Critique of Current Social Vulnerability Indices

and Opportunities for Improvement

BY

APOSTOLIS SAMBANIS
B.S., University of Illinois at Chicago, Chicago, 2007
M.S., University of Illinois at Chicago, Chicago, 2012

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2016

Chicago, Illinois

Defense Committee:

Dr. Michael Cailas, Chair and Advisor, Environmental and Occupational Health Sciences
Dr. Sage Kim, Health Policy and Administration
Dr. Kristin Osiecki, University of Illinois at Springfield
Dr. John Canar, Environmental and Occupational Health Sciences
Dr. David Ibrahim, Environmental and Occupational Health Sciences
Dr. Edward Mensah, Health Policy and Administration

ACKNOWLEDGMENTS

I would like to begin by thanking God, my wife and children for providing me the strength, love, and patience to complete this thesis. I have been helped by so many people and offices that I am not sure where to begin. I would like to extend my deepest thanks to my advisor, Dr. Michael Cailas. He has been a great asset and without him I would not have the tools I needed to get me to this point. I would also like to thank the other members of my dissertation committee: Drs. Sage Kim, Christos Takoudis, Edward Mensah, Kristin Osiecki, John Canar, and David Ibrahim. In addition, Eric Berman with Federal Emergency Management Agency with his help on the disaster management data.

I am truly blessed and this has been an enlightening experience. I have been ameliorated by completing this work and hope this contribution of scientific literature may better the world we live in.

TABLE OF CONTENTS

LIST OF TABLES

TABLE **PAGE**

LIST OF FIGURES

FIGURE **PAGE**

LIST OF ABBREVATIONS

ATSDR	Agency for Toxic Substances and Disease Registry
CART	Classification And Regression Trees
CDC	Centers For Disease Control and Prevention
CF	Classification Failure
CHAID	Chi-squared Automatic Interaction Detector
CMSA	Consolidated Metropolitan Statistical Area
DT	Decision Tree
ER	Event risk
FC	False Classification
FEMA	Federal Emergency Management Agency
FIPS	Federal Information Processing Standard
GIS	Geographic Information Systems
Hazus®-MH	Hazards United States Multi Hazard
H	Hazard
IPCC	Intergovernmental Panel on Climate Change
MSA	Metropolitan Statistical Area
OR	Outcome Risk
PCA	Principal Component Analysis
PR	Percentile Rank
SV	Social Vulnerability
SVI	Social Vulnerability Index
U.S.	United States

SUMMARY

Between 1980 and 2013, the United States (U.S.) experienced 151 weather related disasters causing approximately $1 billion in overall damages with total costs exceeding $1 trillion. Social vulnerability (SV) is a widely used concept that aims to assess the differences in the susceptibility to disasters, losses, and coping and recovery abilities of communities. The SV of populations at risk of disasters in the majority of cases is expressed as an index (SVI) which has the potential to be used for deriving proactive plans that will protect communities and assist them to rebound from emergency situations.

The majority of indices aiming to assess SV are derived with a composite model based on principal component analysis or percentile ranks. Only a few studies have attempted to assess existing SVI in terms of their relation to potential losses from disasters; these assessments found a limited predictive performance in terms of identifying potentially high risk areas.

We argue and demonstrate that the current methodologies for deriving SVI may not capture the qualitatively differentiating nature of vulnerability of communities in geographic areas and do not provide a practical and reliable planning tool. Our study proposes a paradigm shift by considering SV to disasters as a classification issue and, consequently, by introducing classification modeling and performance assessment techniques which are likely to provide a different perspective on attributes influencing SV as well as a reliable approach to identify potentially high risk areas. To demonstrate the potentials of this approach historical U.S. Census and hurricane loss data from the FEMA Hazus® program were used for the Houston metropolitan area.

I. INTRODUCTION AND PROBLEM STATEMENT

Between 1980 and 2013, the United States (U.S.) experienced 151 weather related disasters causing approximately $1 billion in overall damages with total costs exceeding $1 trillion (Smith and Matthew, 2015). There are numerous, complex interactions during a disaster which can have an uneven effect on vulnerable populations (Clark et al., 1998; Flanagan et al., 2011; Juntunen, 2006) with potentially devastating economic, environmental, health, mental, and social consequences (Flanagan et al., 2011; Sherrieb et al., 2010). Identifying hazardous exposures and vulnerability involves looking at the biophysical risk in relation to a social response within a specific geographic domain to better understand disaster risk management and mitigation from a geographical and social context (Cutter et al., 2003).

The University of California, Los Angeles Center for Public Health and Disasters, as part of their Hazard Risk Assessment Instrument (2006), developed a formula as a standardized approach to hazard and risk assessment and public health impacts in a cooperation agreement S1038-19/20 from the Centers For Disease Control and Prevention (CDC) (see Equation 1 in Appendix A). The definitions of the components for the formula include risk as the likelihood or expectation of loss; hazard as a condition posing the threat of harm; vulnerability as the extent to which persons or things are likely to be affected; and resources as those assets in place that will diminish the effects of the hazard (Flannigan et al., 2011).

The CDC defines social vulnerability (SV) as the resilience of communities when confronted by external stresses on human health, stresses such as natural or human-caused disasters, or disease outbreaks (CDC, 2015). Social vulnerability recognizes that individuals have different susceptibility to disasters creating differential loss and coping abilities (Wu et al., 2002) and measures the sensitivity and ability of a population to withstand, respond and recover from natural disasters (Clark et al., 1998).

A relative vulnerability score between different groups, Census areas and geographic areas includes variables that are representative of a characteristic of a system to provide information regarding the susceptibility, coping capacity and resilience of a system that is impacted by an ill-defined event associated with a hazard of natural origin (Birkmann, 2006). A relative social vulnerability score can be a single or an aggregation of several collective variables (Birkmann, 2006), including race or ethnicity, age, income, gender, education, language, household structure, house ownership, the type of social networks, and neighborhood characteristics (Clark et al., 1998; Cutter et al., 2003; Flanagan et al., 2011). It is postulated that these chosen SV indicator variables will help with the development of preventive and mitigating strategies and recovery actions (Clark et al., 1998; Cutter et al., 2003). However, estimating social vulnerability of a community is multifactorial, complex, and sometimes ill-defined making the indirect identification with indicators a necessity (Birkmann, 2006).

According to Juntunen, Social Vulnerability Index (SVI) applications have been used in a disaster management context since the 1970s (2005). Recently, the Centers for Disease Control and Prevention (CDC) Agency for Toxic Substances and Disease Registry (ATSDR), and Flannigan et al. (2011), have created a standardized SVI methodology. This SVI uses U.S. Census variables at the tract level to help local officials identify communities that may need support in preparing for hazards, or recovering from disaster (Flannigan et al., 2011). The SVI ranks each tract on 14 social factors, including poverty, lack of vehicle access, and crowded housing. These variables are then grouped into 4 related themes including socioeconomic status, household composition, minority status/language, and housing/transportation.

The justification for choosing variables within the 4 SVI themes is based on previous research findings. Racial disparity is often linked to social, economic and political marginalization (Cutter et al., 2003) with structural conditions created by discrimination contributing to how minority perceive risk, prepare for disaster, and access resources (Cutter et al., 2003; Elliott and Pais, 2006; Thomas et al., 2013). Families with children and elderly as members of their households tend to struggle more and lack resilience as a result of increased financial burdens and mobility restraints (Cutter et al., 2003; Thomas et al., 2013). Linguistic isolation is also a concern because immigrants or foreign workers can be unfamiliar with the local culture and local hazards making it more difficult for them to respond to emergencies (Thomas et al., 2013). Mobile homes are the dominant form of housing in rural areas and they have a greater potential to be damaged during a disaster when compared to other types of housing (Cutter et al., 2003). In addition, renters do not have the freedom to prepare their rental houses for disasters and are reliant on landlords to pay for mitigation expenses (Thomas et al., 2013). Wealthy people have the ability to absorb the losses and recover from disasters; meanwhile, greater losses and slower recovery occur in low-income and impoverished areas (Cutter et al., 2003).

Aforementioned studies of social vulnerability and disaster risk management have frequently underlined classifying the reasons why vulnerability occurs; however few publications have concentrated on describing those reasons. The need to identify those reasons within a disaster context is significant, especially due to frequent hazard occurrences and disaster results experienced by vulnerable population groups (such as low-income and minorities) (Crowder and Downey, 2010; Finch et al., 2010; Van Zandt et al., 2012). As of late, climate mitigation and adaptation have expanded considerably because of rising public concern associated with the effects of changing climate (Blanco et al., 2009). The damage created by large scale hurricanes (e.g., Katrina, Ike, etc.) are the latest cases that emphasized the susceptibilities of coastal areas and have reintroduced demands for preventive disaster mitigation and emergency response planning. Large amounts of money are disbursed to reestablish the local markets affected by these disasters and to mitigate for further impacts of climatic change. However, funding has not been devoted to the social vulnerability created by disasters. More specifically, funding has not been allocated for assistance with housing and resources for the vulnerable populations. It is often the case that vulnerable populations are located in highly hazardous areas exposed to such disasters (Mohai and Saha, 2006; Wisner et al., 2004), exacerbating, thus, the whole vulnerability of these areas. A current prerequisite in this field of research is to assess the underlying reasons why vulnerability occurs and to recognize the reasons that make the vulnerable population less able to evade areas with high hazards.

There are limitations when using predetermined indices because they do not discern which theme is contributing to the overall vulnerability or variables to a particular theme (Fekete, 2012). There is the possibility of omitting key Census variables that are better indicators of vulnerability or quickly altering demographic composition of various small-area populations in the estimate of population instead of using official Census data. Some of the approaches used to assess vulnerability tend to assign specific weights to indicators based on their relevance and correlation to the topic in question. There are no reference data for weights or final vulnerability scores (Birkmann, 2006) and many studies tend to avoid them "…*the authors simply average the eleven indicators of social vulnerability without imposing weights*." (Cutter et al., 2009). Finally, these parameters and indicators do not occur individually but in combination leading to amplified vulnerability scores (Morrow, 1999; Van de Vliert, 2013).

SVI applied in a disaster management framework needs to refine the formula used to calculate risk and methodologies used to determine the index (Flanagan et al., 2011). The two predominant approaches used are percentile rank (PR) and principal component analysis (PCA). The percentile rank approach was completed recently by Centers for Disease of Control and Prevention (2010) and Flannigan et al. (2011). To derive the SVI using the percentile rank approach, each of the selected 15 Census variables, except per capita income, was ranked from highest to lowest across all Census tracts in the United States. Per capita income was ranked from lowest to highest due to the fact that, unlike the other variables, a higher value indicates lesser vulnerability. A percentile rank was then calculated for each Census tract over each of these variables. Percentile ranks were calculated by using the formula in equation 2 in Appendix A. In addition, a tract-level percentile rank was calculated for each of the 4 domains (i.e., socioeconomic status, household composition, minority status and language, and housing and transportation) based on a comprehensive sum of the percentile ranks of the variables constituting that domain. Lastly, an overall percentile rank for each tract was calculated as the sum of the domain percentile rankings. This process of percentile ranking for all variables, for each domain, and for an overall SVI was then repeated.

A fundamental supposition firmly held in this project is that indices related to vulnerability are, in the final analysis, a classification issue. Review of the literature corroborates this supposition, since the majority of the research in this field aims to generate the spatial distribution of classified regions.

To investigate the potentials and limitations of the PR and PCA approaches for studying and identifying social vulnerability characteristics, the current study selected the Houston metropolitan statistical area (MSA) which is a historically known high risk area for hurricanes, tropical storms, and flooding with racial minority and low income groups (Davidson and Anderton, 2000; Lopez, 2002; Perlin et al., 2001). Increasing growth of population alongside the United States coasts in recent decades has amplified the proportion of vulnerable populations exposed to weather patterns common to hurricanes (i.e., high winds, waves, and storm surge flooding) (Burby, 1998; Deyle et al., 2008; Godschalk et al., 1999).

A number of authors have identified challenges and limitations related to the applicability of SVI (Yoon, 2012; Wolf et al., 2013). After a few decades of existence, it remains unknown if these indices are being adopted by policymakers to understand the disaster risk, vulnerability, and resilience of their communities and plan accordingly; though, they are becoming widely available (Hazards and Vulnerabilities Research Institute, 2016 and CDC, 2015).

From the early 1970s and the introduction of the concept of the hazardousness of a place (Hewitt, 1971) in the, then known as, disaster research field, there was a rapid progression leading to many "*fuzzy definitions and divergent themes*" (Cutter, 1996) and the "*The Vulnerability Paradox*" (Cutter et al., 2003). This trend is continuing at present and more definitions are being introduced especially in relation to the concept of vulnerability (Blaikie et al., 2014). One possible explanation of this state is that the concept of vulnerability remains a captive to various "*epistemological orientations and subsequent methodological practices*" (Cutter, 1996); though, it is likely that the perception of "*vulnerability as an inherent pre-existing condition of human systems, irrespective of the (natural) hazard of interest*" incites the proliferation of such definitions and themes. The reliance of vulnerability on other concepts, which are discipline dependent as well, such as hazard, risk, adaptive capacity, susceptibility, and resilience, propagates this conceptual and methodological ambiguity (Kienberger et al., 2009; Fekete, 2012; Brassett and Vaughan-Williams, 2015).

The majority of SVI studies strongly imply that the proposed indices have a relationship with hazard and risk (e.g., UCLA, 2008; Cutter et al., 2003). In the disaster risk management literature, it is common to encounter this as depicted in the commonly applied Risk equation (Equation 3 Appendix A; UNISDR, 2009). In this equation Risk "*probability of an event and its negative consequences*" (or a loss/event that occurs with a given probability) is a function of Hazard (H) "*a potential event*" and Vulnerability "*the degree of susceptibility of the elements exposed to the event*" (Cardona et al., 2012). As previously discussed, the hazard and vulnerability concepts are discipline dependent. A "hazard" is "*a dangerous phenomenon, substance, human activity or condition* (not realized yet) *that may cause loss of life, injury or other health impacts, property damage, loss of livelihoods and services, social and economic disruption, or environmental damage*" (UNISDR, 2009). ISDR defines vulnerability as the "*set of conditions and processes resulting from physical, social, economic, and environmental factors, which increase the susceptibility of a community to the impact of hazards.*" (ISDR, 2004). This definition requires, for completeness, the concept of susceptibility (S) and indirectly that of adaptive capacity (AC) which can be considered as a function of social capacity and resilience (Kienberger et al., 2009). With these additions and for one particular place, time, and hazard, Equation 3 becomes Equation 4 (in Appendix A) and reveals the full extent of the "*conceptual and methodological ambiguity*" which is propagated in this field.

Estimating the risk for one particular place, time, and hazard becomes a major challenge due to the data-demanding nature of the approach. In addition, as with the majority of published studies, a partial estimation of one constituent of Equation 3 is demonstrated usually in the form of the spatial distribution of vulnerability; e.g., "*classified in 10 classes ranging from low (0) to high (1) vulnerability*" (Kienberger et al., 2009). A number of studies postulated different conceptual frameworks that consider vulnerability as a pre-existing condition, as a tempered response, and as a hazard of place (Cutter, 1996). With each framework, new terms were introduced or familiar old ones were redefined; for example, the hazards-of-place model of vulnerability (Hewitt, 1971; Cutter et al., 1996 and 2003) and the hazard potential, biophysical, social, and place vulnerability terms (Cutter et al., 2000). From the multitude of studies that the applied framework (Schmidtlein et al., 2008), focus is given to the social vulnerability constituent derived as a composite social vulnerability index score with (in most cases) an additive model which included selected principal component scores and standardized socioeconomic variables (e.g., Cutter, 2003). Again, the final outcome is the spatial distribution of social vulnerability classified at, for example, "*5 levels with the most vulnerable counties being those that have a standard deviation score above 1*" (Cutter et al., 2003).

From the numerous SVI studies and publications, only a few of them made an attempt to assess the proposed indices in terms of their relation to losses from disasters and their predictive performance. This substantiates the propensity of SVI to be considered as a preexisting, almost esoteric, characteristic condition of communities. The ones that made the attempt found limited predictive performance results as evident from the following indicative statements:

- *"we realize that the SoVI is not a perfect construct and more refinements are necessary. This is very clear based on the lack of correlation with presidential disaster declarations, which may be a function of the SoVI, but is more likely a function of the frequency and location of disaster events as well as the political process involved in the declaration process itself."* (Cutter et al., 2003)

- "*We cannot say with certainty that an association between tract-level elderly SVI value and mortality exists in this example because we do not have all the data required to do a complete quantitative analysis*." (Flanagan et al., 2011)
- *"There were no statistically significant correlations between social vulnerability and disaster losses (p > 0.05), indicating the impact of disasters was also related to the intensity of hazards and exposure. Disaster relief funds allocated to each province of China depended more on its disaster severity than the regional integrated social vulnerability over the past decade*." (Zhou et al., 2014).

The above review indicates the need to establish a methodology for assessing SVI in terms of their ability to identify actual vulnerable areas to disasters.

A. Major Objectives and Research Foci

The main purpose of this thesis is to establish a new path in vulnerability to disasters studies by considering vulnerability as a classification issue. This paradigm shift offers a broad range of practical and effective tools for assessing existing models as well as developing new classification models with demographic variables as predictors and loses (or risk) as the target dependent variable.

The present thesis is organized in four major research foci, which aim to answer the following questions:

1. Is social vulnerability a classification issue?
2. In the context of classification, is there the ability to assess the predictive performance of current or newly proposed SVI techniques?
3. Are existing methodologies reliable, useful, and feasible in terms of input data-demand and predictive performance?

4. Is the proposed classification-based methodology for deriving SVI to disasters reliable, useful, and feasible?

Listed below is the approach we propose to answer these questions:

- To provide empirical evidence from published studies to substantiate the fact that current social vulnerability index (SVI) methodologies implicitly use the derived indices as a classification tool to identify vulnerable areas.
- To focus on the measurement and assessment techniques of the social vulnerability research agenda and propose a methodology that has the ability to assess the predictive performance of these techniques and subsequently of the proposed indices. This methodology will be applied in the Houston Metropolitan area and two Census data bases will be used (1980 and 2000) as well as two disaster events and their losses (1983 and 2008).
- To provide a preliminary assessment of the performance of the prevailing existing techniques (i.e., PR and PCA) based on the application of the proposed predictive performance metrics.
- To introduce a new and practical framework for identifying classes of vulnerable areas based on their demographic characteristics as well as occurrences of losses related to disasters.

B. Study Area

An MSA can roughly be defined as an area with a substantial population center (a county), and adjacent areas (counties) having a high degree of economic homogeneity, where economic integration is usually measured by commuting patterns (Van Geffen, 2003). A metropolitan area is called a Consolidated Metropolitan Statistical Area (CMSA) if it meets requirements of a

MSA, if it has a population of 1 million or more people, if the component areas are recognized as primary metropolitan statistical areas, and if local opinion favors this designation. Prior to 2000, the Houston MSA included the following counties: Brazoria, Chambers, Fort Bend, Galveston, Harris, Liberty, Montgomery, and Waller. Thus to be consistent for the time periods we used those counties in both the 1980 and 2000 analyses.

C. Demographic Characteristics of Study Area

The city of Houston was founded by the Allen Brothers in 1836 close to the Buffalo Bayou's banks along the Gulf of Mexico (Smith, G.P., 2016). The 2010 U.S. Census showed a population of approximately 2.1 million people making it the fourth largest city in the U.S. after New York, Los Angeles and Chicago in terms of population (City of Houston, 2014).

The Houston-Galveston-Brazoria consolidated metropolitan statistical area, or Houston CMSA, was ranked by Forbes as the tenth fastest growing city in the U.S. in the year 2013-2014. It was also ranked the fifth largest metropolitan area in the U.S. and the metropolitan area with the largest numeric increase in 2013 by the U.S. Census (City of Houston, 2014). The Houston metropolitan area consists of eight counties: Brazoria, Chambers, Fort Bend, Galveston, Harris, Liberty, Montgomery and Waller, as shown in Figure 1 (Federal Highway Administration, 2015). It covers, approximately, 8,778 square miles and according to the 2010 U.S. Census, the total population in all eight counties that make up the Houston metropolitan area is, approximately, 6 million people with Harris County being the most populous one.

According to the U.S. Census of 2000, the city of Houston had a population density of 3,371 per square mile which puts it fourth among Texas' cities, after Garland, Arlington, and Dallas (U.S. Census, 2000). The 2010 American Community Survey shows that the minority

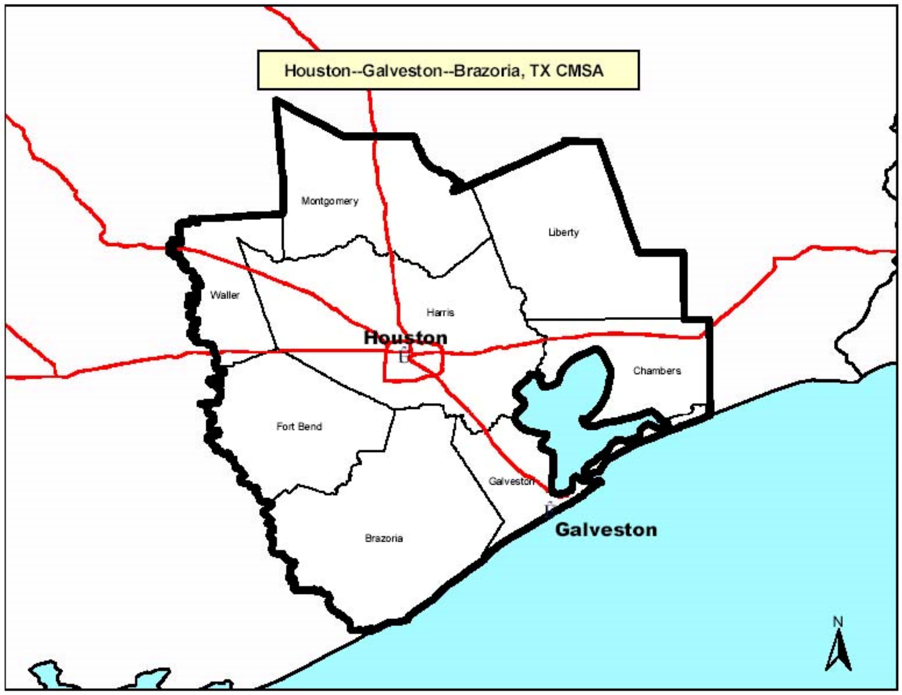


Figure 1. Houston Consolidated Metropolitan Statistical Area

population is about one-fourth the total population of the U.S., with black Americans making up almost half of the minority population (Thomas et al., 2013). The percentage of whites in the city of Houston was 49.3% and 50.5% in the years 2000 and 2010 respectively, while the percentage of blacks was 25.3% and 23.7% in the years 2000 and 2010 respectively. The percentage of Hispanics and Latinos has grown in the city of Houston from 37.4% to 43.8% between the years 2000 and 2010 (U.S. Census, 2000 and 2010). According to 2014 data, there are around 50% non-white minorities and a high percentage of citizens born outside the United States (City of Houston disparities data report, 2008; City of Houston eGovernment Center, 2014). When we compare Houston to the national averages, Houston has a higher percentage population of blacks or African Americans. The percentage of Caucasians in the U.S. was 75.1% of the total population in the year 2000 and 72.4% in the year 2010, while the percentage of blacks and

African Americans was 12.3% and 12% in the years 2000 and 2010 respectively. The percentage of Hispanics and Latinos was only 12.3% in 2000 and 16.3% in 2010 (U.S. Census, 2000 and 2010).

As shown in Table I the percentage of single female householders with children less than 18 years in the city of Houston was 8.8% in 2000 and 8.9% in 2010, which was slightly higher than the national average of 7.2% for both years (U.S. Census, 2000 and 2010). On the other hand, the percentage of renter occupied housing units in the city of Houston was higher than the national average at 54.2% in 2000 and 54.6% in 2010 compared to the national data at 33.8% in 2000 and 34.9% in 2010 (U.S. Census, 2000 and 2010).

In regard to vulnerable population groups in the Houston metropolitan area, a population analysis comparing 2000 and 2010 of the Houston-Galveston coastal area showed an overall population increase of 15%, while the socially vulnerable population percentage including elderly, low income, and Hispanic minority groups, increased from 51% to 56% (Dolan and Messen, 2012). The elderly represented 9% of the total population with a 24% increase between the years 2000 and 2010 (Dolan and Messen, 2012). The Hispanic population represented 42% of the total population, which is considerably higher than the national percentage of 16% (Dolan and Messen, 2012).

People living below poverty threshold occupy 15% of the total population in Houston compared to the national average of 9.2% (Dolan and Messen, 2012; U.S. Census, 2000).

D. Loss Occurrences and Sensitivity of Study Area

Galveston was hit by the Galveston Hurricane during the first year of the 20th century with severe tides causing 8,000 deaths and $30 million in damages (National Hurricane Center, 2015).

TABLE I

U.S. CENSUS DATA FOR THE CITY OF HOUSTON COMPARED TO THE NATIONAL AVERAGE FOR THE YEAR 2000 AND 2010

Subject	The city of Houston data in 2000 (%)	The city of Houston data in 2010 (%)	United States data in 2000 (%)	United States data in 2010 (%)
White	49.3	50.5	75.1	72.4
African American	25.3	23.7	12.3	12
Hispanic or Latino	37.4	43.8	12.5	16.3
65 years and over	8.4	9	12.4	13
Under 19 years old	30.39	28.67	28.59	26.96
Female householder with own children under 18 years	8.8	8.9	7.2	7.2
Renter occupied housing units	54.2	54.6	33.8	34.9

Hurricane Alicia, a category 3 hurricane, hit the Texas coast at Galveston Island's western end in 1983 and caused 21 deaths and $2 billion in damages along its path (National Hurricane Center, 2015). In 2001, tropical storm Allison hit Freeport, Texas bringing heavy rain falls and flood along its path. Heading Northwest through Louisiana, Mississippi, Alabama, Georgia, South Carolina and North Carolina with Houston, Texas being heavily hit by rain and flooding, the storm resulted in an estimated $5 billion in damages and has taken 41 lives (National Hurricane Center, 2015). Hurricane Rita hit southeastern Texas and southwestern Louisiana in 2005 causing damages of around $10 billion and 7 deaths (National Hurricane Center, 2015).

On September 1, 2008, a new tropical depression was formed off the shores of Africa and halfway into the Atlantic Ocean. Later that day, Ike, the ninth tropical storm of the season, had formed over the Atlantic. By September 4, 2008, the tropical storm had grown into a category 4 hurricane (National Hurricane Center, 2015). It ravaged several nations between September 4 and 13 including Turks and Caicos, the Bahamas, Cuba, Haiti and the Dominican Republic until the early hours of September 13 where it made landfall at Galveston, Texas as a category 2 (as shown in Table II) hurricane with winds of 110 mph (National Hurricane Center, 2015; Spence et al., 2011). Ike caused wide spread destruction, damage and death with an estimated $19.3 billion in damages and 114 deaths making it the third costliest Hurricane in the United States history after Katrina and Andrews (Spence et al., 2011).

According to the National Oceanic and Atmospheric Administration's National Climatic Data Center, the gulf coast in the state of Texas is considered one of the contributors to the "billion dollar disaster list" (Dolan and Messen, 2012; U.S. department of commerce, 2014). The National Hurricane Center (2015) lists the following recent events in the Houston metropolitan area:

- 1983 Hurricane Alicia – 21 deaths and $2 billion in damages
- 2001 Tropical Storm Allison – 5 deaths and $5 billion in damages
- 2005 Hurricane Rita – 7 deaths and $10 billion in damages
- 2008 Hurricane Ike – 114 deaths and 19.3 billion in damages.

Hurricane Ike, the latest major hurricane to hit the Texas coast has caused an estimated $19.3 billion in total cost and damages (National Hurricane Center, 2015). Due to the significance of these events, our study will focus on analyzing loss data from Hurricanes Alicia and Ike.

E. Empirical Evidence of Vulnerability as a Classification Issue

The first research focus of this thesis will be addressed by a critical review of the major SVI publications in this field. Specifically, the review will aim to identify empirical evidence, in the form of the eventual and practical usage of the derived index, to corroborate the major supposition postulated in this study. The majority of published studies used Geographic

TABLE II

SAFFIR-SIMPSON HURRICANE WIND SCALE

Hurricane category	Sustained wind	Damages
1	74-95 mph	Very dangerous winds will produce some damage
2	96-110 mph	Extremely dangerous winds will cause extensive damage
3	111-129 mph	Devastating damage
4	130-156	Catastrophic damage
5	≥157 mph	Catastrophic damage

Information Systems (GIS) packages, such as ESRI ArcGIS, that have a standard feature known as class ranges and breaks (Bunting et al., 2014). These features define the amount of data that falls into each class and the appearance of the map. There are two main components in a GIS classification scheme: the number of classes into which the data is to be organized and the method by which classes are assigned. The number of classes is dependent on the objective of the analysis. The rules by which the data are assigned to a class, however, require an explanation. Listed below are the standard ways in which data can be assigned to classes with the use of GIS packages (De Smith et al., 2007):

Manual: Create classes manually if you are looking for features that meet a specific criterion or if you are comparing features to specific, meaningful values. To do this, you would manually specify the upper and lower limit for each class.

Defined Intervals: Defined interval allows you to specify an interval size used to define a series of classes with the same value range.

Equal Interval: The range of possible values is divided into equal-sized intervals. Because there are usually fewer endpoints at the extremes, there are fewer values in the extreme classes. This option is useful to highlight changes in the extremes. It is probably best applied to familiar data ranges such as percentages or temperature.

Quantile: The range of possible values is divided into unequal-sized intervals so that the number of values is the same in each class. Classes at the extremes and middle have the same number of values. Because the intervals are generally wider at the extremes, this option is useful to highlight changes in the middle values of the distribution.

Natural Breaks (Jenks): Natural breaks classes are based on natural groupings inherent in the data. Class breaks are identified that best group similar values and that maximize the differences between classes. The features are divided into classes whose boundaries are set where there are relatively big differences in the data values. Natural breaks are data-specific classifications and not useful for comparing multiple maps built from different underlying information.

Standard Deviation: The standard deviation classification method shows you how much a feature's attribute value varies from the mean.

Geometric Intervals: This classification scheme creates class breaks based on class intervals that have a geometrical series. The geometric coefficient in this classifier can change once (to its inverse) to optimize the class ranges. The algorithm creates geometric intervals by minimizing the sum of squares of the number of elements in each class. This ensures that each class range has approximately the same number of values with each class and that the change between intervals is fairly consistent.

This algorithm was specifically designed to accommodate continuous data. It is a compromise method between equal interval, Natural Breaks (Jenks), and quantile. It creates a balance between highlighting changes in the middle values and the extreme values, thereby producing a result that is visually appealing and cartographically comprehensive.

Table III summarizes information from seven published studies, covering a span of more than ten years, in terms of the major SVI derivation technique and the eventual and practical usage of the derived index. These studies should be considered as an indicator since many of the authors published other, similar methodological, studies that are not listed in Table III to avoid repetition (e.g., Cutter et. al., 1996, 2000, 2007, etc.). As seen in Table III, use of the derived index as an implicit classification tool to visualize vulnerability is noticeable, providing, thus, sufficient empirical evidence to substantiate the major supposition postulated in this study. This supposition is further substantiated if SVI published reports from national and international agencies and organizations are taken into account (e.g., McCarthy, 2001; Briguglio, 2003; Parry et al., 2007, etc.).

TABLE III

PREVIOUS STUDIES CLASSIFICATION METHODS

Author	**Derivation Methodology**	**Map Number of Vulnerability Classes (and Type)**
Cutter et al. (2003)	Principal Component Analysis	Five (Standard Deviation)
Chakraborty et al. (2005)	Maximum value transformation (ratio of value)	Five (Defined interval)
Flanagan et al. (2011)	Percentile Rank	Three (Equal Interval)
Rygel et al. (2006)	Principal Component Analysis	Five (Standard Deviation)
Schmidtlein et al. (2008)	Principal Component Analysis	Five (Standard Deviation)
Yoon (2012)	Percentile Rank and Principal Component Analysis	Five (Standard Deviation)
Zhou et al. (2014)	Principal Component Analysis	Four (Standard Deviation)

II. METHODS AND MATERIALS

A. Data Sources and Variable Selection

The selection of variables, the methods of aggregation (including weighting assignment), the choice of the scale of analysis, and the extent of the research area all have implications on the derivation of the social vulnerability index and its relevance to the condition of the community it pertains to represent (Fekete, 2012). The examination of social vulnerability commonly emphasizes political, social, economic, and institutional elements that impact various social groups' vulnerability to disaster risk exposures (Tate et al., 2010). Due, likely, to changing environmental, economic, and anthropogenic effects, a growing trend in the U.S. is a migration of humans into hazardous areas (Cutter et al., 2007), and there is little research to show how social vulnerability changes in time in these densely populated areas. In, addition, previous studies of social vulnerability in metropolitan areas have focused, mainly, on disaster impacts and recovery operation outcomes (Van Zandt et al., 2012; Zhang and Peacock, 2009) or susceptibility to natural hazards (Maantay and Maroko, 2009; Zahran et al., 2008), however, the generative process and changing pattern of social vulnerability has yet to be studied. For this study, U.S. Census variables for 1980 and 2000 were exported at the Census tract level for the Houston Metropolitan Statistical Area (MSA) using the Social Explorer® program at the Census tract level. With the use of Microsoft Excel the Census variables were filtered to utilize only compatible variables available between the two sets of Censuses. In addition, since the Census data is from different time periods, the Census tract relationships were joined in Microsoft Access by selecting the appropriate crosswalk files for a specific Census year from the Longitudinal Tract Data Base (Logan et al., 2012). Then, by exporting the common Federal

Information Processing Standard (FIPS) Census tract codes and aggregating the data by weight, a common set of records was derived for the two periods.

A range of approaches have been utilized to quantify social vulnerability at various geographic scales and with the use of a variety of variables, however, an optimum scale or a group of variables has yet to be defined (Fekete, 2012). The only consensus in the scientific literature is on the multidimensionality and complexity of social vulnerability. For instance, social vulnerability is frequently derived based on gender (Enarson and Morrow, 1998; Enarson et al., 2007), race and ethnicity (Fothergill et al., 1999; Peacock et al., 1997), poverty (Fothergill and Peek, 2004; Long, 2007), and age (Anderson, 2005; Smith et al., 2009). In particular, poverty and housing are important influences in defining a household's capability to endure socio-economic stresses in metropolitan settings (Moser, 1998; Sanderson, 2000). Following a similar method applied by Schmidtlein et al. (2008) built environment variables were removed from the analysis to concentrate more explicitly on the characteristics of the populations themselves that contributed to vulnerability. By taking into account all of the above and the prerequisite to have common 1980 and 2000 Census variables, the 15 variables listed in Table IV were used for the current study.

B. <u>Loss and Disaster Data</u>

In order to provide a proper answer to the research foci listed in Section A of the Introduction and Problem Statement, related to the vulnerability of the Houston MSA, natural hazards and losses are required to be introduced to serve as a target variable for the proposed performance assessment metrics. Hazard mitigation reduces and/or eliminates long-term risk to people and property resulting from hazardous exposures and effects (Godschalk, 2003). The elimination of this risk differentiates hazard mitigation from immediate activities in disaster preparedness

TABLE IV

1980 AND 2000 CENSUS VARIABLES USED

Category	**Census Variables**
Housing Characteristics	% Occupied Housing Units: Renter Occupied, % Housing Units: 5 or more, % Housing Units: Mobile home or trailer, etc.
Children	% Under 5 years
Elderly	% 65 years and over
Race – African American	% African American
Race – Hispanic	% Hispanic or Latino
Female Head Household with children (under18 years old)	% Households with one or more people under 18 years: Female householder, no husband present
Institutionalized Persons	% Population in group quarters: Institutionalized Population
Education - Less than High School Degree	% Population 25 years and over: Less Than High School
Unemployed	% Population In Labor Force 16 Years And Over: Unemployed
Household Income	Median household income In 1979/1999 Dollars
Below Poverty	% Below Poverty
Mobility	% No Vehicle Available
Social Welfare Recipient	% social welfare
Housing Value	Median value
Occupation Type	Bottom quantile occupation type (e.g. Manual material moving, etc.)

and response (Godschalk, 2003). The Federal Emergency Management Agency (FEMA) Hazus® software package, a nationally applicable standardized methodology, is a disaster mitigation strategic tool that uses geographic information systems (GIS) to evaluate potential physical, economic, and social impacts and losses caused by earthquakes, floods, and hurricanes (Hazus®, 2004). The Hazus® risk assessment approach includes five steps: identifying hazards, profiling hazards, creating assets inventory, estimating losses and considering mitigation options ("Hazus®®-MH risk assessment" Hazus®, 2004). It graphically illustrates the limits of identified high risk locations due to earthquake, hurricane, and floods. Users can then visualize the spatial relationships between populations and other more permanently fixed geographic assets or resources for the specific hazard being modeled, a crucial function in the pre-disaster planning process.

To study Hazus® applicability to SVI in the the Houston Study Area we used historical data from Hurricane Alicia and Ike data which struck in 1983 and 2008, respectively. Utilizing Hazus® after modeled landfall of Hurricane Alicia and Ike you can generate reports that include economic and employment loss from various building types (residential, commercial, industrial, agricultural, government, education, and religious), displacement costs (rental, wage, etc), and debris generated (tree blow down, brick and concrete debris) which determines factors that take a toll on those displaced. The three sets of Hazus data utilized for this study were:

1. number of displaced household,
2. number of short term shelters required, and
3. total building loss in thousands of dollars.

These varialbles were transformed to percent fractional ranks and a Loss Index was created with an additive model which will serve as the target variables.

C. <u>SV Index Methodologies to be Assessed</u>

Yoon (2012) attempted to classify the methodologies used to derive SVI into two categories:

- "*a deductive approach based on a theoretical understanding of relationships*", and
- "*an inductive approach based on statistical relationships*" stating that "*this approach differs from the deductive approach in that it includes <u>all possible variables</u> mentioned by literatures to assess social vulnerability.*"

In the current study we will avoid such a categorization, since, in reality, the "*theoretical understanding of relationships*" is a sought after objective that requires variables and a statistical methodology in order to be assessed. To paraphrase a classic definition of deductive logic (Wrenn, 2016): "*deductive arguments* (approaches) *are usually limited to inferences that follow from definitions, mathematics and rules of formal logic.*" The inductive approach categorization is rendered invalid since it is, practically, impossible to encompass "*all possible variables*" associated with social vulnerability. In practice, the so-called inductive approach involves variable selection/justification as well (Berrang-Ford et al., 2015).

In practice, SVI are based on a range of selected variables, which are transformed and then aggregated (Flanagan et al., 2011). Another common approach is the application of data analytic techniques (such as principle components analysis) in order to reduce dimensionality and use a few of the components representing the original variables to derive the SVI with an additive model (Demšar et al., 2013; Abson et al., 2011). Since the original variables of any given social vulnerability index are often measured in different units (e.g., dollars, number of residents, etc.), a standardization procedure is necessary to convert the unit of measurements into a comparable range of values. Percentile Rank (PR) and Min–Max rescaling techniques are commonly used to make the original variables unitless and comparable (see Equations 5 and 6 in Appendix A).

Principal Component Analysis (PCA) is a multivariate statistical technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (Jolliffe, 2014). The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components. The principal components are orthogonal because they are the eigenvectors of the covariance (or correlation) matrix, which is symmetric (Fukunaga, 2013). In addition, expert judgment, is currently a critical element in the subjective interpretation of the components generated by these prevailing social vulnerability index methodologies (Fekete, 2012).

The goal of using principal component analysis is to aggregate the original Census variables (e.g., nxp matrix) into a few groups, in reality the principle components (i.e., the characteristic p vectors). A common practice is to select a few only components for PCA:

> "*Originally, more than 250 variables were collected, but after testing for multicollinearity among the variables, a subset of 85 raw and computed variables was derived. After all the computations and normalization of data (to percentages, per capita, or density functions), 42 independent variables were used in the statistical analyses (Table II)….. The primary statistical procedure used to reduce the data was factor analysis, specifically, principal components analysis. The use of a reductionist technique such as factor analysis allows for a robust and consistent set of variables that can be monitored over time to assess any changes in overall vulnerability. The technique also facilitates*

> *replication of the variables at other spatial scales, thus making data compilation more efficient. A total of 11 factors* (i.e., principle components) *was produced, which explained 76.4 percent of the variance among all counties."* (Cutter et al., 2003).

The normal practice is to then label – subjectively – each one of the selected principle component based on the magnitude of the coefficients of each component vector (e.g., the coefficients of each characteristic vector). A source of confusion in these studies and publications is the terminology used for the transformed original variables (e.g., principle components) and the individual transformed original observations (e.g., principle component scores; pc-scores). In addition, some studies (see above quote) refer to PCA as factor analysis and adopt the factor analysis terminology (i.e., loadings, common factors, etc.), for the purpose of this study we will adopt the standard PCA terminology and avoid this confusion of these different techniques (Cureton and D'Agostino, 2013). The pc-scores derived from the PCA model with the use of the correlation matrix (default option for most statistical packages) are standardized (e.g., unit variance). In the SVI literature these pc-scores are used to derive the index with a simple additive model (i.e., by adding the scores of the selected principle components) (Yoon, 2012); or by weighing each pc-score and then adding (Inostroza et al., 2016). The weighing of the pc-scores is arbitrary since, in reality, alters the mathematical relationship between the original individual observations and the corresponding pc-scores. (Cutter et al., 2003; Finch et al., 2010; Fekete, 2012).

D. Proposed Classification Methodology

Decision Trees (DT) are considered to be a popular approach for deriving classification models and in this study their applicability in the social vulnerability field of research will be demonstrated. To the best of our knowledge, this is the first application of the DT approach for exploring the potentials of analyzing social vulnerability within a classification framework. Decision tree induction is a well-known and effective classification technique extensively used in several domains. Its major field of application is the data mining and analytics fields where it is used to explore data structures and induce the tree and its rules that will be used to make predictions. In the context of SV studies the prediction from a classification model could be a vulnerability category (i.e., severity class) based on actual instances of losses which are placed in categories or classes. As stated above, this study considers vulnerability as a classification issue and, consequently, the use of these techniques should be expected to provide valuable information on variables influencing vulnerability as well as a reliable mechanism to identify potentially high risk areas.

The PR, PCA, and DT approaches will be implemented with the use of the computer program known as IBM® SPSS® Modeler 16.0 (Larose, 2014). Modeler is an "*extensive predictive analytics platform that is designed to bring predictive intelligence to decisions by providing a range of advanced algorithms and techniques that include text analytics, entity analytics, decision management and optimization.*" (Larose, 2014).

E. Proposed Predictive (Classification) Performance Metrics

The primary objective of the second research focus question is to concentrate on the "*measurement and assessment techniques*" part of the social vulnerability research agenda (Cutter, 1996) and propose a methodology that has the ability to assess the predictive performance of these techniques. In addition to the second research focus question, this project will introduce an approach to generate indices that are aiming to have an optimum classification performance (i.e., to identify regions that are highly vulnerable avoiding underestimation and overestimation of risks). We believe that this approach will create new and more adoptable indices that are usable by policy makers.

Conceptually, indices related to vulnerability are, in the final analysis, a classification issue. The relevant literature corroborates this supposition since the majority of the research in this field aims to generate the spatial distribution of regions classified in terms of vulnerability (Cutter and Finch, 2008); e.g., 5 levels with the "most vulnerable" counties being those that have a standard deviation score above 1 (Cutter et al., 2003). Empirical evidence substantiating vulnerability being a classification issue was presented above (see Section E: Empirical Evidence of Vulnerability as a Classification Issue).

Each record, $i = 1, 2, .., N$, usually a geographic area (tract, county, etc.), is characterized by a number, $j=1, 2,.., p$, of socioeconomic, built environment, environmental, etc., attributes (Schmidtlein et al., 2008). The main objective of SVI studies is to derive from these p attributes an aggregate vulnerability index (i.e. social vulnerability, biophysical vulnerability, etc.) based on a transformation of the original variables (i.e. standardization, min-max, etc.), or with the application of dimensionality reducing techniques such as principle components (Yoon, 2012). Eventually the index is converted to a class label in order to depict the severity of vulnerability on an ordinal scale (i.e., areas which are less or most vulnerable; see also Section E Empirical Evidence of Vulnerability as a Classification Issue). Flanagan et al. (2011) converted the original variables to a percentile rank (ordinal) scale and used an additive model to derive a summary index of social vulnerability. A similar approach was adopted for the Houston MSA by Dada (2015).

A major deficiency in the SVI field of research is the lack of a methodology to validate the predictive performance of the proposed indices. Wolf et al. (2013) used the terms "*significance* *and relevance*" (i.e., "*the index needs to be able to identify hotspots of vulnerability where evidence indicates an elevated vulnerability compared to the surrounding area.*" and "*the hotspots of vulnerability need to match with hotspots of hazard-related mortality and morbidity*") to signify, essentially, the predictive performance of SVIs. In this study these terms will be avoided and the predictive (or classification) performance term will be used accompanied by relevant performance metrics.

For one particular place, time, and hazard, hotspots of vulnerability are essentially areas where the risk (for loss, damage, etc.) is at the highest level compared to other areas of the geographic study range. To accomplish this prerequisite of performance, i.e., reliable hotspot identification, an ordinal scale has to be adopted for the proposed index. The index as a predictor must identify areas of vulnerability that, indeed, are vulnerable based on evidence. For this context the evidence can represent "*expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between natural or human-induced hazards and vulnerable conditions*" (ISDR, 2004). Thus, a predictive performance metric can be derived which will compare the predicted classification of vulnerability to an actual target classification of losses based on realized (or estimated) disaster evidence. In the few studies that did an assessment of performance, this comparison takes place at a ratio scale (i.e., correlation coefficient and significance of regression, Yoon, 2012); what is proposed in this study is to use an ordinal scale for both components of the comparison. A first exploratory attempt with this methodology was made by Bakhsh (2015) which used a coincidence matrix approach to assess the predictive performance of a heat vulnerability index for the state of Missouri.

A critical component of this approach is the transformation of the "predictor" (i.e., SV index) and target variables into m classes which represent the severity of the target event or incidence (e.g., low, mid-low, mid-high, and high mortality). This is accomplished by using a binning methodology with equal counts per bin (if the total records are even) which creates m new nominal class fields based on the values of one or more existing continuous (numeric range) fields. The equal counts approach was selected due to the lack of reliable thresholds for both comparison components. The proposed binning approach (depicted in Appendix B) is in line with the sequential scheme of color differentiation which is typically used to represent differences in the severity of the SV being mapped (depicted in Appendix C). In the majority of cases an ordinal scale (e.g., least vulnerable to most vulnerable) is used which is represented by an appropriate color gradation (i.e., usually from light to dark). Generally, map readers will not be able to tell the difference between more than six or seven levels of color value, especially in the complicated context of the map itself (McGranahan, 1989).

The performance comparison is achieved by using a m x m confusion (or error) matrix (Lewis and Brown, 2001). A major advantage of this approach is that this performance comparison matrix can be used to define specific classification performance metrics. For a given geographic scale, which defines the area of interest, in this case tract, we define the following classification performance metrics: The numbers of correctly classified areas which occur when their instance class (i.e., Target classification of losses) matches the predicted class (i.e., the diagonal elements of the confusion matrix) provide an overall classification (i.e., the diagonal elements of the confusion matrix) provide an overall classification performance measure. The sum of these matching classes divided by the total number of areas, N, yields the Overall Classification Performance (OCP) rate. This metric is similar to the overall classification accuracy, which is used in the remote sensing (Foody, 2002) and data

		Target classification of losses				
		1	2	…	M	Σ
Predicted classification of vulnerability	1	C_{11}	C_{12}	…	C_{1m}	C_{1+}
	2	C_{21}	C_{22}			C_{2+}
	⋮					
	M	C_{m1}			C_{mm}	
	Σ	C_{+1}	C_{+2}			m

Figure 2. Error or confusion matrix (or, in this study, performance comparison matrix)

mining fields (Chen et al., 1996, and Bhardwaj and Pal, 2012). Besides the OCP, or overall accuracy, classification accuracy of individual classes provide valuable performance information. For this purpose, we propose the following metrics:

If the SVI is used to allocate resources, a misclassification of a highly vulnerable area into a non-vulnerable class is likely to have grave consequences this is known as Classification Failure (CF). To quantify this failure the number of areas in the highest instance class (i.e., m) predicted to belong in the lowest (i.e., 1) is used (e.g., C_{1m}). Conceptually, this failure metric can include selected elements of the upper triangular performance comparison matrix which represent the region of risk underestimation; e.g., C_{1m}, C_{1m-1}, C_{2m}, For the purpose of this study, the one element definition will be used (1m) in combination with all the upper triangular elements which define an underestimation range. The classification failure number can be expressed as a rate by dividing it with the total number of areas, N to yield:

If the SVI is used to allocate resources, a misclassification of a non-vulnerable area into a highly vulnerable class is likely to result in a waste of valuable recourses similar to a false alarm incidence this is known as False Classification (FC). To quantify this potential waste, the number of areas in the lowest instance class (i.e., 1) predicted to belong in the highest (i.e., m) are used (i.e., C_{m1}). Conceptually, this false alarm metric can include more elements of the lower triangular predictive performance matrix; which represent the region of risk overestimation, e.g., C_{m1}, C_{m2}, C_{m-1}, etc. For the purpose of this study the one element definition will be used (C_{m1}) for FC. The false classification number can be expressed as a rate by dividing it with the total number of areas, N.

The two off diagonal sections of the predictive performance matrix provide overall predictive performance indicators of risk over/under estimation. The sum of all the areas in the underestimation range divided by N provides the Overall Underestimation Rate (OUR). In a similar fashion, the Overall Overestimation Rate (OOR) is derived.

In the Hazus® Hurricane Model the hazard can be specified as either a single historical storm scenario, user-defined storm scenario, or as a complete probabilistic analysis (Hazus®, 2014). For this study we used the historical hurricane scenarios; however, future research could be completed utilizing user defined and complete probabilistic analysis scenarios to review effects on loss estimates for the region. The creation of potential loss scenarios offers a great opportunity for expanded use of this predictive performance methodology since indices can be assessed based on the generated losses of the area of interest.

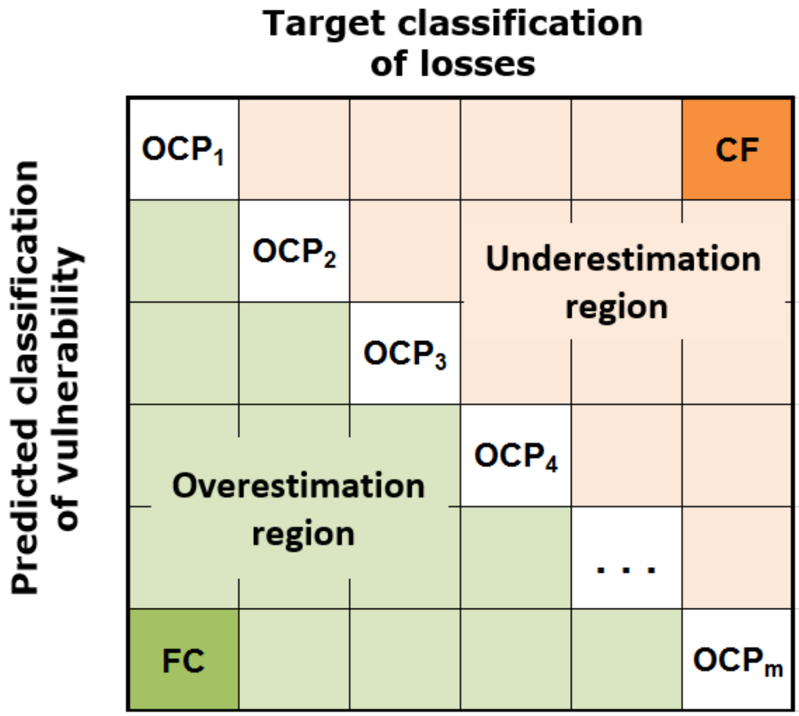


Figure 3. Proposed approach for predicted classification of vulnerability and target classification of losses

F. <u>Loss Classification Model</u>

The previous sections demonstrated that the SV indices, in the final analysis, have been used to classify vulnerability and visualize its spatial distribution. In the context of a classification model approach for vulnerability, which is proposed in this study, the objective can be the assignment of a tract to a certain vulnerability class based on its features and previous manifested disaster losses. A classification model can have an exploratory as well as a predictive use. For example, it will be useful for an emergency management department to have a descriptive table of features defining the characteristics of a high risk Census tract in its jurisdiction. This table can be in the form of a rule set which is a common output from many classification algorithms. The classification model can also be used to predict the class label (i.e., severity of loss) of a Census tract based on its attributes alone.

The fourth research focus of this project was to develop a disaster loss classification methodology and assess its performance in comparison to the percentile rank methodology and that based on principal components. The underlying criteria of this methodology proposed in this study are:

The methodology needs to be reliable. The classification methodology needs to be able to identify (i.e., predict) hotspots of vulnerability that correspond to actual disaster losses. In addition, it is vital that the methodology does not classify wrong hotspots by giving them a very low vulnerability classification (i.e., CF metric). The proposed classification performance metrics quantify this criterion since a reliable classification model must have a low OUR, OOR, CF, and FC, and a high OCP.

The methodology needs to be useful. The complexity of some methodologies places them out of the reach of emergency management administrators and small-scale departments. The approach proposed by Flanagan (2011) offers a viable and attractive alternative for wide spread use; however, its reliability remains to be evaluated.

The methodology needs to be useful. The attributes used for the input data set "*need to be cheap, reliable, recent, routine, and at a sufficient spatial resolution*" (Wolf et al., 2013). This criterion is related to usefulness since small scale emergency management departments will not be able to perform the preliminary data collection and preparation tasks without the support of specialized experts.

The input data set is a collection of geographic area records, in this study at a tract level, i = 1, 2, .., N. Each record is characterized by a number, j=1, 2,.., p, of socioeconomic, built environment, environmental, etc., attributes. The main objective of SVI studies is to derive from these attributes an aggregate vulnerability index (social vulnerability, biophysical vulnerability, etc.) based on a transformation of the original variables (standardization, min-max, etc.,) or with the application of dimensionality reducing techniques such as principle components (Yoon, 2012). A classification model requires not only an input attribute set (e.g., socioeconomic attribute data) but also a target attribute that has to be in a categorical scale (nominal or ordinal). For the purpose of this study the target attribute set will be risk, as a class label, for each record representing "*expected losses (deaths, injuries, property, livelihoods, economic activity disrupted or environment damaged) resulting from interactions between natural or human-induced hazards and vulnerable conditions*" (ISDR, 2004; Cardona, 2005).

To create this target attribute data for the Houston Study Area, historical Hurricane Alicia (1983) and Ike (2008) loss data retrieved from Hazus® listed below were used (see Material and Methods section):

- The number of displaced household,
- The number of short term shelters required, and
- The total building loss in thousands of dollars.

These variables were converted to Percentage Fractional Ranks (i.e., each rank is divided by the number of records with valid values and multiplied by 100) and a simple additive model was used to create the Loss Index for the two events.

In the Hazus® Hurricane Model the hazard can be specified as either a single historical or user-defined storm scenario or as a complete probabilistic analysis. For this study, we used the historical hurricane scenarios; however, future research could be completed utilizing user defined and complete probabilistic analysis scenarios to review effects on loss estimates for the region.

G. Proposed Decision Tree Classification Approach

Decision tree (DT) algorithms are supervised learning algorithms which recursively partition the input data based on its attributes, until some stopping limit is reached (Larose, 2014). As shown in Figure 4, this recursive partitioning gives rise to a tree-like formation. DT are popular tools for classification and prediction that are gaining popularity in many fields. (An introduction of this methodology is given in the Data Mining Lecture notes taught by Dr. Cailas.) In this section, we will cover and discuss the modifications and methodological additions that were introduced in order to apply DT classification to the social vulnerability research field. A basic premise that makes the DT approach attractive is that "*DT methods are exploratory (not inferential) and non-parametric since they do not require assumptions about the data distribution, scale, and model*"; in addition, these methods can easily deal with missing data which is a common characteristic of real-world data sets values as well as categorical attributes.

A decision tree starts from the root node and contains internal nodes and leaf (terminal) nodes, all internal nodes have two or more child nodes. The root and internal nodes contain splits, which are the building blocks of the tree formation. The split at each node is described by a decision that depends on one selected feature of an attribute A (e.g., Income > $40,000). The feature for A is selected among all possible ones, and the split is selected among all possible splits, with the objective to minimizing the heterogeneity of the resulting subsamples forwarded to the child nodes. The aim is that the final partitions (terminal leaves of the tree) are homogeneous with respect to the classes.

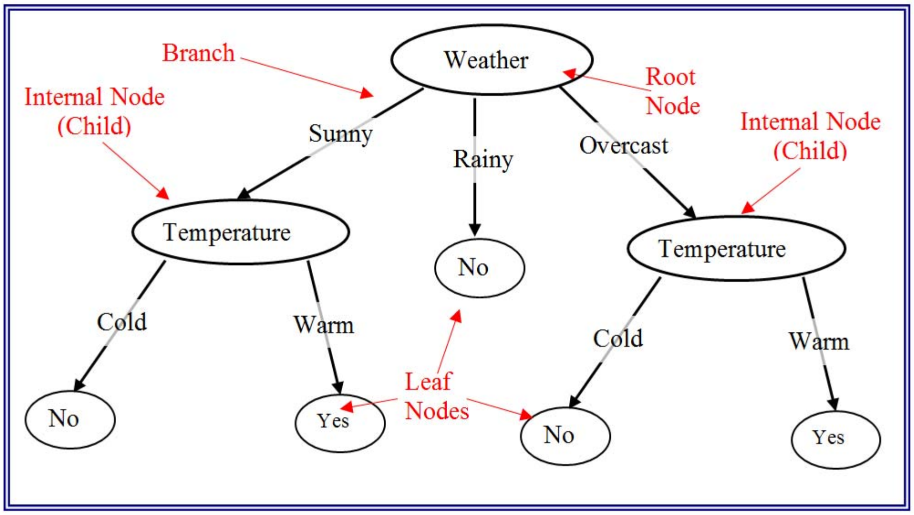


Figure 4. Schematic of a simple decision tree

The criterion for choosing the best splitting rule varies from algorithm to algorithm and the measures they apply. For this project three well-known DT algorithms will be use to explore the applicability of DT in the vulnerability research field and identify advantages and limitations. These algorithms are:

- The CHAID (Chi-squared Automatic Interaction Detector) algorithm proposed by Kass (1980) and the Exhaustive CHAID by Biggs et al. (1991). These algorithms are known to generate relatively small trees that are easy to interpret (Briggs et al., 1991).
- The CART (Classification And Regression Trees) algorithm proposed by Breiman et al. (1984). This algorithm will be used to explore the attributes (independent variables) that are likely the best predictors for the target (dependent) attribute (Sauvé, 2014).
- The C5.0 algorithm which is the most recent version of the ID3.0/C4.5 algorithms developed by Quinlan (1986 and 1993); the improvements are documented by Pang and Gong (2009).

This algorithm will be used to explore the predictive performance of this approach for classifying vulnerability.

The selected algorithms have fundamental differences since C5.0 and CART algorithms use impurity measures to choose the splitting attribute and the split value(s) (Pang and Gong 2009), whereas CHAID uses the chi-square or the F statistic for splitting. The C5.0 uses the gain ratio whereas the CART uses the Gini coefficient for impurity measurements (Quinlan, 1993; Breiman et al., 1984; Larose, 2014).

III. RESULTS AND DISCUSSION

A. Results

The three different aggregated social vulnerability scores created by the three methodologies were compared by using the confusion matrix. The confusion matrix shows the pattern of matches between each generated field and its target field for categorical targets. A table is displayed with rows defined by actual (loss) values and columns defined by predicted values, with the number of records having that pattern in each cell. This is useful for identifying systematic errors in prediction. If there is more than one generated field related to the same output field but produced by different models, the cases where these fields agree and disagree are counted and the totals are displayed. For the cases where they agree, another set of correct/wrong statistics is displayed. The following comparisons were made with each type of methodology (PFR, PCA, and DT):

TABLE V

OVERALL CONFUSION MATRIX COMPARISONS

Columns	Rows
1980 U.S. Census SVI	1983 Loss index
2000 U.S. Census SVI	2008 Loss index

1. **Percentile Rank**

SVI based on Percentile (fractional) ranks in Modeler is an additive model. Each rank is divided by the number of records with valid values and multiplied by 100. Percentage fractional ranks fall in the range of 1–100. All percentage fractional ranks calculated for each of the variables were sorted from high to low with exception of the housing value and household income variables were sorted low to high.

Displayed in Appendix B is the Percentile Rank Confusion Matrix outputs from Modeler for 1980 U.S. Census SVI and 1983 Loss data total percentage. The total percentage calculated was 23.3% for this output. The total percentage calculated was 23.7% for the 2000 U.S. Census SVI and 2008 loss data. Adding the past 1983 hurricane Alicia loss data improved the total percentage calculated to a total of 25.9% (small improvement) from 23.25%.

2. **Principal Component Analysis**

Principal Component Analysis results yield PCs that can be interpreted by review of the Communalities and Components. Communalities is the proportion of each variable's variance that can be explained by the principal components (Bruin, 2006). Table VI lists communalities for both the 1980 and 2000 U.S. Census side by side.

The values in 1980 and 2000 extraction columns indicate the proportion of each variable's variance that can be explained by the principal components. Variables with high values are well represented in the common space, while variables with low values are not well represented. Tables VII and VIII contain component loadings, which are the correlations between the variable and the component. Since these are correlations, possible values range from -1 to +1. To be conservative when interpreting the results in Table VII and VIII correlations are taken as an absolute value and values that are 0.5 or less are considered low correlations that are probably

TABLE VI

1980 AND 2000 U.S. CENSUS PRINCIPAL COMPONENT ANALYSIS COMMUNALITIES

Category	**1980 U.S. Census Communalities**	**2000 U.S. Census Communalities**
Housing Characteristics	0.587	0.547
Children	0.85	0.779
Elderly	0.695	0.764
Race – African American	0.734	0.809
Race – Hispanic	0.566	0.873
Female Head Household with children (under18 years old)	0.869	0.723
Institutionalized Persons	0.628	0.779
Education - Less than High School Degree	0.908	0.891
Unemployed	0.83	0.567
Household Income	0.88	0.839
Below Poverty	0.936	0.903
Mobility	0.87	0.762
Social Welfare Recipient	0.841	0.815
Housing Value	0.824	0.796
Occupation Type	0.923	0.878

not meaningful. Using the 1980 U.S. Census data, the PCA led to selection of three components and explained 79.6% of the variance. Using the 2000 U.S. Census data, the PCA led to selection of four components and explained 78.1% of the variance. For the 1980 data set, the interpretation of Components 1, 2, 3, and 4 led to Socioeconomic, Wealth, Institutionalized, and Race interpretations, respectively. For the 2000 data set, the interpretation of Components 1, 2, 3, and 4 led to Socioeconomic, Wealth, Race, and Institutionalized interpretations, respectively. The set of components derived from both PCAs had broadly similar subject interpretations.

To help improve interpretation and predictability of the PCA components, a VariMax

TABLE VII

1980 U.S. CENSUS PRINCIPAL COMPONENT ANALYSIS COMPONENT LOADINGS

Component Matrix[a]

	Component			
	1	2	3	4
Housing Characteristics	0.742	0.12	0.146	0.07
Children	0.862	0.213	-0.248	0.053
Elderly	0.794	0.233	0.102	0.025
Race – African American	0.695	-0.478	0.152	-0.381
Race – Hispanic	0.601	0.099	-0.441	0.563
Female Head Household with children (under18 years old)	0.888	-0.234	0.162	-0.222
Institutionalized Persons	0.203	-0.009	0.766	0.568
Education - Less than High School Degree	0.931	-0.023	-0.203	0.146
Unemployed	0.904	-0.107	-0.039	-0.012
Household Income	0.57	0.744	0.03	-0.157
Below Poverty	0.922	-0.293	-0.016	-0.013
Mobility	0.826	-0.422	0.1	-0.003
Social Welfare Recipient	0.908	0.126	-0.003	0.019
Housing Value	0.446	0.758	0.225	-0.267
Occupation Type	0.95	-0.121	-0.077	-0.03

a) Extraction Method: Principal Component Analysis with no rotation

TABLE VIII

2000 U.S. CENSUS PRINCIPAL COMPONENT ANALYSIS COMPONENT LOADINGS

Component Matrix[a]

	Component			
	1	2	3	4
Housing Characteristics	.546	-.058	.195	.456
Children	.727	.206	.447	-.089
Elderly	.453	.687	-.277	-.097
Race – African American	.608	-.091	-.656	-.023
Race – Hispanic	.661	-.093	.653	-.003
Female Head Household with children (under18 years old)	.821	-.065	-.189	-.098
Institutionalized Persons	.050	-.022	-.168	.865
Education - Less than High School Degree	.888	-.096	.304	-.023
Unemployed	.744	-.036	-.086	.066
Household Income	-.195	.875	.166	.092
Below Poverty	.943	-.103	-.033	.040
Mobility	.834	-.090	-.213	.116
Social Welfare Recipient	.634	.569	-.225	-.195
Housing Value	-.166	.849	.096	.194
Occupation Type	.933	-.025	.019	-.084

a) Extraction Method: Principal Component Analysis with no rotation.

rotation was performed which is a rotation that tends to load each variable highly on just one component (Tabachnick and Fidell, 2013).

Additionally, as proposed by Schmidtlein et al. to simplify the interpretation of the components, the absolute value of the pc-scores were used (2008). Displayed in Appendix B is the PCA Coincidence Matrix outputs from Modeler for 1980 and 2000 U.S. Census and 2008 Disaster. The 1980 U.S. Census coincidence matrix total percentage calculated was 26.6% when compared to the 1983 loss data. When completing a coincidence matrix with VariMax rotation of the PCA components the total percentage improved by one percentage point. After applying absolute values to the pc-scores, the total percentage improved to 35.7%. The 2000 U.S. Census coincidence matrix total percentage calculated was 27.1% when compared to the 2008 loss data. When completing a coincidence matrix with VariMax rotation of the components the total percentage improved by one percentage point. After applying absolute values to the pc-scores the total percentage improved to 33.1%.

3. Predictive Performace of Percentile F Rank and Principal Component Analysis

To assess the performance of the existing methodologies during the 2000 and 1980 period the same input data sets were used (15 Census variables). The target variable was the loss index created by the 3 Hazus® generated variables representing actual losses. The results are summarized in Table IX.

The predictive performance of the existing methodologies seems to be problematic since in most of the examined cases the overall underestimation rate (OUR) is higher than the overall classification performance (OCP) one. The PFR methodology is likely to be unstable since the uniform ranking (i.e., all variables rank the same) performs better that the logical conceptual

TABLE IX

PREDICTIVE PERFORMANCE OF PRINCIPAL COMPONENT ANALYSIS AND PERCENTILE F RANK

	CF	OUR	FC	OOR	OCP
PFR 80 (additive/ u.rank)	6.4%	42.3%	8.9%	37.9%	19.9%
PFR 80 (c_rank)	6.2%	43.1%	9.9%	38.2%	18.6%
PR 00 (additive/ u.rank)	7.1%	40.4%	9.0%	38.5%	21.1%
PFR 00 (c_rank)	8.6%	41.1%	9.9%	40.6%	18.3%
PFR 80+ 83 losses Pred 08	2.8%	37.6%	4.9%	33.2%	29.2%
PCA 80 Vmx	4.8%	40.1%	7.6%	35.8%	24.1%
PCA 80 Vmx (-1/1st)	3.9%	38.9%	6.2%	34.6%	26.6%
PCA 80 NoV (-1/1st)	3.0%	32.9%	3.3%	32.0%	35.1%
PCA 00 NoV (-1/1st)	3.4%	34.7%	4.8%	32.1%	33.1%
PCA 00 Vmx (-1/1st)	1.9%	37.4%	4.5%	33.0%	29.7%
PCA 00 NoV	2.6%	38.0%	4.7%	34.8%	27.1%
Notes	u.rank= uniform rank c.rank =conceptual rank Vmx= varimax rotation NoV= No varimax rotation -1/1st = -1 the 1st Principal Component PFR 80 + 83 losses Pred 08 = PFR model based on 1980 data with the addition of 1983 loss index to predict 2008				

ranking (i.e., high to low or low to high based on the vulnerability contribution of each variable). The graph below (Figure 5) summarizes the predictive performance of the traditional SV index derivation approaches and underlines their relative high failure rates, especially that of the PFR methodology. Underestimation is a critical performance assessment metric since it is directly related to potential losses of human lives. The proposed classification failure (CF) metric elaborates on this aspect of performance. In the next section, the impact of failure will be further highlighted by adding the potential human risk dimension.

4. Decision Tree Approach

To explore the applicability of DT in the vulnerability research field and identify practical advantages and limitations, the CHAID, CART, and C5.0 algorithms have been applied at a basic level without the implementation of algorithm improvement modifications (e.g., boosting, pruning, twoing, etc.) in order to obtain comparable results. These applications are presented in Table X. As seen, the C5.0 has the best classification performance (above 80% for the 1980 and 2000 classification models). This performance comes at a complexity cost since this level requires close to 400 nodes compared to the close to 50 nodes for the CHAID and CART algorithms. Overfitting is an issue with the C5.0 algorithm (see Appendix A); however, there are modifications that can reduce this “complexity” which are far beyond the scope of this paper (Pandya and Pandya, 2015). For demonstration purposes the C5.0 algorithm was applied with the significant only variables (i.e., 5 and 6 variables; identified by the importance metric; SPSS-Modeler) and not all the 15 attributes. Even these reduced classification models have an above 70% overall performance and low CF rates (see Table X). A prominent finding is the comparison of the predictive performances between the traditional SV approaches (i.e., PCA and PFR) and the classification performance of the DT algorithms. This is demonstrated in Figure 6.

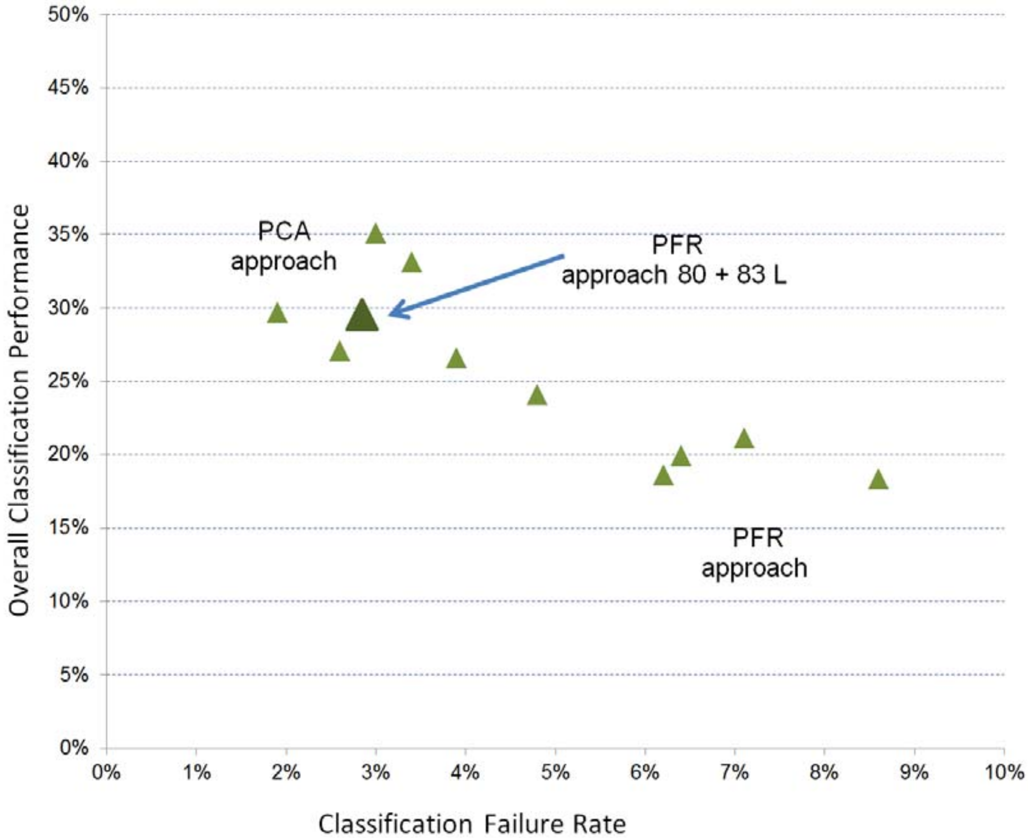


Figure 5. Predictive performance of current techniques

The CF metric can be used to identify the real dimensions of failure. For example, a 2.2% CF rate of the classification model derived by the CART algorithm, which was applied to the 1980 15 socioeconomic attributes with the 1983 loss index as a target, means that: 23 Census tracts experienced losses at the highest severity level (4) and the classification algorithm misclassified them by assigning them to the lowest level of vulnerability (1). As shown in Table XI, this signifies that, approximately, 74,000 residents of Houston (4.6% of the 1980 overall population) were located in high losses areas, which were classified as non-vulnerable.

5. Exploration and Attribute Selection

Attribute selection is a major issue that, tentatively, can be resolved by relying on the extended literature; however, a methodology for variable selection has not yet been established. Reliability remains an issue and further research is warranted. The difference in decision tree attribute selection is described herein. The selected algorithms have fundamental differences since C5.0 and CART algorithms use impurity measures to choose the splitting attribute and the split value(s), whereas CHAID uses the chi-square or the F statistic for splitting (Hapfelmeier, 2016). The C5.0 uses the gain ratio whereas the CART uses the Gini coefficient for impurity measurements (Quinlan, 1993; Breiman et al., 1984).

6. Rule Set and Decision Tree Interpretation

Displayed in Figure 7, is an example of the CHAID two-branch decision tree. As you can see this is as simplistic as you can make it for interpretation. This CHAID two-branch decision tree accounts for 43.1 % of the overall classification performance.

To explore decision tree interpretation, we expand on the branch on the housing characteristics branch; this represents neighbors greater than 14.67% of social welfare recipients have a housing characteristics (% Occupied Housing Units: Renter Occupied, % Housing

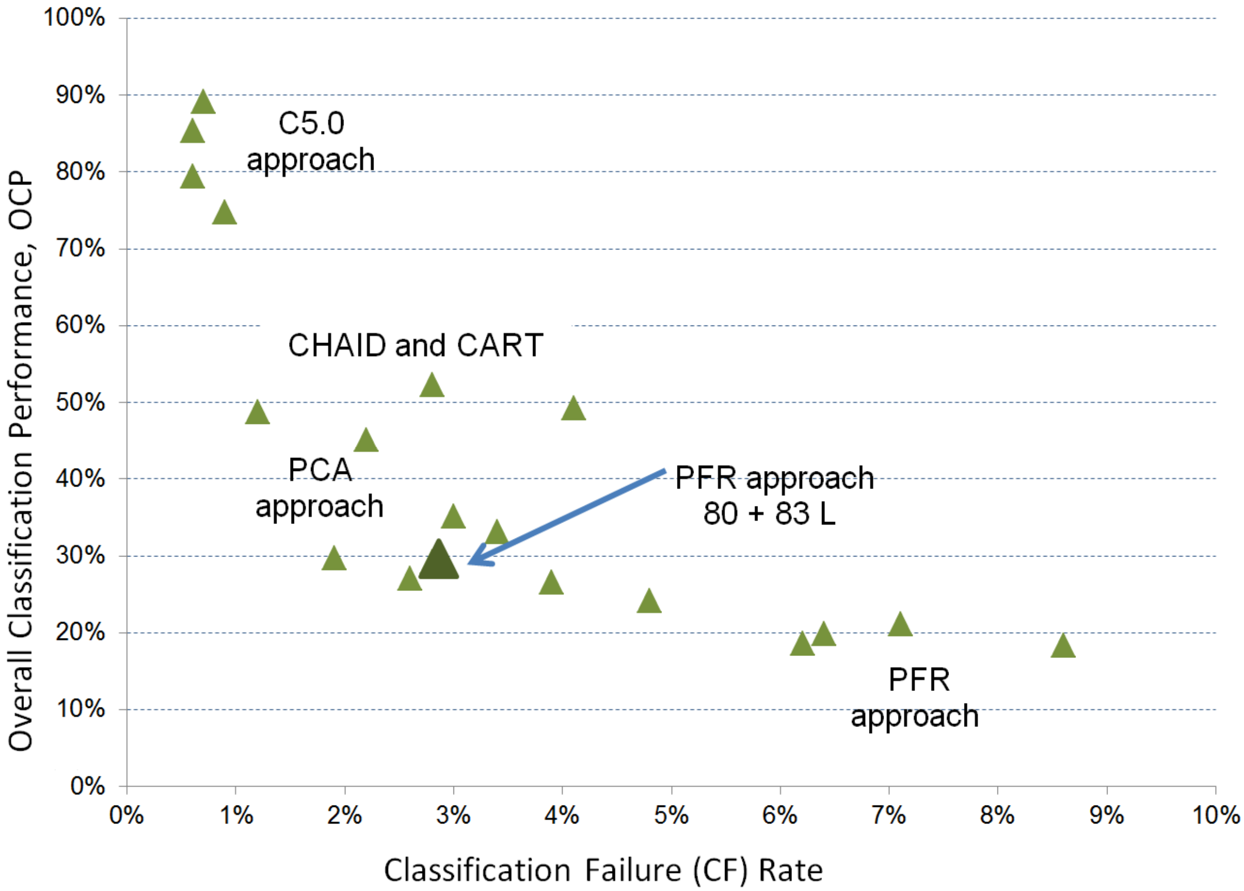


Figure 6. Decision tree predictive performance comparison to current methodologies

TABLE X

PREDICTIVE PERFORMANCE RESULTS FOR CLASSIFICATION MODELS

	CF	OUR	FC	OOR	OCP
C5.00/08	0.7%	5.6%	0.19%	5.4%	89.1%
C5.00(5)/08	0.9%	10.5%	0.28%	14.9%	74.7%
CHAID.00/08	2.8%	26.6%	0.94%	21.2%	52.3%
CART.00/08	1.2%	24.4%	2.5%	26.9%	48.7%
C5.80/83	0.6%	8.6%	0.94%	6.0%	85.4%
C5.80(6)/83	0.6%	11.3%	0.66%	9.3%	79.4%
CHAID.80/83	4.1%	28.9%	2.7%	21.8%	49.3%
CART.80/83	2.2%	28.5%	1.5%	26.4%	45.1%
Notes	C5.00/08 = C5.0 algorithm/2000 Census data / 2008 loss C5.00 (5)/08 = C5.0 algorithm/2000 Census data model with 5 attributes/ 2008 loss CHAID.00/08 = CHAID algorithm/2000 Census data / 2008 loss CART.00/08 = CART algorithm/2000 Census data / 2008 loss C5.80/83= C5.0 algorithm/1980 Census data / 1983 loss C5.80(6)/83= C5.0 algorithm/1980 Census data model with 5 attributes/ 1983 loss CHAID.80/83= CHAID algorithm/1980 Census data / 1983 loss CART.80/83= CART algorithm/1980 Census data / 1983 loss				

Units: 5 or more, % Housing Units: Mobile home or trailer, etc.) split into three categories. The p value for the calculated based on the chi square value and degrees of freedom the probability is 0.01 meaning that there is only a 1% chance that this deviation is due to chance alone and therefore, other factors must be involved.

7. **Validation**

A typical way to validate the performance of classification models is to apply them for predicting a validation set. For demonstration purposes, the initial DT classification models developed (i.e., trained) with the 1980/1983 attributes were used to predict the 2008 losses as a worst case validation scenario. As summarized in Table XII, the performance metrics and raises interesting questions about the usefulness of the C5.0 algorithm as a reliable predictive tool. As discussed earlier, a low classification failure and high overall classification performance is important in measuring predictive performance. All three decision tree models had a classification failure of 6.5% or less and the overall classification performance of between 37.3% and 53.9%. These results are much better than the fractional percentile rank and principal component analysis methodologies applied validation (results not shown). As a reminder, as depicted in Table IX, the overall classification performance for fractional percentile rank and principal component analysis was less than 29.2% and 35.1%, respectively.

A. **Summary Discussion of Results and Findings**

Vulnerability studies is a multidisciplinary field with management of natural hazards, climate change, social sciences investigators working on models that have helped explain and further develop the field of study (Ionescu et al., 2009). Concerning vulnerable population groups, it has been found that the most vulnerable counties are concentrated in the southern part of the U.S.

TABLE XI

THE IDENTITY OF CLASSIFICATION FAILURE FOR 1980 DATA

FIPS	Total Population	% under 5	% over 65	% no car
48039661200	516	0.5	0.3	0.2
48039661400	846	0.6	0.4	0.3
48157674502	776	6.4	5.3	11.1
48157675500	1654	16.7	16.1	0.0
48167720700	3248	7.4	7.5	7.9
48167721500	2510	6.9	6.1	1.5
48201241500	5194	28.6	16.8	7.2
48201310100	5567	20.4	22.2	73.5
48201312500	3349	7.2	9.6	40.4
48201312600	6224	11.0	14.0	44.3
48201313800	3987	8.7	10.6	26.4
48201330800	1690	24.2	25.9	7.3
48201331400	2104	13.1	5.7	43.2
48201333902	2290	7.5	3.2	0.0
48201410100	6929	10.7	10.9	59.0
48201412200	4651	4.6	19.7	8.2
48201431700	3240	3.2	19.0	1.3
48201520500	7153	28.1	19.1	21.9
48201543200	4640	10.6	8.8	0.7
48201551900	400	0.3	0.1	0.0
48201555402	3106	6.8	14.6	3.1
48339690602	125	0.2	0.1	0.0
48339692300	3837	8.5	3.9	0.6
Median CF Sample	3,240	7.5	9.6	7.2
Median (all tracts)	3,924	4.2	2.6	1.3

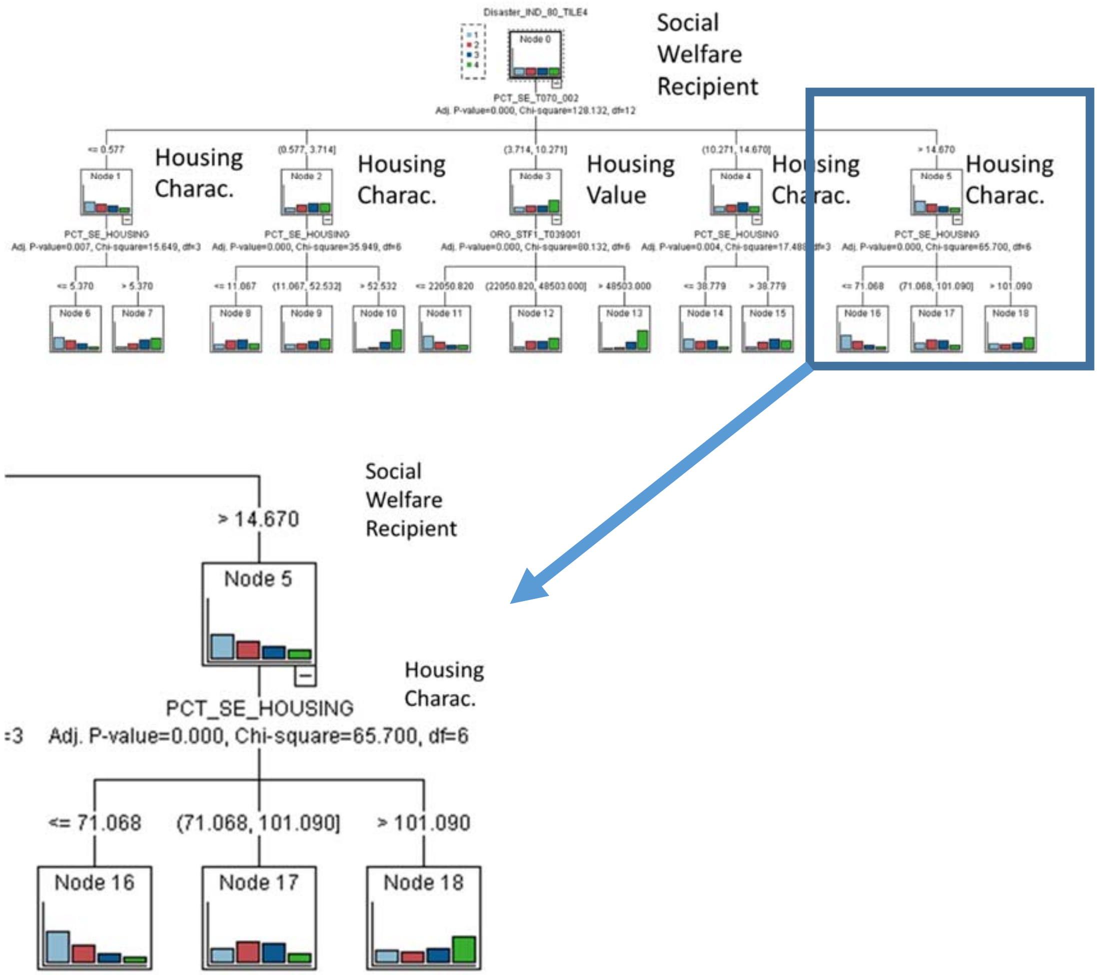


Figure 7. CHAID decision tree results

TABLE XII

VALIDATION OF CLASSIFICATION MODELS

Validation of classification models					
	CF	OUR	FC	OOR	OCP
C5.0.80.p08	2.2%	22.5%	1.3%	23.6%	53.9%
CHAID.80.p08	6.5%	37.5%	2.1%	22.0%	40.5%
CART.80.p08	3.0%	31.8%	1.9%	30.9%	37.3%

because of racial and ethnic disparities and fast population growth (Cutter, 2006). In the research of climate change for instance, vulnerability is frequently assumed to mean hazard exposure, susceptibility to that hazard, and the emergency response to a hazard (Parry et al., 2007). Risk is understood to be outside of common society structures in the prevailing versions of vulnerability (McLaughlin and Dietz, 2008) and can be estimated for a static area and hazard. However, what has not been documented well, the capability of adaptation, hazard exposure and long-term damages after a disaster which can vary over time (Turner et al., 2003).

Vulnerability is always defined in relation to a specified time period, system, hazard, or range of hazards (IPCC, 2014). The term hazard refers specifically to physical manifestations of climatic extreme events or change, such as floods, droughts, heatwaves, etc. A disaster (e.g., Hurricanes Alicia and Ike) is an outcome of a hazard, facilitated by the properties of the human system that is exposed to and affected by the hazard. Ideally, as stated in the Intergovernmental Panel on Climate Change (IPCC) Third Assessment Report in 2001, "*the vulnerability of a human system can be determined by the nature of the physical hazard(s) to which it is exposed, the likelihood or frequency of occurrence of the hazard(s), the extent of human exposure to* *hazard, and the system's sensitivity to the impacts of the hazard(s).*" This composite vulnerability, which is a function of hazard, exposure, and sensitivity, is usually referred to as biophysical vulnerability (Based on IPCC report, 2001). For human systems (e.g., communities, etc.), vulnerability is viewed as an inherent property of the system stemming from its internal characteristics is usually referred to as social vulnerability. Subsequently, SVI is considered a pre-existing condition of human systems, irrespective of the natural hazard of interest. The place vulnerability (Hazard-of-place framework) is the interaction of these two vulnerabilities. Event risk (ER) as the "*risk of occurrence of any particular hazard (H) or extreme event*" and outcome risk (OR) as "*the risk of a particular outcome*" or "*integrates both the characteristics of a system and the chance of the occurrence of an event that jointly results in losses.*" (Sarewitz et al., 2003). SVI describes inherent characteristics of a system that create the potential for harm and independent of the probabilistic risk of occurrence (ER) of any particular hazard or extreme event. Derivation of indicators and indices "*quantitative measures intended to represent a characteristic or a parameter of a system of interest*" (Cutter et al. 2008). The objective for using indices is to provide a quick and consistent method for depicting vulnerability and recognizing issues that may need to be addressed. In terms of the present study, SVI have to be reliable quantitative measures intended to represent the inherent multidimensional characteristics of a human system that create the potential for harm. Issues related to the creation of a SVI:

1. Which variables (attributes) to include as the base for the index?
2. How should these variables be "combined" to derive the index?
3. Is the SVI reliable?

Variable selection (Issue 1) is a major issue that, tentatively, can be resolved by relying on the extended literature; however, a methodology for variable selection has not yet been established. Climate change adaptation plans tend to exaggerate future variability of weather, which may lead to undesirable outcomes (Macintosh, 2013). Research specifies that exposure to a hazard and vulnerability, or the differences in features for example socioeconomic, race, and household composition (*inter alia*), matters when forecasting the effects of natural hazards (Highfield et al., 2014). Adaptation methods that disregard the social undercurrents of a metropolitan area can overlook population at risk to these natural disasters. This becomes increasingly difficult in a multi natural hazard metropolitan area where regulations and guidelines can supplementary to climate adaptation and mitigation, and the vulnerable population can not simply find it reasonable to dwell in places which are non-hazardous.

This standpoint highlights the socioeconomic features that guide a community's capability to mitigate and adapt to a hazard event (Cutter et al., 2003; Laska and Morrow, 2006; Peacock et al., 1997) and is frequently described using individual characteristics (e.g., age, race, etc.). This approach sets disasters and their impacts within a wider social contexts (Wisner et al., 2004) and highlights social factors that affect the proneness of various of groups to harm (Cutter et al., 2003). Studies of social vulnerability have comprehensively documented the inconsistent impacts of hazardous events on socially vulnerable population groups (Cutter, 1996; Cutter et al., 2003; Fothergill and Peek, 2004; Highfield et al., 2014; Peacock et al., 2007; Zahran et al., 2008). To answer Issue 2, since areas of vulnerability differ from place to place and over time in the same region, the SVI was calculated for each Census tract and decadal period, separately. Figures located in Appendix C, show vulnerability maps for the study area using the three different derivation methods (i.e., PFR, PCA, and DT).

As depicted in this study there are implications to sensitivity when changes in index construction to create vulnerability indices. At present, the adequacy of the representation of vulnerability produced by the index and depicted in maps can only be corroborated by local expert knowledge of the community. Currently, the traditional SVI need to be coupled with expert guidance and additional, local, information to ensure that the representations of vulnerability produced are reasonable and consistent with locally based geographic knowledge of the study area (Flanagan et al., 2011). The importance of expert judgment in the index creation process is not limited to validation of vulnerability representation. Expert judgment is also a critical element in the subjective interpretation of the components generated by the PCA. These components must be interpreted to determine whether the pc-scores are assigned a positive, negative or absolute value before they are combined to create the index. The literature suggests future research on the SVI algorithm could be designed to assess the impact of changes in the interpretation of components on the final index and provide more concrete guidance to this crucial element. However, using the proposed predictive performance metrics researchers can begin to measure the reliability of SVI derivation methodologies in terms of their ability to identify vulnerable areas prone to loss; thus, limiting the reliance on expert judgement information. Furthermore, the predictive performance of the new decision tree methodology for deriving vulnerable areas outperforms significantly the current methodologies (i.e., PR and PCA).

Using the historical data, the probabilities of disaster occurrence have been found for each country in the world and for each kind of disaster. These probabilities have been used to build the Decision tree models which will use as a forecasting tool to predict the probability of disaster and all the variables. Furthermore, Decision trees are usually used to provision decision-making in an uncertain situation. These forecasts are used by the various international and national humanitarian organizations in emergency logistics planning. This leads to better coordination of search and rescue activities and efficient evacuation of injured people. Furthermore, overall health conditions of everyone in the affected area depend on the timely availability of commodities such as food shelter and medicine. Issue 3 is discussed in detail in the section below.

a. **<u>Responses to the Research Foci</u>**

As mentioned previously, the following research foci were pursued:

1. Is social vulnerability a classification issue?
2. In the context of classification, is there the ability to assess the predictive performance of current or newly proposed SVI derivation techniques?
3. Are existing SVI derivation methodologies reliable, useful, and feasible?
4. Is the newly proposed classification based methodology reliable, useful and feasible?

Based on the new classification framework proposed in this study and the derived methodologies for performance assessment and modeling these critical for the SV research field questions were answered. The first research focus question, as presented in the empirical evidence section, was answered by examining the ultimate and practical usage of the derived index. In the majority of cases, indeed, the derived indices were used implicitly as a classification tool (Cutter et al., 2003; Chakraborty et al., 2005; Flanagan et al., 2011; Schmidtlein et al., 2008; Yoon, 2012; see also Table III). The relevant literature of national and international organizations and agencies corroborates this supposition since the majority of the published reports aim to generate the spatial distribution of regions classified in terms of vulnerability.

In response to the second research focus question, the present study developed classification predictive performance metrics, which were applied for assessing the performance of the current major methodologies (i.e., PR and PCA). The proposed predictive performance metrics are able to identify critical misclassified hotspots (i.e., CF metric) and quantify the extent of identification error. In addition, with the use of the proposed metrics a comprehensive assessment of SVI derivation approaches can be conducted and assessed based on quantifiable criteria (i.e., a reliable SVI model must have a low OUR, OOR, CF and FC and a high OCP). A major breakthrough in this field of study is the introduction of the risk/loss estimates derived with the application of Hazus®. These estimates and the proposed metrics provide a reliable tool to assess proposed SVI based on actual or simulated loss data.

In response to the third research focus question, the predictive performance of the existing methodologies seems to be problematic since in most of the examined cases the OUR is higher than the OCP one. The PFR methodology is likely to be unstable since the uniform ranking (i.e., all variables rank the same) performs better that the logical conceptual ranking (i.e., high to low or low to high based on the vulnerability contribution of each variable). As demonstration in the current study the existing percentile rank methodology may be useful because it is a viable and attractive alternative when compared to the complexity of principal component analysis especially when it comes to interpretation of the results. Lastly, since small scale emergency management departments will not be able to perform the preliminary data collection and the required data preparation tasks without the support of specialized experts they will not be able to derive SVI based on the principal component analysis approach.

In response to the fourth research focus question, the new proposed decision tree approach is shown to be extremely reliable. As shown in Table IX, the CF is 4.1% or below and above 45.1% or above OCP which is higher than existing methodologies. The attributes used for the input data set "*need to be cheap, reliable, recent, routine, and at a sufficient spatial resolution*"

(Wolf et al., 2013). This criterion is related to usefulness since small scale emergency management departments will not be able to perform the preliminary data collection and preparation tasks without the support of specialized experts. The decision tree algorithm can be built into a software application so emergency departments can take disaster loss target data (easily exported from existing software like Hazus®) and import local Census data of their area of evaluation; in addition, there is no large database limitations.

A further application of the proposed DT analysis approach is the use of the rule sets that are generated. These rule sets can be applied directly with a database access language (e.g., SQL or by using a simple queries) so that counties falling into a particular category (i.e., hotspot) may be identified without the use of complicated modeling techniques and expensive experts.

IV. SUMMARY OF MAJOR CONCLUSIONS AND FINDINGS

This study demonstrated that social vulnerability studies are, in the final analysis, a classification issue. Conducting SV studies with the use of classification modeling techniques has a number of advantages and overcomes some of the limitations of traditional techniques especially in terms of developing classification models that offer a reliable predictive performance.

The predictive performance metrics proposed in this study provide a valuable tool for assessing social vulnerability indices. This study demonstrated that with the use of the Hazus® software a loss index can be established which could be used to assess the indices with actual historical, or simulated, disaster losses. The proposed predictive performance metrics in combination with the Hazus® derived loss index have the potential to establish a standard of assessing social vulnerability indices in terms of reliability and usefulness.

The creation of potential loss scenarios with the use of the Hazus® software offers a great opportunity for expanded use of this predictive performance methodology since indices can be assessed based on the generated losses of the area of interest.

The introduction of classification models offers an attractive venue for developing reliable and useful techniques to identify areas at high risk. In addition, they are likely to offer practical tools to emergency management agencies for identifying and exploring vulnerable (to losses) sections of their cities.

V. LIMITATIONS

1. The spatial dimension of the SVI was not explored.
2. One only major urban region was used (Houston MSA).
3. Variable selection was limited to tract level Census data and two time periods.

VI. FURTHER RESEARCH

1. Optimization of C5.0 (e.g., pruning) and comparison of performance with other DT methodologies.
2. Overfitting and underfitting issues in relation to CF metric.
3. Stability and sensitivity of DT models.
4. Usefulness of rule sets for emergency management personnel with SQL implementation.
5. Sensitivity of binning approach.
6. Application of the DT approach to other areas with different characteristics.
7. Application to simulated losses.

.

APPENDICES

Appendix A
EQUATIONS

$$\text{Risk} = \text{Hazard} * (\text{Vulnerability} - \text{Resources})$$

Equation 1. Risk Equation - Retrieved from UCLA documentation (2008)

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

where N = the total number of data points, and all sequences of ties are assigned the smallest of the corresponding ranks

Equation 2. Percentile Rank Equation - Retrieved from Flanagan et al. (2011)

$$\text{Risk} = \text{Hazard x Vulnerability or Risk} = f\,(\text{Hazard, Vulnerability})$$

Equation 3. Risk Equation - Retrieved from UNISDR documentation (2009)

$$\text{Risk} = f\ [\text{Hazard}, (\text{Vulnerability}= f\,(S, (AC= f\,(SC,R))))]$$

Equation 4. Risk Equation - Retrieved from UNISDR documentation (2009)

$$Z = (\text{score} - \text{mean})/\text{standard deviation}$$

Equation 5. Percentile Rank Equation - Retrieved from Yoon (2012)

Appendix A (continued)

$$V_{\mathrm{i}} = \frac{X_{\mathrm{i}} - X_{\min}}{X_{\max} - X_{\min}}$$

Where Xmax is the maximum value
Where Xmin is the minimium value

Equation 6. Min-Max Rescaling Equation - Retrieved from Yoon (2012)

$$\text{OCP} = \sum_{i=1}^{n} C_{i,i} \,/\text{N}$$

Equation 7. Overall Classification Performance (OCP)

$$\text{CF} = C_{1,n} \,/\text{N}$$

Equation 8. Classification Failure (CF)

$$\text{FC} = C_{n,1} \,/\text{N}$$

Equation 9. False Classification (FC)

Appendix B
MODLER OUTPUTS

Binning method: Tiles (equal count)

Binned field: SVI_IND

Tile: 4

Bins will be created using the values shown in the table

Bin	Lower	Upper
1	>= 324.64220001	<= 607.81323332
2	> 607.81323332	<= 759.68316435
3	> 759.68316435	<= 903.87353297
4	> 903.87353297	<= 1135.28862102

Binning Approach

DisVul_IND_TILE4

SVI_PCA_3 all_TILE4		1	2	3	4	Total
1	Count	71	80	73	42	266
	Row %	26.692	30.075	27.444	15.789	100
	Column %	26.792	30.075	27.444	15.849	25.047
	Total %	6.685	7.533	6.874	3.955	25.047
2	Count	68	58	70	70	266
	Row %	25.564	21.805	26.316	26.316	100
	Column %	25.660	21.805	26.316	26.415	25.047
	Total %	6.403	5.461	6.591	6.591	25.047
3	Count	69	76	61	60	266
	Row %	25.940	28.571	22.932	22.556	100
	Column %	26.038	28.571	22.932	22.642	25.047
	Total %	6.497	7.156	5.744	5.650	25.047
4	Count	57	52	62	93	264
	Row %	21.591	19.697	23.485	35.227	100
	Column %	21.509	19.549	23.308	35.094	24.859
	Total %	5.367	4.896	5.838	8.757	24.859
Total	Count	265	266	266	265	1062
	Row %	24.953	25.047	25.047	24.953	100
	Column %	100	100	100	100	100
	Total %	24.953	25.047	25.047	24.953	100

PCA Coincidence of 80 data / 83 disaster

Appendix B (continued)

		DisVul_IND_TILE4				
SVI_PCA_3 all_TILE4		1	2	3	4	Total
1	Count	73	80	67	46	266
	Row %	27.444	30.075	25.188	17.293	100
	Column %	27.547	30.075	25.188	17.358	25.047
	Total %	6.874	7.533	6.309	4.331	25.047
2	Count	56	65	81	64	266
	Row %	21.053	24.436	30.451	24.060	100
	Column %	21.132	24.436	30.451	24.151	25.047
	Total %	5.273	6.121	7.627	6.026	25.047
3	Count	64	65	63	74	266
	Row %	24.060	24.436	23.684	27.820	100
	Column %	24.151	24.436	23.684	27.925	25.047
	Total %	6.026	6.121	5.932	6.968	25.047
4	Count	72	56	55	81	264
	Row %	27.273	21.212	20.833	30.682	100
	Column %	27.170	21.053	20.677	30.566	24.859
	Total %	6.780	5.273	5.179	7.627	24.859
Total	Count	265	266	266	265	1062
	Row %	24.953	25.047	25.047	24.953	100
	Column %	100	100	100	100	100
	Total %	24.953	25.047	25.047	24.953	100

PCA Coincidence of 80 data / 83 disaster with VariMax Rotations

		DisVul_IND_TILE4				
SVI_PCA_3 all_TILE4		1	2	3	4	Total
1	Count	110	79	51	26	266
	Row %	41.353	29.699	19.173	9.774	100
	Column %	41.509	29.699	19.173	9.811	25.047
	Total %	10.358	7.439	4.802	2.448	25.047
2	Count	67	79	75	45	266
	Row %	25.188	29.699	28.195	16.917	100
	Column %	25.283	29.699	28.195	16.981	25.047
	Total %	6.309	7.439	7.062	4.237	25.047
3	Count	47	53	79	87	266
	Row %	17.669	19.925	29.699	32.707	100
	Column %	17.736	19.925	29.699	32.830	25.047
	Total %	4.426	4.991	7.439	8.192	25.047
4	Count	41	55	61	107	264
	Row %	15.530	20.833	23.106	40.530	100
	Column %	15.472	20.677	22.932	40.377	24.859
	Total %	3.861	5.179	5.744	10.075	24.859
Total	Count	265	266	266	265	1062
	Row %	24.953	25.047	25.047	24.953	100
	Column %	100	100	100	100	100
	Total %	24.953	25.047	25.047	24.953	100

PCA Coincidence of 80 data / 83 disaster with Absolute Value

Appendix B (continued)

		DisVul_IND_TILE4				
SVI_PCA_4 all_TILE4		1	2	3	4	Total
1	Count	82	93	63	28	266
	Row %	30.827	34.962	23.684	10.526	100
	Column %	30.827	34.962	23.684	10.606	25.047
	Total %	7.721	8.757	5.932	2.637	25.047
2	Count	74	53	68	71	266
	Row %	27.820	19.925	25.564	26.692	100
	Column %	27.820	19.925	25.564	26.894	25.047
	Total %	6.968	4.991	6.403	6.685	25.047
3	Count	60	56	69	81	266
	Row %	22.556	21.053	25.940	30.451	100
	Column %	22.556	21.053	25.940	30.682	25.047
	Total %	5.650	5.273	6.497	7.627	25.047
4	Count	50	64	66	84	264
	Row %	18.939	24.242	25.000	31.818	100
	Column %	18.797	24.060	24.812	31.818	24.859
	Total %	4.708	6.026	6.215	7.910	24.859
Total	Count	266	266	266	264	1062
	Row %	25.047	25.047	25.047	24.859	100
	Column %	100	100	100	100	100
	Total %	25.047	25.047	25.047	24.859	100

PCA Coincidence of 00 data / 08 disaster

		DisVul_IND_TILE4				
SVI_PCA_4 all_TILE4		1	2	3	4	Total
1	Count	66	90	68	42	266
	Row %	24.812	33.835	25.564	15.789	100
	Column %	24.812	33.835	25.564	15.909	25.047
	Total %	6.215	8.475	6.403	3.955	25.047
2	Count	65	57	70	74	266
	Row %	24.436	21.429	26.316	27.820	100
	Column %	24.436	21.429	26.316	28.030	25.047
	Total %	6.121	5.367	6.591	6.968	25.047
3	Count	76	55	62	73	266
	Row %	28.571	20.677	23.308	27.444	100
	Column %	28.571	20.677	23.308	27.652	25.047
	Total %	7.156	5.179	5.838	6.874	25.047
4	Count	59	64	66	75	264
	Row %	22.348	24.242	25.000	28.409	100
	Column %	22.180	24.060	24.812	28.409	24.859
	Total %	5.556	6.026	6.215	7.062	24.859
Total	Count	266	266	266	264	1062
	Row %	25.047	25.047	25.047	24.859	100
	Column %	100	100	100	100	100
	Total %	25.047	25.047	25.047	24.859	100

PCA Coincidence of 00 data / 08 disaster with VariMax Rotations

Appendix B (continued)

		DisVul_IND_TILE4				
SVI_PCA_4 all_TILE4		1	2	3	4	Total
1	Count	97	70	63	36	266
	Row %	36.466	26.316	23.684	13.534	100
	Column %	36.466	26.316	23.684	13.636	25.047
	Total %	9.134	6.591	5.932	3.390	25.047
2	Count	69	82	68	47	266
	Row %	25.940	30.827	25.564	17.669	100
	Column %	25.940	30.827	25.564	17.803	25.047
	Total %	6.497	7.721	6.403	4.426	25.047
3	Count	49	55	77	85	266
	Row %	18.421	20.677	28.947	31.955	100
	Column %	18.421	20.677	28.947	32.197	25.047
	Total %	4.614	5.179	7.250	8.004	25.047
4	Count	51	59	58	96	264
	Row %	19.318	22.348	21.970	36.364	100
	Column %	19.173	22.180	21.805	36.364	24.859
	Total %	4.802	5.556	5.461	9.040	24.859
Total	Count	266	266	266	264	1062
	Row %	25.047	25.047	25.047	24.859	100
	Column %	100	100	100	100	100
	Total %	25.047	25.047	25.047	24.859	100

PCA Coincidence of 00 data / 08 disaster with Absolute Values

		DisVul_IND_TILE4				
SVI_PercRank_IND_TILE4		1	2	3	4	Total
1	Count	69	88	62	47	266
	Row %	25.940	33.083	23.308	17.669	100
	Column %	25.940	33.083	23.308	17.803	25.047
	Total %	6.497	8.286	5.838	4.426	25.047
2	Count	59	58	71	78	266
	Row %	22.180	21.805	26.692	29.323	100
	Column %	22.180	21.805	26.692	29.545	25.047
	Total %	5.556	5.461	6.685	7.345	25.047
3	Count	75	58	57	76	266
	Row %	28.195	21.805	21.429	28.571	100
	Column %	28.195	21.805	21.429	28.788	25.047
	Total %	7.062	5.461	5.367	7.156	25.047
4	Count	63	62	76	63	264
	Row %	23.864	23.485	28.788	23.864	100
	Column %	23.684	23.308	28.571	23.864	24.859
	Total %	5.932	5.838	7.156	5.932	24.859
Total	Count	266	266	266	264	1062
	Row %	25.047	25.047	25.047	24.859	100
	Column %	100	100	100	100	100
	Total %	25.047	25.047	25.047	24.859	100

Percentile Rank Coincidence of 00 data / 08 disaster

Appendix B (continued)

Results for output field DisVul_IND_TILE4

Comparing $C-DisVul_IND_TILE4 with DisVul_IND_TILE4

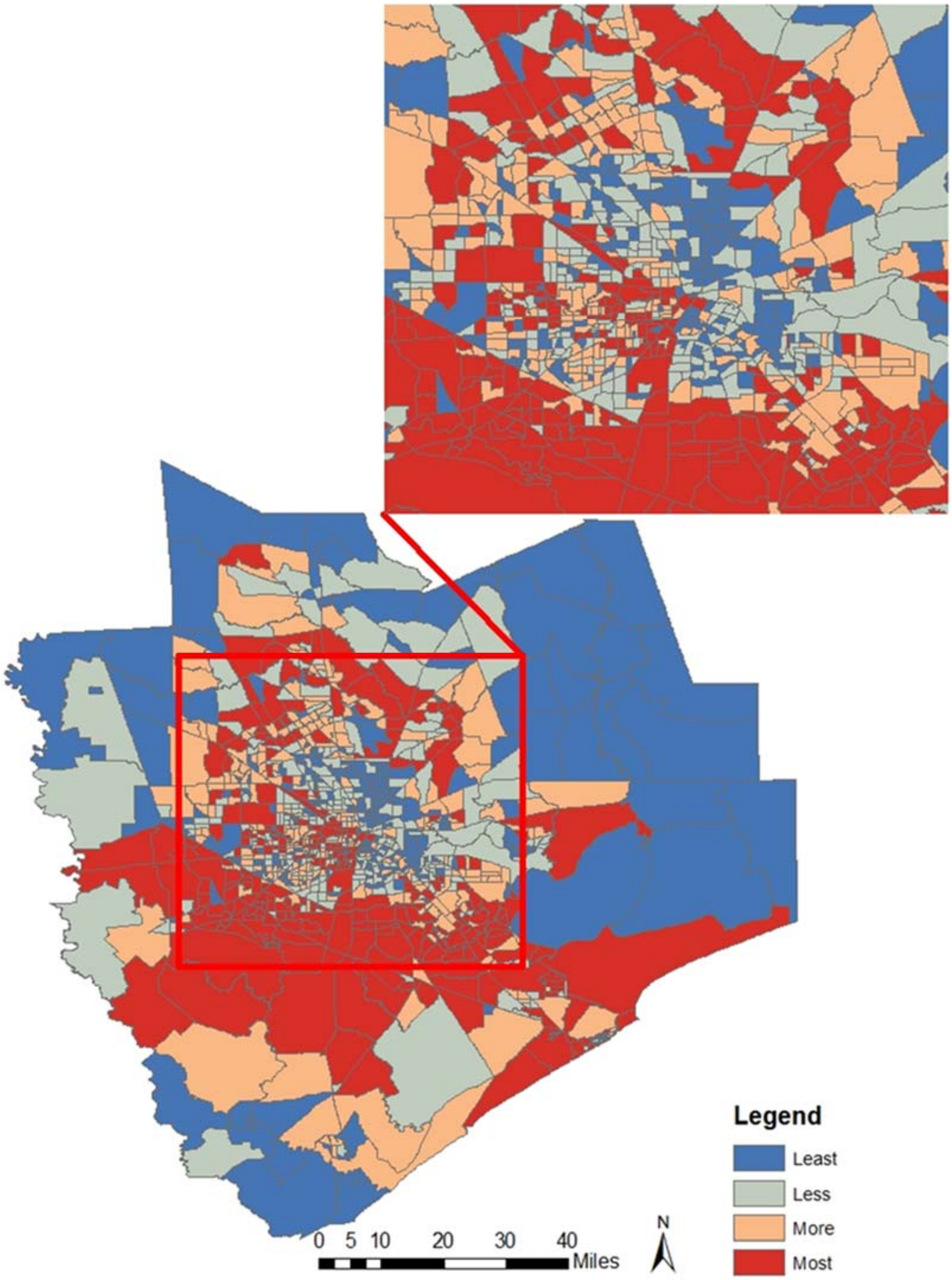
Correct	946	89.08%
Wrong	116	10.92%
Total	1,062	

Coincidence Matrix for $C-DisVul_IND_TILE4 (rows show actuals)

	1	**2**	**3**	**4**
1	252	7	5	2
2	8	237	13	8
3	18	13	213	22
4	7	4	9	244

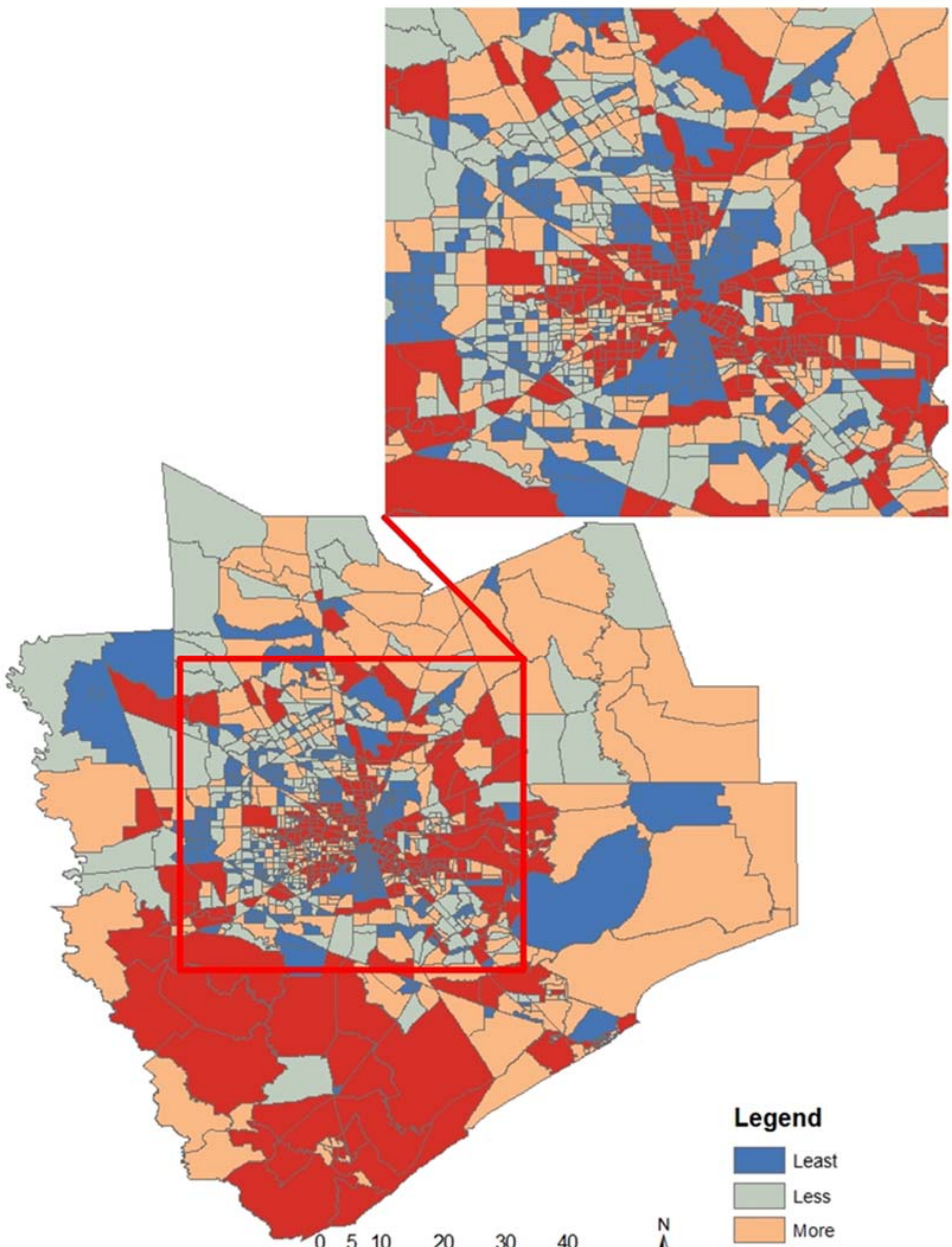
Decision Tree Four Tile Predictive Performance

Appendix C
MAPS



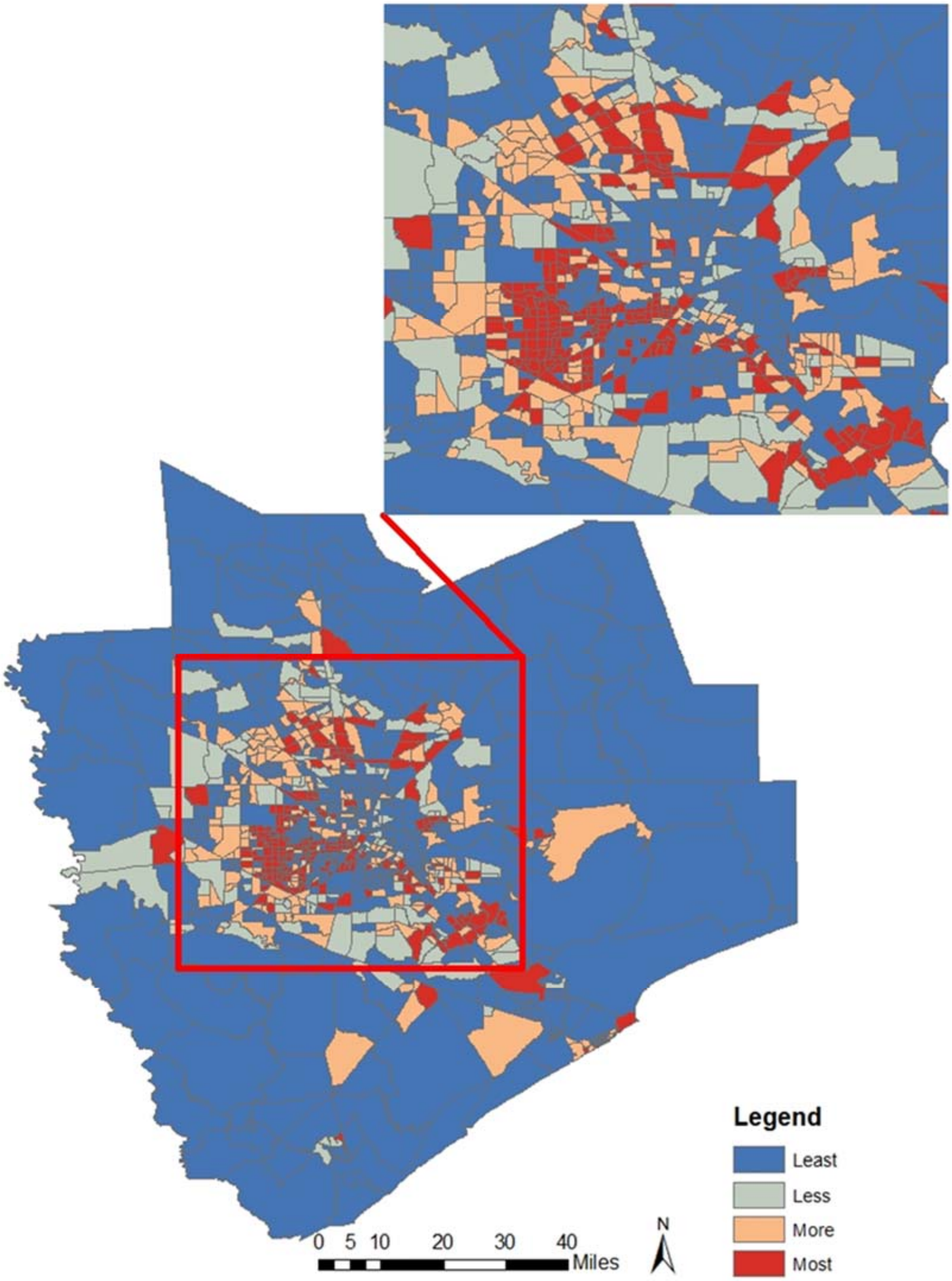
1983 Loss Map

Appendix C (continued)



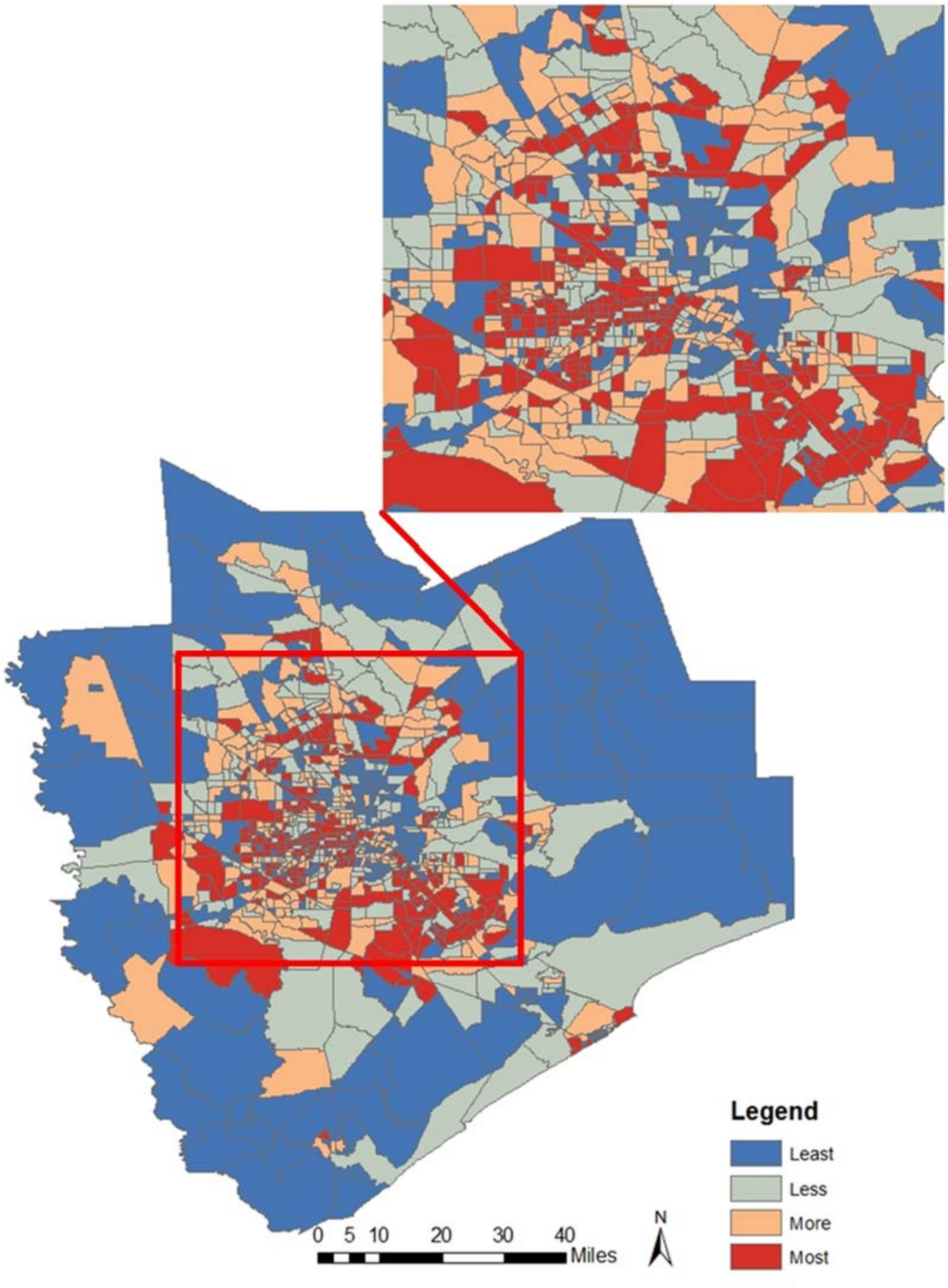
1980 PCA Map

Appendix C (continued)



1980 CHAID DT Map

Appendix C (continued)



1980 C50 DT Map

CITED LITERATURE

Abson, D. J., Dougill, A. J., & Stringer, L. C. (2012). Using principal component analysis for information-rich socio-ecological vulnerability mapping in Southern Africa. *Applied Geography*, *35*(1), 515-524.

Anderson, W. A. (2005). Bringing Children into Focus on the Social Science Disaster Research Agenda. *International Journal of Mass Emergencies and Disasters*, *23*(3), 159–175.

Asante-Duah, Kofi Public Health Risk Assessment for Human Exposure to Chemicals. *Environmental Pollution*, 2002 Vol 6.

Bakhsh, H. (2015). *Mapping Social Vulnerability and Exposure Parameters to Extreme Heat Events in Missouri* (Doctoral dissertation, Rice University).

Barnett, J., & O'Neill, S. (2010). Maladaptation. *Global Environmental Change*, *20*, 211–213. doi:10.1016/j.gloenvcha.2009.11.004

Barsley, W., De Young, C & Brugère, C. 2013. Vulnerability assessment methodologies: an annotated bibliography for climate change and the fisheries and aquaculture sector. *FAO Fisheries and Aquaculture Circular No. 1083*. Rome, FAO. 43 pp

Berrang-Ford, L., Pearce, T., & Ford, J. D. (2015). Systematic review approaches for climate change adaptation research. *Regional Environmental Change*, *15*(5), 755-769.

Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Bigss, D., Ville, B., and Suen, E. (1991). A Method of Choosing Multiway Partitions for Classification and Decision Trees. *Journal of Applied Statistics*, 18, 1, 49-62.

Birkmann, J. (2006). *Measuring Vulnerability to Natural Hazards: Towards Disaster Resilient Societies*: United Nations University Press.

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (1994). *At Risk: Natural Hazards, People's Vulnerability and Disasters* (1st ed.). London: Routledge.

Blaikie, P., Cannon, T., Davis, I., & Wisner, B. (2014). *At risk: natural hazards, people's vulnerability and disasters*. Routledge.

Blanchard-Boehm, R. D. (1998). Understanding Public Response to Increased Risk from Natural Hazards. *International Journals of Mass Emergencies and Disasters*, *16*(3), 247–278.

Blanco, H., Alberti, M., Olshansky, R., Chang, S., Wheeler, S. M., Randolph, J., Watson, V. (2009). Shaken, shrinking, hot, impoverished and informal: Emerging research agendas in planning. *Progress in Planning*, *72*(4), 195–250. doi:10.1016/j.progress.2009.09.001

Brassett, J., & Vaughan-Williams, N. (2015). Security and the performative politics of resilience: Critical infrastructure protection and humanitarian emergency preparedness. *Security Dialogue,* 46(1), 32-50.

Breiman L., Friedman J., Olshen R., and Stone C.. Classification and Regression Trees. Wadsworth Int. Group, 1984.

Briguglio, L. (2003). The vulnerability index and small island developing states: A review of conceptual and methodological issues. In*AIMS Regional Preparatory Meeting on the Ten Year Review of the Barbados Programme of Action: Praia, Cape Verde*.

Bruin, J. 2006. Stata Programs for Data Analysis. UCLA: Statistical Consulting Group. Retrieved January 6, 2014, from http://www.ats.ucla.edu/stat/stata/ado/analysis/.

Bunting, P., Clewley, D., Lucas, R. M., & Gillingham, S. (2014). The remote sensing and GIS software library (RSGISLib). *Computers & geosciences*, *62*, 216-226.

Burby, R. J. (1998). *Cooperating with nature: Confronting natural hazards with land use planning for sustainable communities*. Washington, DC: Joseph Henry Press.

Cardona, O.D., M.K. van Aalst, J. Birkmann, M. Fordham, G. McGregor, R. Perez, R.S. Pulwarty, E.L.F. Schipper, and B.T. Sinh, 2012: Determinants of risk: exposure and vulnerability. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation [Field, C.B., V. Barros, T.F. Stocker, D. Qin, D.J. Dokken, K.L. Ebi, M.D. Mastrandrea, K.J. Mach, G.-K. Plattner, S.K. Allen, M. Tignor, and P.M. Midgley (eds.)]. *A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC)*. Cambridge University Press, Cambridge, UK, and New York, NY, USA, pp. 65-108.

Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. Social Vulnerability Index 2010 Database Texas. http://svi.cdc.gov/SVIDataToolsDownload.html. Accessed on May 20, 2015.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*, *8*(6), 866-883.

City of Houston Official Website (2014). http://www.houstontx.gov/abouthouston/houstonfacts.html. Accessed on January 1, 2016.

Clark, G., Moser, S., Ratick, S., Dow, K., Meyer, W., Emani, S., . . . Schwarz, H. (1998). Assessing the Vulnerability of Coastal Communities to Extreme Storms: The Case of Revere, MA., USA. *Mitigation and Adaptation Strategies for Global Change, 3*(1), 59-82. doi: 10.1023/A:1009609710795

Crowder, K., & Downey, L. (2010). Interneighborhood migration, race, and environmental hazards: modeling microlevel processes of environmental inequality. *AJS; American Journal of Sociology*, *115*, 1110–1149. doi:10.1086/649576

Cureton, E. E., & D'Agostino, R. B. (2013). *Factor analysis: An applied approach*. Psychology Press.

Cutter, S. L. (1996). Vulnerability to environmental hazards. *Progress in Human Geography*, *20*, 529–539. doi:10.1177/030913259602000407

Cutter, S. L., Mitchell, J. T., & Scott, M. S. (2000). Revealing the vulnerability of people and places: a case study of Georgetown County, South Carolina. *Annals of the association of American Geographers*, 90(4), 713-737.

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social Vulnerability to Environmental Hazards. *Social Science Quarterly, 84*(2), 242-261. doi: 10.1111/1540-6237.8402002

Cutter, S. L., Johnson, L. A., Finch, C., & Berry, M. (2007). The U.S. Hurricane Coasts: Increasingly Vulnerable? *Environment: Science and Policy for Sustainable Development*, *49*(7), 8–21. doi:10.3200/ENVT.49.7.8-21

Cutter, S. L., & Finch, C. (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), 2301–2306. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2268131/

Cutter, S. L., Emrich, C. T., Webb, J. J., & Morath, D. (2009). Social vulnerability to climate variability hazards: a review of the literature. *Final Report to Oxfam America*, *5*.

Dada, D. M. (2015). An Investigation of Social Vulnerability and TRI-reporting Facilities in the Houston Metropolitan Area. (Doctoral dissertation, University of Illinois at Chicago).

Davidson, P., & Anderton, D. L. (2000). Demographics of dumping. II: A national environmental equity survey and the distribution of hazardous materials handlers. *Demography, 37*(4), 461-466.

Demšar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: an overview. *Annals of the Association of American Geographers*, *103*(1), 106-128.

De Smith, M. J., Goodchild, M. F., & Longley, P. (2007). *Geospatial analysis: a comprehensive guide to principles, techniques and software tools*. Troubador Publishing Ltd.

Deyle, R. E., Chapin, T. S., & Baker, E. J. (2008). The Proof of the Planning Is in the Platting: An Evaluation of Florida‘s Hurricane Exposure Mitigation Planning Mandate. *Journal of the American Planning Association*, *74*(3), 349–370. doi:10.1080/01944360802229612

Dolan, G., & Messen, D. (2012). Social vulnerability: An emergency managers' planning tool. *J Emerg.Manag., 10*(3), 161-169.

Eakin, H., & Luers, A. L. (2006). Assessing the Vulnerability of Social-Environmental Systems. *Annual Review of Environment and Resources*, *31*(1), 365–394. doi:10.1146/annurev.energy.30.050504.144352

Elliott, J. R., & Pais, J. (2006). Race, class, and Hurricane Katrina: Social differences in human responses to disaster. *Social Science Research, 35*(2), 295-321. doi: http://dx.doi.org/10.1016/j.ssresearch.2006.02.003

Emrich, C. T., & Cutter, S. L. (2011). Social vulnerability to climate-sensitive hazards in the southern United States. *Weather, Climate, and Society*, 3(3), 193-208.

Enarson, E., Fothergill, A., & Peek, L. (2007). Gender and Disaster : Foundations and Directions. In E. L. Quarantelli, R. Dynes, & H. Rodriguez (Eds.), *Handbook of Disaster Research* (pp. 130–146). Springer New York. doi:10.1007/978-0-387-32353-4_8

Enarson, E., & Morrow, B. H. (1998). *The gendered terrain of disaster: through women's eyes*. Miami, FL: Florida International University. Retrieved from http://www.ops.org.bo/textocompleto/ide18245.pdf

Eriksen, S., & Kelly, P. (2007). Developing credible vulnerability indicators for climate adaptation policy assessment. *Mitigation and Adaptation Strategies for Global …*, *12*(4), 495–524. doi:10.1007/s11027-006-3460-6

Federal Emergency Management Agency. (2014). Disaster Declarations by Year. Retrieved February 26, 2014, from http://www.fema.gov/disasters/grid/year?field_disaster_type_term_tid_1=6840.

Federal Highway Administration. (2015). Journey to Work Profile: Houston--Galveston--Brazoria, TX CMSA. Retrieved February 2, 2015 from https://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/journey_to_work/jtw8.cfm.

Fekete, A. (2012). Spatial disaster vulnerability and risk assessments: challenges in their quality and acceptance. *Natural hazards*, 61(3), 1161-1178.

Finch, C., Emrich, C. T., & Cutter, S. L. (2010). Disaster disparities and differential recovery in New Orleans. *Population and Environment*, *31*(4), 179–202. doi:10.1007/s11111-009-0099-8

Flanagan, B. E., Gregory, E. W., Hallisey, E. J., Heitgerd, J. L., & Lewis, B. (2011). A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management*, *8*(1). doi: 10.2202/1547-7355.1792

Foody, G. M. (2002). Status of land cover classification accuracy assessment. *Remote sensing of environment*, *80*(1), 185-201.

Fothergill, A., Maestas, E. G., & Darlington, J. D. (1999). Race, ethnicity and disasters in the United States: a review of the literature. *Disasters*, *23*(2), 156–173. doi:10.1111/1467-7717.00111

Fothergill, A. (2004). *Heads Above Water: Gender, Class and Family in the Grand Forks Flood*. Albany, NY: State University of New York Press,.

Fothergill, A., & Peek, L. (2004). Poverty and disasters in the United States: A review of recent sociological findings. *Natural Hazards*, *32*(1), 89–110.

Godschalk, D. R., Beatly, T., Berke, P., Brower, D. J., & Kaiser, E. J. (1999). *Natural hazard mitigation: Recasting disaster policy and planning*. Washington, DC: Island Press.

Godschalk, D. R. (2003). Urban Hazard Mitigation: Creating Resilient Cities. *Natural Hazards Review, 4*(3), 136-143.

Handmer, J. W., Dovers, S., & Downing, T. E. (1999). Societal Vulnerability to Climate Change and Variability. *Mitigation and Adaptation Strategies for Global Change*, *4*, 267–281. doi:10.1023/a:1009611621048

Hapfelmeier, A. (2016). Incremental Linear Model Trees on Big Data (Doctoral dissertation, Johannes Gutenberg Universität Mainz).

Hazards and Vulnerabilities Research Institute, Department of Geology, University of South Carolina, Social Vulnerability Index: Frequently Asked Questions, http://webra.cas.sc.edu/ hvri/products/sovifaq.aspx#score, Accessed March 2, 2016 (2016).

Hazus®®-MH risk assessment and user group series. Using HAZUS-MH for risk assessment. How-To Guide. (2004). Retrieved March, 20, 2014, from http://www.fema.gov/pdf/plan/prevent/hazus/fema433.pdf.

Hewitt, K. (1997). *Regions of risk : a geographical introduction to disasters*. *Themes in resource management.* (p. xx, 389 p.).

Highfield, W. E., Peacock, W. G., & Van Zandt, S. (2014). Mitigation Planning: Why Hazard Exposure, Structural Vulnerability, and Social Vulnerability Matter. *Journal of Planning Education and Research*, *34*, 287–300. doi:10.1177/0739456X14531828

Intergovernmental Panel on Climate Change (IPCC) (2001) Climate change 2001: Impacts, Adaptation and Vulnerability, Summary for Policymakers, WMO.

Inostroza, L., Palme, M., & de la Barrera, F. (2016). A Heat Vulnerability Index: Spatial Patterns of Exposure, Sensitivity and Adaptive Capacity for Santiago de Chile. *PloS one*, *11*(9).

Ionescu, C., Klein, R. J. T., Hinkel, J., Kumar, K. S. k., & Klein, R. (2009). Towards a Formal Framework of Vulnerability to Climate Change. *Environmental Modeling & Assessment*, *14*(1), 1–16. doi:10.1007/s10666-008-9179-x

Jolliffe, I. (2014). Principal Component Analysis. Wiley StatsRef: Statistics Reference Online.

Juntunen, L. (2006). Addressing Social Vulnerability to Hazards. *Disaster Safety Review, 4*(2), 3-10.

Kasperson, R. E., Dow, K., Archer, E., Caceres, D., Downing, T., Elmqvist, T., … Ziervogel, G. (2005). Vulnerable Peoples and Places. In R. Hassan, R. Scholes, & N. Ash (Eds.), *Ecosystems and Human Wellbeing: Current State and Trends* (p. Vol. 1 143–164). Washington, DC: Island Press.

Kass, G. V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 20, 2, 119-127

Larose, D. T. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

Laska, S., & Morrow, B. H. (2006). Social Vulnerabilities and Hurricane Katrina: An Unnatural Disaster in New Orleans. *Marine Technology Society Journal*, *40*(4), 16–26. doi:10.4031/002533206787353123

Lewis, H. G., & Brown, M. (2001). A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing*, *22*(16), 3223-3235.

Logan, J.R., Xu, Z. and Stults, B. 2012. "Interpolating U.S. Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database" *Professional Geographer,* 66,3: 412-420.

Long, A. P. (2007). Poverty is the new prostitution: race, poverty, and public housing in post-Katrina New Orleans. *The Journal of American History*, *94*(3), 795–803. doi:doi: 10.2307/25095141

Lopez, R. (2002). Segregation and black/white differences in exposure to air toxics in 1990. *Environ Health Perspect, 110 Suppl 2*, 289-295.

Maantay, J. A., Maroko, A. R., & Herrmann, C. (2007). Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartography and Geographic Information Science*, *34*(2), 77–102. doi:10.1559/152304007781002190

Macintosh, A. (2013). Coastal climate hazards and urban planning: How planning responses can lead to maladaptation. *Mitigation and Adaptation Strategies for Global Change*, *18*, 1035–1055. doi:10.1007/s11027-012-9406-2

McCarthy, J. J. (2001). *Climate change 2001: impacts, adaptation, and vulnerability: contribution of Working Group II to the third assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.

McLaughlin, P., & Dietz, T. (2008). Structure, agency and environment: Toward an integrated perspective on vulnerability. *Global Environmental Change*, *18*, 99–111. doi:10.1016/j.gloenvcha.2007.05.003

Mileti, D. S., & Darlington, J. D. (1997). The Role of Searching in Shaping Reactions to Earthquake Risk Information. *Social Problems*. doi:10.1525/sp.1997.44.1.03x0214f

Mohai, P., & Saha, R. (2006). Reassessing racial and socioeconomic disparities in environmental justice research. *Demography*, *43*(2), 383–399. doi:10.1353/dem.2006.0017

Morrow, B. H. (1999). Identifying and mapping community vulnerability. *Disasters, 23*(1), 1-18.

Moser, C. O. N. (1998). The asset vulnerability framework: Reassessing urban poverty reduction strategies. *World Development*, *26*(1), 1–19. doi:10.1016/S0305-750X(97)10015-8

National Hurricane Center / National Ocenaic and Atmospheric Administration. Educational Resources. http://www.nhc.noaa.gov/outreach/. Accessed on May 20, 2015.

Norris, F., Sherrieb, K., & Galea, S. (2010). Prevalence and consequences of disaster-related illness and injury from Hurricane Ike. *Rehabil Psychol, 55*(3), 221-230. doi: 10.1037/a0020195

Orlove, B. (2009). The past, the present and some possible futures of adaptation. In W. N. Adger, I. Lorenzoni, & K. O'Brien (Eds.), *Adapting to Climate Change Thresholds Values Governance* (pp. 131–163). Cambridge, MA: Cambridge University Press.

Pandya, R., & Pandya, J. (2015). C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, *117*(16).

Pang, S. L., & Gong, J. Z. (2009). C5. 0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice*, *29*(12), 94-104.

Parry, M. L., Canziani, O. F., Palutikof, J. P., van der Linden, P. J., & Hanson, C. E. (2007). *Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate*

Change (Vol. 37, p. 2407). Cambridge, MA: Cambridge University Press. doi:10.2134/jeq2008.0015br

Peacock, W. G., Morrow, B., & Gladwin, H. (1997). *Hurricane Andrew: Ethnicity, Gender and the Sociology of Disasters*. New York, NY: Routledge.

Peacock, W. G. (2003). Hurricane Mitigation Status and Factors Influencing Mitigation Status among Florida's Single-Family Homeowners. *Natural Hazards Review*. doi:10.1061/(ASCE)1527-6988(2003)4:3(149)

Perlin, S. A., Wong, D., & Sexton, K. (2001). Residential proximity to industrial sources of air pollution: interrelationships among race, poverty, and age. *J Air Waste Manag Assoc, 51*(3), 406-421.

Perry, R. W., Greene, M., & Mushkatel, A. H. (1986). *American minority citizens in disaster*. Battelle Human Affairs Research Centers.

Perry, R. W., & Nelson, L. (1991). Ethnicity and hazard information dissemination. *Environmental Management*, *15*(4), 581–587. doi:10.1007/BF02394747

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Quinlan, J.R., Induction of decision trees, *Machine Learning* 1, 81-106, 1986.

Quinlan, J.R., Simplifying decision trees, *International Journal of Man-Machine Studies*, 27, 221-234, 1987.

Russell, L. A., Goltz, J. D., & Bourque, L. B. (1995). Preparedness and Hazard Mitigation Actions before and after Two Earthquakes. *Environment and Behavior*. doi:10.1177/0013916595276002

Rygel, L., O'sullivan, D., & Yarnal, B. (2006). A method for constructing a social vulnerability index: an application to hurricane storm surges in a developed country. *Mitigation and adaptation strategies for global change*,*11*(3), 741-764.

Safavian, S. R., & Landgrebe, D. (1990). A survey of decision tree classifier methodology.

Sanderson, D. (2000). Cities, Disasters and Livelihoods. *Environment and Urbanization*, *12*(2), 93–102. doi:10.1057/palgrave.rm.8240068

Sarewitz, D., Pielke, R., & Keykhah, M. (2003). Vulnerability and risk: some thoughts from a political and policy perspective. *Risk analysis*, 23(4), 805-810.

Sauve, M., & Tuleau-Malot, C. (2014). Variable selection through CART.*ESAIM: Probability and Statistics*, *18*, 770-798.

Schmidtlein, M. C., Deutsch, R. C., Piegorsch, W. W., & Cutter, S. L. (2008). A sensitivity analysis of the social vulnerability index. *Risk Analysis*, *28*(4), 1099-1114.

Sherrieb, K., Norris, F. H., & Galea, S. (2010). Measuring capacities for community resilience. *Social Indicators Research*, 99(2), 227-247.

Smit, B., & Wandel, J. (2006). Adaptation, adaptive capacity and vulnerability. *Global Environmental Change*, *16*, 282–292. doi:10.1016/j.gloenvcha.2006.03.008

Smith, G. P. (2016). Remembrances of the Past, Concerns for the Future, and the Potential Resilience of a Southern Coastal Town. *Southern Cultures*, *22*(2), 64-87.

Smith, A. B., & Matthews, J. L. (2015). Quantifying uncertainty and variable sensitivity within the U.S. billion-dollar weather and climate disaster cost estimates. *Natural Hazards*, *77*(3), 1829-1851.

Smith, S. M., Tremethick, M. J., Johnson, P., & Gorski, J. (2009). Disaster planning and response: considering the needs of the frail elderly. *International Journal of Emergency Management*, *6*(1), 1–13. doi:10.1504/IJEM.2009.025170

Tabachnick, B. G., & Fidell, L. S. (2013). Using Multivariate Statistics, 6th International edition (cover) edn.

Tate, E., Cutter, S. L., & Berry, M. (2010). Integrated multihazard mapping. *Environment and Planning B: Planning and Design*, *37*(4), 646–663. doi:10.1068/b35157

Thomas, D. S. K., Phillips, B. D., Lovekamp, W. E., & Fothergill, A. (2013). *Social Vulnerability to Disasters, Second Edition*: Taylor & Francis.

Turner, B. L., Kasperson, R. E., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., … Schiller, A. (2003). A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences of the United States of America*, *100*, 8074–8079. doi:10.1073/pnas.1231335100

United Nations International Strategy for Disaster Reduction (UNISDR): Global Assessment Report on Disaster Risk Reduction. United Nations, Geneva, Switzerland, 2009. ISBN/ISSN: 9789211320282, 207 pp http://www.preventionweb.net/english/ hyogo/gar/report/index.php?id=1130&pid:34&pih:2.

University California of Los Angeles (UCLA) Center for Public Health and Disasters. January 2006. Hazard Risk Assessment Instrument (HRAI).

U.S. Census. (2000). Summary File 2; generated by Apostolis Sambanis; using American FactFinder; Retrieved March,15, 2014, from http:// http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_00_SF2_DP1&prodType=table

U.S. Census. (2010). Summary File 1; generated by Apostolis Sambanis; using American FactFinder; Retrieved March,15, 2014, from http:// http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_SF1_SF1DP1&prodType=table

U.S. Census. (2012). Nearly 1 in 5 People Have a Disability in the U.S., Census Bureau Reports Retrieved February 21, 2015, from https://www.census.gov/newsroom/releases/archives/miscellaneous/cb12-134.html

U.S. Census. (2013). Ten Fastest Growing Metro Areas. Retrieved August, 29, 2014, from https://www.census.gov/newsroom/releases/pdf/CB14-51_countymetropopest2013tables.pdf

U.S. Department of Commerce. (2014). National Oceanic and Atmosphere Administration: Billion Dollar Weather Disasters 1980-2013. Retrieved March 28, 2014, from http://www.ncdc.noaa.gov/billions/

Van de Vliert, E. (2013). Climato-economic habitats support patterns of human needs, stresses, and freedoms. Behavioral and Brain Sciences, 36(05), 465-480.

Van Geffen, J. (2003). Metropolitan statistical area changes. *Texas Labor Market*.

Van Zandt, S., Peacock, W. G., Henry, D. W., Grover, H., Highfield, W. E., & Brody, S. D. (2012). Mapping social vulnerability to enhance housing and neighborhood resilience. *Housing Policy Debate*, *22*(1), 29–55. doi:10.1080/10511482.2011.624528

Wisner, B., Blaikie, P., Cannon, T., & Davis, I. (2004). *At risk: natural hazards, people's vulnerability, and disasters* (2nd ed.). London: Routledge.

Wolf, T., & McGregor, G. (2013). The development of a heat wave vulnerability index for London, United Kingdom. *Weather and Climate Extremes*, *1*, 59-68.

Wrenn, C. B. (2003). Naturalistic epistemology. The Internet Encyclopedia of Philosophy, http://www.iep.utm.edu/, Accessed: October 7, 2016.

Wu, S.-Y., Yarnal, B., & Fisher, A. (2002). Vulnerability of coastal communities to sea-level rise: a case study of Cape May County, New Jersey, USA. *Climate Research, 22*(3), 255-270. doi: 10.3354/cr022255

Yoon, DK. Assessment of social vulnerability to natural disasters: a comparative study. *Nat Hazards*, 2012; 63:823–843.

Zahran, S., Brody, S. D., Peacock, W. G., Vedlitz, A., & Grover, H. (2008). Social vulnerability and the natural and built environment: A model of flood casualties in Texas. *Disasters*, *32*(4), 537–560. doi:10.1111/j.1467-7717.2008.01054.x

Zhang, Y., & Peacock, W. G. (2009). Planning for Housing Recovery? Lessons Learned From Hurricane Andrew. *Journal of the American Planning Association*, *76*(1), 5–24. doi:10.1080/01944360903294556

Zhou, Y., Li, N., Wu, W., & Wu, J. (2014). Assessment of provincial social vulnerability to natural disasters in China. *Natural hazards*, *71*(3), 2165-2186.

VITA

NAME:	Apostolis Stefanos Sambanis
EDUCATION:	B.S., Earth and Environmental Science, University of Illinois at Chicago, Chicago, Illinois, 2007
	M.S., Public Health, University of Illinois at Chicago, Chicago, Illinois, 2012
TEACHING:	Environmental and Occupational Health Science Division, University of Illinois at Chicago, 2011-2016
EXPERIENCE:	HSE Generalist, Honeywell UOP, July 2013 to Present
	Regulatory Administrator, True Value Company, November 2011 to July 2013
	Project Geologist, Weaver Boos Consultants, July 2008 to November 2011
	Staff Geologist, Environmental Resources Management, December 2006 to July 2008
	Laboratory Technician, University of Illinois at Chicago Organic Geochemistry Laboratory, August 2004 to December 2006