensitive attributes in discretized form.

3.4. Social-environmental characteristics

Additional variables were collected to examine if they can help predict rescue requests during hurricane flooding. For the environmental variables, we collected the 2016 land cover data from the U.S. Geological Survey (USGS) and Multi-Resolution Land Characteristics Consortium (MRLC) National Land Cover Database (NLCD) (Yang *et al* 2018). We then calculated the land cover type area percentage by block group using the Tabulate Area tool from Zonal toolset in ArcGIS Desktop (Esri 2021). Similarly, mean elevation was tabulated from the National Elevation Dataset (NED) Digital Elevation Model (DEM) in one arc-second (approximately 30 m) resolution (U.S. Geological Survey (USGS) 2017). The percentage of flood zone area per block group was tabulated from Federal Emergency Management Agency (FEMA) National Flood Hazard Layer (NFHL) (FEMA 2020). Road length per block group area was tabulated from the U.S. Census (2016) TIGER/Line 'All Roads' shapefiles, and percent of road length in flood zone was tabulated by overlaying roads shapefiles with flood zone polygons. Finally, data for the seven socioeconomic variables, which have been used frequently in previous vulnerability and resilience studies, were acquired from the ACS 2015 5-year estimates (table 3) (Cai *et al* 2018, Lam *et al* 2016, Wang *et al* 2023).

This data collection resulted in a dataset with socioeconomic and environmental predictors, as well as SVI indices for all 3,017 block groups of the Houston MSA. The target variable represented by the rescue request rate had values in 534 block groups. We thus made a subset of the 534 block groups containing rescue requests for training the model (531 after removing the three cases with missing values in the independent variables).

For the fairness testing, we chose nine variables (the SVI, its sub-indices and digital access variables) as sensitive attributes, hypothesizing that they were the likely barriers to access social media to request for rescue. Since fairness tests are concerned with binary sensitive attributes, SVI and its themes were discretized by splitting the cases at the 0.25 quantile and assigning the 0–0.25 group as privileged (low vulnerability). For the digital access variables, the block groups were also split at the 0.25 quantile with the high adoption group assigned as privileged.

3.5. Comparison data: Boston housing value dataset

We used an independent dataset to provide a reference point for our algorithm fairness testing. This dataset was Boston Housing Value from Harrison and Rubinfeld (1978), which was corrected for a few minor errors and augmented with census tract-level coordinates (Pace and Gilley 1997; accessible from the statlib index). The dataset consists of 506 cases and 14 variables, one of which is the target variable 'Median value of owneroccupied homes in \$1000's' ($MEDV$), and the remaining 13 variables are used as predictors. Among them are B and $LSTAT$. B is the value of $1000(Bk-0.63)^2$, where Bk is the proportion of black people, and $LSTAT$ is percent of the population with 'lower status' (adults without some high school education and proportion of male workers classified as laborers). We considered B and $LSTAT$ as sensitive attributes, because housing values are likely unfair on the basis of race and socioeconomic status. The remaining variables are related to environment and infrastructure. For the B variable, we split the set at $B=360$, with values above 360 being the privileged group, as they reflect zero to low proportion of black people per census tract. For $LSTAT$ we used the dataset mean ($LSTAT=12.65$) as the split, with the values below the mean being the privileged group.

4. Analysis methods

4.1. Random Forest Regression and variable selection

We used the *ranger* package in R for variable selection and Random Forest model fitting (Breiman 2001; Wright and Ziegler 2017). Random Forest is a supervised ML method that creates many decision trees (i.e., an ensemble or a forest) with their individual predictions being used in a voting scheme to make final predictions (Esri 2022). A random subset of the training data and a random subset of explanatory variables are used in each tree and then the votes from all decision trees are considered. This way the overfitting associated with individual decision trees is addressed. The mean square error and R^2 are obtained from the out-of-bag data (data excluded in the training step) to evaluate the model performance.

We adopted the modeling steps recommended in Altmann *et al* (2010) to fit Random Forest with relevant features. First, Random Forest is fit with all 25 potential predictors (table 3), and corrected impurity importance for each predictor is estimated (Nembrini *et al* 2018). The impurity importance can be used to rank predictors using Gini coefficients, and it reflects the number of times a predictor variable is responsible for a split (individual decision within a tree) and the impact of that split per the number of trees. Corrected impurity importance is modified to be compatible with further statistical testing for importance significance.

Then, permutation importance as in Altmann *et al* (2010) is computed by permuting the response vector (target variable) several times and producing a vector of importance values for each predictor across all permutations, from which p-values are retrieved. Predictors with significant p-values are then retained in a more parsimonious model. We used 100 permutations, as a larger number of permutations did not change the order of returned p-values. Finally, the Random Forest model with a traditional impurity measure and significant predictors selected in the previous step was used for model explanations and fairness evaluations.

4.2. Model diagnostics

4.2.1. Dataset level exploration

To address RQ1, which is to explain the predictors' influence on model predictions, we employed a dataset-level exploration. We utilized Partial Dependence (PD) plots that show the expected values of model predictions as a function of a selected explanatory variable while holding other variables constant (Goldstein *et al* 2015). They are derived by averaging many (or possibly all) instances of ceteris-paribus (CP) profiles, which are plots that show the change of a predicted value induced by a change of a single explanatory variable for a single observation (Biecek 2018).

4.2.2. Regression fairness

To address RQ2, which is to examine whether the Random Forest model exhibits bias, we utilized the fairness criteria of *independence*, *separation*, and *sufficiency*, adapted for regression by Steinberg *et al* (2020) and implemented in the R package *fairmodels* (Wiśniewski and Biecek 2022).

Referring to definitions in section 2, for the regression case we denote Y as the continuous dependent variable (ground truth) and S as the continuous predicted values (in place of binary Y and predicted class $\hat{Y}$). To define a continuous measure of the degree to which the fairness criteria are satisfied, Steinberg *et al* (2020) transforms them as ratios:

$$ratio_{ind} = \frac{P(S\mid A=1)}{P(S\ \mid A=0)},$$

$$ratio_{sep} = \frac{P(S\mid Y,\ A=1)}{P(S\mid Y,\ A=0)},$$

$$ratio_{suff} = \frac{P(Y\mid S,\ A=1)}{P(Y\mid S,\ A=0)}$$

The model that satisfies a given criterion will have its ratio close to 1. Steinberg *et al* (2020) then transforms each ratio using Bayes' Theorem so that they can be estimated using the density ratios. The density ratios are obtained in expectation, by approximating them as an empirical average over the data with the outputs of the probabilistic classifiers $\rho(a|s)$, $\rho(a|y,\ s)$, and $\rho(a|y)$. Using $\rho(a|s)$, $ratio_{ind}$ is determined by how much more

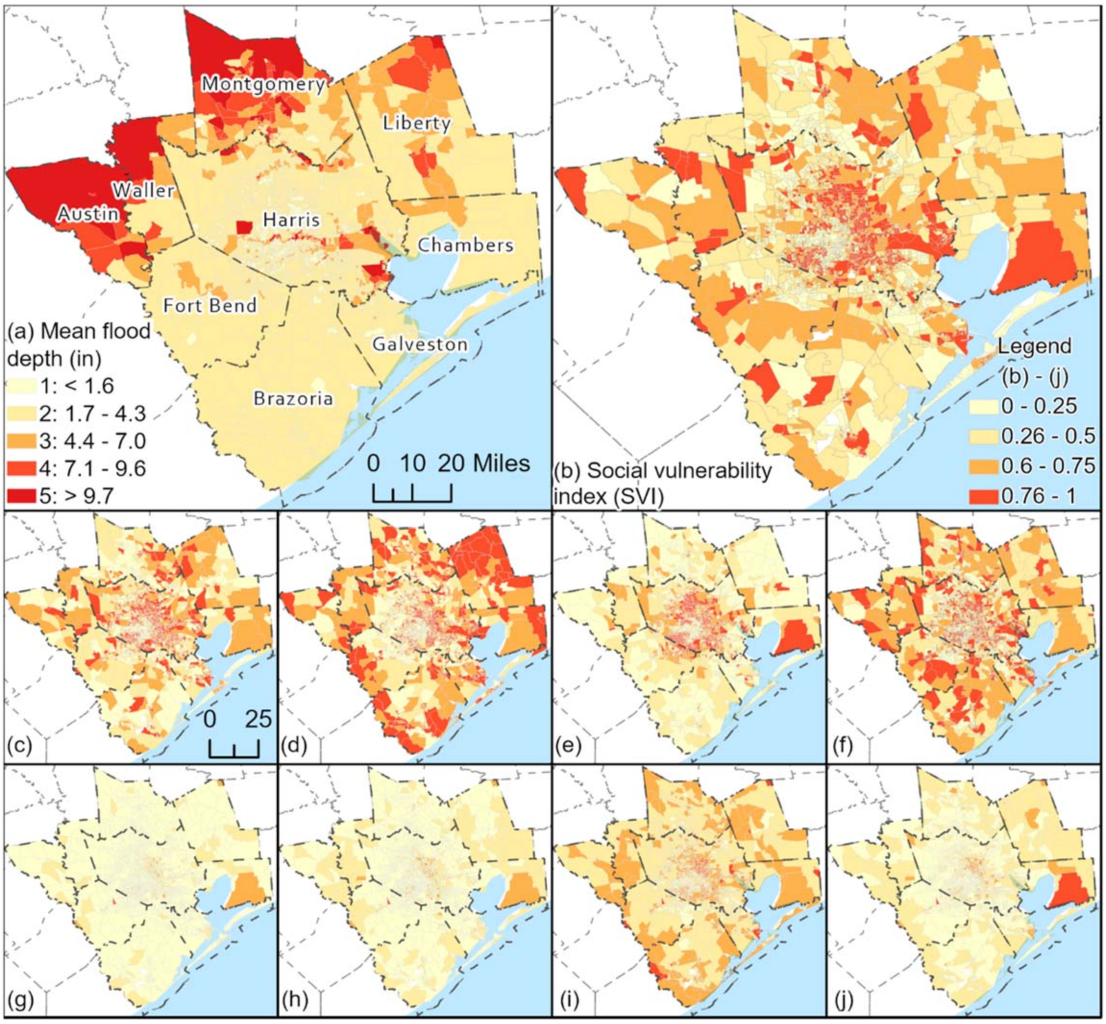

Figure 2. Study area, its physical (a) and social (b) vulnerability and sensitive attributes by block group: (c) SVI Theme 1 Socioeconomic Status, (d) SVI Theme 2 Household composition & Disability, (e) SVI Theme 3 Minority Status & Language, (f) SVI Theme 4 Housing and Transportation, (g) % households with no computer or device, (h) % households with no Internet, (i)% households with cellular data plan (reversed), (j) % households with smartphone (reversed). Darker shade is higher vulnerability.

predictive S is of A, over the base rate distribution of A. The *separation* ratio uses $\rho(a|y, s)$ and $\rho(a|y)$ to find the additional predictive power that the joint distribution of Y and S has in predicting A over the marginal distribution of Y. Similarly, *sufficiency* ratio is derived from $\rho(a|y, s)$ and $\rho(a|s)$ to show how much more predictive the joint distribution of Y and S is of A over the marginal distribution of S.

5. Results

5.1. Spatial patterns of rescue requests, vulnerability, and digital access

We mapped the spatial patterns of key variables to provide exploratory analysis of their real-world representation and aid the analysis of model fairness (figure 2).

Figure 2(a) depicts the mean flood depth per block group during Hurricane Harvey. These flood depths were measured by USGS using the high-water marks method and were tabulated from raster data into block groups by the authors using the ArcGIS Zonal Statistics tool (FEMA 2018; Esri 2021). In most cases the pattern of flood depth coincides with the location of major rivers and water bodies in the study area (figure 3). On the contrary, the pattern of social vulnerability represented by SVI (figure 2(b)) is scattered across the urban Harris County (the city of Houston, TX) and the neighboring suburban counties.

In terms of the individual SVI themes, the socioeconomic theme (figure 2(c)) appears very similar to the overall SVI (figure 2(b)), whereas the household composition and disability theme (figure 2(d)) shows many more vulnerable block groups located farther from the urban core, potentially due to many families caring for children and older or disabled adults, settling farther from the urban areas. The minority status and language

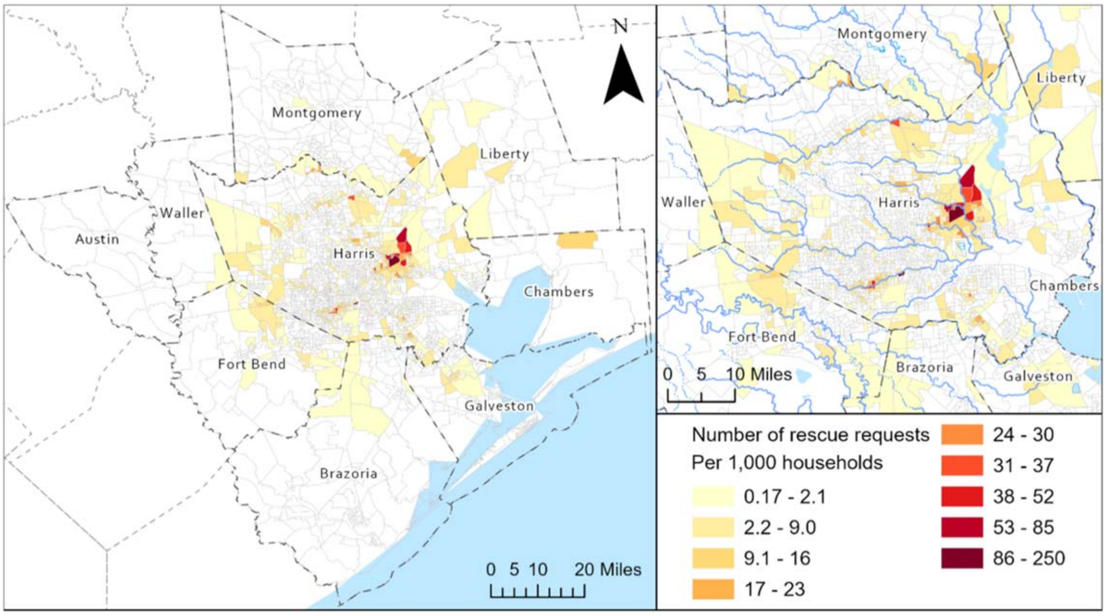

Figure 3. Combined social media and volunteer rescue requests per 1,000 households.

barrier theme points to most of the vulnerable populations concentrated within Harris County (figure 2(e)). Finally, the housing type and transportation theme shows an evenly scattered pattern across the study area (figure 2(f)).

As for the digital access variables (figures 2(g)–(j)), lack of access to cellular data plan is overall higher than the values of the other three variables (figure 2(i)). No cellular data connection may become a barrier for social media access during a major disaster when other means of communication go offline. Other digital disparities, like lack of access to Internet, computer, or a device, and specifically a smartphone, are less pronounced. The patterns of the four variables are similar, with minor differences in intensity. We note that the patterns of digital access resemble the pattern of SVI Theme 2 Household composition and disability (figure 2(d)), which could suggest a lesser adoption of technology by vulnerable age and disability groups.

Figure 3 shows the spatial pattern of the target variable, which is combined social media and volunteer-collected rescue requests per 1,000 households in each block group. Despite most of the study area has been flooded, we found rescue requests through social media in only 534 out of 3,017 block groups.

The map shows that rescue requests were mostly from Harris County with high concentration in its central part (figure 3). Adjacent counties to the north of Harris County (Liberty, Montgomery, and Waller) received high level of flooding but with few social media rescue requests (figure 3). Overall, we do not observe a striking mismatch between the rescue request locations (figure 3) and social vulnerability or access to digital technology (figurex 2(b)–(g)). We found that many of the rescue locations were from socially vulnerable block groups with limited access to smartphones or internet. This is likely due to many of the requests being posted on behalf of elderly or disabled family members. However, we also found that many socially vulnerable block groups had very few or no rescue requests. We conducted further modeling and fairness measurement to test whether this mismatch will lead to AI bias, when predicting rescue request rates.

5.2. Random forest model of rescue requests

The results of the Random Forest model fitting were as follows. First, the model with all 25 variables (table 3) was fit and p-values as a measure of variable importance were estimated with 100 permutations, resulting in a model with 70.28 MSE (mean squared error) and 0.11 out-of-bag R^2. Next, the model was fit with the seven variables that had p-value < 0.05 (table 4). This led to an improvement in model performance with 64.47 MSE and 0.19 out-of-bag R^2. Table 4 lists the predictor variables in the order from highest to lowest importance, along with their importance scores and numbers of block groups excluded from the test set due to their corresponding values being out of the training set's range. The values in the Importance column are the sum of the Gini coefficients from all the trees for each variable listed.

Five out of the seven significant predictors were features of the physical environment, such as percent of roads in flood zone, percent of flood zone area, mean elevation, and percent of forested and wetland areas. Only two selected features were related to the social environment, and they were percent of renter occupied housing

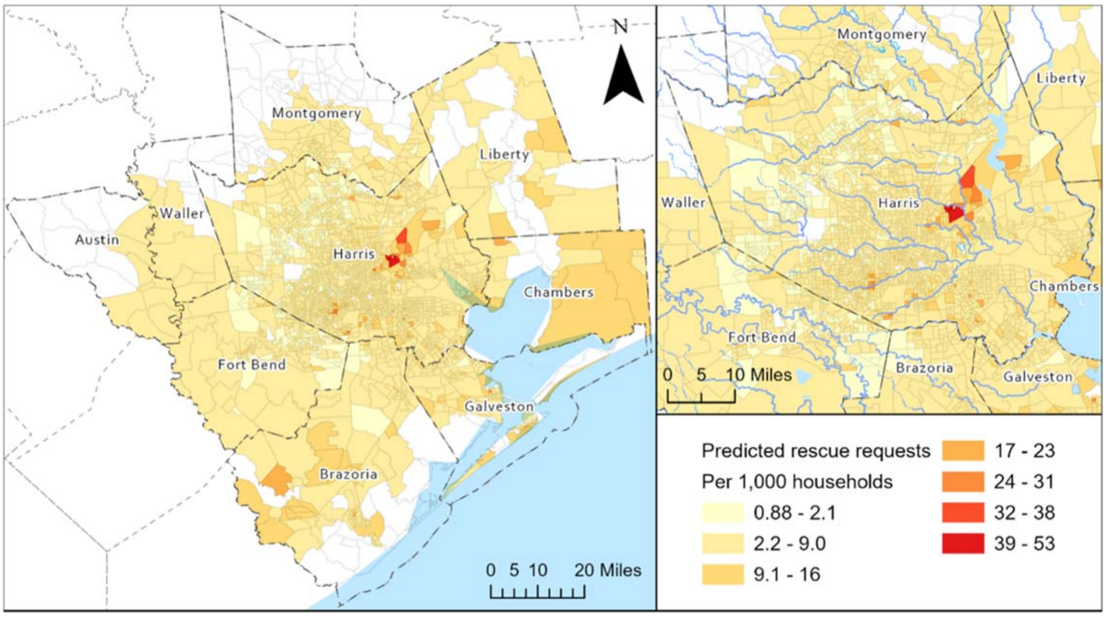

Figure 4. Predicted rescue request values (per 1,000 households), excluding 137 out of 3,017 cases for which predictors' values were out-of-range of the training data.

Table 4. Random forest variable importance and number of predictors' cases out of range.

Variable	Importance	< Min	> Max
% of renter occupied housing	6,987.5	0	0
% of roads in flood zone	6,298.8	0	0
% of flood zone area	5,292.3	0	0
% of wetland area	3,715.5	0	37
% of herbaceous, forested, and shrub area	5,139.7	0	29
Mean Elevation	5,288.4	5	70
% of households with no computer or device	4,863.4	0	3

units and percent of households with no computer or device. Overall, the selected variables are reasonable predictors of rescue needs. The predicted values of rescue requests produced by the final model are mapped in figure 4.

5.3. Rescue request model explorations

To better understand the pattern of individual variables' contributions to the model predictions, we conducted dataset-level explorations using the partial dependence (PD) plots (figure 5). These plots show that the contributions of the predictors are not monotonically increasing or decreasing. For example, as the percent of herbaceous and forested area increases, the predicted rescue request rate increases until it reaches the level of 40% where the rate of predictions' growth flattens. This shows that herbaceous cover of the block group contributes to the rescue request rate, likely due to the lack of roads and inability to pass flooded forested areas, which would leave people stranded.

We observe a similar but less steep upward growth in predicted rescue rate with respect to the percent of flood zone area where a notable peak exists when it reaches 100%. Percent of road length in flood zone behaves similarly to the percent of flood zone area, but in a step-like pattern, with the lower predicted rescue rates occuring when the road length in flood zone is under 50%. The variable percent of wetland cover produces a similar step-like prediction pattern, but with a steep increase that occurs at the 20%–25% of wetland area. Overall, all three variables show how flood-prone the area is, and two of them suggest a threshold effect. That is, when a block group has more than 50% of roads in flood zone or more than 20% wetland cover, it will result in a much higher rescue request rate. On the contrary, mean elevation has a negative relationship with flood hazard (percent of flood zone area and percent of road length in flood zone), and its impact on the predicted rescue rate

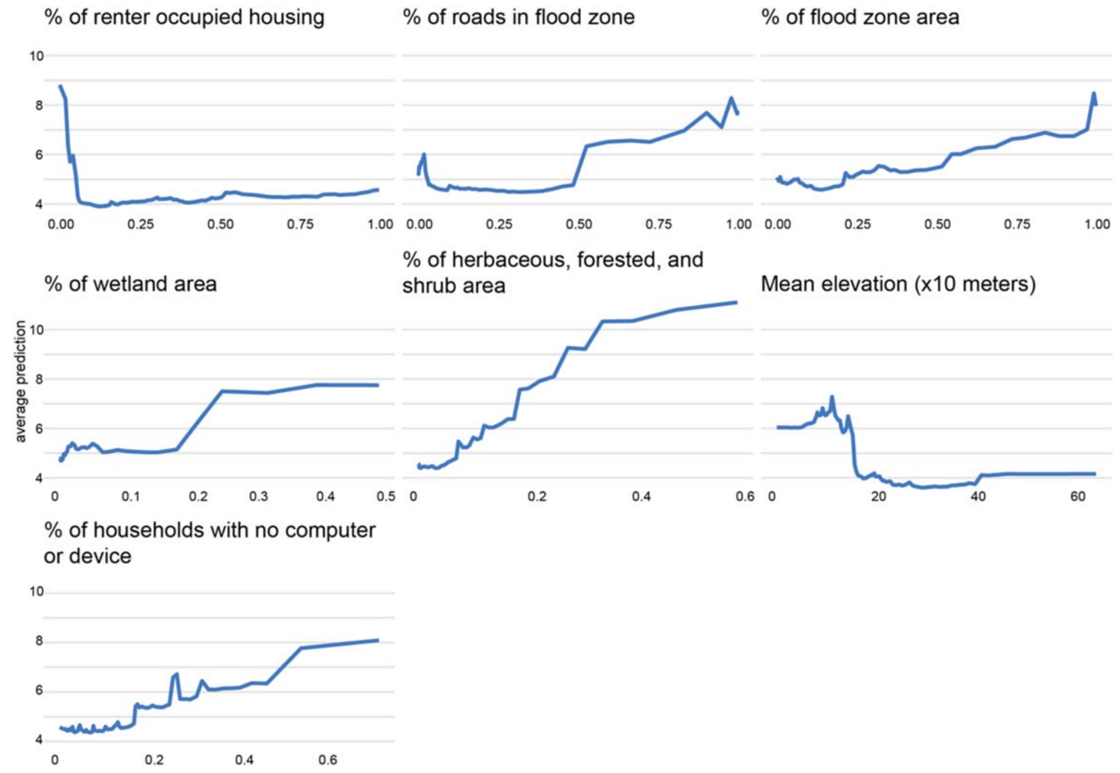

Figure 5. Partial dependence plots (in the order of variable importance from left to right). Y-axes represent average rescue rate predictions, and X-axes are values of each predictor.

is most pronounced at elevations lower than 180 meters. While higher values of the features related to flood hazard and environmental vulnerability produce higher rescue request predictions, the housing and digital access variables show an unexpected pattern of contribution. On one hand, the block groups with the lowest percent of renter occupied units get the highest predicted rescue rates, pointing to a potential social disparity due to home ownership. On the other hand, we found that higher rates of social media rescue requests are predicted for the block groups with less access to technology (percent of households with no computer or device). The PD plots (figure 5) demonstrate that complex patterns of variable contributions to ML models can be observed and analyzed. These patterns may have implications for understanding the real rescue request needs during a hazardous event.

5.4. Rescue request model fairness measurements

As stated before, we assigned the group with SVI and its four themes in the lowest 0.25 quantile as privileged (the least vulnerable in the study area). Similarly, based on the four digital connectivity variables the group within the 0.25 quantile highest access is privileged. The bounds of the privileged groups in each sensitive attribute, and the calculated *independence*, *separation*, and *sufficiency* ratios are shown in table 5. In sum, we found no unfairness in the prediction model based on social vulnerability characteristics or digital access variables, therefore our hypothesis that the model is unfair is not accepted.

However, we note a slight deviation from complete fairness in the *sufficiency* value for the SVI Theme 4 Housing and Transportation variable (table 5). This implies a slightly uneven distribution of error between the two groups, or that Y and S are better at predicting A than just S (i.e., S is less sufficient). However, the *sufficiency* ratio is still close to 1, and there is no discernable difference in the pattern of predictions between the groups (figure 6(a)). A related attribute, percent of renter-occupied housing, is a predictor in the final model and shows a notable difference in predicted rescue rate between its low and high values in the PD profile (figure 5).

5.5. Alternative dataset fairness measurement: Boston housing data

We applied the same modeling and fairness testing approach to a well-studied Boston housing dataset to compare the fairness ratios with the rescue model. We tested variable significance using the permutation importance measure on the Boston data and found that none of the variables needed to be excluded, thus, the entire dataset was used. Fitting Random Forest on the Boston housing data with all 13 predictors resulted in out-of-bag MSE of 10.06 and 0.88 out-of-bag R^2. Fairness ratios for the protected attribute B were 11.45 (*independence*),

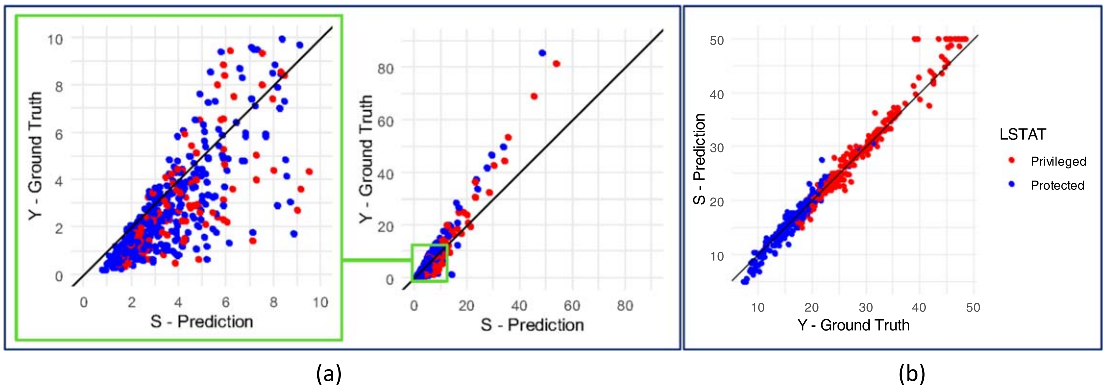

Figure 6. Predictions plotted against ground truth of (a) rescue rates (red dots belong to privileged group, and blue to protected group according to SVI Theme 4 Housing and Transportation) and (b) housing prices from Boston dataset, demonstrating an unfair model (privileged cases have higher housing prices than non-privileged).

Table 5. Algorithm fairness metrics per binary sensitive feature.

Sensitive attribute	Privileged Min	Privileged Max	$ratio_{ind}$	$ratio_{sep}$	$ratio_{suff}$
SVI	0.0003	0.265[a]	1.002	1.002	1.001
SVI Theme 1	0.003	0.255[a]	1.006	1.006	1.002
SVI Theme 2	0.005	0.238[a]	1.002	1.002	1.021
SVI Theme 3	0	0.324[a]	1.001	1.001	1.002
SVI Theme 4	0	0.245[a]	1.002	1.004	1.116
% No computer or device	0	0.028[a]	1.001	1.001	1.003
% No Internet	0	0.055[a]	1.002	1.002	1.0003
% With cellular data plan	0.647[a]	0.902	1.005	1.005	1.0007
% With smartphone	0.857[a]	1	1.006	1.005	1.001

[a] Dataset's 0.25 quantile.

1.28 (*separation*), and 1.09 (*sufficiency*), which are above the fairness threshold of 1.25. For $LSTAT$ the results were even more striking, with $ratio_{ind} = 889, 292.7$, $ratio_{sep} = 1, 769.66$, and $ratio_{suff} = 9.62$ (figure 6(b)).

To mitigate the model bias, we removed the sensitive attributes B (represents the proportion of black people) and $LSTAT$ (represents th population with lower status) from the set of model predictors. As a result, Random Forest performance decreased slightly (MSE 14.16; R^2 0.83), but unfairness persisted, with only a slight improvement in fairness ratios. For the protected attribute B, the *independence* ratio was 9.87, *separation*—1.07, and 1.00 *sufficiency*. For $LSTAT$, the *independence* ratio became 112,024, *separation* reduced to 2.13, and *sufficiency* to 1.00. The model remains unfair, and further fairness mitigation beyond removing the sensitive attributes from training data is needed. By contrast, the rescue model appears fair.

6. Discussion

This study aimed to examine if bias exists in AI algorithms used in disaster management, specifically social media rescue models. The Random Forest rescue model and its explorations provide useful information on physical and social vulnerabilities that put residents at risk at getting stranded due to flooding, thus needing urgent assistance. This study demonstrates how the model can be evaluated for algorithm bias using the three criteria—*independence*, *separation*, and *sufficiency*, which is a novel contribution to the disaster risk management literature.

We did not find that the rescue model violates *independence*, *separation*, or *sufficiency* for any of the nine sensitive attributes tested. However, some data biases may be difficult to evaluate due to the demographic features being inferred from the flood victims' locations, rather than directly based on their identity. Similarly, the data may be incomplete or missing, which poses a limitation, despite that we made the best possible effort to find every social media or volunteer-collected rescue request from Houston MSA. Further research, such as approximating rescue needs from indirect indicators, may be conducted to identify how many people needed but could not reach rescue volunteers because of the potential barriers to doing so. We adopt Steinberg *et al* (2020) density ratio implementation to calculate the three fairness metrics *independence*, *separation*, and *sufficiency* for our regression task. Experiments using other implementations of these three metrics may lead to varying results.

ML and AI are often employed to overcome limitations of multivariate statistics, for example, its required assumptions for linear models. It should be noted that the fairness criteria used in this study are applicable to evaluate fairness of the linear models, because linear models can also be unfair. In fact, modern AI fairness literature has drawn many ideas from the fairness criteria developed for linear models in the 1960s and 1970s, primarily in the fields of education testing and psychometrics (Barocas *et al* 2019). Despite that, the urgency and necessity to consider fairness of AI models arises from its rising popularity and availability, its ability to capture complex interactions from big data (whereas linear statistics assume no collinearity), and difficulty of explaining its internal structure (Barocas *et al* 2019).

An additional research need arises at the intersection of fair AI and geospatial AI. AI fairness usually considers an individual as the subject of AI decisions, whereas in geo-AI the subject is an areal unit in which a group of people may be residents, or owners of the property, or have some other relation. While we demonstrate an application of regression fairness metrics to two geospatial datasets, this aspect is largely unexplored in AI fairness literature. For example, more research is needed on how notions of group and individual fairness translate from traditional to spatial applications, where an individual (a single data instance) represents a group of people, and a group of cases is a group of groups.

7. Conclusion

The objective of this study was to analyze if potential biases exist towards protected groups in a Random Forest regression of social media rescue requests from Hurricane Harvey. Nine sensitive attributes tabulated at the census block group level were considered, including the Social Vulnerability Index (SVI), its four sub-indices, and rates of access to any computer or device, a smartphone, Internet, and cellular data plan, as they are considered as likely barriers to access rescue through social media.

The study addresses two related research questions. To answer RQ1, seven variables were found to be significant predictors of the rescue request rates. In the order of importance, they were percent of renter occupied housing units, percent of roads in flood zone, percent of flood zone area, percent of wetland cover, percent of herbaceous, forested and shrub cover, mean elevation, and percent of households with no computer or device. In general, more rescue requests are expected in the block groups with higher physical or environmental vulnerability. On the other hand, higher home ownership rates (lower percent of renter occupied housing) and lower rate of access to a computer or device per household are associated with higher rescue needs. The Partial Dependence plots show non-linear relationships, which help further explain the complex relationships between rescue request rates and the seven predictor variables.

Regarding RQ2, we found that the model learnt from the social media rescue data satisfied the three fairness criteria—*independence*, *separation*, and *sufficiency ratios*—for all of the nine sensitive attributes, which were social vulnerability and digital access variables. Our hypothesis that higher social vulnerability and lower digital access are associated with lower predicted social media rescue rates, which leads to unfair predictions based on these characteristics, is rejected. All three ratios for all assessed sensitive attributes are close to 1.0, indicating that the model is fair. We found one small deviation from perfect fairness, as reflected by the *sufficiency* ratio, for the variable of SVI Theme 4 Housing and Transportation vulnerability.

Our study of the rescue AI model contributes new insights on the physical and social vulnerabilities that put residents at risk of getting stranded due to flooding, and thus needing urgent assistance. Moreover, by providing an explanation of the Random Forest model and evaluating its fairness, we demonstrate how the same methodology of fairness evaluation could be applied in future AI modeling in disaster research and help reduce *allocation* or *quality-of-service harms* that could deepen social and geographical disparities.

Acknowledgments

This article is based on work supported by grants from the National Science Foundation of the United States (under Grant Numbers 1927513, 1620451, 1762600). Any opinions, fundings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agency.

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

ORCID iDs

Volodymyr V Mihunov https://orcid.org/0000-0002-1490-6124

References

Altmann A, Toloşi L, Sander O and Lengauer T 2010 Permutation importance: a corrected feature importance measure *Bioinformatics* **26** 1340–7

Arrazolo S 2017 Harvey Rescue Doc Accessed August 2019 from (https://data.world/sya/harvey-rescue-doc)

Barocas S, Hardt M and Narayanan A 2019 Fairness and Machine Learning: Limitations and Opportunities. Web (https://fairmlbook.org)

Behl S, Rao A, Aggarwal S, Chadha S and Pannu H S 2021 Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises *Int. J. Disaster Risk Reduct.* **55** 102101

Biecek P 2018 DALEX: explainers for complex predictive models in R *Journal of Machine Learning Research* **19** 1–5

Bird S *et al* 2020 Fairlearn: A toolkit for assessing and improving fairness in AI *Microsoft.* (https://microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/)

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Cai H *et al* 2018 A synthesis of disaster resilience measurement methods and indices *Int. J. Disaster Risk Reduct.* **31** 844–55

Choi E Y, Kanthawala S, Kim Y S and Lee H Y 2022 Urban/rural digital divide exists in older adults: does it vary by racial/ethnic groups *J Appl Gerontol* **41** 1348–56

Chouldechova A 2017 Fair prediction with disparate impact: a study of bias in recidivism prediction instruments *Big Data* **5:2** 153–163

Cutter S L, Boruff B J and Shirley W L 2003 Social vulnerability to environmental hazards* *Social Science Quarterly* **84** 242–61

Dargin J S, Fan C and Mostafavi A 2021 Vulnerable populations and social media use in disasters: uncovering the digital divide in three major U.S. hurricanes *Int. J. Disaster Risk Reduct.* **54** 102043

Dwork C, Hardt M, Pitassi T, Reingold O and Zemel R 2012 Fairness through awareness *Proc. of the 3rd Innovations in Theoretical Computer Science Conf. on - ITCS '12, 2012-01-01* (ACM Press) (https://doi.org/10.1145/2090236.2090255)

Esri 2021 *How the zonal statistics tools work* Accessed January 2022 from (https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/how-zonal-statistics-works.htm)

Esri 2022 *How Forest-based Classification and Regression works* Accessed January 2022 from (https://pro.arcgis.com/en/pro-app/2.8/tool-reference/spatial-statistics/how-forest-works.htm)

Federal Emergency Management Agency (FEMA) 2018 Harvey flood depths grid Accessed August 2020

Federal Emergency Management Agency (FEMA) 2020 National Flood Hazard Layer (NFHL) Accessed August from (https://msc.fema.gov/portal/advanceSearch)

Fitzsimons J, Al Ali A, Osborne M and Roberts S 2019 A General Framework for Fair Regression *Entropy* **21** 741

Fjeld J, Achten N, Hilligoss H, Nagy A and Srikumar M 2020 Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI *Berkman Klein Center Research* Publication No. 2020-1 SSRN Electronic Journal (https://doi.org/10.2139/ssrn.3518482)

Flanagan B E, Gregory E W, Hallisey E J, Heitgerd J L and Lewis B 2011 A Social Vulnerability Index for Disaster Management *Journal of Homeland Security and Emergency Management* **8** 0000102202154773551792

Flanagan B E, Hallisey E J, Adams E and Lavery A 2018 Measuring community vulnerability to natural and anthropogenic hazards: the centers for disease control and prevention's social vulnerability index *J. Environ. Health* **80** 34–6

Gevaert C M, Carman M, Rosman B, Georgiadou Y and Soden R 2021 Fairness and accountability of AI in disaster risk management: Opportunities and challenges *Patterns* **2** 100363

Global Facility for Disaster Reduction and Recovery (GFDRR) 2018 *Machine Learning for Disaster Risk Management.* (Washington, DC: GFDRR)

Goldstein A, Kapelner A, Bleich J and Pitkin E 2015 Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation *Journal of Computational and Graphical Statistics* **24** 44–65

Harrison D and Rubinfeld D L 1978 Hedonic housing prices and the demand for clean air *Journal of Environmental Economics and Management* **5** 81–102 Original data

Lam N S N, Reams M, Li K, Li C and Mata L P 2016 Measuring community resilience to coastal hazards along the northern gulf of mexico *Nat Hazards Rev* **17** 04015013

Lepri B, Oliver N, Letouzé E, Pentland A and Vinck P 2018 Fair, transparent, and accountable algorithmic decision-making processes *Philosophy & Technology* **31** 611–27

Li Z *et al* 2022 Social vulnerability and rurality associated with higher severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection-induced seroprevalence: a nationwide blood donor study-united states, July 2020-June 2021 *Clin Infect Dis* **75** e133–43

Mihunov V V, Jafari N H, Wang K, Lam N S N and Govender D 2022 Disaster impacts surveillance from social media with topic modeling and feature extraction: case of hurricane harvey *International Journal of Disaster Risk Science.* **13** 729–742

Mihunov V V, Lam N S N, Zou L, Wang Z and Wang K 2020 Use of twitter in disaster rescue: lessons learned from hurricane harvey *Int. J. Digital Earth* **13:12** 1–13

Mittelstadt B, Russell C and Wachter S 2019 Explaining explanations in AI *Proc. of the Conf. on Fairness, Accountability, and Transparency, 2019-01-29* (ACM) (https://doi.org/10.1145/3287560.3287574)

Morrow B H 1999 Identifying and mapping community vulnerability *Disasters* **23** 1–18

Nembrini S, König I R and Wright M N 2018 The revival of the Gini importance *Bioinformatics* **34** 3711–8

Niculescu-Mizil A and Caruana R 2005 Predicting good probabilities with supervised learning *Proc. of the 22nd Int. Conf. on Machine Learning* (https://doi.org/10.1145/1102351.1102430)

Pace R K and Gilley O W 1997 Using the spatial configuration of the data to improve estimation *Journal of the Real Estate Finance and Economics* **14** 333–40

Pestre G, Letouzé E and Zagheni E 2020 The ABCDE of big data: assessing biases in call-detail records for development estimates *The World Bank Economic Review* **34** S89–97

Ramesh B *et al* 2022 Flooding and emergency department visits: effect modification by the CDC/ATSDR Social Vulnerability Index *Int. J. Disaster Risk Reduct.* **76** 102986

Shang J, Sun M and Lam N S N 2020 List-wise fairness criterion for point processes *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining, 2020-08-23* (ACM) (https://doi.org/10.1145/3394486.3403246)

Steinberg D, Reid A and O'Callaghan S 2020 Fairness measures for regression via probabilistic classification *Ethics of Data Science Conf. (EDSC)* (Sydney, Australia)

Suresh H and Guttag J 2021 A framework for understanding sources of harm throughout the machine learning life cycle *Equity and Access in Algorithms, Mechanisms, and Optimization, 2021-10-05* (ACM) (https://doi.org/10.1145/3465416.3483305)

U.S. Census Bureau 2016 TIGER/Line Shapefiles (machine-readable data files) All Roads (https://census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.2016.html) Accessed August 2020

U.S. Geological Survey (USGS) 2017 1 Arc-second Digital Elevation Models (DEMs) - USGS National Map 3D Elevation Program (3DEP) Downloadable Data Collection: U.S. Geological Survey. (https://sciencebase.gov/catalog/item/4f70aa71e4b058caae3f8de1) Accessed August 20202020

Wang K, Lam N S N and Mihunov V 2023 Correlating twitter use with disaster resilience at two spatial scales: a case study of hurricane sandy *Ann. Gis* **29** 1–20

Wang Z *et al* 2022 A machine learning approach for detecting rescue requests from social media *ISPRS International Journal of Geo-Information* **11** (11) 570

Wang Z, Lam N S N, Obradovich N and Ye X 2019 Are vulnerable communities digitally left behind in social responses to natural disasters? An evidence from Hurricane Sandy with Twitter data *Appl. Geogr.* **108** 1–8

Watson K M, Harwell G R, Wallace D S, Welborn T L, Stengel V G and McDowell J S 2018 *Characterization of peak streamflows and flood inundation of selected areas in southeastern Texas and southwestern Louisiana from the August and September 2017 flood resulting from Hurricane Harvey* (Reston, VA: U.S. Geological Survey) (*Scientific Investigations Report*) 2018-5070 (https://doi.org/10.3133/sir20185070)

Wiśniewski J and Biecek P 2022 fairmodels: a flexible tool for bias detection, visualization, and mitigation in binary classification models *The R Journal* **14** 227–43

Wright M N and Ziegler A 2017 Ranger: a fast implementation of random forests for high dimensional data in C++ and R *Journal of Statistical Software* **77** 1–17

Yang L, Jin S, Danielson P *et al* 2018 A new generation of the united states national land cover database: requirements, research priorities, design, and implementation strategies *ISPRS J. Photogramm. Remote Sens.* **146** 108–23

Yu M, Yang C and Li Y 2018 Big data in natural disaster management: a review *Geosciences* **8** 165

Zadrozny B and Elkan C 2002 Transforming classifier scores into accurate multiclass probability estimates *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (New York, NY, United States: Association for Computing Machinery) pp 694–9

Zhou B *et al* 2022 VictimFinder: Harvesting rescue requests in disaster response from social media with BERT *Comput. Environ. Urban Syst.* **95** 101824

Zou L *et al* 2018 Social and geographical disparities in Twitter use during Hurricane Harvey *Int. J. Digital Earth* **12** 1300–18