GEOGRAPHICAL DISPARITIES OF FLASH FLOOD VULNERABILITY IN TEXAS

A Thesis

by

MOXUAN LI

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Lei Zou
Committee Members,	Burak Güneralp
	Nasir Gharaibeh
Head of Department,	David Cairns

August 2022

Major Subject: Geography

ABSTRACT

As one of the most dangerous natural disasters, flash floods account for 52% of economic losses and over 70% of fatalities and injuries caused by flood-related disasters. There is an urgent need to evaluate community-based flash flood vulnerability, identify its driving factors, and develop mitigation strategies in different communities to reduce damages from future events. However, most precedent analyses of flash flood vulnerability rely on subjectively selected social variables and the developed models lack validation. This project aims to fill this gap by developing a framework to assess vulnerability at the community-level using historical flash flood data and determine the socioeconomic and environmental factors that affect vulnerability. Flash flood records collected from the National Oceanic and Atmospheric Administration (NOAA) and SHELDUS database, socioeconomic data derived from U.S. Census Bureau, as well as terrain data derived from GMTED2010 dataset via USGS and Multi-Resolution Land Characteristics (MRLC) Consortium were utilized in this project. First, this study statistically analyzed the location, frequency, and damage of flash flood events in Texas at block-group-level Second, we defined and calculated a flash flood vulnerability index based on the average damage (sum of property and crops damage) per capita per event. Third, social variables used in previous study on vulnerability assessment are collected, and their correlation with the derived vulnerability index was examined. The results could support further analysis of natural disaster risk assessment and monitoring and assist disaster mitigation and responding.

Keywords: Flash flood, vulnerability, spatial analysis, risk assessment, disaster mitigation

DEDICATION

To my mother, father, grandfather, and grandmother, as well as my academic advisor and instructors during my college study.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Lei Zou, who is my academic advisor during my Master's Program and guidesd me throughout this project. I would also like to acknowledge Dr. Burak Güneralp and Nasir Gharaibeh fir joining my committee and support this research.

I would also like to show my deep appreciation to my colleagues in Gear Lab for not only our collaboration on study and analysis but also our friendship developed during the past one year, as well as the encouragement I got from them. My life during Master's program was terribly influenced by COVID-19 pandemic, while fortunately I meet my colleagues in Gear Lab and they supported me a lot on both academics and daily life.

In addition, I would specifically acknowledge the support provided by Mr. Mingzheng Yang from Gear Lab, since he offered tremendous help in this project, especially the data analysis conducted in chapter III and chapter IV. He instructed me on the use of different learning algorithms and models, as well as the interpretation of the outputs of different models. Without his help, it would be way more challenging for me to finish this thesis project.

Besides, I would like to appreciate Dr. Xiao Li from Texas A&M Transportation Institute (TTI). Although he didn't provide direct help on this thesis project, the intern opportunity he provided to me significanly enhanced my research ability and writing skills.

In the end, I would like to thank the Texas A&M University Graduate and Professional School to allow me to construct this LaTeX thesis template.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work is supported by a thesis committee consisting of Professor Lei Zou and Burak Güneralp from Department of Geography, and Professor Nasir Gharaibeh from Zachry Department of Civil & Environmental Engineering.

The flash flood data analyzed for Chapter III was derived from National Oceanic and Atmospheric Administration (NOAA) and the Spatial Hazard Events and Losses Database for the United States (SHELDUS), the socio-economic data was derived from NHGIS. The analyses depicted in Chapter III and IV were conducted with the support from Mr. Mingzheng Yang, a Ph.D. student from the Department of Geography at Texas A&M University.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

No other outside source of funding was provided.

NOMENCLATURE

VIM	Vulnerability Inference Measurement
RVIM	Raw VIM
BG	Block Group
CN	County
NOAA	National Oceanic and Atmospheric Administration
NCDC	National Climatic Data Center
SHELDUS	Spatial Hazard Events and Losses Database for the United States
LR	Linear Regression
MLR	Multinomial Logistic Regression
RF	Random Forest
SS	Stepwise Selection

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Flash Flood

Flood, “an overflow of water onto normally dry land”, has been one of the most common forms of natural disasters in recent decades (NWS 2022b). According to the NWS, flood is the deadliest form of all weather-related hazards in the United States other than heat, accounting for around 19 % of all weather-related fatalities during the past 30 years (NWS 2020).

Specifically, flash floods are a unique form of floods that feature their rapidity. Flash floods usually occur within 2 hours and are caused by certain events like heavy rainfall or snowmelt (NWS 2022a). Monitoring and predicting flash flood frequencies and intensities are challenging because of such rapidity. Consequently, flash floods have caused devastating damage to local communities as people have little time to foresee, prepare for, and respond to those events promptly. Flash flood accounts for the majority of the damage caused by flood-related events (Ashley and Ashley 2008). In the past 20 years, flash flood events caused approximately 72% of flood-relevant fatalities, 72% of flood-induced injuries, and 52% of flood-related economic losses in the U.S., making them one of the most lethal natural disasters (NWS 2019).

1.2 Flash Flood Vulnerability

Due to the difficulty of monitoring flash floods and their subsequent hazardous impacts, investigations on flash flood threats and damages have attracted the attention of governments and researchers, all aiming at identifying pathways to mitigate the risk of flash floods. According to a report published by the Intergovernmental Penal on Climate Change (IPCC), disaster risk should be estimated by three dimensions: (1) hazard, the natural event threatening in the targeted area and its possible occurrence; (2) exposure, buildings/properties/humans that are in the area where hazards may occur; and (3) vulnerability representing the propensity of exposed elements to suffer adverse effects when impacted by hazard events (IPCC 2012). Accurate disaster risk evaluation relies on finding suitable methods and indicators to quantify the three dimensions.

Among the three aspects, the hazard frequency and intensity are capricious since they depend on the terrain, climate, and other natural conditions. Exposure, on the other hand, could be solved via long-term urban planning and short-term activities, e.g., evacuation. As a result, vulnerability becomes the focus that communities could target by actions such as improving the drainage system and enhancing building strength to get immediate improvement. Analyzing the vulnerability to flash floods will shed light on how communities suffer from this disaster and thus establish proper measures to reduce the potential damage caused by future events.

One critical issue in flash flood vulnerability assessment is that the definition of "vulnerability", varies among studies. Cutter and Finch (2008) define vulnerability as "a measure of both the sensitivity of a population to natural hazards and its ability to respond and recover from the impacts of hazards" in an analysis of the spatial changes in vulnerability in the United States at the county level (Cutter and Finch 2008). This definition is a composite of two processes, sensitivity and recovery, and has been widely adopted in numerous investigations (Kirby et al. 2019; Tate 2012). In more recent year, Lam et al. established the Resilience Inference Measurement (RIM) Model to conceptualize and evaluate the disaster resilience process. This model defines vulnerability as "the latent relationship between exposure and damage" (Lam et al. 2016). By selecting quantitative indicators of exposure and damage to explore their disparate relationships, the RIM model has been applied to evaluate community resilience and vulnerability to coastal hazards (Cai et al. 2016), earthquakes (Li et al. 2016), and drought (Mihunov et al. 2018; Mihunov et al. 2019). Lam's definition of vulnerability provides a self-validated measurement framework that can be seamlessly customized for diverse disaster types. Therefore, this thesis adopts its vulnerability concept and framework for flash flood vulnerability assessment.

1.3 Research Questions and Objectives

To understand the root causes of flash flood vulnerability, this project aims to answer the following **research questions**:

1. What are the spatial patterns of flash flood occurrence and economic impact?

2. Which communities are more vulnerable to flash flood events?

3. How can communities reduce flash flood vulnerability in hazard-prone areas?

To answer these questions, it is necessary to establish a model to quantify the spatial disparities in vulnerability. Using Texas as the study area, this research developed a framework to assess the spatial patterns of vulnerability to flash floods at two scales: block-group and county. The **research objectives** are three-fold:

1. Develop a vulnerability inference measurement (VIM) framework to quantify community vulnerability to flash floods using multiple databases.

2. Examine the spatial disparities of flash floods in Texas at multiple administrative scales, i.e., county and block-group.

3. Identify potential contributors of flash flood vulnerability at different scales and inform decision and policy making for risk mitigation.

Historical flash flood records were derived from the Storm Events Database provided by National Oceanic and Atmospheric Administration (NOAA) and the Spatial Hazard Events and Losses Database for the United States (SHELDUS). Socioeconomic data supplied by U.S. Census Bureau and Digital Elevation Model (DEM) data obtained from U.S. Geological Survey were also utilized. First, this study statistically analyzed the time, duration, frequency, and damage of flash flood events at the block group and county levels. Second, we defined and calculated flash flood vulnerability using a newly established Vulnerability Inference Measurement (VIM) model to reveal its spatial patterns at the two scales in Texas. Third, potential contributors of flash flood vulnerability were identified through multiple statistical and learning models. The results of this project could support further investigations of natural disaster risk assessment and monitoring as well as assisting disaster mitigation and responding.

1.4 Literature Review

A lot of researches have been conducted around natural disasters. Some studies analyzed the disaster management and evaluated their performance in disaster resilience, and some other studies aims to quantify the social resilience to disasters using a series of social factors.

In 2012, Berke established a model to evaluate the reliability of State Hazard Mitigation Plan under Disaster Mitigation Act (DMA), which is designed to reduce disaster damages, in building community resilience. This study lists 6 principles that are closely aligned with DMA requirements for preparing state mitigation plans, including 4 internal principles (goal, fact base, mitigation policies, implementation and monitoring) and 2 external principles (interorganizational coordination, participation), to evaluate the plan. The researchers also strategically quantified the plans through: (1) collecting samples of State Hazard Mitigation Plan from 30 coastal states; (2) the sampled plans were then coded/scored based on the six principles of plan quality; (3) each plan were independently examined by 2 of the 4 coders on the coding team; and (4) the index score of each principle was calculated for each plan. The final index score of the plan quality was calculated by summing the scores for each of the items and then dividing the sum by the total number of items combined. The differences of derived index scores represent the differences of quality of those plans, and by tracing back the score of each principles the root reasons could be revealed and thus help governments to improve the mitigation plan (Berke, Smith, and Lyles 2012).

In recent years, more frameworks on quantifying social resilience to disasters emerges. Cutter et al. (2014) established a model to evaluate resilience from 6 aspects, including social resilience, economic resilience, community capital, institutional resilience, housing/infrastructure, as well as environmental resilience. They collected 49 social variables and scored each aspect based on those variables from 0-6 and summed all the 6 scores up to conclude the final score for resilience. This method has some limitations. First, the weighting method is subjective and lacks validation. Second, although the variables of each aspect are synthesized using Principal Component Analysis (PCA), the final index were calculated directly by summing up the scores of these six aspects and the difference of weights among these aspects were ignored. Third, as the research team disclaimed, “the final score is not an absolute measurement of community resilience for a single location, but rather a relative value in which multiple places can be compared” (Cutter, Ash, and Emrich 2014), thereby we cannot extract more information from the resulted scores other than which community has higher resilience. Nevertheless, this study demonstrates the potential of synthesizing different socioeconomic data for the quantification of disaster resilience. Since vulnerability is defined as a part of resilience in this investigation, similar methods could be utilized in our model.

Khajehei (2020) raised an advanced way to quantify the vulnerability to flash floods. Similar to Cutter’s method, they collected a series of data from different categories, including 32 variables of (1) demographic socioeconomic status, (2) race and ethnicity, (3) age, (4) gender, (5) housing and transportation, and (6) industrial economy. They normalized all variables of socio-economic data to z-scores and ran a more advanced PCA algorithm, Probabilistic Principal Component Analysis (PPCA), to synthesize these variables (Khajehei et al. 2020). Comparing to Cutter’s analysis, this one is specifically designed for flash flood vulnerability.

Cai’s analysis (2018), on the other hand, summarized the most frequently used socioeconomic variables in the assessment of disaster resilience. They reviewed 174 academic articles related to disaster resilience analysis, and classified the commonly used socioeconomic variables into 7 different categories: (1) community, (2) infrastructure, (3) individual/household, (4) urban system, (5) economy, (6) others (government, ecosystem, social-ecological system, etc.), and (7) not specified (Cai et al. 2018). Combining Cai’s conclusions and Khajehei’s model, we would be able to identify the most frequently used socioeconomic indices for flash flood vulnerability inference, which are detailed in Section 2.3.

In rencent year, more advanced techniques are introduced in the quantitative assessment of disaster risk. Sarker et al. (2020) did a literature review and summarized several methods that use big data, including satellite imagery, aerial imagery and videos, wireless sensor web network, and LiDAR, etc., in the assessment of disaster resilience, especially its capability and potential on mitigating disaster risks. They indicate that the analysis of various big data can improve disaster resilience and management through multiple avenues: (1) in the preparedness phase, big data could help detect and monitor disasters, and thus giving people early warning to help them be prepared for upcoming disasters and reduce damages; (2) in the mitigation phase, big data could accurately assess disaster risk and forecast upcoming disaster events; and (3) in the response and recovery phases, big data could help recovery teams to collect data about current and cascading damage caused by disaster events, and thus support post-disaster management. However, difficulties exist in analyzing big data, such as the infrastructure for big data collection, and the requirements of expert teams with enough technical capacities, etc. Therefore, there is still a lot of improvement that needs to be realized before we make big data a reliable channel to assess disaster resilience (Sarker et al. 2020).

Based on the literature review, it is worth noticing that most of the current progress on evaluating flash flood vulnerability is based on subjectively selected socioeconomic variables while lacking either a connection to the actual influences caused by flash flood, or validation on the factor selections. Another critical fact revealed in Cai's report is that "only 10.3% of all the 174 articles conducted empirical validation of their proposed resilience indices" (Cai et al. 2018), which further proves that we still lack reliable root understanding to the cause of disaster risks. Therefore, establishing the connection between the concept of "vulnerability" and the actual effects caused by flash flood hazards and thus developing a quantitative method to evaluate vulnerability would be critical for understanding the driving forces behind.

2. STUDY AREA AND DATA

This project covers all the flash flood events that took place within Texas during 2005 to 2020. The major dataset used for this project include Texas geographic data, flash flood data, and contributor data.

2.1 Texas Geographic Data

The Texas geographic data includes the area unit boundaries in Texas at two geographic levels: county and block-group. Each of them is stored in a shapefile derived from the U.S. Census database via National Historical Geographic Information System (NHGIS; Figures 2.1).

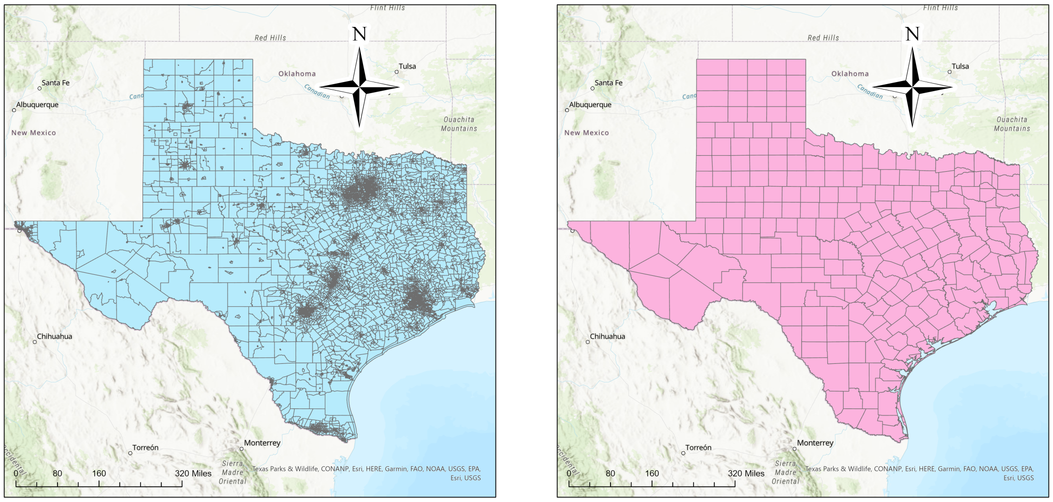

Figure 2.1: Texas Block Group Boundaries (Left) and Texas County Boundaries (Right).

2.2 Flash Flood Data

The flash flood data have two sources. The raw flash flood event records were derived from National Climatic Data Center (NCDC) Storm Event Database developed by NOAA, which contains various attributes for each storm record, i.e., time, coordinates, injuries, fatalities, damages, causes, etc. In addition, the aggregated flash flood occurrence and damage data were derived from SHELDUS database, which summarized various types of weather-related natural disasters recorded in NCDC database and their influences, including economic losses, fatalities, injuries, and durations, etc., at county level by month and by disaster categories. The economic losses were adjusted considering the 2020 inflation rate to enable temporal comparison. These two datasets for flash flood records were used to quantify vulnerability at different spatial scales. Figures 2.2 shows two maps for the county-level aggregated flash flood frequency and damage calculated using NCDC dataset provided by NOAA, and Figure 2.3 shows maps for frequency and damage calculated using data derived from SHELDUS. We can notice that the overall distribution of damage is very close between the map generated using NCDC dataset and SHELDUS dataset, while the aggregated frequency is much higher in the map generated using NCDC data than in the map generated using SHELDUS data. This inconsistency could be caused by the different methods used in data fusion and spatial analysis.

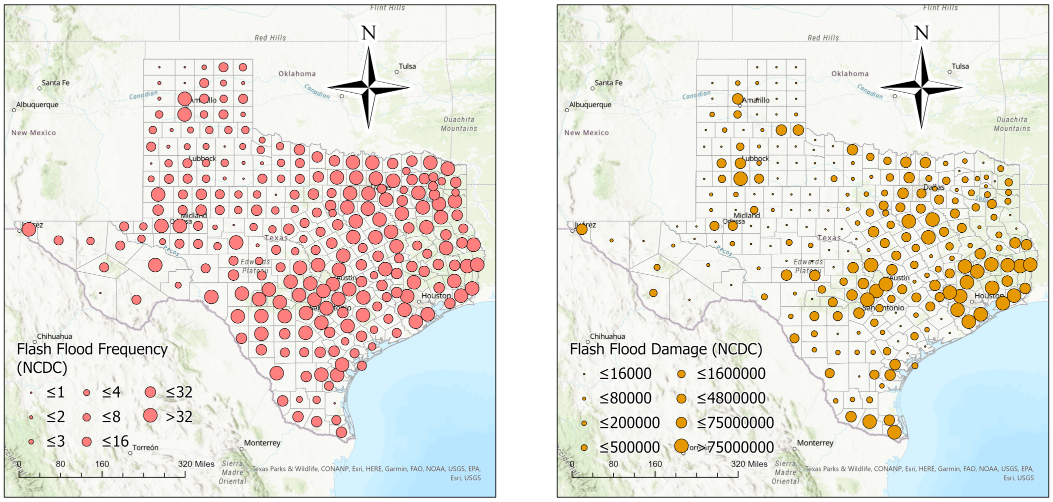

Figure 2.2: Aggregated Flash Flood Frequency (Left) and Damage (Right) by County calculated Using NCDC Dataset (2005-2019).

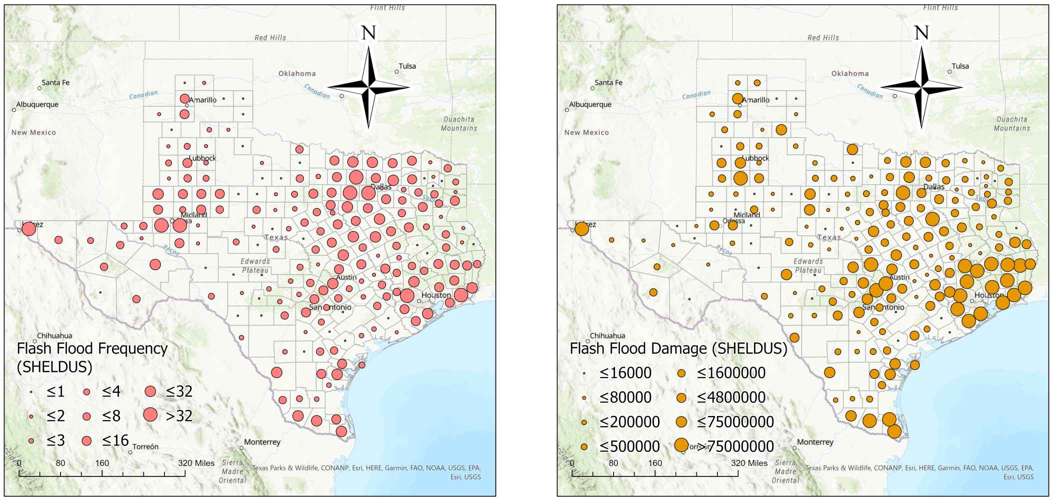

Figure 2.3: Aggregated Flash Flood Frequency (Left) and Damage (Right) by County calculated Using SHELDUS Dataset (2005-2019).

2.3 Contributor Data

The contributor data were used for the correlation analysis, which aims to identify what factors, and in which ways, influence the vulnerability. These factors include two categories: socio-economic status and terrain conditions. The variable selection is introduced in the following sections. A complete list of factors for correlation analysis and their definitions are shown in Table 2.1.

Factor Category	Factor field name	Definition
Socioeconomic Variables	PCT_POVERTY	Percentage of population in poverty
	PER_CAPITA_INCOME	Per capita income
	MED_HSHD_INCOME	Median household income
	Median_HSHD_Value	Median household value
	PCT_Under_EDU	Percentage of population aged 25 years or older with less than 12th grade education
	Median_Gross_Rent	Median gross rent
	PCT_Extractive	Percentage employment in extractive industries
	PCT_Over_200K	Percentage of households earning greater than US $200,000 annually
	PCT_Service	Percentage employment in service industry
	UNEMPLOYMENT_RATE	Percentage civilian unemployment
	PCT_ASIAN	Percentage Asian
	PCT_AFRICAN_AMERICAN	Percentage Black or African American
	PCT_HSHD_ENG_LMT	percentage of limited English speaking households
	PCT_HISPANIC	Percentage Hispanic
	MEDIAN_AGE	Median age
	PCT_KID_OLD	Percentage of population under 5 years or 65 and older
	PCT_UNDER18	Percentage of population under 18 years old
	PCT_FEMALE	Percentage female
	PCT_FEMALE_LABOR	Percentage female participation in labor force
	PCT_FEMALE_HSHD	Percentage female-headed households
	HSHD_SIZE	Average household size
	PCT_MOBILE_HOMES	Percentage mobile homes
	PCT_HOUSING_NO_CAR	Percentage of housing units with no cars
	PCT_RENTER	Percentage renters
	PCT_UNOCCUPIED_HOUSING	Percentage unoccupied housing units
Terrain Conditions	MEAN_ELE	Average elevation
	MEAN_SLP	Average slope
	MEAN_Impervious	Average impervious rate
	MEDIAN_Impervious	Median impervious rate

Table 2.1: List of Potential Contributors for Correlation Analysis.

2.3.1 Socio-Economic Status

Khajehei (2020) provides a list of socio-economic factors that can be used in the inference of social vulnerability to flash flood. Based on variable definitions and data availability, we adopted the variables selected in the abovementioned literature and added some variables that could reflect additional aspects of socio-economic status. The detailed modification to the original variable list is shown and explained below:

- Removed:
  - Percentage of households receiving social security: data unavailable.
  - Percentage Native American: data only available at county level.
  - Percentage of population living in nursing and skilled-nursing facilities: data unavailable.
  - Industrial Economy (whole category): variables not clearly defined.
- Changed:
  - Percentage speaking English as a second language with limited English proficiency: changed to “percentage of limited English speaking households” due to the data unavailability (A "limited English speaking household" is one in which no member 14 years old and over (1) speaks only English or (2) speaks a non-English language and speaks English "very well." In other words, all members 14 years old and over have at least some difficulty with English. (Bureau 2022)).
  - People per unit: changed to “Average household size” due to the data unavailability (people per unit is not clearly defined so no data source were found, while average household size is the count of people by household and data are available on NHGIS).
- Added:

- Median household income: another variable that reflects “income” other than “Per capita income” and “Percentage of households earning greater than US $200,000 anually”

In addition, Population data was also collected but it was not used as a potential contributor, but as a input parameter for VIM index calculation. All data in the adapted variable list were derived from the US Census Database via NHGIS.

2.3.2 Terrain Conditions

The data for terrain conditions include elevation and impervious rate data (Figure 2.4). The raw elevation data were the DEM files derived from the GMTED2010 dataset via USGS and converted into slope data. The raw impervious data were derived from Multi-Resolution Land Characteristics (MRLC) Consortium. All the three data (DEM, slope, and impervious rate) were processed through zonal statistics to compute the mean and median values for each spatial unit (block group and county levels).

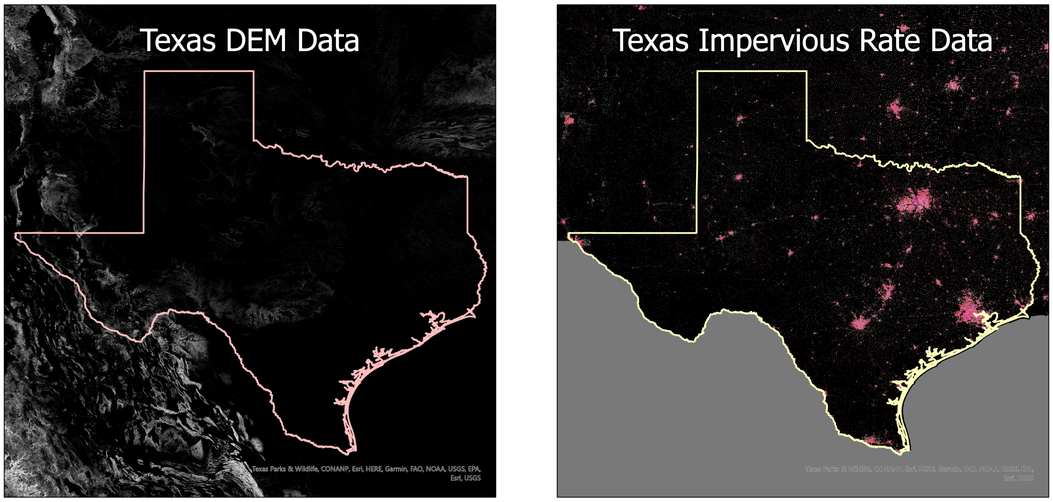
Texas DEM Data
Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, FAO, NOAA, USGS, EPA, Esri, USGS
Texas Impervious Rate Data
Texas Parks & Wildlife, CONANP, Esri, HERE, Garmin, FAO, NOAA, USGS, EPA, Esri, USGS

Figure 2.4: Texas DEM Data Derived from USGS (Left) and Impervious Rate Data Derived from MRLC (Right).

2.4 Variable Filtering

A total of 29 potential contributors were selected in the initial collection. The potential collinearity existed within the selected variables could lead to overfitting in the final model. Therefore, a correlation analysis was conducted to remove potentially correlated variables before inputting them in regression and machine learning models. Specifically, the Pearson correlation between each pair of variables in the initial list was tested. If the correlation coefficient of a certain pair of variables was greater than 0.7, we examined the correlation between each variable and the VIM Index and kept the one with a higher correlation. An example of three highly correlated variables are shown in Figure 2.5. The correlation test resulted in 19 variables in the final list, which will be utilized in the subsequent modeling process (Table 3.2).

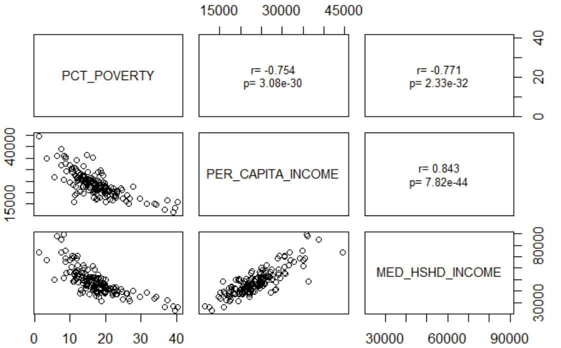

Figure 2.5: Three Highly Correlated Variables: PCT_POVERTY, PER_CAPITA_INCOME, and MED_HSHD_INCOME.

Factor Category	Factor field name	Definition
Socioeconomic Variables	MED_HSHD_INCOME	Median household income
	PCT_Under_EDU	Percentage of population aged 25 years or older with less than 12th grade education
	PCT_Extractive	Percentage employment in extractive industries
	PCT_Service	Percentage employment in service industry
	UNEMPLOYMENT_RATE	Percentage civilian unemployment
	PCT_ASIAN	Percentage Asian
	PCT_AFRICAN_AMERICAN	Percentage Black or African American
	MEDIAN_AGE	Median age
	PCT_KID_OLD	Percentage of population under 5 years or 65 and older
	PCT_UNDER18	Percentage of population under 18 years old
	PCT_FEMALE	Percentage female
	PCT_FEMALE_LABOR	Percentage female participation in labor force
	HSHD_SIZE	Average household size
	PCT_MOBILE_HOMES	Percentage mobile homes
	PCT_HOUSING_NO_CAR	Percentage of housing units with no cars
	PCT_RENTER	Percentage renters
	PCT_UNOCCUPIED_HOUSING	Percentage unoccupied housing units
Terrain Conditions	MEAN_SLP	Average slope
	MEAN_Impervious	Average impervious rate

Table 2.2: Final List of Potential Contributors for Correlation Analysis after Inter-correlation Test.

3. RESEARCH DESIGN

Based on the study objectives discussed in previous sections, our research could be divided into two modules: (1) vulnerability index establishment (objectives 1&2) and (2) contributor exploration (objective 3).

3.1 Vulnerability Index Establishment

For objective 1, we establish a model to quantify the vulnerability to flash flood, called the Vulnerability Inference Measurement (VIM) model. Based on Susan Cutter and Nina Lam's analysis, this study defines vulnerability as "the measure of sensitivity which turns the exposure to actual damage when disasters take place". Therefore, the main concept of the VIM model is to use the average economic damage fall on each individual caused by flash floods of the same level of threats in the given area to evaluate vulnerability. Specifically, vulnerability is defined as Damage per Capita per Event, as presented below:

$$RawVIM = \frac{TotalDamage}{Population * EventFrequency}$$

To accomplish the calculation, we need the historical flash flood records. And considering both the precision and the level of detail of our output, we decided to run the model at two different social scales: block group level and county level.

3.1.1 VIM Index at Block Group Level

Since no aggregated data for flash floor records at BG level could be found, the first step is to accomplish data aggregation. The NCDC Storm Event Database has very detailed information for various natural disasters, and we use the flash flood records in this database to complete this part of our research.

The NCDC data provide us the crop damage and property damage of each event, so the total damage could be easily concluded by summing these two values up. However, this damage value is for each whole event, while one flash flood event may not be within a single area unit, which in this case is a block group. To solve this issue, first we need the coordinate data, including the starting point and ending point, to identify the affected area of each event, and estimate the influence of each affected block group based on the ratio of that part to the whole affected region. We established a set of procedures to estimate the partial damage in each area unit: for each event, we (1) link the starting point and ending point to create a straight line which simulates the route of this event, (2) split the line by the boundaries of the block groups it crosses, (3) get the ratio of the length of each splitted part of this line to the total length of the whole line, and (4) calculate the partial damage of each part by multiplying the total damage to the ratio we got in step (3) (as shown below).

$$Damage_a = \frac{Length_a}{TotalLength * EventDamage}$$

This equation is used to convert the total amount of economic damage of each single event into the partial amount for each area unit for each event.

Next, we would use spatial join to (1) count the total number of events in each block group as frequency by block group (F_{BG}), and (2) sum up the partial damages of all these events in each block group as the aggregated damage by block group (AD_{BG}). And thereby, with the population by block group (P_{BG}) data derived from US Census database, the raw VIM index at block group level ($RVIM_{BG}$) could be calculated using the equation below:

$$RVIM_{BG} = \frac{AD_{BG}}{P_{BG} * F_{BG}}$$

Since the output of this equation may not be normally distributed values, we may need a series of operations like logarithm or square root to standardize it and achieve the VIM Index.

The overall workflow for the establishment of VIM index at block group level is shown in figure 3.1.

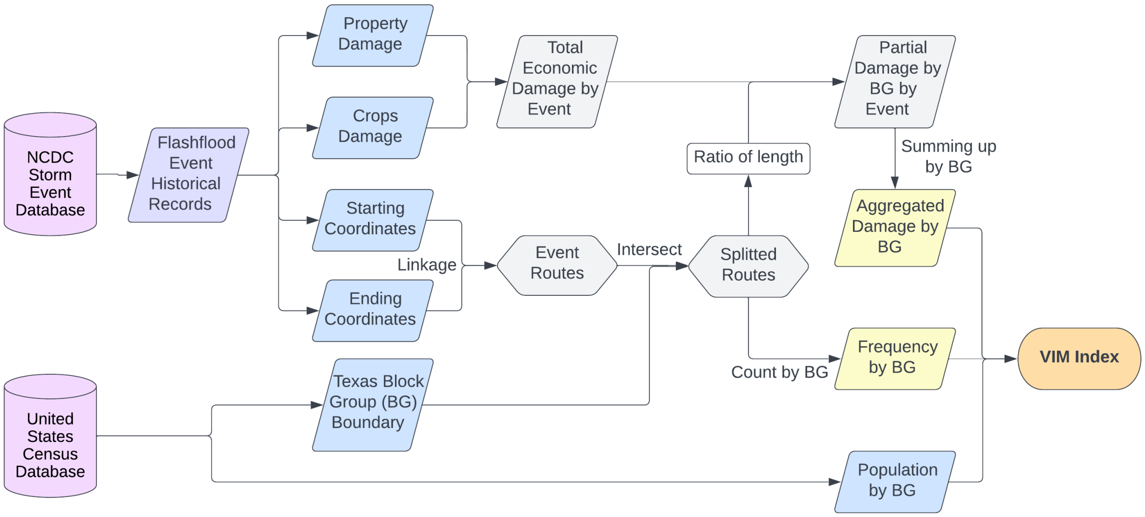

Figure 3.1: Workflow for VIM Index Calculation at Block Group Level.

3.1.2 VIM Index at County Level

Probably because county is a much larger social level and a more frequently used unit for various studies, there has been an aggregated database for natural disasters at county level provided by SHELDUS. We derived the flash flood record during 2011 to 2020 from this database for the VIM Index quantification at county level.

The damage data in SHELDUS database include six categories (CEMHS 2022):

1. CropDmg: Damage to crop in U.S. dollars (current year).

2. CropDmg(Adj): Damage to crop in adjusted U.S. dollars (selected base year).

3. CropDmgPerCapita: Damage to crop in adjusted U.S. dollars (base: 2015) divided by the annual county population; per capita calculations are based on current population.

4. PropertyDmg: Damage to property in U.S. dollars (current year).

5. PropertyDmg(Adj): Damage to property in adjusted U.S. dollars (selected base year).

6. PropertyDmgPerCapita: Damage to property in adjusted U.S. dollars (base: 2015) divided by the annual county population; per capita calculations are based on current population. current population.

In this research, we are going to use the CropDmg(Adj) and PropertyDmg(Adj) for damage calculation. Since the records in this database are already aggregated by month and county, we can simply achieve the aggregated damage by county (AD_C) with the following steps: (1) summing up the CropDmg(Adj) and PropertyDmg(Adj) values for each record as total damage (by month by county), (2) group the records by county, and (3) sum up all total damages for each county.

Meanwhile, the Records field in this dataset represents the frequency of flash flood events in the given month and given county, so when we group the records by county we can also sum up the Records field to achieve the frequency by county (F_C). Similar to the methods for block group level analysis, the raw VIM index at county level ($RVIM_C$) could also be calculated using the equation below:

$$RVIM_C = \frac{AD_C}{P_C * F_C}$$

And this value also needs to be standardized to finally achieve the VIM index at county level. The whole workflow for VIM index calculation at county level is shown in figure 3.2.

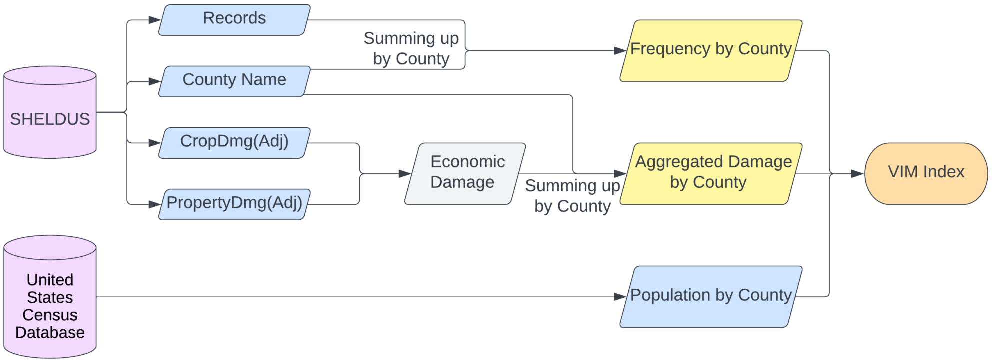

Figure 3.2: Workflow for VIM Index Calculation at County Level.

3.2 Contributor Exploration

After concluding the VIM index, we need to identify the contributors, or in other words, what factors may lead to high vulnerability. We have two categories of factors: socio-economic variables and terrain conditions. Using both kinds of factors and run correlation analysis between those variables and the concluded VIM index, we could identify the potential contributors behind.

The regression analysis would be the major method for identifying the contributors. In this study we utilized three models for regression, including Linear Regression (LR) model, Multinomial Logistic Regression (MLR) model, and Random Forest (RF) model.

We've concluded VIM index to represent the level of vulnerability in previous modules, but that is a continuous parameter, which cannot be used for MLR model and is not very suitable for RF model (the model would work but the performance would be very bad). Therefore, we categorized the VIM index before we ran MLR and RF model. In addition, for LR model and MLR model, Stepwise Selection (SS) was used to determine the best combination of variables which could lead to the best model performance.

3.2.1 Linear Regression (LR) Model

When there is a continuous dependent variable and a series of predictors, LR Model, which builds up a linear combination of those variables, is the simplest way to establish the regression function (Su, Yan, and Tsai 2012). In our research, we aim to explore the potential correlation or association between our concluded VIM index, which is a continuous variable, and a list of potential contributors. Therefore, LR model is one of the most ideal methods to provide the most straightforward result for our analysis. By checking the function of finally concluded LR model, we can identify how those variables are correlated, or associated with the vulnerability of the society to flash flood.

3.2.2 Multinomial Logistic Regression (MLR) Model

LR model is a very functional method to explore the correlation between explanatories and response, but for vulnerability to flash flood, which is very complex, such a model established based on simply linear correlation could not be sufficient to fully explore the relations and associations behind, and the use of continuous index might be over specific and could make the model unable to reach a good result. Therefore, Logistic Regression is introduced in this study. By dividing the continuous index into categories, we could generalize the response variable and thus enable the model to identify the correlation between explanatory and response. On the other hand, if the response variable is too general, there could be too much information be eliminated or missed. Therefore, to find a balance between specificity and generality, we decided to classify the vulnerability into three categories (will be detailly discussed in section 3.2.4) instead of two categories (which is more commonly used in Logistic Regression model). Because of this, we need to use MLR model, which is a "maximum likelihood estimator" for study involves polychotomous dependent variable (Kwak and Clayton-Matthews 2002), to identify the contributors.

3.2.3 Random Forest (RF) model

Random forest is a more advanced learning algorithm which is able to identify the prediction rules, or in other words the correlation and association, behind the response variable and a series of explanatory variables, and rank all explanatories based on their importance in the final model (Boulesteix et al. 2012). Given the complexity of vulnerability, RF could be a very powerful tool to explore the contributors. Similar to MLR model, RF may not perform very well on dealing with continuous numeric values, so we first need to use the classified vulnerability categories for RF model as well. In addition, the RF model cannot use Stepwise to select variables, so we ran it with two sets of variables. The first variable list is the same as the variable list used in MLR model, which is selected using SS (RF1 model); the other variable list is the whole list of all 19 variables (RF2 model).

3.2.4 Vulnerability Classification

We've concluded a vulnerability index, which can represent the level of vulnerability. However, this is a continuous variable, which would make it difficult for model to have very good performance. Therefore, we classified the vulnerability of each area unit into three ordinal categories, “Low”, “Median”, and “High”, depending on their VIM index: if the VIM of a given area unit is over 0.5 standard deviation higher than the mean VIM, it is classified as "High"; if the VIM in a given area unit is below 0.5 standard deviation lower than the mean VIM, it is classified as "Low"; if the VIM of a given area unit is within 0.5 standard deviation from the mean VIM, it is classified as "Medium" (as shown in Figure 3.3).

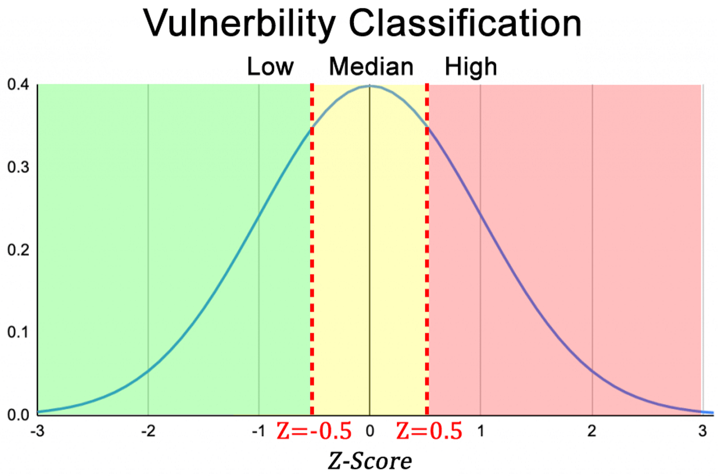

Figure 3.3: Vulnerability Classification.

3.2.5 Stepwise Selection (SS) Algorithm

For LR model and MLR model, we do have a list of variables that are selected based on both existing analysis and intercorrelation test, yet it doesn’t mean that using all those variables in the model could lead to the best results. Instead, just like many other regression models, the selection of subset of explanatories could significantly influence the performance of our models. Therefore, we introduced SS Algorithm in this study.

4. RESULTS AND DISCUSSION

4.1 VIM Index result

4.1.1 VIM at Block Group Level

Using the methods introduced in previous chapter, we first concluded the raw VIM index. As shown in figure 4.1, from the histogram on the left we can find that the raw VIM values are extremely right skewed and we need to convert it into normal distribution. After attempts we find that using a logarithm transformation as shown below could standardize it and generate values ranging from 0 to 8 for this data set. A histogram for the normalized VIM is also included in figure 4.1.

$$VIM_{BG} = log_{10}(RVIM_{BG} * 10000 + 1)$$

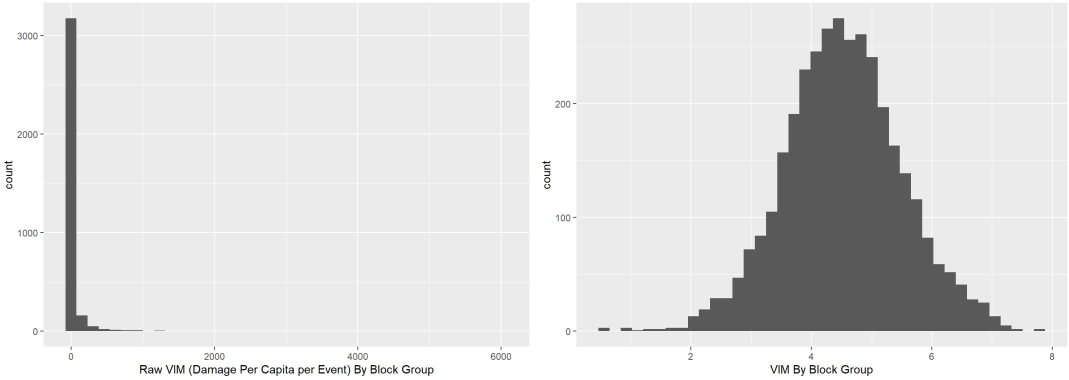

Figure 4.1: Histograms for Raw VIM Index (Left) and Normalized VIM (Right) by Block Group.

The spatial distribution of VIM in Texas and block group level is shown in figure 4.2. Due to the large proportion of area units with no flash flood records, there are large blank areas on this map, which makes the overall pattern hard to tell. Especially for some of the east areas of Texas, where the block groups are very small around big cities like Houston and Dallas, such discontinuousness makes it very hard to tell what average level of vulnerability is in those places. On the other hand, we can still use the resulting data, including both $RVIM$ and VIM, to identify the influence from different contributors in subsequent analysis.

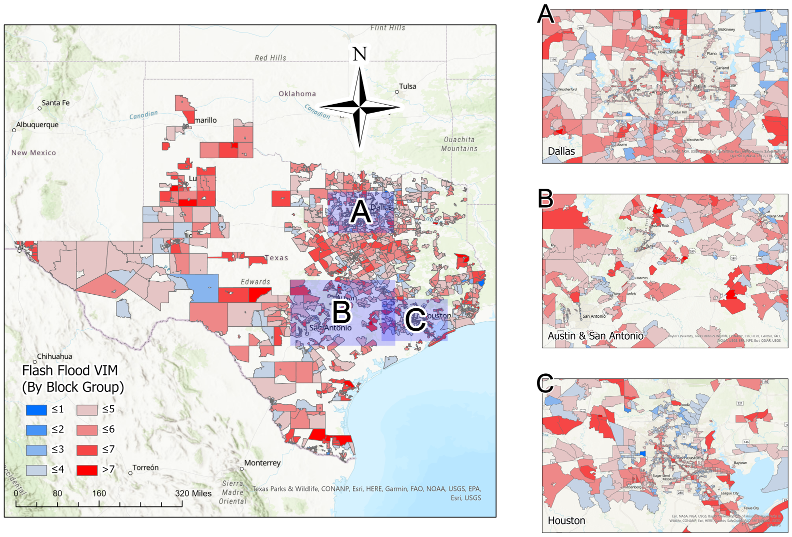

Figure 4.2: Spatial Distribution of VIM Index at BG Level (A. Dallas; B. Austin and San Antonio; C. Houston)

4.1.2 VIM at County Level

Similarly, we have concluded raw VIM by county with extremely right skewed distribution (as shown in figure 4.3). Using an equation shown below we can standardize the indices into values ranging from 0 to 8.

$$VIM_C = log_{10}(RVIM_C * 10000 + 1)$$

However, this time the resulting VIM are still not in a perfect normal distribution (also shown in figure 4.3). Instead, not only it is slightly right-skewed, but also the overall shape formed by the histograms is very unsmooth.

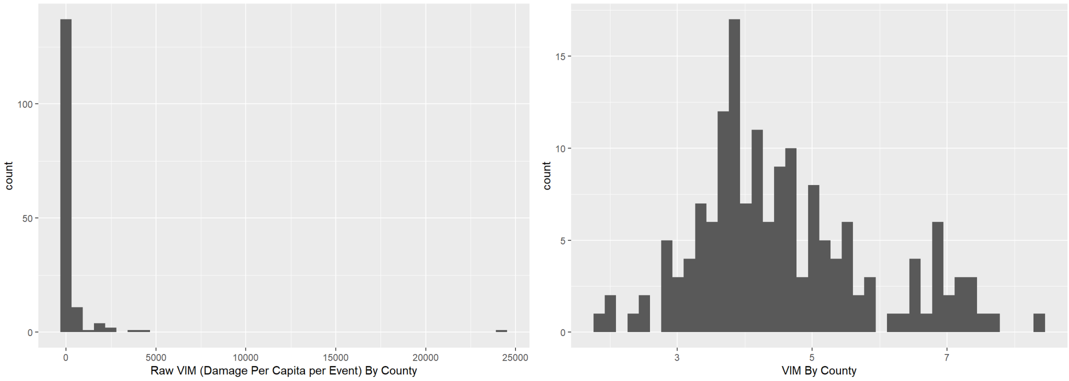

Figure 4.3: Histograms for Raw VIM Index (Left) and Normalized VIM (Right) by County.

On the other hand, the map for VIM by county (figure 4.4) provided us a much better representation of vulnerability's spatial pattern than the maps at block group level. Although the gaps still exist, the overall distribution could be observed more directly. The east area of Texas, especially Houston and its surrounding areas, have obviously higher vulnerability than most other counties of Texas, meaning these counties would more likely to experience high damage when flash floods take place there.

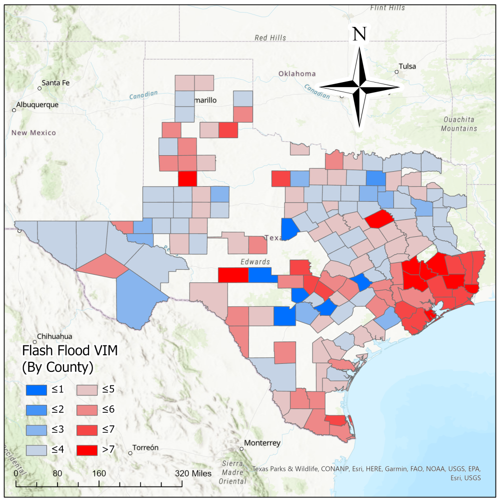

Figure 4.4: Spatial Distribution of VIM Index at County Level.

4.2 Vulnerability Contributors Exploration

4.2.1 Model Performance

The variable selection and performance of all three models for block group level is shown in Table 4.1, and for county level is shown in Table 4.2. For the analysis of each social scale, we used 2 combinations of variables for RF model, that the first one is the same variable list used in MLR model selected via Stepwise, and the second one is the combination of all 19 variables in the list.

Model Name	Algorithm	Variables in final model	Model performance
LR	Linear Regression	MED_HSHD_INCOME PCT_Under_EDU PCT_Service PCT_ ASIAN PCT_AFRICAN_AMERICAN PCT_UNDER18 PCT_MOBILE_HOMES PCT_UNOCCUPIED_HOUSING MEAN_Impervious	Adjusted R^2 = 0.06 Cor = 0.25
MLR	Multinomial Logistic Regression	MED_HSHD_INCOME PCT_Under_EDU PCT_Extractive PCT_Service PCT_ASIAN PCT_AFRICAN_AMERICAN PCT_UNDER18 PCT_MOBILE_HOMES PCT_UNOCCUPIED_HOUSING MEAN_Impervious	Accuracy = 42.78%
RF1	Random Forest	Same Variable List as MLR Model	OOB estimate of error rate: 42.18% (Accuracy = 57.82%)
RF2	Random Forest	Whole Variable List	OOB estimate of error rate: 41.91% (Accuracy = 58.09%)

Table 4.1: Model Summary for BG Level Analysis.

Model Name	Algorithm	Variables in final model	Model performance
LR	Linear Regression	PCT_Under_EDU PCT_Service UNEMPLOYMENT_RATE PCT_ASIAN PCT_FEMALE PCT_FEMALE_LABOR MEAN_SLP MEAN_Impervious	Adjusted R^2 = 0.11 Cor = 0.39
MLR	Multinomial Logistic Regression	PCT_AFRICAN_AMERICAN PCT_KID_OLD PCT_FEMALE PCT_HOUSING_NO_CAR PCT_UNOCCUPIED_HOUSING MEAN_Impervious	Accuracy = 53.16%
RF1	Random Forest	Same Variable List as MLR Model	OOB estimate of error rate: 41.15% (Accuracy = 58.85%)
RF2	Random Forest	Whole Variable List	OOB estimate of error rate: 43.19% (Accuracy = 56.81%)

Table 4.2: Model Summary for County Level Analysis.

For linear correlation modeling for the continuous VIM index, LR model provided a function with 6.05% variance explained (which is represented by adjusted R^2) for BG level analysis using 9 variables, and 10.8% variance explained for county level analysis using 8 variables. For both model, percentage of population aged 25 years or older with less than 12th grade education, percentage employment in service industry, percentage Asian, as well as average impervious rate, are selected as the explanatories

Among the three model for categories vulnerability, RF model performs better than MLR model at both BG level and county level. For RF model using same variable list as the MLR model, it reached the accuracy of 57.82% at BG level using 10 variables, and 58.85% at county level using 6 variables; for RF model using the whole variable list, it reached the accuracy of 58.09% at BG level, and 56.81% at county level; For MLR model, it only achieved accuracy of 42.78% at BG level using 10 variables, and 53.16% at county level using 6 variables.

In addition, for LR, MLR, and RF using SS variable list, the county level models perform better than BG level model, while for RF using the whole variable list, the BG level model performs better.

4.2.2 Interpretation of Correlation and Association

In total we have 8 different models with different selections of variables, and even the same variables could appear to have different types of correlation/association with the vulnerability. To identify the consistency of each variable among different model, we classify the correlation/association into 6 classes based on our observations (shown in Table 4.3). Since some of those correlations can only be observed in certain models due to the different complexity of them, we rank them into two levels. Level 1 means in can be observed in all models, and level 2 means it can only be observed in MLR and RF models.

Class	Description	Level
Positive	When the given variable **increases**, the vulnerability tends to **increases**; when the give variable **decreases**, the vulnerability tend to **decreases**.	1
Negative	When the given variable **increases**, the vulnerability tends to **decreases**; when the give variable **decreases**, the vulnerability tends to **increases**.	1
Divergent	When the given variable **increases**, probability of **both high and low** vulnerability increases; when the give variable **decreases**, probability of **median** vulnerability increases.	2
Convergent	When the given variable **decreases**, probability of **both high and low** vulnerability increases; when the give variable **increases**, probability of **median** vulnerability increases.	2
Concave-up/ upwards	When the given variable **increases**, the vulnerability **decreases at first**; after this variable reaches a certain point and **continues increasing**, the vulnerability **turns to increase**.	2
Concave-down/ downwards	When the given variable **increases**, the vulnerability **increases at first**; after this variable reaches a certain point and **continues increasing**, the vulnerability **turns to decrease**.	2

Table 4.3: Classification of Correlation or Association.

4.2.2.1 LR Model

The performance of the LR model is shown in Figure 4.5. We can tell that, for both BG level model and county level model, the estimated VIM index falls in a much narrower range than the actual VIM index we observed. For BG level model most estimated VIM falls within 4 and 5, and for county level model most estimated VIM falls within 3.5 and 5.5, while actually the VIM ranges from 0 to 8 for both social scales. This is a more direct representation of the low value of adjusted R2, and indicates that the LR model itself is very limited on predicting the VIM index using the given variables in this research.

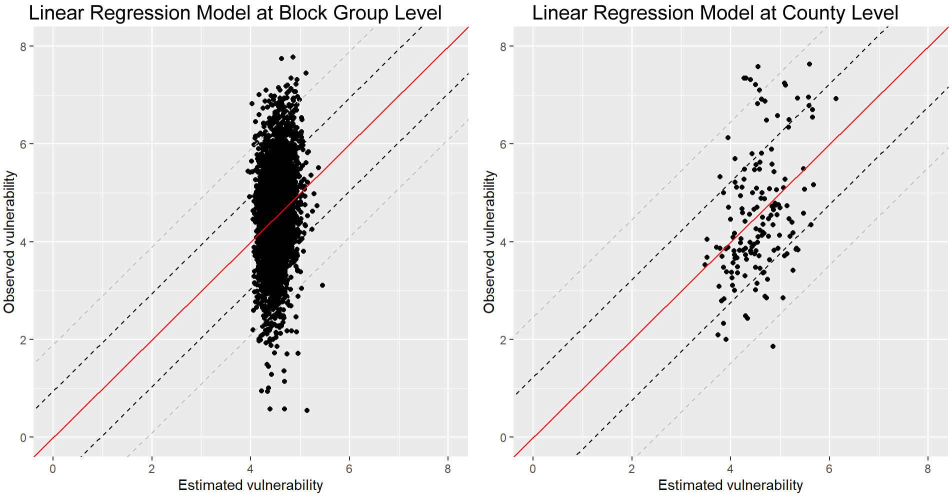

Figure 4.5: Estimated VIM vs. Observed VIM of LR Model at BG Level (left) and County Level (Right).

The coefficients for LR model at BG level are shown in Table 4.4. In this model, 6 variables have positive correlation with vulnerability, including “MED_HSHD_INCOME”, “PCT_Under_EDU”, “PCT_Service”, “PCT_ASIAN”, “PCT_MOBILE_HOMES”, and “PCT_UOCCUPIED_HOUSING”. On the other hand, “PCT_AFRICAN_AMERICAN”, “PCT_UNDER18”, and “MEAN_Impervious” has negative correlation with vulnerability.

The coefficients for LR model at county level are shown in Table 4.5. “PCT_Service”, “UN-

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.527e+00	1.251e-01	36.182	< 2e-16
MED_HSHD_INCOME	2.547e-06	7.532e-07	3.381	0.000731
PCT_Under_EDU	3.836e-03	1.908e-03	2.010	0.044506
PCT_Service	1.189e-02	2.303e-03	5.163	2.56e-07
PCT_ASIAN	5.207e-03	2.607e-03	1.997	0.045890
PCT_AFRICAN_AMERICAN	-3.163e-03	1.209e-03	-2.617	0.008922
PCT_UNDER18	-1.212e-02	3.129e-03	-3.873	0.000109
PCT_MOBILE_HOMES	5.587e-03	1.327e-03	4.212	2.60e-05
PCT_UOCCUPIED_HOUSING	7.189e-03	1.598e-03	4.497	7.11e-06
MEAN_Impervious	-6.385e-03	1.143e-03	-5.584	2.53e-08

Table 4.4: Coefficients for LR model at BG level.

EMPLOYMENT_RATE", "PCT_ASIAN", and "PCT_FEMALE" have positive correlation with the vulnerability. This implies that the low level of education, high occupation in service industry, high percentage of Asian population, and high percentage of female population, could lead to, or associate with high vulnerability. On the other hand, "PCT_Under_EDU", "PCT_FEMALE_LABOR", "MEAN_Slope", and "MEAN_Impervious" have negative correlation with vulnerability, which indicates that high percentage of population without enough education (low level of education), high percentage of female labor, dramatic rise or fall of the land surface (slope), as well as high level of impervious rate could associate with low vulnerability.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.93153	2.80724	1.044	0.29805
PCT_Under_EDU	-0.03748	0.01535	-2.442	0.01576
PCT_Service	0.06630	0.03142	2.110	0.03653
UNEMPLOYMENT_RATE	6.23681	2.07475	3.006	0.00311
PCT_ASIAN	0.14999	0.05723	2.621	0.00968
PCT_FEMALE	0.06406	0.04371	1.466	0.14484
PCT_FEMALE_LABOR	-0.08567	0.03953	-2.167	0.03180
MEAN_SLP	-2.33962	1.32520	-1.765	0.07953
MEAN_Impervious	-0.02555	0.01251	-2.042	0.04289

Table 4.5: Coefficients for LR model at County level.

The inconsistency between the two models is at the correlation between education and vulnerability. In BG model, the high percentage of population without enough education, or in other words, the low level of education, in a certain area are more likely to associate with high vulnerability to flash flood, but in the county model, areas with low levels of education are more likely to have low vulnerability.

In addition, by comparing the p-values in the two models, it's easily to find that all variables in BG-level LR model have relatively important correlation with VIM ($p < 0.05$), yet the correlation of some variables in county-level LR model is not important enough ($p > 0.05$).

To provide a more thorough understanding of the correlation of each variable to VIM index, a complete table for the results of Pearson's correlation test for all variables in table 2.1 is provided (Table 4.6).

Variable	p-value (BG)	cor (BG)	p-value (CN)	cor (CN)
MED_HSHD_INCOME	0.0101 *	0.0437	0.2309	-0.0958
PCT_Under_EDU	0.7865	-0.0046	0.6262	-0.0390
PCT_Extractive	0.0694	0.0309	0.9659	0.0034
PCT_Service	6.64e-10 ***	0.1047	0.2806	0.0864
UNEMPLOYMENT_RATE	0.2623	-0.0191	0.0037 **	0.2294
PCT_ASIAN	0.8748	-0.0027	0.9481	0.0052
PCT_AFRICAN_AMERICAN	2.34e-06 ***	-0.0802	0.423	0.0642
MEDIAN_AGE	0.5883	0.0092	0.03787 *	0.1654
PCT_KID_OLD	2.26e-06 ***	0.0803	0.1840	0.1062
PCT_UNDER18	3.42e-07 ***	-0.0865	0.0633	-0.1481
PCT_FEMALE	0.9066	0.0020	0.2332	-0.0954
PCT_FEMALE_LABOR	0.8239	0.0038	0.5009	-0.0539
PCT_FEMALE_HSHD	0.0019 **	-0.0529	0.9713	0.0029
HSHD_SIZE	0.0206 *	-0.0394	0.5127	-0.0524
PCT_MOBILE_HOMES	< 2.2e-16 ***	0.1457	0.0516	0.1552
PCT_HOUSING_NO_CAR	0.0036 **	-0.0495	0.8230	-0.0179
PCT_RENTER	9.81e-10 ***	-0.1036	0.0187 *	-0.1869
PCT_UNOCCUPIED_HOUSING	1.16e-12 ***	0.1204	0.0059 **	0.2181
MEAN_SLP	0.0058 **	0.0469	0.2708	-0.0881
MEAN_Impervious	< 2.2e-16 ***	-0.1704	0.0438 *	-0.1607

Table 4.6: Pearson's Correlation Test for the Whole Variable List (the number of asterisks indicates the level of significance of the correlation).

4.2.2.2 MLR model

The effect of each variable selected via SS on the MLR model at BG level is shown in figure 4.6. We can find that: for “MED_HSHD_INCOME”, “PCT_Under_EDU”, “PCT_Service”, “PCT_ASIAN”, “PCT_MOBILE_HOMES”, and “PCT_UOCCUPIED_HOUSING”, the correlation/association with vulnerability is positive; for “PCT_UNDER18”, “MEAN_Impervious”, and “PCT_AFRICAN_AMERICAN”, the correlation/association is negative correlation or association with vulnerability; for “PCT_Extractive”, the correlation/association is divergent.

It is markable that for “PCT_AFRICAN_AMERICAN”, its correlation with vulnerability is hard to tell since it has insignificant influence on the probability of low vulnerability; yet when it increases, the probability of median vulnerability will increase and the probability of high vulnerability decreases, which should also be considered as a kind of negative correlation or association.

The effect of each variable selected via SS on the MLR model at county level is shown in figure 4.7. We can find that: for “PCT_AFRICAN_AMERICAN” and “PCT_UNOCCUPIED_HOUSING”, the correlation/association with vulnerability is positive; for “PCT_HOUSING_NO_CAR” and “MEAN_Impervious”, the correlation/association is negative; for “PCT_KID_OLD”, the correlation/association is convergent; for “PCT_FEMALE”, the correlation/association is divergent.

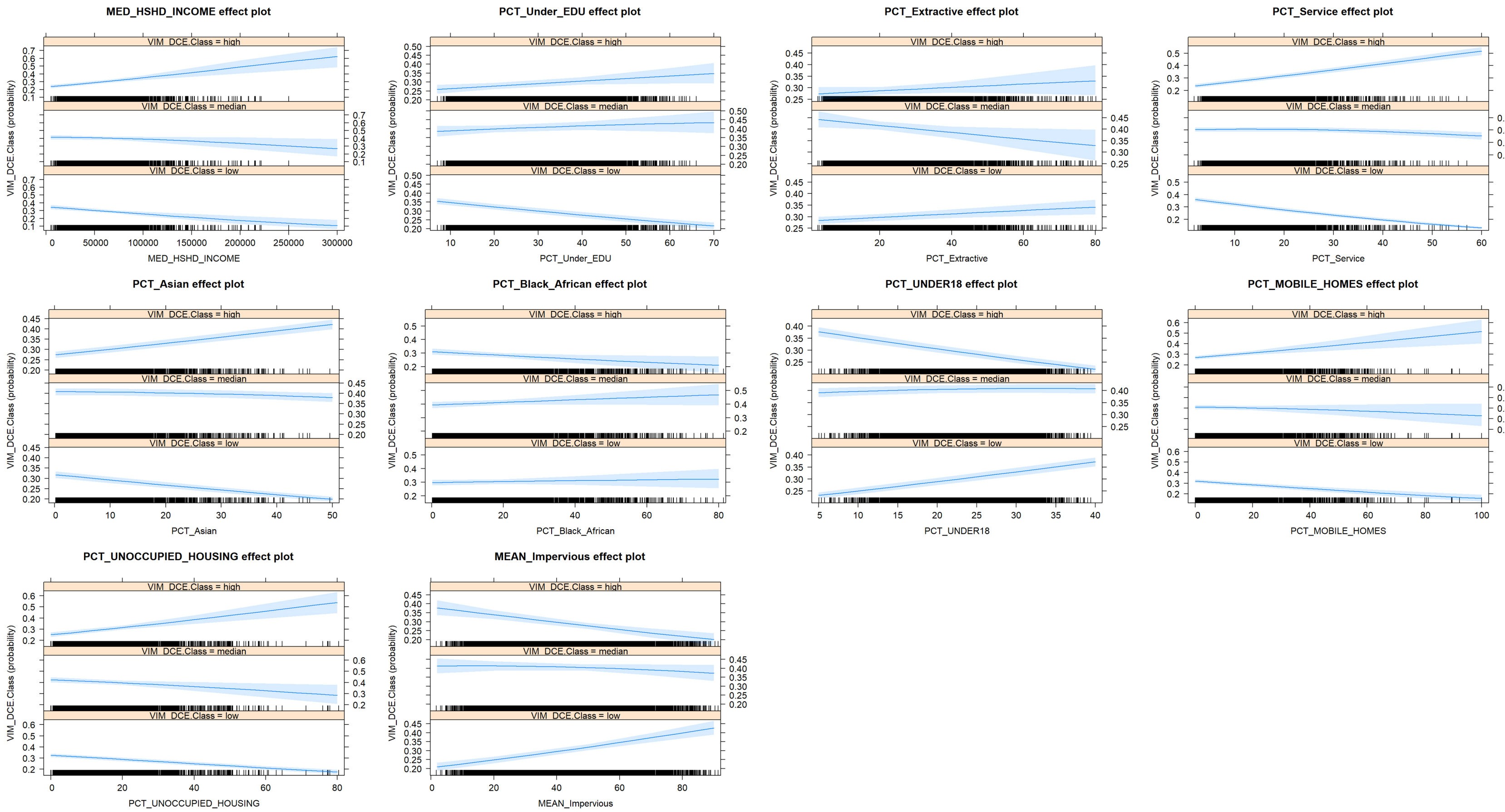

Figure 4.6: Effect of each selected variable on MLR model at BG level.

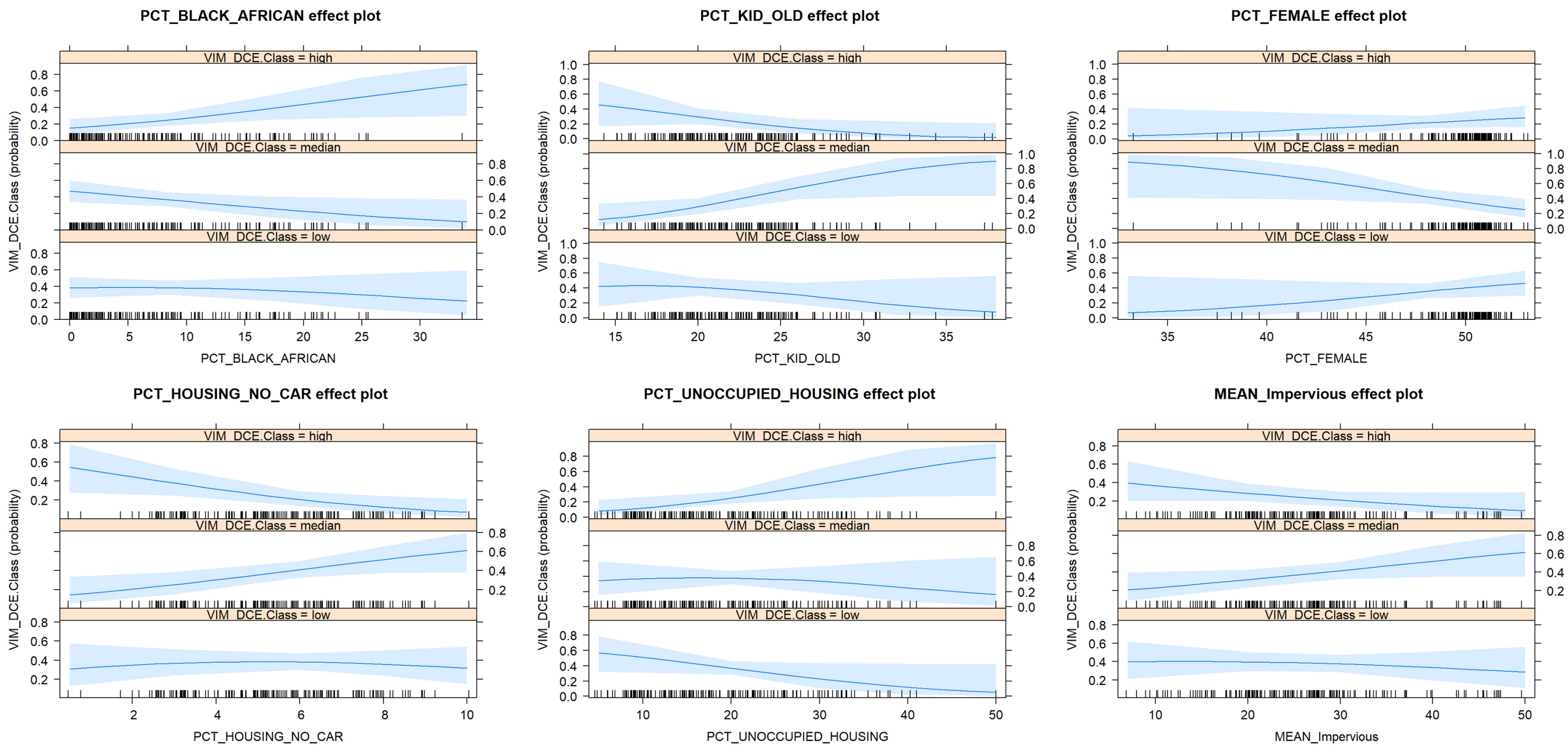

Figure 4.7: Effect of each selected variable on MLR model at County level.

4.2.2.3 RF1 Model

The partial dependence of each selected variable in the BG-level RF1 model is shown in Figure 4.8. For “PCT_Service”, “PCT_ASIAN”, “MED_HSHS_INCOME”, “PCT_UNOCCUPIED_ HOUSING”, and “PCT_MOBILE_HOMES”, the correlation/association with vulnerability is positive; for “PCT_AFRICAN_AMERICAN” and “PCT_UNDER18” , the correlation/association is negative; and for “PCT_Extractive”, the correlation is divergent; for “MEAN_Impervious”, the correlation/association appears to be concave-up, and the peak is at around MEAN_Impervious = 50; similarly, the correlation/association of “PCT_UNDER_EDU” with vulnerability is also comcave-up, the peak is at around PCT_UNDER_EDU = 20, and after it gets over 30, the vulnerability don’t have significant change anymore.

The partial dependence of each selected variable in the county-level RF1 model is shown in Figure 4.9. For “PCT_AFRICAN_AMERICAN” and “PCT_UNOCCUPIED_HOUSING”, the correlation/association with vulnerability appears to be positive; for “PCT_FEMALE”, the correlation/association is negative; for “PCT_KID_OLD” and “PCT_HOUSING_NO_CAR”, the correlation/association is converngent; and for “MEAN_Impervious”, being consistent with BG level RF1 model, the correlation/association is concave-up, while the peak is at around MEAN_Impervious = 30.

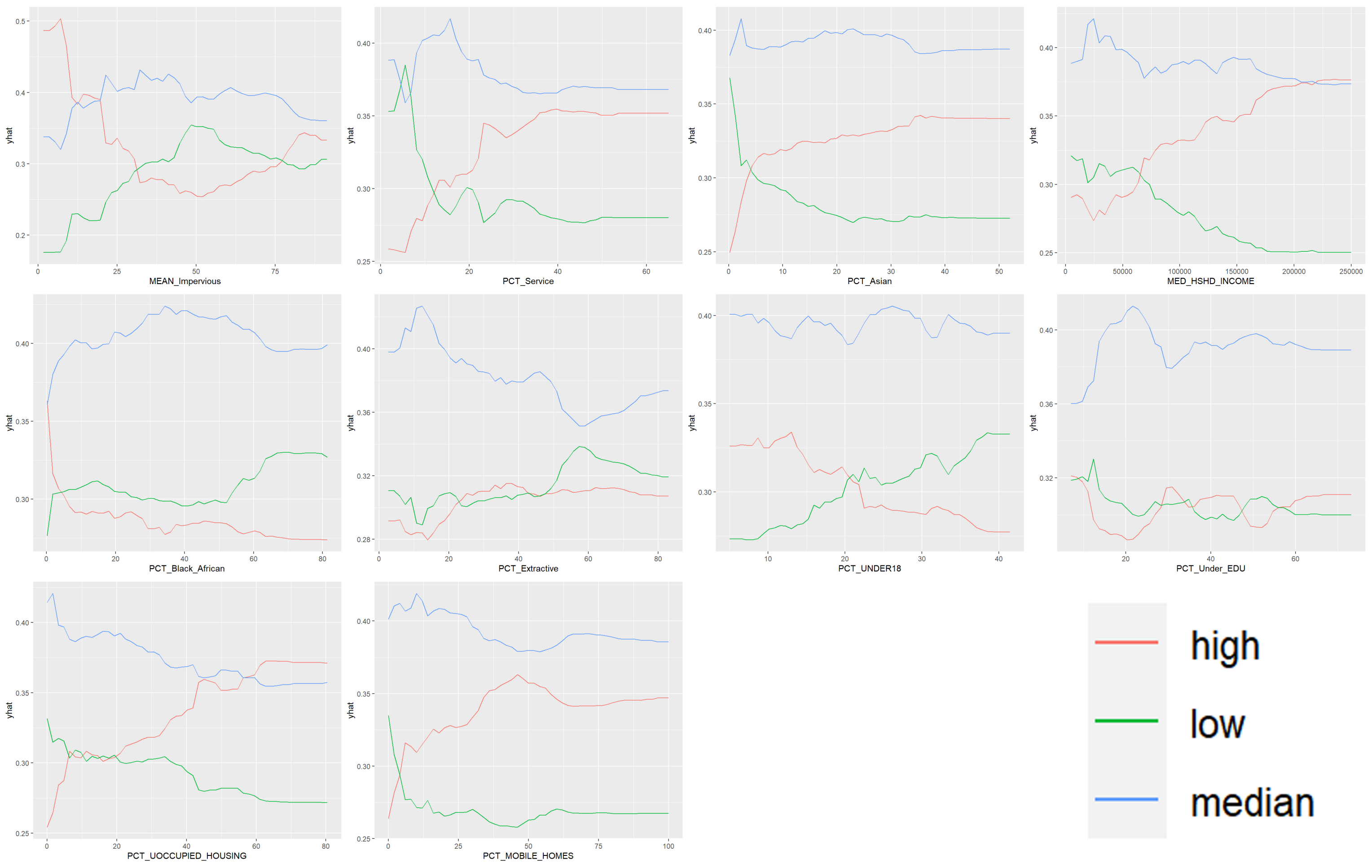

Figure 4.8: Dependence of each variable in BG-level RF1 model.

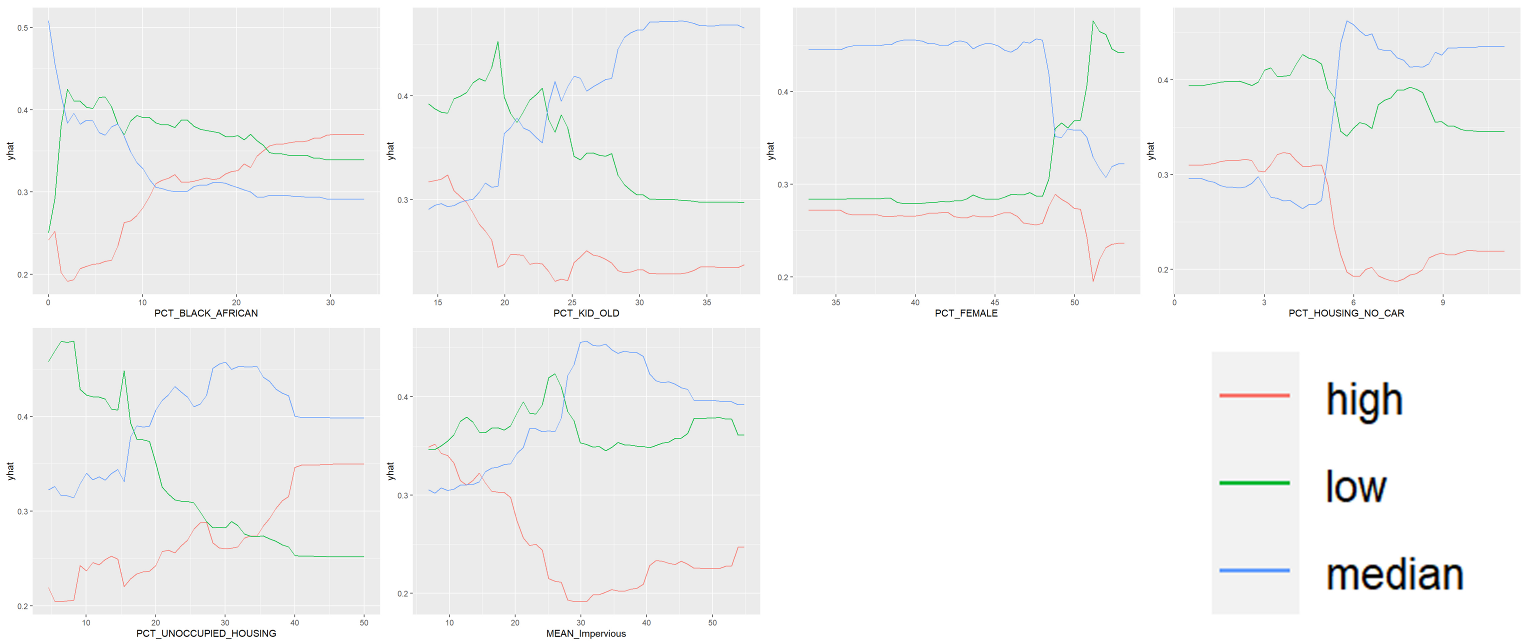

Figure 4.9: Dependence of each variable in county-level RF1 model.

4.2.2.4 RF2 Model

RF2 model uses the whole list of 19 variables. To better focus on the variables that are highly correlated with vulnerability, here we only represent the partial dependence chart for the 10 variables with the highest importance in each model and treat them as the selected variables. The partial dependence of each selected variable in the BG-level RF2 model is shown in Figure 4.10. For “PCT_Service”, “PCT_ASIAN”, “PCT_KID_OLD”, and “MED_HSHD_INCOME”, the correlation/association with vulnerability is positive; for “PCT_AFRICAN_AMERICAN” and “PCT_UNDER18”, the correlation/association is negative; for "PCT_Extractive”, being consistent with the BG-level MLR and RF1 models, the correlation/association is divergent; similar to RF1 model, for “MEAN_Impervous”, the correlation/association is concave-up with peak at around MEAN_Impervious = 50; “MEAM_SLP” and “UNEMPLOYMENT_RATE” don’t have significant influence to the probability of high vulnerability, yet they form concave-down curves for medium vulnerability and concave-up curves for low vulnerability, so the overall pattern of the vulnerability should also be concave-down, and the peak is at around MEAN_SLP = 0.2 and UNEMPLOYMENT_RATE = 15, respectively.

The partial dependence of each selected variable in the county-level RF2 model is shown in Figure 4.11. For “PCT_UNOCCUPIED_HOUSING”, “PCT_AFRICAN_AMERICAN”, and “UNEMPLOYMENT_RATE”, the correlation/association with vulnerability is positive; for “MEAN_Impervious”, “PCT_HOUSING_NO_CAR”, “PCT_FEMALE”, “PCT_RENTER”, and “MEAN_SLP”, the correlation/association is negative; for “PCT_KID_OLD”, the correlation/association is convergent; for “PCT_Extractive”, the correlation/association is concave-down.

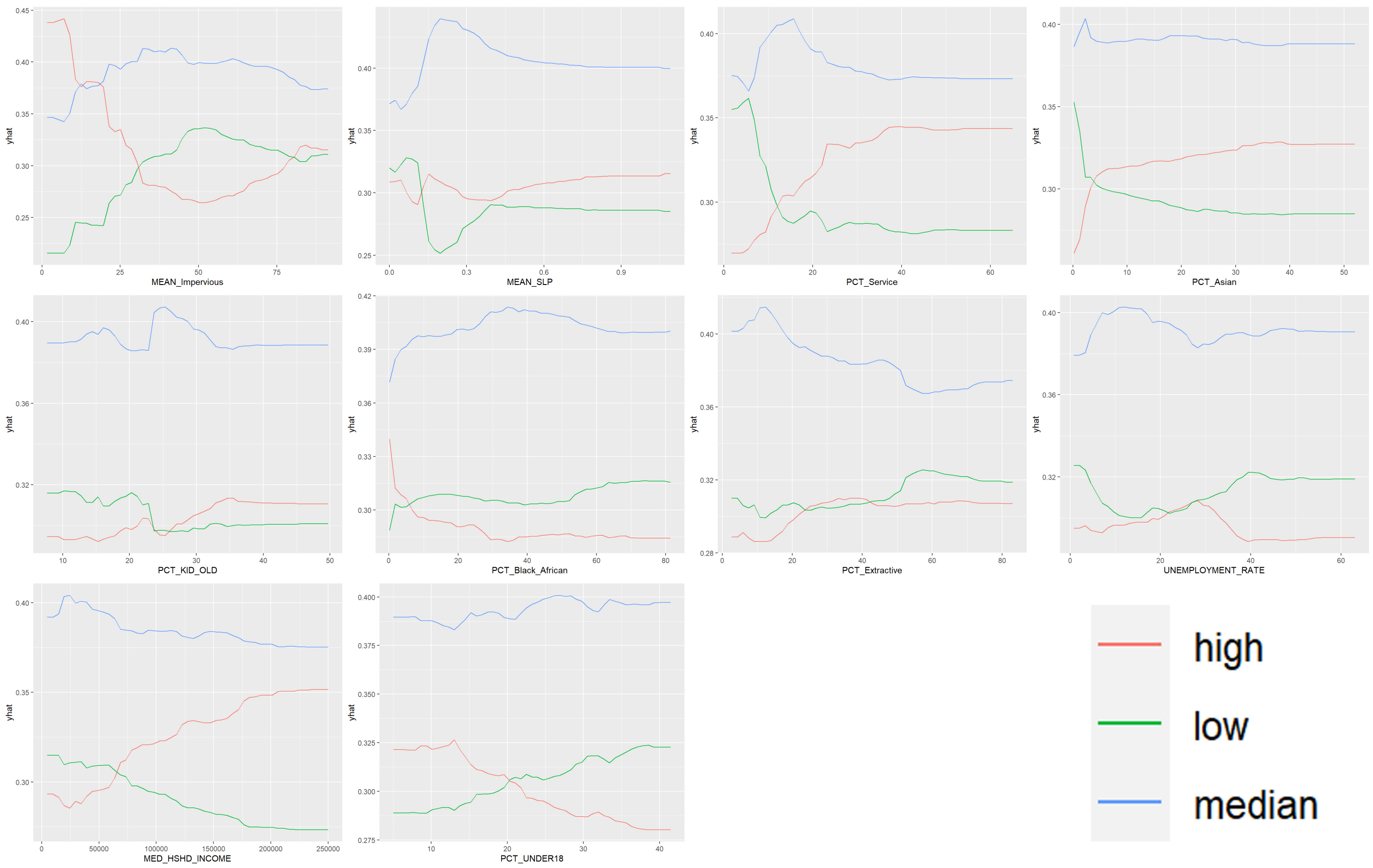

Figure 4.10: Dependence of each variable in BG-level RF2 model.

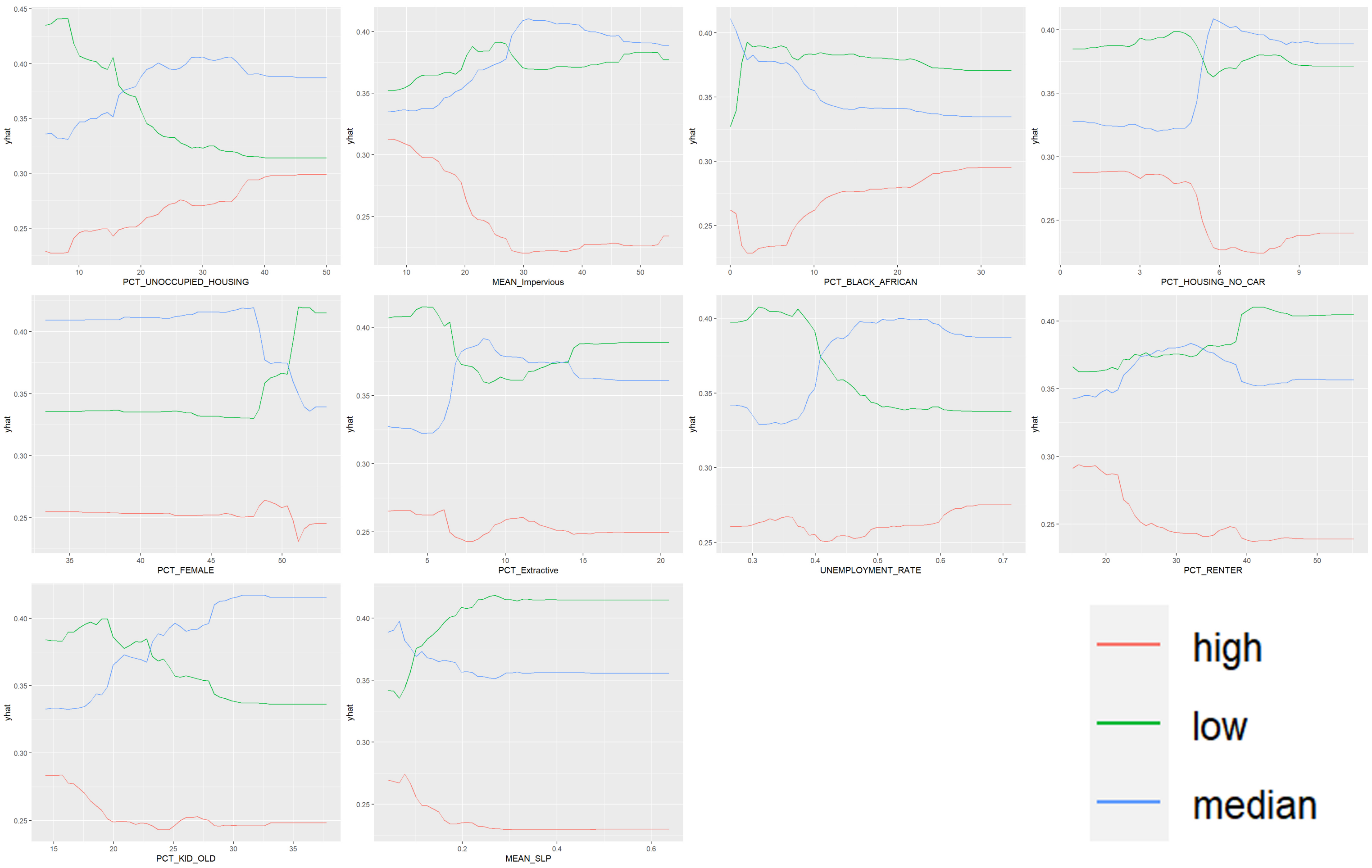

Figure 4.11: Dependence of each variable in county-level RF2 model.

4.2.2.5 Summary of Correlation or Association

A summary of all detected potential correlation/association is shown in Table 4.7. Among these models, there are 6 sets of variable selections (there are 8 models, while RF1 and MLR model share the same set of variables; also, for RF2 models we only consider the 10 variables with highest importance as the selected variables). We consider the variables appear in 3 or more of those variable selections and have consistent correlation/association with vulnerability as the confident variables, and they include (shown in Table 4.6): MED_HSHD_INCOME (positive), PCT_Service (positive), PCT_ASIAN (positive), PCT_UNDER18 (negative), and PCT_UNOCCUPIED_HOUSING (positive). Based on those observations we can conclude that the potential contributors or associated variables of high vulnerability in the given area includes: (1) high level of (household) income, (2) high percentage of employment in service industry, (3) high percentage of Asian in the whole population, (4) low proportion of population of under 18 years old, and (5) large proportion of unoccupied housing.

Besides, MEAN_Impervious are included in all variable selections, but inconsistency exists among the models, that in 5 models it has negative correlation/association with vulnerability, while in the other 3 models it has concave-up correlation/association. However, by observing the effect and partial dependence figures of this variable, we noticed that in the models it has concave-up correlation/association, the highest vulnerability in the increasing phase is lower than in the decreasing phase, so the overall trend is still decreasing. Therefore, we can assume that such inconsistency is mainly because the limitation of our models in exploring such complex correlation, and we still include MEAN_Impervious as a confident contributor that has concave-up correlation with vulnerability, which means either too low or too high impervious rate could associate with high vulnerability.

In addition, PCT_AFRICAN_AMERICAN appeared frequently in the resulted models, yet its influence to vulnerability is inconsistent at different level: in BG-level models it appeared to have negative correlation, while in county-level models it appeared to have positive correlation. By checking the dependence charts of the RF models we noticed that this variable only influence vulnerability when the value is relatively low, and have less influence to vulnerability after it reaches 10%, which might be a possible reason for the inconsistency at different social scales. Whatever the reason is, such an inconsistency makes it invalid to be used as an indicator for vulnerability.

Model	LR (BG)	LR (CN)	MLR (BG)	MLR (CN)	RF1 (BG)	RF1 (CN)	RF2 (BG)	RF2 (CN)
MED_HSHD_INCOME	Positive		Positive		Positive		Positive	
PCT_Under_EDU	Positive	Negative	Positive		Upwards			
PCT_Extractive			Divergent		Divergent		Divergent	Downwards
PCT_Service	Positive	Positive	Positive		Positive		Positive	
UNEMPLOYMENT_RATE		Positive					Downwards	Positive
PCT_ASIAN	Positive	Positive	Positive		Positive		Positive	
PCT_AFRICAN_AMERICAN	Negative		Negative	Positive	Negative	Positive	Negative	Positive
MEDIAN_AGE								
PCT_KID_OLD				Negative		Convergent	Positive	Convergent
PCT_UNDER18	Negative		Negative		Negative		Negative	
PCT_FEMALE		Positive		Convergent		Negative		Negative
PCT_FEMALE_LABOR		Negative						
HSHD_SIZE								
PCT_MOBILE_HOMES	Positive		Positive		Positive			
PCT_HOUSING_NO_CAR				Negative		Convergent		Negative
PCT_RENTER								Negative
PCT_UNOCCUPIED_HOUSING	Positive		Positive	Positive	Positive	Positive		Positive
MEAN_SLP		Negative					Downwards	Negative
MEAN_Impervious	Negative	Negative	Negative	Negative	Upwards	Upwards	Upwards	Negative

Table 4.7: Coefficients for LR model at County level.

5. CONCLUSION

In this research, a VIM framework was established and it uses historical flash flood records as the main input data, combined with administrative boundary data and population data, to calculate an index that can be used to represent the level of vulnerability to flash flood in the given area.

In the case study in Texas, although the block-group level analysis failed to provide a straightforward visualization of the spatial distribution of vulnerability due to the large gaps in the maps caused by areas with no flash flood records, a county-level VIM index indicates that the coastal area along the Gulf of Mexico, especially Houston and the surrounding areas, are more susceptible to flash flood damages.

Through the regression analysis, a series of models are utilized to identify the potential correlation or associations between potential contributors with the quantified vulnerability and established 8 different models based on 3 algorithms to predict vulnerability using a series of socio-economic or geographical conditions. Among the 8 models we developed, the county-level Random Forest model using 6 variables achieved the highest performance (accuracy = 58.85%), and other 3 random forest models also achieved accuracy of over 56%.

In addition to directly predicting the vulnerability, the models we established also revealed some potential correlation or association between those variables and vulnerability. By checking the coefficients of LR models, effect charts of MLR models, as well as the partial dependence charts of RF models, we concluded that: the high level in medium household income, high proportion of employment in service industry, high proportion of Asian population, low proportion of population under 18 years old, as well as large amount of unoccupied housing, could contribute to high vulnerability; impervious rate could also be an critical factor, and either too high or too low impervious rate could associate with high vulnerability and find a balance point would be useful for vulnerability reduction.

However, the current analysis is run in a general way with some very basic variables. To explain the root cause of our detected correlations or associations and in what way our community can reduce the vulnerability to flash flood, more detailed analysis would be necessary. For example, the correlation between impervious rate and vulnerability may involves a lot of other variables, like the level of urbanization, development of infrastructure (e.g. drainage system), etc. Those potential hidden variables are not included in current analysis and if we only look at the overall correlation between the variables we selected with vulnerability, we still lack the root understanding on how we could reduce vulnerability. And that's why in the research we are more likely to call those variables "contributors" instead of "causes" or "driving factors".

Overall, the result of this research could help people understand what areas in Texas are more susceptible to flash floods, and what conditions may lead to or be associated with high vulnerability. These are not only helpful for governments or social leaders to make new policies, but also instructive for the public to be prepared for potentially upcoming flash flood disasters.

REFERENCES

Ashley, Sharon T., and Walker S. Ashley. 2008. "Flood fatalities in the United States." *Journal of Applied Meteorology and Climatology* 47:805–818.

Berke, Philip, Gavin Smith, and Ward Lyles. 2012. "Planning for Resiliency: Evaluation of State Hazard Mitigation Plans under the Disaster Mitigation Act." *Natural Hazards Review* 13 (2): 139–149. ISSN: 1527-6988. https://doi.org/10.1061/(asce)nh.1527-6996.0000063.

Boulesteix, Anne Laure, Silke Janitza, Jochen Kruppa, and Inke R. König. 2012. "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2 (6): 493–507. ISSN: 19424795. https://doi.org/10.1002/widm.1072.

Bureau, US Census. 2022. "Frequently Asked Questions (FAQs) About Language Use." (accessed: 05.30.2022). https://www.census.gov/topics/population/language-use/about/faqs.html.

Cai, Heng, Nina S.N. Lam, Yi Qiang, Lei Zou, Rachel M. Correll, and Volodymyr Mihunov. 2018. "A synthesis of disaster resilience measurement methods and indices." *International Journal of Disaster Risk Reduction* 31 (October): 844–855. ISSN: 22124209. https://doi.org/10.1016/j.ijdrr.2018.07.015.

Cai, Heng, Nina S.N. Lam, Lei Zou, Yi Qiang, and Kenan Li. 2016. "Assessing community resilience to coastal hazards in the Lower Mississippi River Basin." *Water (Switzerland)* 8 (2). ISSN: 20734441. https://doi.org/10.3390/w8020046.

CEMHS. 2022. "Spatial Hazard Events and Losses Database for the United States, Version 20.0 [Online Database]." (accessed: 04.19.2022).

Cutter, Susan L., Kevin D. Ash, and Christopher T. Emrich. 2014. "The geographies of community disaster resilience." *Global Environmental Change* 29 (November): 65–77. ISSN: 09593780. https://doi.org/10.1016/j.gloenvcha.2014.08.005.

Cutter, Susan L., and Christina Finch. 2008. "Temporal and spatial changes in social vulnerability to natural hazards." *Proceedings of the National Academy of Sciences of the United States of America* 105 (7). ISSN: 00278424. https://doi.org/10.1073/pnas.0710375105.

IPCC. 2012. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation.* https://www.ipcc.ch/site/assets/uploads/2018/03/SREX_Full_Report-1.pdf.

Khajehei, Sepideh, Ali Ahmadalipour, Wanyun Shao, and Hamid Moradkhani. 2020. "A Place-based Assessment of Flash Flood Hazard and Vulnerability in the Contiguous United States." *Scientific Reports* 10 (1). ISSN: 20452322. https://doi.org/10.1038/s41598-019-57349-z.

Kirby, Ryan H, Margaret A Reams, Nina S N Lam, Lei Zou, Gerben G J Dekker, and D Q P Fundter. 2019. "Assessing Social Vulnerability to Flood Hazards in the Dutch Province of Zeeland." *International Journal of Disaster Risk Science* 10 (2): 233–243. ISSN: 2192-6395. https://doi.org/10.1007/s13753-019-0222-0. https://doi.org/10.1007/s13753-019-0222-0.

Kwak, Chanyeong, and Alan Clayton-Matthews. 2002. "Multinomial Logistic Regression." *Nursing Research* 51 (6). ISSN: 0029-6562. https://journals.lww.com/nursingresearchonline/Fulltext/2002/11000/Multinomial_Logistic_Regression.9.aspx.

Lam, Nina S. N., Margaret Reams, Kenan Li, Chi Li, and Lillian P. Mata. 2016. "Measuring Community Resilience to Coastal Hazards along the Northern Gulf of Mexico." *Natural Hazards Review* 17 (1): 04015013. ISSN: 1527-6988. https://doi.org/10.1061/(asce)nh.1527-6996.0000193.

Li, Xiaolu, Nina Lam, Yi Qiang, Kenan Li, Lirong Yin, Shan Liu, and Wenfeng Zheng. 2016. "Measuring County Resilience After the 2008 Wenchuan Earthquake." *International Journal of Disaster Risk Science* 7 (4): 393–412. ISSN: 21926395. https://doi.org/10.1007/s13753-016-0109-2.

Mihunov, Volodymyr V., Nina S. N. Lam, Lei Zou, Robert V. Rohli, Nazla Bushra, Margaret A. Reams, and Jennifer E. Argote. 2018. "Community Resilience to Drought Hazard in the South-Central United States." *Annals of the American Association of Geographers* 108 (March): 739–755. https://doi.org/10.1080/24694452.2017.1372177. https://doi.org/10.1080/24694452.2017.1372177.

Mihunov, Volodymyr V., Nina S.N. Lam, Robert V. Rohli, and Lei Zou. 2019. "Emerging disparities in community resilience to drought hazard in south-central United States." *International Journal of Disaster Risk Reduction* 41:101302. ISSN: 2212-4209. https://doi.org/https://doi.org/10.1016/j.ijdrr.2019.101302. https://www.sciencedirect.com/science/article/pii/S2212420918313189.

NWS. 2019. "NWS Preliminary US Flood Fatality Statistics (2019)." (accessed: 05.04.2022). https://www.weather.gov/arx/usflood.

———. 2020. "Weather Related Fatality and Injury Statistics." (accessed: 05.04.2022). https://www.weather.gov/hazstat/.

———. 2022a. "Flash Flooding Definition." (accessed: 05.04.2022). https://www.weather.gov/phi/FlashFloodingDefinition.

———. 2022b. "Flood and flash flood definitions." (accessed: 05.04.2022). https://www.weather.gov/mrx/flood_and_flash.

Sarker, Md Nazirul Islam, Yang Peng, Cheng Yiran, and Roger C. Shouse. 2020. "Disaster resilience through big data: Way to environmental sustainability." *International Journal of Disaster Risk Reduction* 51. ISSN: 22124209. https://doi.org/10.1016/j.ijdrr.2020.101769.

Su, Xiaogang, Xin Yan, and Chih Ling Tsai. 2012. "Linear regression." *Wiley Interdisciplinary Reviews: Computational Statistics* 4 (3): 275–294. ISSN: 19395108. https://doi.org/10.1002/wics.1198.

Tate, Eric. 2012. "Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis." *Natural Hazards* 63 (2): 325–347. ISSN: 1573-0840. https://doi.org/10.1007/s11069-012-0152-2. https://doi.org/10.1007/s11069-012-0152-2.