Social vulnerability in the path of hurricanes in the Southeast US

Polina Berezina, Shengxi Gui, Yang Li and Zhihao Wang

The Ohio State University

STAT 6560 Applied Multivariate Analysis

Autumn 2019

1 INTRODUCTION

Natural disasters impact human lives and change functioning of local, and international civic order. The concept of a place becomes central to disaster studies: what types of places and to what degree are impacted by natural disasters? Scholars have traditionally focused on physical places, but since the last decade there has been a growing interest and need in studying society as a place and most important stakeholder in disaster vulnerability. Researchers like S. Cutter (Cutter, Boruff, & Shirley, 2003) pioneered studies of social vulnerability.

This paper focuses on a specific type of a natural disaster – hurricanes. The Southeast region of the US has been most historically devastated with hurricane disasters. Studying social vulnerability of this region in the context of highly-destructive hurricane within the last decade is imperative.

Nonprofit organizations, including Oxfam and Direct Relief, have been releasing analysis and mappings of social vulnerability in the US to different natural disasters (Direct Relief, n.d.; Direct Relief, n.d.; Oxfam, n.d.). Considering the importance of this topic for decision-making and funding distribution, as well as academic and institutional efforts to improve methodology, this research aims to enhance assessment of communities' social vulnerability through a principal component analysis of US Census data and creation of a holistic index of social vulnerability. We also explore NOAA Historical Hurricane Tracks dataset as a proxy of biophysical vulnerability to later compare coincidence of highly impacted areas with vulnerable communities. To the authors' best knowledge, these two datasets have not yet been integrated in the disaster studies and vulnerability literature.

This paper is structured as follows. Section 1 is an introduction to the general problem investigated and the proposed approach. Section 2 provides the data collection methods and pre-processing, as well as introduces research methods. Section 3 outlines methodology, in particular, principal component analysis and creation of a social vulnerability index. Section 4 presents analysis workflow. Section 5 summarizes the results, and section 6 concludes with general remarks and future research directions.

2 DATA

2.1 Study area

We implement our study on the Southeastern United States, including Mississippi, Alabama, Florida, Georgia and South Carolina (Figure 1), which are dominated by the humid subtropical climate and suffer high risk to hurricane attacks because of their adjacency to the North Atlantic Ocean as well as the Gulf of Mexico. The states we focused on can be regarded as a sample of the social vulnerability of coastal region due to hurricanes, but are not able to represent the inland region of the continental U.S. because the impact of hurricanes will decrease with time and distance after landfall.

It has also been shown that communities of the Southeast of United States suffer particularly because of unequal disaster aid distribution and marginalization (Simpson, 2003). As a unit of study, we chose U.S. census counties. There were 486 counties in total in the chosen study area. This might raise a concern of the modifiable areal unit problem (MAUP) (Cressie, 1996). However in the context of this study, a larger unit, like a county, in comparison to a U.S. Census block or tract would allow for faster processing and more reliable estimation of biophysical vulnerability index.

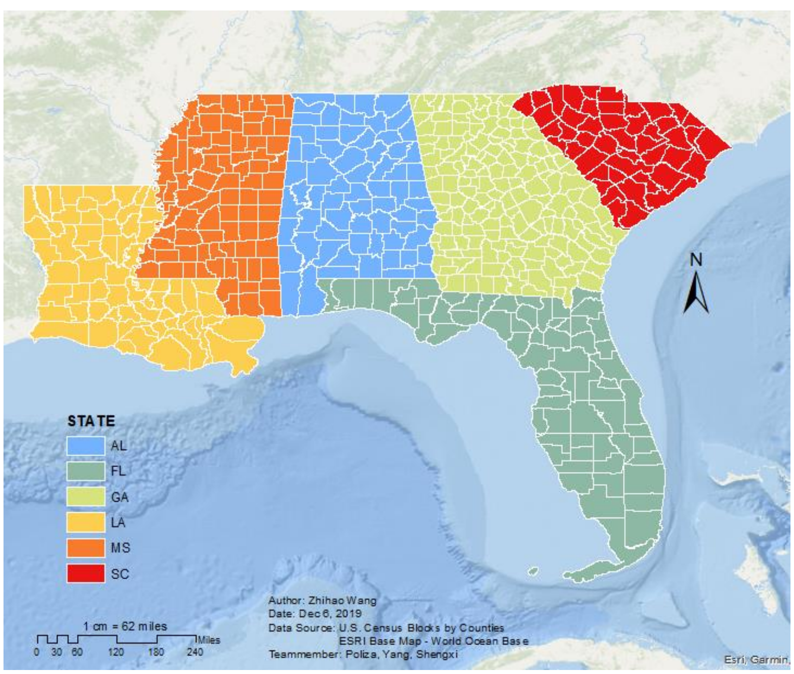

Figure 1 Study Area

2.2 Social vulnerability

A general consensus exists within the social science community about some of the major factors that influence social vulnerability. Cutter et al. (2003) gives a detailed review of the prevailing characteristics which describe the social condition. Social vulnerability data on Southern United States were collected on county level from the United States Census Bureau data portal (American FactFinder, n.d.).

Table 1 gives a list of social vulnerability concepts as well as the related variables which indicates the social condition to some extent. There are no outliers in our raw data, but data gap does exist due to data missing. Therefore, we preprocessed the data by merge variables according to their GeoID field, which generated a table with counties having full dataset for all interested variables. The variables are listed in Table 2.

Table 1 Social Vulnerability Concepts and Metrics (2015)

Concept	Description	Variables
Gender	Women might have to go through a more difficult recovery because of sector-specific employment, lower wages, and family care responsibilities.	1. Sex_M 2. Sex_F
Age	Extremes of the age spectrum affect the movement out of harm's way. Parents lose time and money caring for children when daycare facilities are affected; elderly may have mobility constraints or mobility concerns increasing the burden of care and lack of resilience	1. Age_1 2. Age_2 3. Age_3 4. Age_4 5. Age_5 6. Age_6 7. Age_7 8. Age_8 9. Age_9
The Disability	The disabilities are more vulnerable to natural disturbances because of their physical inconvenience or mental disorders.	1. Dis_T_M 2. Dis_P_M 3. Dis_T_F 4. Dis_P_F

Education Level	Education is closely related to socioeconomic status. Higher educational attainment referring to a greater lifetime earnings. Lower education constrains the ability to understand warning information and access to recovery information.	1. Edu 2. Edu_p
Socioeconomic status (income, health insurance)	The ability to absorb losses and enhance resilience to hazard impacts.	1. Hhold_inc 2. Fmly_inc 3. Insur_M 4. Insur_P_M 5. Insur_F 6. Insur_P_F
Residential property	The value, quality, and density of residential construction affects potential losses and recovery. Expensive homes on the coast are costly to replace; mobile homes are easily destroyed and less resilient to hazards.	1. Value 2. Units 3. Rent
Occupation	Some occupations, especially those involving resource extraction, may be severely impacted by a hazard event.	1. MBF 2. CES 3. ELCAM 4. HT
Infrastructure and lifelines	Loss of sewers, bridges, water, communications, and transportation infrastructure compounds potential disaster losses.	1. Plumbing 2. Kitchen 3. Tele 4. Utility 5. Gas 6. Elec
Population growth	Counties experiencing rapid growth lack available quality housing, and the social services network may not have had time to adjust to increased populations. New migrants may not speak the language and not be familiar with bureaucracies for obtaining relief or recovery information, all of which increase vulnerability growth	1. Pop_growth
Poverty	Poverty is an important indicator of the resilience to natural disasters.	1. Uneply_rate 2. Hhold_lt10k 3. Hhold_lt10k_P 4. Fmly_lt10k 5. Fmly_lt10k_P 6. Value_lt50k 7. Cost_M_lt200

Table 2 Variable Names and Descriptions

Name	Description
Sex_M	Male
Sex_F	Female
Age_1	Population Estimates - 2015 - Both Sexes; Under 18 years
Age_2	Population Estimates - 2015 - Male; Under 18 years
Age_3	Population Estimates - 2015 - Female; Under 18 years
Age_4	Population Estimates - 2015 - Both Sexes; 18 to 64 years
Age_5	Population Estimates - 2015 - Male; 18 to 64 years
Age_6	Population Estimates - 2015 - Female; 18 to 64 years
Age_7	Population Estimates - 2015 - Both Sexes; 65 years and over
Age_8	Population Estimates - 2015 - Male; 65 years and over
Age_9	Population Estimates - 2015 - Female; 65 years and over
Dis_T_M	With a disability; Estimate; Sex - Male
Dis_P_M	Percent with a disability; Estimate; Sex - Male
Dis_T_F	With a disability; Estimate; Sex - Female
Dis_P_F	Percent with a disability; Estimate; Sex - Female
Edu	Total; Estimate; Population enrolled in college or graduate school
Edu_p	Percent; Estimate; Population enrolled in college or graduate school
Insur_M	Insured; Estimate; Sex - Male
Insur_P_M	Percent Insured; Estimate; Sex - Male
Insur_F	Insured; Estimate; Sex - Female
Insur_P_F	Percent Insured; Estimate; Sex - Female
Units	Housing Unit Estimate - 2015
Uneply_rate	Percent; Employment Status - Civilian labor force - Unemployment Rate
Hhold_inc	Estimate; Income and Benefits (in 2015 inflation-adjusted dollars) - Total households - Mean household income (dollars)
Fmly_inc	Estimate; Income and Benefits (in 2015 inflation-adjusted dollars) - Families - Mean family income (dollars)
MBF	Total; Estimate; Management, business, science, and arts occupations: - Management, business, and financial occupations:

CES	Total; Estimate; Management, business, science, and arts occupations: - Computer, engineering, and science occupations:
ELCAM	Total; Estimate; Management, business, science, and arts occupations: - Education, legal, community service, arts, and media occupations:
HT	Total; Estimate; Management, business, science, and arts occupations: - Healthcare practitioner and technical occupations:
Plumbing	Occupied housing units; Estimate; Complete Facilities - With complete plumbing facilities
Kitchen	Occupied housing units; Estimate; Complete Facilities - With complete kitchen facilities
Tele	Occupied housing units; Estimate; Telephone Service Available - With telephone service
Utility	Occupied housing units; Estimate; House Heating Fuel - Utility gas
Gas	Occupied housing units; Estimate; House Heating Fuel - Bottled, tank, or LP gas
Elec	Occupied housing units; Estimate; House Heating Fuel - Electricity
Pop_growth	Annual Estimates of the Components of Population Change - Total Population Change
Hhold_lt10k	Estimate; Income and Benefits (in 2015 inflation-adjusted dollars) - Total households - Less than $10,000
Hhold_lt10k_P	Percent; Income and Benefits (in 2015 inflation-adjusted dollars) - Total households - Less than $10,000
Fmly_lt10k	Estimate; Income and Benefits (in 2015 inflation-adjusted dollars) - Families - Less than $10,000
Fmly_lt10k_P	Percent; Income and Benefits (in 2015 inflation-adjusted dollars) - Families - Less than $10,000
White	Estimate; Race - One race - White
Blk	Estimate; Race - One race - Black or African American
Indian	Estimate; Race - One race - American Indian and Alaska Native
Asian	Estimate; Race - One race - Asian
Rent	Estimate; Median gross rent -- - Total
Value_lt50k	Estimate; Value - Less than $50,000
Value	Owner-occupied housing units with a mortgage; Estimate; Value - Median (dollars)
Cost_M_lt200	Owner-occupied housing units with a mortgage; Estimate; Monthly Housing Costs - Less than $200

2.3 Biophysical vulnerability

The Southeast United States is particularly at risk of extreme hurricanes and tropical storms, which need to be considered to evaluate biophysical vulnerability. The National Oceanic and

Atmospheric Administration (NOAA) released a dataset of historical hurricane track called International Best Track Archive for Climate Stewardship (IBTrACS), which records observations from all the Regional Specialized Meteorological Centers and other international centers and individuals.

This is spatial data, which includes lines and points of historical hurricane tracks from 1980 to 2019. We focus on the spatial analysis based on lines since 1980. Because census data are up to 2015, hurricane historical tracks after 2016 are discarded.. By selecting tracks around the study area (LA, MS, AL, GA, SC, FL, for hurricane with landfall < 120 km), 2065 tracks have been utilized to analysis (each track represent a 3 hours observation). We selected hurricane category as a key attribute of to represent the impact of hurricanes, or in other words, biophysical vulnerability of each county over the 1980-2015 period.

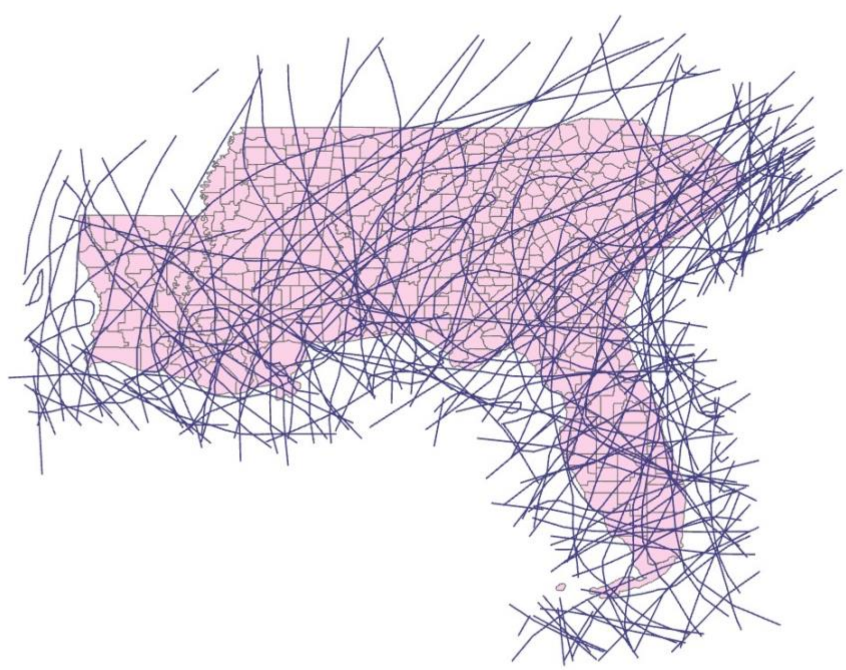

Figure 2 Hurricane track lines since 1980 and counties of the Southeastern U.S.A.

Table 3 Hurricane category description

Variable name in file	Value	Description
USA_SSHS (W: wind speed, unit: knot)	-4	Post-tropical
	-3	Miscellaneous disturbances
	-2	Subtropical
	-1	Tropical depression [W <34 kts]
	0	Tropical storm [34<W<64]
	1	Category 1 [64<=W<83]
	2	Category 2 [83<=W<96]
	3	Category 3 [96<=W<113]
	4	Category 4 [113<=W<137]
	5	Category 5 [W >= 137]

3 METHODS

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is used to examine the social vulnerability. Since a large number of social variables are collected in our analysis, PCA, as a primary statistical procedure for reducing data dimension, allows the use of a small set of variables to explain most information. Specifically, much of the variability in the original variables can be accounted for by a small set of principal components, which are the linear combination of the original variables. Relying on the eigenvectors, PCA geometrically represent the selection of a new coordinate system obtained by rotating the original system. Thus, the PCA outputs are more likely to reveal relationships that were not previously suspected.

Due to the complexity of the social variables, varimax rotation can also be used to clarify relationships among principal components. Specifically, varimax rotation is intended to best represent the shared variance by adjusting the baseline or orthogonal axis relative to the coordinance of data points. Thus, as a complementary technique, varimax rotation does not involve a change in the relative locations of data points.

3.2 Index of social vulnerability

There are multiple ways to generate a composite index of social vulnerability for each county using principal components. In this study, due to a lack of validation data, we decided to proceed with creating a score by multiplying each variable value with its PC loading, weighting each component scores by its proportion of variance explained and then summing the result. An alternative approach could be weighting each component equally, which in our case, would mean assigning a weight of 0.1 to each component, instead of using their proportion of explained variance.

The resulting scores need to be further standardized by creating z-scores and assigning specific z-score thresholds to vulnerability classes based on their distribution and visual analysis. Section 4 will cover in details this classification scheme.

3.3 Index of biophysical vulnerability

Research indicated that the average diameter of hurricane force winds is 100 miles (Willoughby, 2007). Former work (Emrich & Cutter, 2011) defined the hurricane wind impact areas as 50 miles on either side of the linear historic hurricane track. After preprocessing the tracks, we created a spatial buffer and computed the amount of land area within the hurricane wind impact zone for each county. The amount of land in the hurricane wind zone divided by the total land area in the county produced the percentage of land area affected. We used ArcGIS to perform buffer and spatial intersection, and then calculate the level of hurricane impact for counties in southeastern U.S. For each county:

$$R = \frac{\Sigma Area_{hurricane}}{Area_{county}}$$

where $Area_{hurricane}$ means the area historically impacted by hurricanes, $Area_{county}$ means the area of that county, and R is the ratio of these two areas, representing a hurricane influence, or impact score, for that county. And the index of biophysical vulnerability is calculated by using normalization of the ratio *R*. For each *R*:

$$BI = \frac{R_i - R_{min}}{R_{max} - R_{min}}$$

Where *BI* is the index of biophysical vulnerability with the range of [0, 1]. Hurricane wind impact areas are defined by hurricane category (from -4 to 5). Hurricane with a high category number has great impact area. The amount of land in the hurricane wind zone divided by the total land area in the county produced the percentage of land area affected. The relationship between hurricane category and buffer radius is shown in table 4.

Table 4 Hurricane category and a corresponding impact radius used for buffer

Category	**Buffer radius (km)**
-4~0	40
1	50
2	60
3	80
4	100
5	120

4 ANALYSIS

4.1 Principal Component Analysis on Social Vulnerability

To analyze social vulnerability, the first step is to perform PCA on the pre-processed census variables, total 47. The covariance matrix is computed and serves as the input in PCA function. Varimax rotation is then added to the model for further comparison. The R package psych::principal() is used to run all the experiments in this section.

4.1.1 PCA using the covariance matrix without rotation

Table 5 Loadings for Principal Components Based on Covariance Matrix without Rotation

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Sex_M	0.987	0.148								
Sex_F	0.987	0.150								
Age_1	0.982	0.151								
Age_2	0.982	0.151								
Age_3	0.983	0.151								
Age_4	0.983	0.160								
Age_5	0.983	0.160								
Age_6	0.984	0.159								
Age_7	0.953	0.102								0.126
Age_8	0.947									0.145
Age_9	0.957	0.115								0.112
Dis_T_M	0.979									
Dis_P_M	-0.491	0.394	-0.363	0.472	0.134			-0.104		0.102
Dis_T_F	0.981	0.130								
Dis_P_F	-0.513	0.383	-0.275	0.536	0.192			-0.111		
Edu	0.951	0.159								-0.106
Edu_p	0.317	-0.197	0.409	-0.262	-0.204	0.429	0.398	-0.229	0.210	-0.182
Insur_M	0.991	0.127								
Insur_P_M		-0.426	0.481	0.583			0.248		-0.175	
Insur_F	0.990	0.130								
Insur_P_F		-0.414	0.553	0.516		0.117	0.257		-0.131	0.148
Units	0.988	0.113								
Uneply_rate	-0.292	0.633		-0.280		0.224		-0.169	-0.153	0.442
Hhold_inc	0.601	-0.720			-0.114			0.117		0.156
Fmly_inc	0.589	-0.708			-0.153			0.108	0.127	0.169
MBF	0.978	0.150								
CES	0.944									
ELCAM	0.981	0.145								
HT	0.984	0.120								
Plumbing	0.235	-0.348	0.205	-0.203	0.715			0.113		
Kitchen		-0.196	0.255	-0.260	0.777	-0.122	0.137	0.159		
Tele	0.117	-0.386	0.203	0.162	0.293	0.130	-0.302	-0.659	0.289	0.106
Utility	-0.187	0.130	0.797	-0.129	-0.128	-0.169	-0.363		-0.240	
Gas	-0.489	0.475	0.175	0.311		-0.158		0.258	0.503	
Elec	0.424	-0.367	-0.700		0.133	0.214	0.267			
Pop_growth	0.897		-0.110							0.127
Hhold_lt10k	0.959	0.201								
Hhold_lt10k_P	-0.409	0.712	0.223	-0.236		0.191	0.275			0.101
Fmly_lt10k	0.960	0.206								
Fmly_lt10k_P	-0.369	0.702	0.179	-0.325			0.122			0.131
White	0.979	0.125								
Blk	0.892	0.213	0.161							-0.111
Indian	0.826						-0.116			
Asian	0.906	0.155							0.107	
Rent	0.710	-0.502	-0.135	-0.248	-0.130					0.144
Value_lt50k	-0.519	0.690			0.109				0.208	
Value	0.561	-0.654		-0.143	-0.202					0.220
Cost_M_lt200				0.108		0.795	-0.432	0.372		

From the above loadings (Table 5) what we can understand is, the first loading vector places approximately equal weight on gender, age, education, total disabilities for both gender, total health insurance for both gender, housing units, four types occupations, households or family with income less than 10,000 dollars per year, and white race; with much less weight on infrastructure except for kitchen facilities, the disability in percent, households or family with income less than 10,000 dollars in percent, etc; with no contribution from health insurance in percent, with or without complete kitchen facilities and monthly housing cost. The second loading vector places most of it weight on unemployment rate, households or family income, as well as the housing value, and much less but equal weight on the rest of the variables. Hence, this component roughly corresponds to the level of economic status. Overall, we see the rest PC emphasizes the contribution from the disability, infrastructure rather than gender, age, occupation, etc.

Table 6 Principal Components of Covariance Matrix without Rotation

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS loadings	29.236	5.512	2.424	1.817	1.477	1.089	0.932	0.864	0.647	0.572
Proportion Var	0.609	0.115	0.050	0.038	0.031	0.023	0.019	0.018	0.013	0.012
Cumulative Var	0.609	0.724	0.774	0.812	0.843	0.866	0.885	0.903	0.917	0.929

For the PCA analysis based on covariance matrix, it can be concluded that PC1 explains 60.9% of the total variance, which means that nearly two-thirds of the information in the dataset (50 variables) can be encapsulated by just that one Principal Component. PC2 explains 11.5% of the variance, following with PC3 to PC 10 with no more than 5% percent contribution each.

Summary also yields cumulative proportion of the principal components. For the cumulative proportion, PC1 to PC 6 explains more than 85% variance, adding with PC 7 and PC 8, more than 90% of the variety of the social attributes could be included.

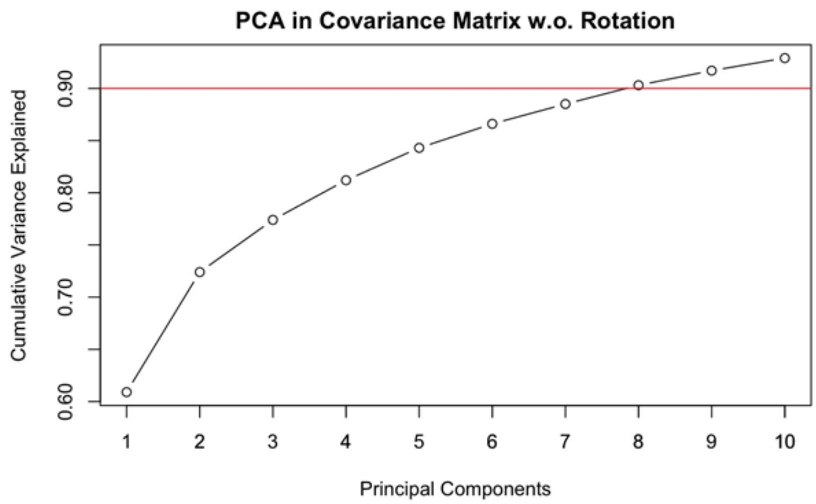

Figure 3 Cumulative Variance in Covariance Matrix without Rotation

4.1.2 PCA using covariance matrix with rotation

Similar to the previous loading results, which are computed without matrix rotation, the first principal component from correlation matrix has large positive associations with gender, age, education, total disabilities for both gender, total health insurance for both gender, housing units, four types of occupations, households or family with income less than 10,000 dollars per year, white race; with much less positive weight on education level in percent, households or family income, infrastructure except for utility service and gas, as well as rent and housing values; with negative weight on the disability in percent, health insurance in percent, households or family with income less than 10,000 dollars in percent, infrastructure, the disability in percent, health insurance in percent, households or family with income less than 10,000 dollars in percent and housing value less than 50,000 US dollars.

Table 7 Loadings for Principal Components Based on Covariance Matrix with Rotation

	RC1	RC2	RC3	RC4	RC5	RC6	RC7	RC8	RC9	RC10
Sex_M	0.982	0.162								
Sex_F	0.983	0.159								
Age_1	0.980	0.153								
Age_2	0.980	0.153								
Age_3	0.980	0.153								
Age_4	0.983	0.145								
Age_5	0.982	0.145								
Age_6	0.983	0.146								
Age_7	0.934	0.215	0.110							
Age_8	0.922	0.232	0.124							
Age_9	0.943	0.201								
Dis_T_M	0.961	0.188					0.107			
Dis_P_M	-0.313	-0.609	0.350	0.127	-0.214	-0.364				
Dis_T_F	0.975	0.158								
Dis_P_F	-0.324	-0.659	0.291	0.169	-0.169	-0.330	-0.113	0.133		
Edu	0.949	0.118				0.205				
Edu_p	0.221	0.189		0.157		0.872		0.101		
Insur_M	0.980	0.179								
Insur_P_M	-0.106	0.165	-0.105	0.892					-0.121	
Insur_F	0.981	0.175								
Insur_P_F	-0.116	0.161	-0.133	0.886						
Units	0.971	0.201								
Uneply_rate	-0.106	-0.530	-0.139	-0.208					0.714	
Hhold_inc	0.346	0.884		0.146						
Fmly_inc	0.339	0.863		0.201						
MBF	0.976	0.168								
CES	0.920	0.216				0.109				
ELCAM	0.976	0.158								
HT	0.974	0.179								
Plumbing	0.124	0.190			0.832		0.107	0.119	-0.127	
Kitchen					0.913					
Tele		0.199		0.111	0.123			0.935		
Utility	-0.108	-0.101	-0.944	0.127						
Gas	-0.287	-0.461	-0.128		-0.115		-0.778			
Elec	0.248	0.343	0.785	-0.140			0.383			
Pop_growth	0.851	0.281	0.167							
Hhold_lt10k	0.973									
Hhold_lt10k_P	-0.189	-0.668	-0.161	-0.105		0.329	-0.180	-0.257	0.425	
Fmly_lt10k	0.975									
Fmly_lt10k_P	-0.161	-0.604	-0.231	-0.233		0.179	-0.145	-0.269	0.433	
White	0.968	0.185								
Blk	0.916		-0.107			0.136				
Indian	0.803	0.191								0.150
Asian	0.906	0.120				0.118				
Rent	0.483	0.754	0.156	-0.111		0.110	0.142			
Value_lt50k	-0.273	-0.704	-0.117	-0.152			-0.411		0.177	
Value	0.308	0.853	0.154			0.102				
Cost_M_lt200										0.991

Table 8 Principal Components of Covariance Matrix with Rotation

	RC1	RC2	RC3	RC4	RC5	RC6	RC7	RC8	RC9	RC10
SS loadings	26.682	6.541	2.055	1.941	1.684	1.348	1.127	1.119	1.042	1.032
Proportion Var	0.556	0.136	0.043	0.040	0.035	0.028	0.023	0.023	0.022	0.021
Cumulative Var	0.556	0.692	0.735	0.775	0.810	0.839	0.862	0.885	0.907	0.929

The first two principal components account for nearly 70% of the total variance; With the contribution from PC 3 to PC 9, more than 90% of the variety of the social attributes could be included. A cumulative graph is plotted to visualize the cumulative Variance explained by each subsequential principal component.

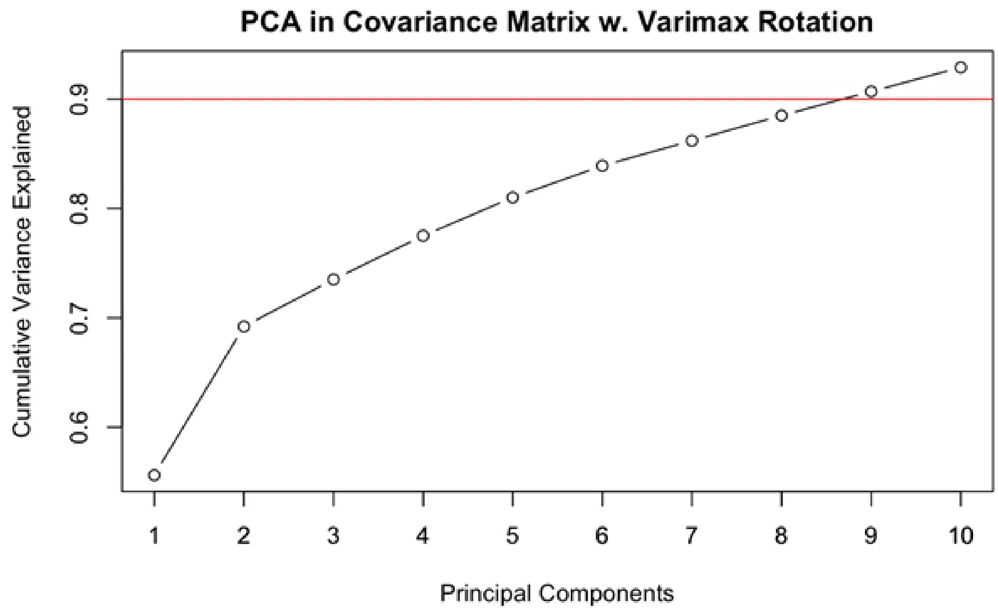

Figure 4 Cumulative Variance in Covariance Matrix with Rotation

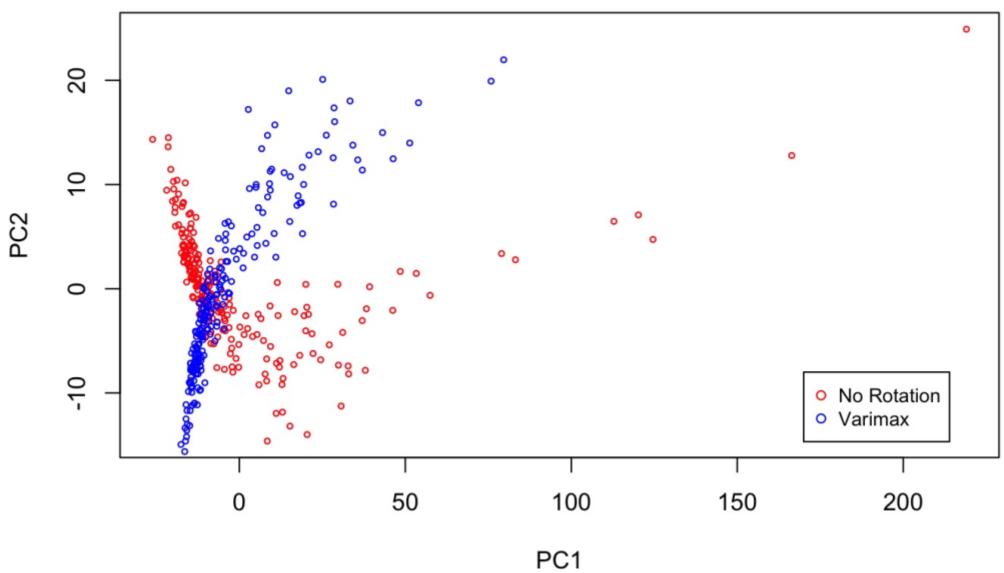

Figure 5 The comparison of the first and second PC before and after matrix rotation

A traditional way to simplify loadings is by rotation. The varimax in R does an orthogonal rotation to the original covariance matrix. It is not obvious from our results (Table 5 and Table 7) that the components are enhanced by rotation. However, from Fig 5, it can be observed that the relationship between the first and second PC after covariance matrix rotation is different from the one without rotation.

4.1.3 PCA using correlation matrix without rotation

As we've normalized the covariance matrix before computing the principal components, leading to the same result as from the correlation matrix, which could be observed from comparing Table 9 with Table 5, and Table 11 with Table 7.

Table 9 Loadings for Principal Components Based on Correlation Matrix without Rotation

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Sex_M	0.987	0.148								
Sex_F	0.987	0.150								
Age_1	0.982	0.151								
Age_2	0.982	0.151								
Age_3	0.983	0.151								
Age_4	0.983	0.160								
Age_5	0.983	0.160								
Age_6	0.984	0.159								
Age_7	0.953	0.102								0.126
Age_8	0.947									0.145
Age_9	0.957	0.115								0.112
Dis_T_M	0.979									
Dis_P_M	-0.491	0.394	-0.363	0.472	0.134			-0.104		0.102
Dis_T_F	0.981	0.130								
Dis_P_F	-0.513	0.383	-0.275	0.536	0.192			-0.111		
Edu	0.951	0.159								-0.106
Edu_p	0.317	-0.197	0.409	-0.262	-0.204	0.429	0.398	-0.229	0.210	-0.182
Insur_M	0.991	0.127								
Insur_P_M		-0.426	0.481	0.583			0.248		-0.175	
Insur_F	0.990	0.130								
Insur_P_F		-0.414	0.553	0.516		0.117	0.257		-0.131	0.148
Units	0.988	0.113								
Uneply_rate	-0.292	0.633		-0.280		0.224		-0.169	-0.153	0.442
Hhold_inc	0.601	-0.720			-0.114			0.117		0.156
Fmly_inc	0.589	-0.708			-0.153			0.108	0.127	0.169
MBF	0.978	0.150								
CES	0.944									
ELCAM	0.981	0.145								
HT	0.984	0.120								
Plumbing	0.235	-0.348	0.205	-0.203	0.715			0.113		
Kitchen		-0.196	0.255	-0.260	0.777	-0.122	0.137	0.159		
Tele	0.117	-0.386	0.203	0.162	0.293	0.130	-0.302	-0.659	0.289	0.106
Utility	-0.187	0.130	0.797	-0.129	-0.128	-0.169	-0.363		-0.240	
Gas	-0.489	0.475	0.175	0.311		-0.158		0.258	0.503	
Elec	0.424	-0.367	-0.700		0.133	0.214	0.267			
Pop_growth	0.897		-0.110							0.127
Hhold_lt10k	0.959	0.201								
Hhold_lt10k_P	-0.409	0.712	0.223	-0.236		0.191	0.275			0.101
Fmly_lt10k	0.960	0.206								
Fmly_lt10k_P	-0.369	0.702	0.179	-0.325			0.122			0.131
White	0.979	0.125								
Blk	0.892	0.213	0.161							-0.111
Indian	0.826						-0.116			
Asian	0.906	0.155							0.107	
Rent	0.710	-0.502	-0.135	-0.248	-0.130					0.144
Value_lt50k	-0.519	0.690			0.109				0.208	
Value	0.561	-0.654		-0.143	-0.202					0.220
Cost_M_lt200				0.108		0.795	-0.432	0.372		

Table 10 Principal Components of Correlation Matrix without Rotation

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
SS loadings	29.236	5.512	2.424	1.817	1.477	1.089	0.932	0.864	0.647	0.572
Proportion Var	0.609	0.115	0.050	0.038	0.031	0.023	0.019	0.018	0.013	0.012
Cumulative Var	0.609	0.724	0.774	0.812	0.843	0.866	0.885	0.903	0.917	0.929

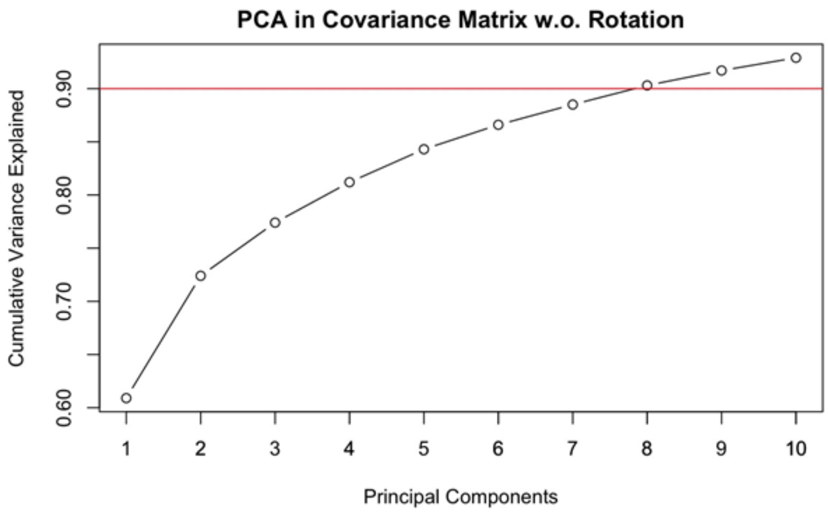

Figure 6 Cumulative Variance in Correlation Matrix without Rotation

4.1.4 PCA using correlation matrix with rotation

Table 11 Loadings for Principal Components Based on Correlation Matrix with Rotation

	RC1	RC2	RC3	RC4	RC5	RC7	RC9	RC8	RC10	RC6
Sex_M	0.982	0.162								
Sex_F	0.983	0.159								
Age_1	0.980	0.153								
Age_2	0.980	0.153								
Age_3	0.980	0.153								
Age_4	0.983	0.145								
Age_5	0.982	0.145								
Age_6	0.983	0.146								
Age_7	0.934	0.215	0.110							
Age_8	0.922	0.232	0.124							
Age_9	0.943	0.201								
Dis_T_M	0.961	0.188					0.107			
Dis_P_M	-0.313	-0.609	0.350	0.127	-0.214	-0.364				
Dis_T_F	0.975	0.158								
Dis_P_F	-0.324	-0.659	0.291	0.169	-0.169	-0.330	-0.113	0.133		
Edu	0.949	0.118				0.205				
Edu_p	0.221	0.189		0.157		0.872		0.101		
Insur_M	0.980	0.179								
Insur_P_M	-0.106	0.165	-0.105	0.892					-0.121	
Insur_F	0.981	0.175								
Insur_P_F	-0.116	0.161	-0.133	0.886						
Units	0.971	0.201								
Uneply_rate	-0.106	-0.530	-0.139	-0.208					0.714	
Hhold_inc	0.346	0.884		0.146						
Fmly_inc	0.339	0.863		0.201						
MBF	0.976	0.168								
CES	0.920	0.216				0.109				
ELCAM	0.976	0.158								
HT	0.974	0.179								
Plumbing	0.124	0.190			0.832		0.107	0.119	-0.127	
Kitchen					0.913					
Tele		0.199		0.111	0.123			0.935		
Utility	-0.108	-0.101	-0.944	0.127						
Gas	-0.287	-0.461	-0.128		-0.115		-0.778			
Elec	0.248	0.343	0.785	-0.140			0.383			
Pop_growth	0.851	0.281	0.167							
Hhold_lt10k	0.973									
Hhold_lt10k_P	-0.189	-0.668	-0.161	-0.105		0.329	-0.180	-0.257	0.425	
Fmly_lt10k	0.975									
Fmly_lt10k_P	-0.161	-0.604	-0.231	-0.233		0.179	-0.145	-0.269	0.433	
White	0.968	0.185								
Blk	0.916		-0.107			0.136				
Indian	0.803	0.191								0.150
Asian	0.906	0.120				0.118				
Rent	0.483	0.754	0.156	-0.111		0.110	0.142			
Value_lt50k	-0.273	-0.704	-0.117	-0.152			-0.411		0.177	
Value	0.308	0.853	0.154			0.102				
Cost_M_lt200										0.991

Table 12 Principal Components of Correlation Matrix with Rotation

	RC1	RC2	RC3	RC4	RC5	RC6	RC7	RC8	RC9	RC10
SS loadings	26.682	6.541	2.055	1.941	1.684	1.348	1.127	1.119	1.042	1.032
Proportion Var	0.556	0.136	0.043	0.040	0.035	0.028	0.023	0.023	0.022	0.021
Cumulative Var	0.556	0.692	0.735	0.775	0.810	0.839	0.862	0.885	0.907	0.929

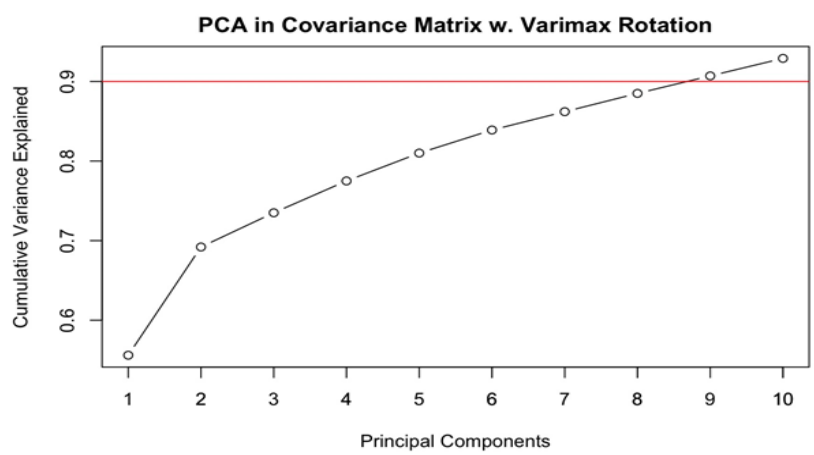

Figure 7 Cumulative Variance in Correlation Matrix with Rotation

4.2 Social vulnerability scores

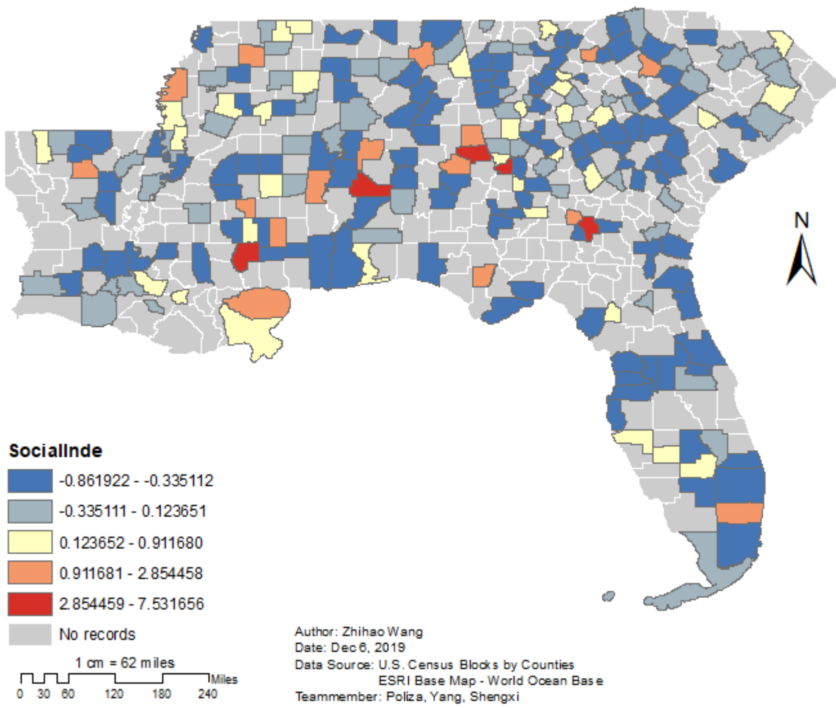

Figure 8 Map of social vulnerability index on a county level

From Fig 8, a few counties with higher social vulnerability (light color region) are randomly located on the central part of southern U.S., which demonstrates a weak relationship between geolocation and the social vulnerability.

Therefore, it is imperative to focus on investigating the relationship between the socioeconomic variables and index scores first. It appears that counties with low income and low number of Non-Hispanic people have a higher social vulnerability score (>0.7), whereas counties with higher income, larger spread of utilities, higher number of Non-Hispanic people, rent pricing have a lower vulnerability score.

Based on this visual analysis, index score distribution and understanding underlying connections to the socioeconomic data, we proceeded by grouping these scores into three classes:

1. Low Social Vulnerability < -0.5,
2. Moderate Social Vulnerability -0.5 – 0.5,

3. High Social Vulnerability > 0.5

4.3 Biophysical vulnerability and hurricane impact scores

To analyze the biophysical vulnerability, the first step is to use ArcGIS software to perform the spatial analysis of buffer based on hurricane tracks, and the range of buffer has shown in table 4. The hurricane observation lines that represent the same hurricane will be merged together. The result of the buffer is shown in figure 9.

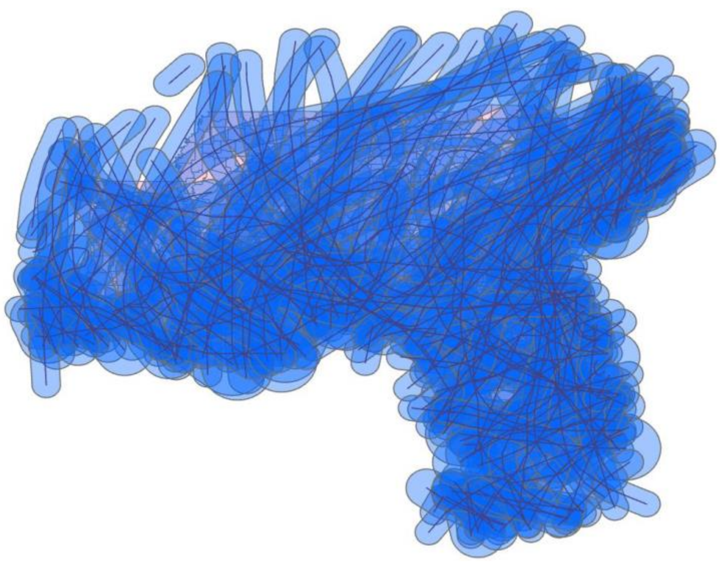

Figure 9 Buffer of each hurricane lines

Combining buffer and county map, we performed spatial analysis of pairwise intersection, to compute the area that individual hurricane influenced.

By summing up the area impacted by every hurricane, we will calculate the ratio of impact by using the formula in section 3.3 and obtain the index of biophysical vulnerability (Figure 10).

From the final result of biophysical vulnerability score, depicted on figure 11, we can observe that the area of South Florida, South Alabama, West Louisiana, and West part of South Carolina have the greatest influence from hurricane from 1980.

Biophysical vulnerability distribution (Figure 11) showed that the counties around the coast will have greater influence. The areas that have been impacted most are three counties in the eastern Florida: Indian River, St. Lucie, and Martin. Moreover, compared with counties near the North Atlantic receive more impact than cities near the Gulf of Mexico.

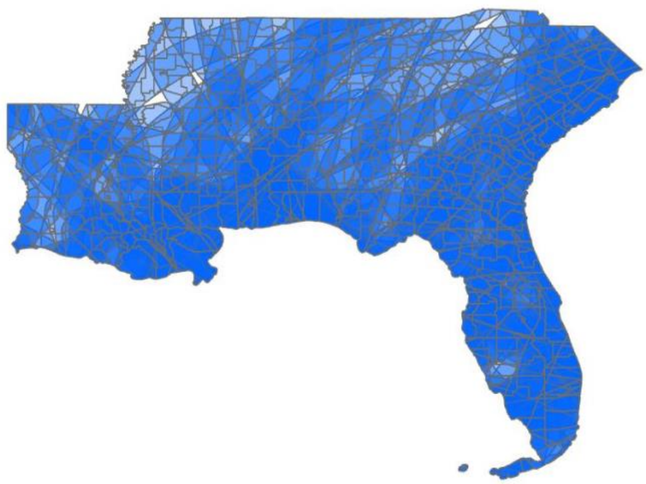

Figure 10 Intersection result of hurricane buffer and county area

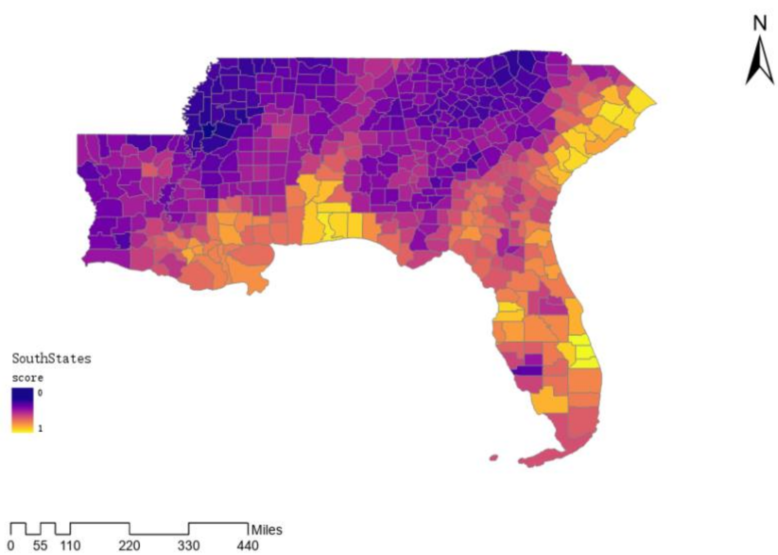

Figure 11 Intersection result of hurricane buffer and county

5 RESULTS

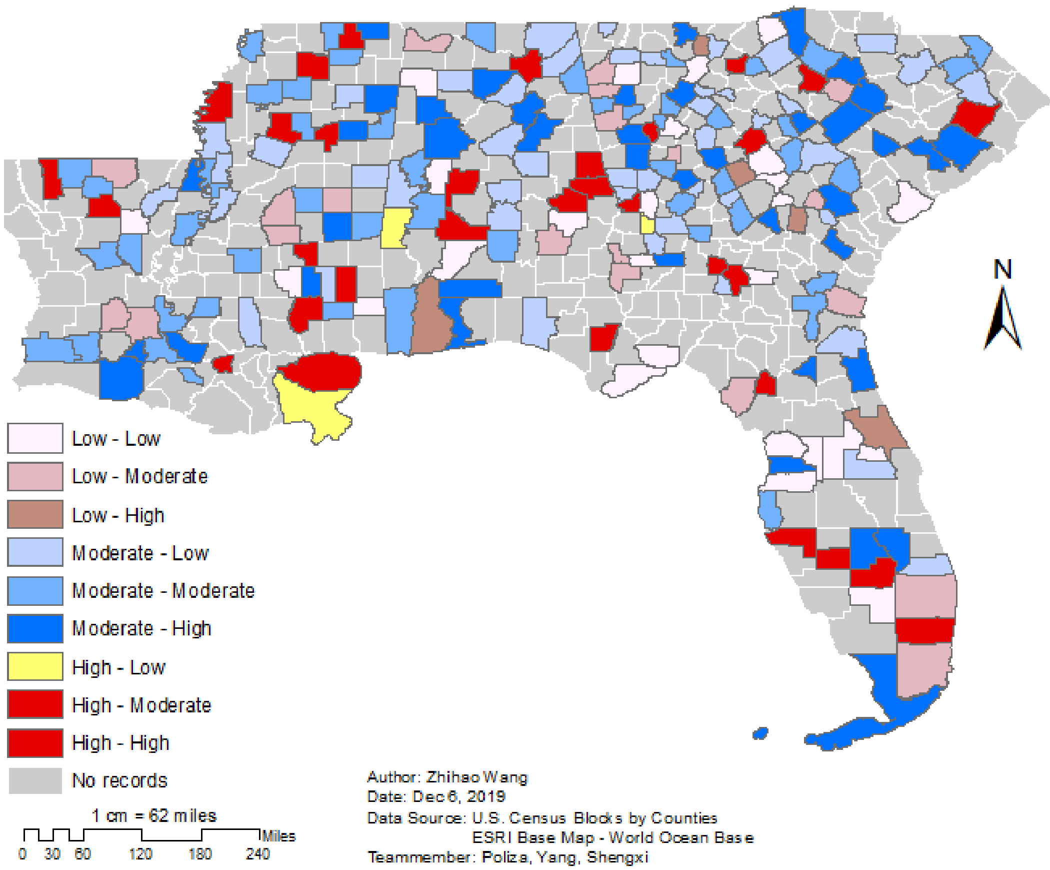

Figure 12 Southeast U.S. social and biophysical vulnerability scores compared per county

The social vulnerability and the biophysical vulnerability not align well with each other. From figure 8, it is clear that the distribution of regions with higher social vulnerability due to hurricanes are mainly inland, which indicates that the vulnerability due to social components are not closely related to the distance with the hurricane landfall point. However, from figure 11, the counties with higher biophysical vulnerability are obviously gathered at the coast, which is intuitive because the intensity of hurricane will decrease along time after landfall due to the loss of latent heat as well as earth surface friction. The joint vulnerability are based on the methodology which divides the social vulnerability and the biophysical vulnerability into 3 levels, respectively, then the total

9 levels of joint vulnerability are extend from low social-low biophysical vulnerability to high social-high biophysical vulnerability. The southeast of Texas and the southern part of Florida experience the higher social and biophysical vulnerability. In contrast, the counties with moderate and lower joint vulnerability scores are randomly spread in higher latitudes, especially in the east part of the study area.

However, this study also has some limitations. The choice of a unit of analysis, county, play an important role in our analysis. While a county-grouped data allowed for faster processing and initial estimations, it is worth further exploring how a different unit of analysis can influence biophysical and social vulnerability estimates and principal components. In addition, decision-makers might be interested in assessing the implications of the modifiable areal unit problem (MAUP) (Cressie, 1996) in the context of this study and best practices the aid distribution on a neighbourhood, local and state levels. Uncertainties also exist in the analysis due to census counties. For example, the population density within a county is not homogeneous so that a population center would be better to assess the biophysical vulnerability (i.e. buffers) than a center of geometries. Lack of detailed data limits our analysis.

The selection of weights for combining different principal components also introduces bias in our analysis. Based on the literature, the proportion of the variance explained in each principal component is used as the weights for that component. There are many other ways to explore, for example, using equal weights or pre-defined weights based on the knowledge.

Since methods used in this study are mainly based on the census data, we believe the entire frame can be applied to any other vulnerability measurement such as wildfire, earthquake, as well as flooding. The essential key part is to add the biophysical information so that the PCA can reveal more underlying patterns. Additionally, weight selection to combine different principal components depends on data, which is beyond this discussion.

6 CONCLUSIONS

The social vulnerability and the biophysical vulnerability are based on loads of variables that measure the socioeconomic status, education level, poverty, building status, the possibility of being attacked by hurricanes, etc. Principal component analysis gives us a robust way to get the most contributed variables and group them into loadings.

It is not intuitional to understand the social vulnerability and biophysical vulnerability due to hurricanes unless they can be observed directly from the map. Our study generates the social vulnerability map and the biophysical vulnerability map, respectively, and also get the joint map which measures the combined vulnerability by 3 categories (Low, Moderate, High), which could be easily perceived by the readers rather than extract information from tedious dataset.

The next step could be combining the social vulnerability and biophysical vulnerability as one index and measuring its spatial and temporal variability, which requires a historical reconstruction of all the variables utilized in this study. Moreover, the study of social and biophysical vulnerability could be projected into the future as long as the analog data could be developed under future scenarios.

REFERENCES

Aksha, S. K., Juran, L., Resler, L. M., & Zhang, Y. (2019). An Analysis of Social Vulnerability to Natural Hazards in Nepal Using a Modified Social Vulnerability Index. International Journal of Disaster Risk Science, 10(1), 103-116.

Summary: The Social Vulnerability Index was implemented in Nepali context to study vulnerable populations and factors that contribute to the disaster risk reduction. Authors used principal component analysis and put reduced principal components into the Nepali geography perspective in the discussion section.

Data Access and Dissemination Systems (DADS). (n.d.). American FactFinder. Retrieved from https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.

Cressie, N. A. (1996). Change of support and the modifiable areal unit problem. Geographical Systems, 3(2-3), 159-180.

Summary: A classic paper in spatial statistics describing the modifiable areal unit problem, its implications on geographical research analysis and results interpretations.

Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. Social science quarterly, 84(2), 242-261.

Summary: The paper used principal components analysis to reduce the socioeconomic data to estimate social vulnerability in the U.S. to natural hazards. It allowed for a robust and consistent set of variables that can be monitored over time to assess any changes in overall vulnerability

Direct Relief. (n.d.). California Wildfires: Social Vulnerability Risk. Retrieved from https://directrelief.maps.arcgis.com/apps/InteractiveLegend/index.html?appid=8d1fc11b7d1e4ac8a1e7ce2a27ef7e09

Direct Relief. (n.d.). Hurricane Harvey (8/17). Retrieved from https://directrelief.maps.arcgis.com/apps/CompareAnalysis/index.html?appid=90beb69bb62e49979b519869f388a0b4

Oxfam. (n.d.). Mapping Social Vulnerability in Southeastern United States and The Gulf Coast. Retrieved from https://policy-practice.oxfamamerica.org/work/poverty-in-the-us/mapping-social-vulnerability-in-southeastern-states-and-the-gulf-coast/

Simpson, R. (2003). Hurricane!: coping with disaster. Washington D.C.: American Geophysical Union.

Summary: A book that covers a range of topics on the science of hurricane tacking and landfall prediction, dissemination of public warning, coping with hurricanes for societies and recent scientific findings on warning systems. It provides a useful insight of hurricane coping “science” and progress in the field.

Wigtil, G., Hammer, R. B., Kline, J. D., Mockrin, M. H., Stewart, S. I., Roper, D., & Radeloff, V. C. (2016). Places where wildfire potential and social vulnerability coincide in the coterminous United States. International journal of wildland fire, 25(8), 896-908.

Summary: Examined the distribution of place vulnerability to wildfire in California by developing neighbourhood-level social vulnerability measures and combining these with existing data characterising biophysical vulnerability to wildfire. The vulnerability index was constructed using census block group data from the US Census, including socioeconomic and demographic variables, and wildfire potential data from the Wildland Fire Potential dataset (USDA Forest Service).

Willoughby, H. E., Rappaport, E. N., & Marks, F. D. (2007). Hurricane forecasting: The state of the art. *Natural Hazards Review*, *8*(3), 45-49.

Summary: This paper provides the information of where have the greatest impact by a hurricane in North America. Besides, it shows the impact area of the hurricane with different categories.

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.

Summary: This paper provides a method to evaluate the impact of the hurricane, like the method to calculate the *ratio*. And this paper provides the method of virtualization of combining natural hazards and social vulnerability.

SUPPLEMENT

1 Biography

Polina Berezina is a master's student at the Department of Geography, the Ohio State University, where she studies natural disaster assessment techniques using very-high resolution satellite imagery. Her academic interests include geospatial analysis, remote sensing, and disaster management. Polina holds a Bachelor's degree in Geology and Geoinformatics from Taras Shevchenko National University of Kyiv.

Shengxi Gui is a second-year master student in Civil, Environmental & Geodetic Engineering. Before coming to OSU, he earned the B.S in Remote sensing in Wuhan University (Wuhan, China). Shengxi's research focus on the method of 3D reconstruction in urban area based on stereo satellite imagery, including image dense matching algorithm, Virtual Reality development, and buildings model fitting.

Yang Li is a second-year PhD student from the Environmental Science Graduate Program (ESGP). She got her bachelor and master degree in Geographic Information System from Wuhan University and Beijing Normal University, respectively. Her previous study was mainly about utilizing remote sensing techniques to explore the relationship between phenology and climate change. Now She is trying to add the economic component to quantify the influence from climate change on biophysical and biochemical forcings.

Zhihao Wang is a second-year master's student at the Department of Geography. Before coming to OSU, he received B.S. of Remote Sensing at Wuhan University (China) and Bachelor of Environmental Studies at the University of Waterloo (Canada) with a Computer Science minor. Zhihao is broadly interested in big satellite image processing, including global land cover and land use classification, change detection, as well as urban/environmental influence on human decision-making process.

2 Code

2.1 Data preprocessing and analysis of US Census data in R

```
# Global Variable
setwd("C:/Users/Yang Li/Desktop/Allfiles")

# gender -----
data.gender_raw <- read.csv('gender.csv')

# Total; Population Estimates (as of July 1) - 2015 - Male;
# Total; Population Estimates (as of July 1) - 2015 - Female;
vars.gender <- c(2,26,27)
data.gender <- data.gender_raw[-1,vars.gender]
```

```
colnames(data.gender)[1] <- 'GEOG_ID'
colnames(data.gender)[2] <- 'Sex_M'
colnames(data.gender)[3] <- 'Sex_F'

write.csv(data.gender, 'data_gender.csv')

# age -----
data.age_raw <- read.csv('age.csv')

# under 18, 18-64, over 64
vars.age <- c(2,481:483, 577:579, 673:675)
data.age <- data.age_raw[-1,vars.age]
colnames(data.age)[1] <- 'GEOG_ID'
for(i in 2:length(vars.age)-1) {
  print(i)
  colnames(data.age)[i+1] <- paste0('Age_', i, seq='')
}

write.csv(data.age, 'data_age.csv')

# disability -----
data.disability_raw <- read.csv('disabilities.csv')

# with a disability and percent with a disibility - male & female
vars.disability <- c(2,12,14,18,20)
data.disability <- data.disability_raw[-1,vars.disability]
colnames(data.disability)[1] <- 'GEOG_ID'
colnames(data.disability)[2] <- 'Dis_T_M'
colnames(data.disability)[3] <- 'Dis_P_M'
colnames(data.disability)[4] <- 'Dis_T_F'
colnames(data.disability)[5] <- 'Dis_P_F'

write.csv(data.disability, 'data_disability.csv')

# education -----
data.edu_raw <- read.csv('education.csv')

# total and percent enrolled in college or grad school
vars.edu <- c(2,112, 114)
data.edu <- data.edu_raw[-1,vars.edu]
colnames(data.edu)[1] <- 'GEOG_ID'
colnames(data.edu)[2] <- 'Edu'
colnames(data.edu)[3] <- 'Edu_p'

write.csv(data.edu, 'data_edu.csv')

# health insurance -----
data.health_raw <- read.csv('health_insurance.csv')

# total and percent insured male, then female
vars.health <- c(2,146,148,156,158)
data.health <- data.health_raw[-1,vars.health]
colnames(data.health)[1] <- 'GEOG_ID'
colnames(data.health)[2] <- 'Insur_M'
colnames(data.health)[3] <- 'Insur_P_M'
colnames(data.health)[4] <- 'Insur_F'
```

```
colnames(data.health)[5] <- 'Insur_P_F'

write.csv(data.health, 'data_health.csv')

# housing units -----
data.units_raw <- read.csv('housing_units.csv')

# Housing Unit Estimate (as of July 1) - 2015
vars.units <- c(2,11)
data.units <- data.units_raw[-1,vars.units]
colnames(data.units)[1] <- 'GEOG_ID'
colnames(data.units)[2] <- 'Units'

write.csv(data.units, 'data_units.csv')

# income -----
data.income_raw <- read.csv('income.csv')

# unemployment rate_percent
# mean household income
# mean family income
vars.income <- c(2,38,252,348)
data.income <- data.income_raw[-1,vars.income]
colnames(data.income)[1] <- 'GEOG_ID'
colnames(data.income)[2] <- 'Uneply_rate'
colnames(data.income)[3] <- 'Hhold_inc'
colnames(data.income)[4] <- 'Fmly_inc'


write.csv(data.income, 'data_income.csv')

# occupation -----
data.occup_raw <- read.csv('occupation.csv')

# Total; Management, business, and financial occupations
# Total; Computer, engineering, and science occupations
# Total; legal, community service, arts, and media occupations
# Total; Healthcare practitioner and technical occupations
vars.occup <- c(2,24,54,94,144)
data.occup <- data.occup_raw[-1,vars.occup]
colnames(data.occup)[1] <- 'GEOG_ID'
colnames(data.occup)[2] <- 'MBF'
colnames(data.occup)[3] <- 'CES'
colnames(data.occup)[4] <- 'ELCAM'
colnames(data.occup)[5] <- 'HT'

write.csv(data.occup, 'data_occup.csv')

# physical characteristics -----
data.phy_raw <- read.csv('physical_cha.csv')

# Occupied housing units; Estimate; COMPLETE FACILITIES - With complete plumbing facilities
# Occupied housing units; Estimate; COMPLETE FACILITIES - With complete kitchen facilities
# Occupied housing units; Estimate; TELEPHONE SERVICE AVAILABLE - With telephone service
# Occupied housing units; Estimate; HOUSE HEATING FUEL - Utility gas
# Occupied housing units; Estimate; HOUSE HEATING FUEL - Bottled, tank, or LP gas
```

```
# Occupied housing units; Estimate; HOUSE HEATING FUEL - Electricity
vars.phy <- c(2,148,154,184,190,196,202)
data.phy <- data.phy_raw[-1,vars.phy]
colnames(data.phy)[1] <- 'GEOG_ID'
colnames(data.phy)[2] <- 'Plumbing'
colnames(data.phy)[3] <- 'Kitchen'
colnames(data.phy)[4] <- 'Tele'
colnames(data.phy)[5] <- 'Utility'
colnames(data.phy)[6] <- 'Gas'
colnames(data.phy)[7] <- 'Elec'

write.csv(data.phy, 'data_phy.csv')

# population -----
data.pop_raw <- read.csv('population_growth.csv')

# Annual Estimates of the Components of Population Change - July 1, 2014 to July 1, 2015 - Total Population
Change [1]
vars.pop <- c(2,11)
data.pop <- data.pop_raw[-1,vars.pop]
colnames(data.pop)[1] <- 'GEOG_ID'
colnames(data.pop)[2] <- 'Pop_growth'

write.csv(data.pop, 'data_pop.csv')

# poverty -----
data.poverty_raw <- read.csv('poverty.csv')

# Estimate; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Total households - Less
than $10,000
# Percent; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Total households - Less
than $10,000
# Estimate; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Total households -
Mean household income (dollars)
# Estimate; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Families - Less than
$10,000
# Percent; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Families - Less than
$10,000
# Estimate; INCOME AND BENEFITS (IN 2015 INFLATION-ADJUSTED DOLLARS) - Families - Mean family
income (dollars)
vars.poverty <- c(2,208,210,304,306)
data.poverty <- data.poverty_raw[-1,vars.poverty]
colnames(data.poverty)[1] <- 'GEOG_ID'
colnames(data.poverty)[2] <- 'Hhold_lt10k'
colnames(data.poverty)[3] <- 'Hhold_lt10k_P'
colnames(data.poverty)[4] <- 'Fmly_lt10k'
colnames(data.poverty)[5] <- 'Fmly_lt10k_P'


write.csv(data.poverty, 'data_poverty.csv')

# race -----
data.race_raw <- read.csv('race.csv')

# Estimate; RACE - One race - White
# Estimate; RACE - One race - Black or African American
```

```
# Estimate; RACE - One race - American Indian and Alaska Native
# Estimate; RACE - One race - Asian
vars.race <- c(2,128,132,136,156)
data.race <- data.race_raw[-1,vars.race]
colnames(data.race)[1] <- 'GEOG_ID'
colnames(data.race)[2] <- 'White'
colnames(data.race)[3] <- 'Blk'
colnames(data.race)[4] <- 'Indian'
colnames(data.race)[5] <- 'Asian'

write.csv(data.race, 'data_race.csv')

# rent -----
data.rent_raw <- read.csv('rent.csv')

# Estimate; Median gross rent -- - Total
vars.rent <- c(2,4)
data.rent <- data.rent_raw[-1,vars.rent]
colnames(data.rent)[1] <- 'GEOG_ID'
colnames(data.rent)[2] <- 'Rent'

write.csv(data.rent, 'data_rent.csv')

# value of house -----
data.home_raw <- read.csv('value_home.csv')

# Owner-occupied housing units with a mortgage; Estimate; VALUE - Less than $50,000
# Owner-occupied housing units with a mortgage; Estimate; VALUE - Median (dollars)
# Owner-occupied housing units with a mortgage; Estimate; MONTHLY HOUSING COSTS - Less than $200
vars.home <- c(2,6,20,60)
data.home <- data.home_raw[-1,vars.home]
colnames(data.home)[1] <- 'GEOG_ID'
colnames(data.home)[2] <- 'Value_lt50k'
colnames(data.home)[3] <- 'Value'
colnames(data.home)[4] <- 'Cost_M_lt200'


write.csv(data.home, 'data_home.csv')

# Aggregate All Data -----

data.all <- merge(data.gender, data.age, by='GEOG_ID')
data.all <- merge(data.all, data.disability, by='GEOG_ID')
data.all <- merge(data.all, data.edu, by='GEOG_ID')
data.all <- merge(data.all, data.health, by='GEOG_ID')
data.all <- merge(data.all, data.units, by='GEOG_ID')
data.all <- merge(data.all, data.income, by='GEOG_ID')
data.all <- merge(data.all, data.occup, by='GEOG_ID')
data.all <- merge(data.all, data.phy, by='GEOG_ID')
data.all <- merge(data.all, data.pop, by='GEOG_ID')
data.all <- merge(data.all, data.poverty, by='GEOG_ID')
data.all <- merge(data.all, data.race, by='GEOG_ID')
data.all <- merge(data.all, data.rent, by='GEOG_ID')
data.all <- merge(data.all, data.home, by='GEOG_ID')

write.csv(data.all, 'data_all.csv')
```

2.2 Calculating hurricane ratio in Matlab

```
county=xlsread('D:\\OSU\\3rd\\multivariate\\project\\buffer\\area.xlsx','Sheet1');
buffer=xlsread('D:\\OSU\\3rd\\multivariate\\project\\buffer\\area.xlsx','Sheet2');
num=max(county(:,1));
area_cou=county(:,20);
area_buf=zeros(num+1,1);
for i=0:num
        temp_buf=find(buffer(:,6)==i);
        area_buf(i+1)=sum(buffer(temp_buf,24));
end
ratio=area_buf./area_cou;
```

2.3 PCA and derivation of social vulnerability index scores in R

```
# set work directory
root <- dirname(rstudioapi::getSourceEditorContext()$path)
setwd(root)

# read data
census.raw <- read.csv("data_all.csv")
census.raw <- census.raw[,-1]
nObs <- nrow(census.raw)

# check missing values
foo <- na.omit(census.raw)
if (nrow(foo) != nObs){
  print('Missing value exists!')
} else {
  rm(foo)
}

# Perform z-score normalization
census.mat <- data.matrix(census.raw[,-1])
census.mat <- scale(census.mat, center=TRUE, scale=TRUE)


# Compute variance and correlation matrix
census.cov <- cov(census.mat)
census.cor <- cor(census.mat)
```

```
# 2. PCA Analysis
## 2.1 Covariance Matrix

Without rotation, 10 loadings.
```{r}
library(psych)
test <- na.omit(census.cov)
```
```

````
pca.cov_woR <- psych::principal(census.cov, nfactors = 10, rotate='none', covar=TRUE)
print(pca.cov_woR$loadings)
```

Cumulative Variance:
```{r}
# use the proportion of eigenvalues to obtain the variance explained
# plot(cumsum(pca.cov_woR$values)/sum(pca.cov_woR$values),
plot(c(0.609, 0.724, 0.774, 0.812, 0.843, 0.866, 0.885, 0.903, 0.917, 0.929),
    type='b',
    xlim=c(1,10),
    main='PCA in Covariance Matrix w.o. Rotation',
    xlab='Principal Components',
    ylab='Cumulative Variance Explained'
)
axis(side=1, at=seq(1,10,1))
abline(h=0.9, col='red')
```

With 'Varimax' rotation, 10 loadings.
```{r}
library(GPArotation)

pca.cov_wR <- psych::principal(census.cov, nfactors = 10, rotate='Varimax', covar=TRUE)
print(pca.cov_wR$loadings)
```

Cumulative Variance:
```{r}
# use the proportion of eigenvalues to obtain the variance explained
#plot(cumsum(pca.cov_wR$values)/sum(pca.cov_wR$values),
plot(c(0.556, 0.692, 0.735, 0.775, 0.810, 0.839, 0.862, 0.885, 0.907, 0.929),
    type='b',
    xlim=c(1,10),
    main='PCA in Covariance Matrix w. Varimax Rotation',
    xlab='Principal Components',
    ylab='Cumulative Variance Explained'
)
axis(side=1, at=seq(1,10,1))
abline(h=0.9, col='red')
```

Plot the first two principal components
```{r}
pca_wor12 <- pca.cov_woR$loadings[,1:2]
pca_wr12 <- pca.cov_wR$loadings[,1:2]

pc_wor <- census.mat %*% pca_wor12
pc_wr <- census.mat %*% pca_wr12

plot(pc_wor, col='red', cex=0.5)
points(pc_wr, col='blue', cex=0.5)
legend(170, -8, legend=c("No Rotation", "Varimax"),
      col=c("red", "blue"), pch=1, cex=0.8)
```
````
````

2.2 Correlation Matrix

After normalization, two matrices are same

Without rotation, 10 loadings.
```{r}
pca.cor_woR <- psych::principal(census.cor, nfactors = 10, rotate='none', covar=FALSE)
print(pca.cor_woR$loadings)
```

Cumulative Variance:
```{r}
plot(cumsum(pca.cor_woR$values)/sum(pca.cor_woR$values),
    type='b',
    xlim=c(1,15),
    main='PCA in Correlation Matrix w.o. Varimax Rotation',
    xlab='Principal Components',
    ylab='Cumulative Variance Explained'
)
axis(side=1, at=seq(1,15,1))
abline(h=0.9, col='red')

```

With 'Varimax' rotation, 10 loadings.
```{r}
pca.cor_wR <- psych::principal(census.cor, nfactors = 10, rotate='Varimax', covar=FALSE)
print(pca.cor_wR$loadings)
```

Cumulative Variance:
```{r}
plot(cumsum(pca.cor_wR$values)/sum(pca.cor_wR$values),
    type='b',
    xlim=c(1,15),
    main='PCA in Correlation Matrix w. Varimax Rotation',
    xlab='Principal Components',
    ylab='Cumulative Variance Explained'
)
axis(side=1, at=seq(1,15,1))
abline(h=0.9, col='red')

```

Link Social and Biophysical Index
```{r}
setwd(root)
posb <- read.csv('index_socvuln.csv')

res_shp <- census.raw[,1:2]
res_shp$tset <- posb[,2]
colnames(res_shp)[3] <- 'SocialIndex'

bio <- read.csv('ratio.csv')
bio <- bio[,c(6,21)]
```

```
colnames(bio)[1] <- 'GEOG_ID'
colnames(bio)[2] <- 'BioIndex'

res_shp <- merge(res_shp, bio, by='GEOG_ID')
res_shp <- res_shp[,-2]
tmp_mean <- mean(data.matrix(res_shp[,3]))
tmp_std <- sd(data.matrix(res_shp[,3]))
res_shp[,3] <- (res_shp[,3] - tmp_mean)/tmp_std

res_shp[,4] <- cut(res_shp$SocialIndex, c(-Inf, -0.5, 0.5, Inf), c(1, 2, 3))
colnames(res_shp)[4] <- 'socialLab'
res_shp[,5] <- cut(res_shp$BioIndex, c(-Inf, -0.5, 0.5, Inf), c(10, 20, 30))
colnames(res_shp)[5] <- 'bioLab'

res_shp[,6] <- paste(res_shp[,4],res_shp[,5], sep="")
colnames(res_shp)[6] <- 'compareLab'


shp_fid <- read.csv('final_shp.csv')
shp_fid <- shp_fid[,c(1,5)]
colnames(shp_fid)[1] <- 'FIDDD'
colnames(shp_fid)[2] <- 'GEOG_ID'

res_shp1 <- merge(res_shp, shp_fid, by='GEOG_ID')

write.csv(res_shp1, 'final_index.csv')
```