Illinois Prevention Research Center

ILLINOIS PRC RESEARCH BRIEF
NO. 114
SEPTEMBER 2019

A New Approach to the Social Vulnerability Indices: Decision Tree-Based Vulnerability Classification Model

APOSTOLIS SAMBANIS,[1] SAGE KIM,[2] KRISTIN OSIECKI,[3] MICHAEL D. CAILAS[1]

Key Findings

- **The Social Vulnerability Index (SVI) has been a critical measure used to identify geographic areas with increased risk of natural and man-made disasters.**
- **SVIs are typically derived using either transformations or principal component analysis of the original variables, which often misclassify area-level vulnerability.**
- **The use of disaster loss classification (DLC) as a target variable in combination with a decision tree algorithm creates a more accurate SVI measure.**
- **Preliminary findings indicate that the DLC offers an innovative approach to predictive performance assessment of SVI.**

AUTHOR AFFILIATIONS

1. Environmental and Occupational Health Sciences, School of Public Health, University of Illinois at Chicago, Chicago, IL.
2. Health Policy and Administration, School of Public Health, University of Illinois at Chicago, Chicago, IL.
3. University of Minnesota, Rochester, Center for Learning Innovation, Rochester, MN.

Summary

The research of climate change examines social vulnerability by looking at hazard exposure, susceptibility to that hazard, and emergency response capacity. The Social Vulnerability Index (SVI), a composite score identifying populations at risk from disasters, is often used to predict vulnerability and plan for community-based disaster prevention and emergency response. However, current methods for deriving SVI may not adequately capture qualitatively different vulnerabilities in different communities. Our study introduces a decision tree-based approach to developing an SVI that captures the heterogeneity of both vulnerable populations and disasters. Furthermore, we demonstrate the importance of incorporating a disaster loss classification (DLC) into estimating social vulnerability to increase the predictive performance of the model.

We utilized decision tree algorithms to create an SVI. Sociodemographic data were retrieved from the U.S. Census for the Houston Metropolitan Statistical Area (MSA) and hurricane loss data for Hurricane Alicia in 1983 were obtained from the Federal Emergency Management Agency (FEMA) HAZUS program. [1]

Findings suggest that the SVI based on the decision tree approach dramatically increased the accuracy of predicting high vulnerability areas. The predictive performance rate was over 77% for the decision tree approach, compared to 35% for the principal component analysis (PCA) method. Our SVI based on decision tree methods can more accurately classify area-level vulnerability to disasters.

Background

Vulnerability is a critical concept in understanding the risk and outcomes of exposure to hazards. Risk as a technical/mathematical term is narrowly defined as value-neutral and free from social, political, cultural, and historical contexts. However, the vulnerability of an area is shaped by socioeconomic context and political practices. Since the early 1980s, scholars have considered vulnerability as a condition of a system that exists before it encounters a hazard, and individual social factors exacerbate the effects of disasters.[2-6] Within this conceptualization, even natural disasters are not quite "natural." [2-4] To understand more fundamental, upstream causes that determine risk and vulnerability to disasters, the concept of social vulnerability was introduced.[7,9] Social vulnerability can be defined as the social, economic, demographic, geographic, and political characteristics that shape the capacity of communities to deal with, respond to, and recover from environmental hazards.

To quantify overall place vulnerability, the social vulnerability index (SVI) has been used.[8,10,11] Two dominant approaches to building a composite score of SVI are: to add the percentile ranks of the original indicator variables, or to utilize in an additive model the principal component scores from

the application of the principle components analysis (PCA) technique to the indicator variables. Indicator variables include some variation of area-level estimates by percent: poverty; welfare recipients; unemployed; racial/ethnic minorities; dependent children; elderly; less than high school education; female headed-households; people without private modes of transport; and institutionalized individuals.

The primary limitation of current SVIs is that the derivation methods used to construct the indices cannot sufficiently account for the multiplicative, non-linear nature of vulnerability. Social science scholars have argued that multiple neighborhood factors of disadvantage are often spatially concentrated, and combined, risk factors produce neighborhood conditions that are beyond the additive effect of discrete risks.[12-14] For example, racial residential segregation intensifies economic difficulty, and this "double jeopardy" puts specific neighborhoods into greater cumulative vulnerability.[15] As one way to explore the concentration effects, scholars have grouped SVI scores into quartiles, and the highest vulnerability areas are compared with the rest of the areas having lower levels of vulnerability.

However, these index measures are not designed to distinguish qualitatively different groups of neighborhoods. This problem of mismatch between types of measurements and types of phenomena that one attempts to measure may stem from how we conceptualize neighborhood context and its effects.

To address the limitation of constructing SVIs, we introduce a new approach, which utilizes a decision tree model to classify area-level vulnerability. A decision tree (DT) model is a method for classification.[16-18] DT methods utilize machine learning algorithms that recursively partition the input data based on their attributes. The primary aim of the partition process is to reach the final partitions of homogenous classes.[19] In this study, we examined the accuracy of an index created using the DT-based derivation approach compared with an SVI-based on the PCA derivation approach. Disaster Loss classification (DCL) was based on the effects of Hurricane Alicia in 1983 on the Houston MSA, which is a known high-risk area for hurricanes, tropical storms, and flooding. In particular, Hurricane Alicia resulted in 21 deaths and $2 billion in damages.[20]

Data and Methods

We examined the Houston MSA including Brazoria, Chambers, Fort Bend, Galveston, Harris, Liberty, Montgomery, and Waller counties. The U.S. Census variables for 1980 were retrieved at the census tract level. To create the SVIs, we used, as a percent of the overall target area characteristic, the following 15 sociodemographic indicator variables: residents living below poverty; welfare recipients; residents with less than high school education; unemployed; median housing value; occupied housing units; renter-occupied homes; children 5 years and younger; elderly 65 years and over; population in group quarters; residents without vehicle available; African Americans; Hispanics; female households with children under 18 years; and median household income. A total of 1,062 census tracts were included in this analysis.

Hurricane related loss data were retrieved from the FEMA HAZUS platform, including the number of displaced households, the number of short-term shelters required, and total building loss in thousands of dollars. FEMA HAZUS is a disaster mitigation strategic tool that uses geographic information systems to evaluate potential physical, economic, and social impacts and losses caused by earthquakes, floods, and hurricanes.[21] High-risk locations are illustrated by geographic boundaries to show the areas of greatest threat visually. Using historical data for Hurricane Alicia retrieved from HAZUS, a disaster loss classification (DLC) score for each census tract area is generated to test the predictive performance of the PCA and DT social vulnerability models.

We then computed two types of SVIs, one derived from the DT model and the other using PCA. For the analyses in this project, the IBM SPSS Modeler 15.0 is applied, and the DT model is derived with the C5.0 algorithm.[22,23] The performance comparison between the DT and PCA approaches for identifying vulnerable areas is achieved

FIGURE 1: **Schematic of metrics derived from performance assessment matrix**

		Social Vulnerability Classification, j					∑
		1	2	..j..		m	
Disaster Loss Classification, i	1	OCP_{11}	c_{12}	…		FC_{1m}	c_{+1}
	2	c_{21}	OCP_{22}	Overestimation, OE			c_{2+}
				OCP_{33}			
	…	Underestimation, UE			…		
	m	CF_{m1}	…	…		OCP_{mm}	
	∑	c_{+1}	c_{2+}	…			N

by using a confusion matrix.[24] The confusion matrix identifies the number of (in)correctly classified areas from two classifiers. In this study, this matrix was used as the performance assessment (PA) matrix for the derived SVIs. The matching areas are contained in the diagonal elements of the PA matrix, c_{ii}, providing an overall classification performance measure (Figure 1). The sum of the matching areas divided by the total number of areas, N, yields an overall classification performance (OCP) rate. [25] Similarly, the off-diagonal elements, c_{ij}, of the PA matrix identify the misclassified areas in reference to the vulnerability based on the actual DLC. For example, in Figure 1, the bottom row of the 1st column indicates the areas with the highest actual disaster loss that were predicted to be the lowest vulnerable

areas by the SVI. This specific type of underestimation error is termed Classification Failure (CF) to emphasize the potential severe consequences. Overall, the cells below the diagonal (OCP) on the PA matrix indicate underestimation error (UE) of vulnerability.

On the other hand, the element on the top right-hand corner indicates the areas that were predicted to have the highest vulnerability but experienced the lowest level of disaster loss. Such extreme overestimation error is called False Classification (FC_{1m}). The elements above the diagonal indicate overestimation error (OE). We pay particular attention to CF, compared with FC, because of its potential significant public health consequences.

For this study, the sum of the areas in the elements above and below the diagonal line divided by N will define, respectively, the overall underestimation error (UE) rate, and the overestimation error (OE) rate. The following equations are used:

$$UE = \frac{1}{N} \times \left[\sum_{j=1}^{m-1} \sum_{i=2}^{m} c_{ij} \right]$$

$$OE = \frac{1}{N} \times \left[\sum_{j=2}^{m} \sum_{i=1}^{m-1} c_{ij} \right]$$

Where: m = the total number of selected categories, usually 4 (quartile).

i,j = the individual row and column, respectively, elements, c_{ij}, of the matrix.

Results and Discussion

The PA matrix shows the pattern of matches between the two classification methods. For this study, the DT model has a better classification performance. The DT model showed a CF of 0.7% and an OCP of 77.1%. On the other hand, the SVI derived by PCA yielded a CF rate of 3.3% and an OCP of 35.0% (Figure 2). To give a sense of the severity of misclassification, for example, a 3.3% CF results in 35 census tracts experiencing disaster losses at the highest severity level while being classified to belong to the lowest level of vulnerability.

While the performance of the DT model is superior to the PCA model, this performance comes at a cost of complexity of the model, since this level of accuracy requires multiple partitions. Overfitting can be an issue with some DT algorithms; however, several modifications are feasible to improve this deficiency.[26] Overall, this study raised a vital measurement issue concerning social vulnerability. We proposed an innovative approach to conduct social vulnerability research, which takes advantage of the fast-developing field of predictive analytics. Our findings show that the use of classification modeling offers a more accurate way to predict vulnerability to potential disasters and hazards. Our approach may contribute to reducing disaster losses by improving the ability to predict and prepare for potential harms.

FIGURE 2: **Performance comparison between decision tree and principal component analysis-based Social Vulnerability Index models for Hurricane Alicia (1983)-related losses in the Houston Metropolitan Statistical Area**

		DT.C5.0 SVI Quartile			
		1	2	3	4
Disaster Loss Quartile	1	225	16	12	13
	2	51	176	17	22
	3	12	17	185	52
	4	7	14	10	233
		CF = 0.7%		UE = 10.5%	
		OCP = 77.1%		OE = 12.4%	

		PCA SVI Quartile			
		1	2	3	4
Disaster Loss Quartile	1	110	65	52	39
	2	69	83	60	54
	3	52	68	77	69
	4	35	50	77	102
		CF = 3.3%		UE = 33.1%	
		OCP = 35.0%		OE = 31.9%	

DT.C5.0 = Decision tree; C5.0 algorithm
PCA = Principal Component Analysis
OCP = Overall Classification Performance rate
SVI = Social Vulnerability Index
CF = Classification Failure rate
UE, OE = Under/Over Estimation rate

References

1. Federal Emergency Management Agency (FEMA). HAZUS. 2019.
2. Dyson ME. Ch 1. Unnatural disasters: Race and poverty. *Come hell or high water: Hurricane Katrina and the color of disaster.* New York, NY: Basic Books; 2006:1-14.
3. Logan J. Ch 12. Unnatural disaster: Social impacts and policy choices after Katrina. In: Bullard R, Wright B, eds. *Race, place, and environmental justice after Hurricane Katrina: Struggles to reclaim, rebuild, and revitalize New Orleans and the Gulf Coast.* Boulder, CO: Westview Press; 2009:249-264.
4. O'Keefe P, Westgate K, Wisner B. Taking the naturalness out of natural disasters. *Nature.* 1976;260:566-567.
5. Ge Y, Dou W, Zhang H. A New Framework for Understanding Urban Social Vulnerability from a Network Perspective. *Sustainability.* 2017;9(1723):1-16.
6. Bergstrand K, Mayer B, Brumback B, Zhang Y. Assessing the Relationship Between Social Vulnerability and Community Resilience to Hazards. *Social Indicators Research.* 2015;122(2):391-409.
7. Burton C, Rufat S, Tate E. Social Vulnerability: Conceptual Foundations and Geospatial Modeling. In: Fuchs S, Thaler T, eds. *Vulnerability and resilience to natural hazards.* Cambridge, UK: Cambridge University Press; 2018.
8. Cutter SL. Vulnerability to environmental hazards. *Progress in Human Geography.* 1996;20(4):529-539.
9. Pelling M. Ch 1. Tracing the roots of urban risk and vulnerability. In: Pelling M, ed. *The vulnerability of cities: Natural disasters and social resilience*. New York, NY: Taylor & Francis; 2003:3-10.
10. Singh SR, Eghdami MR, Singh S. The Concept of Social Vulnerability: A Review from Disasters Perspectives. *International Journal of Interdisciplinary and Multidisciplinary Studies (IJIMS).* 2014;1(6):71-82.
11. Weichselgartner J. Disaster mitigation: the concept of vulnerability revisited. *Disaster Prevention and Management.* 2001;10(2):85-94.
12. Massey DS, Denton NA. *American Apartheid: Segregation and the Making of the Underclass.* Cambridge, MA: Harvard University Press; 1993.
13. Wilson WJ. *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago, IL: University Of Chicago Press; 1987.
14. Galster G, Santiago AM. Do Neighborhood Effects on Low-Income Minority Children Depend on Their Age? Evidence from a Public Housing Natural Experiment. *Housing Policy Debate.* 2017;27(4):584-610.
15. Barber S, Hickson D, Kawachi I, Subramanian S, Earls F. Double-jeopardy: The joint impact of neighborhood disadvantage and low social cohesion on cumulative risk of disease among African American men and women in the Jackson Heart Study. *Social Science & Medicine.* 2016;153:107-115.
16. Song Y-Y. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry.* 2015;27(2):130-135.
17. Lee S, Park I. Application of decision tree model for the ground subsidence hazard mapping near abandoned underground coal mines. *J Environ Manage.* 2013;127:127-166.
18. Linder S, Marko D, Tian Y, Wisniewski T. A Population-Based Approach to Mapping Vulnerability to Diabetes *Int J Environ Res Public Health.* 2018;15(10):1-13.
19. Larose D, Larose C. *Discovering Knowledge in Data.* Hoboken, NJ: Wiley; 2014.
20. National Hurricane Center. National Hurricane Center and Central Pacific Hurricane Center – Hurricanes in History. 2015; https://www.nhc.noaa.gov/outreach/history/.
21. Federal Emergency Management Agency (FEMA). Disasters. 2014; http://www.fema.gov/disasters/grid/year?field_disaster_type_term_tid_1=6840). Accessed 26 February 2018
22. Quinlan J. Induction of decision trees. *Machine Learning.* 1986;1(1):81-106.
23. Devi BR, Rao KN, Setty SP, Rao MN. Disaster prediction system using IBM SPSS data mining tool. *International Journal of Engineering Trends and Technology.* 2013;4:3352-3357.
24. Lewis H, Brown M. A generalized confusion matrix for assessing area estimates from remotely sensed data. *International Journal of Remote Sensing.* 2001;22(16):3223-3235.
25. Chen M-S, Han J, Yu P. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering.* 1996;8(6):866-883.
26. Pandya R. C5. 0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications.* 2015;117(16):18-21.

ACKNOWLEDGMENTS

The authors would like to thank Mr. Eric Berman, HAZUS Program Manager at the Federal Emergency Management Agency (FEMA) for providing data and technical support for this study.

SUGGESTED CITATION

Sambanis A., Kim S., Osiecki K., Cailas M.D. A new approach to the social vulnerability indices: Decision tree-based vulnerability classification model. Research Brief No. 114. Illinois Prevention Research Center, University of Illinois at Chicago. Chicago, IL. September 2019. https://go.uic.edu/SVI-Decision-Tree-Classification