MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2023-017 l April 2023
https://doi.org/10.4054/MPIDR-WP-2023-017

Methodological Improvements in Social Vulnerability Index Construction Reinforce Role of Wealth Across International Contexts

Ronak Paul
Sean Reid
Carolina Coimbra Vieira l coimbravieira@demogr.mpg.de
Chris Wolfe
Yan Zhang
Yuan Zhao
Rumi Chunara

This working paper has been approved for release by: Emilio Zagheni (sekzagheni@demogr.mpg.de),
Head of the Laboratories of Migration and Mobility and Population Dynamics and Sustainable Well-Being.

Methodological Improvements in Social Vulnerability Index Construction Reinforce Role of Wealth Across International Contexts

Ronak Paul[a,1], **Sean Reid**[b,1], **Carolina Coimbra Vieira**[c,1], **Chris Wolfe**[d,1], **Yan Zhang**[e,1], **Yuan Zhao**[f,1], **and Rumi Chunara**[f,2]

[a]**International Institute for Population Sciences**
[b]**University of California Santa Barbara**
[c]**Max Planck Institute for Demographic Research**
[d]**University of Nevada, Reno**
[e]**University of Oxford**
[f]**New York University**
[1]**Authors contributed equally to this work.**
[2]**To whom correspondence should be addressed. E-mail: rumi.chunaranyu.edu**

ABSTRACT

Social conditions shape an individual's response to external hazards. This includes the degree to which an individual or community may respond to natural disasters, economic changes, or global health crises. Social vulnerability is a multi-dimensional measure of these social conditions that defines how such actors may respond to hazards. Factors that compose social vulnerability are theoretically well defined, such as economic status, age, disability, language, ethnicity, and location, which have enabled the creation and validation of social vulnerability indices across many specific locations and outcomes. However, social vulnerability index construction methods generally assume structured, linear relationships amongst input variables and may not capture subtle nonlinear patterns that may better capture the multi-dimensionality of social vulnerability. The advent of global harmonized data sources and novel methodologies in machine learning techniques enables policymakers and hazard researchers to expand the toolkit for delineating patterns of social vulnerability. Across eight countries we leverage these tools and find that wealth-related factors explain the largest variance and most common element in social vulnerability. The relevance of a data-driven approach to variable selection as well as how the constructed index relates to childhood mortality, are used for internal and external validation of the constructed indices. Given the growing nature of hazards that affect multiple environmental, social, and economic aspects of society the consistent relevance of wealth is important to impact resilience to natural and other risks.

Keywords: Social Vulnerability | Principal Component analysis| Autoencoder

INTRODUCTION

Since the original conceptualization of an index to measure social vulnerability Cutter et al. (2003), a range of studies have demonstrated the numerous ways in which social vulnerability is a significant determinant of community outcomes. The relevance of social vulnerability has been studied in relation to exposure to many natural, anthropogenic, and socio-natural hazards Aksha et al. (2019); Cutter and Emrich (2006); Karaye and Horney (2020); Flanagan et al. (2018).

Exposure and recovery from hazards are more challenging for those who experience social vulnerability Fothergill and Peek (2004); Cutter and Emrich (2006). This includes weather and climate hazards, which have cost the United States government and private insurers more than $2.2 trillion dollars since 1980 for Economic Information (2023). Studies of the impact of social vulnerability on the COVID-19 pandemic showed that a percentile increase in the Center for Disease Control and Prevention's social vulnerability index was associated with a 65% increase in COVID-19 case counts in the United States Karaye and Horney (2020). Counties in the highest social vulnerability quartile also have significantly higher mortality for cardiovascular disease (CVD), ischemic heart disease, stroke and hypertension Khan et al. (2021). Originally developed in the American context, the social vulnerability framework has been applied in contexts as varied as Pakistan, Germany, Nepal, and China, each made possible with feasible data, usually through a country's census data or in some cases survey data Aksha et al. (2019); Fekete (2009); Hamidi et al. (2022); Zhang et al.

(2017). Adaptations in new contexts allow for the opportunity to compare social vulnerability across borders to address enduring questions regarding how social vulnerability manifests across different locations Oulahen et al. (2015), as well as informing efforts to mitigate vulnerability to the growing and intertwined nature of global hazards Keim (2008).

Social vulnerability indices have many uses in research and practice. Some work has focused on validation of the predictive capability of social vulnerability to particular outcomes (such as floods, earthquakes, and non-communicable diseases) with varied results Khan et al. (2021); Wallace et al. (2015). Indeed, measuring social vulnerability in relation to specific health outcomes or for places at varying spatial scales are unique challenges. For each context, there are specific data requirements that should be fine tuned to represent the appropriate concepts at the appropriate level. At the same time, the growing nature of intertwined environment, social, health, and other hazards such as climate-driven disease pandemics or health effects motivate the need to better understand social vulnerability and to identify and compare the components that compose social vulnerability in order to strengthen societal resilience to these increasing shocks Keim (2008).

Existing work has largely focused on using structured statistical methods (i.e., parametric models) to create and evaluate social vulnerability models Cutter et al. (2003); Cutter and Finch (2008); Schmidtlein et al. (2008); Goodman et al. (2021). However, given the complex nature of social factors, possible interactions, mediation, and feedback mechanisms, more flexible models have shown promise Zhao et al. (2021). Other methodological challenges include the fact that data resources used are often limited. Variables included are selected by hand (including but not always through expert opinion), to represent specific concepts, as opposed to a data-driven method. To date, the data are largely selected from one location's census, which may not be available to reproduce in another location. These data and method limitations also affect the quality and comparability of developed social vulnerability measures. Recent efforts such as the Integrated Public Use Microdata Series (IPUMS) project Sobek and Ruggles (1999) allow for consistent social vulnerability indices to be created across contexts by merging census data from multiple countries together as well as harmonizing the data such as unifying different numeric classification systems. As census samples were never designed with compatibility in mind and come with challenges of different sampling methods, record layouts, variable coding, and uneven documentation, such efforts have high utility and enable novel analyses.

Creating and comparing standardized indices, such as standard social vulnerability indices can be useful for advancing data collection and analysis. Comparing an index across places enables assessment of the utility of variables across contexts, illumination of gaps such as latent concepts that need better elucidation, or simplification of the measure which can inform new data collection efforts in places without existing data resources Oulahen et al. (2015). Alongside the potential of augmenting data, with sufficient data, novel analytic techniques such as deep learning have been shown to capture non-linear and potentially subtle patterns. This is in contrast to standard statistical methods which have been commonly used for grouping multi-dimensional social vulnerability measures, e.g. linear combinations through principal components analysis Cutter et al. (2003); Cutter and Finch (2008). These new data and analytic resources thus can be used to potentially improve measures, as well as assess if explicating the theoretical concept of vulnerability in different ways (through more data or better pattern recognition) identifies different important factors in vulnerability assessment. Harmonized data and new analytic methods such as deep learning allow further comparisons of how social vulnerability is explicated across contexts and can inform how more data or detecting more subtle patterns reveal insights about the contribution of different dimensions to social vulnerability. Leveraging new harmonized data and analytic methods allow us to assess and validate the social vulnerability index internally as well as across countries. Further, leveraging these data and analytic resources can be used to identify and compare the components that compose social vulnerability that can be used to understand vulnerability and strengthen societal resilience to these increasing shocks Keim (2008).

Here, we produce a standard and comparable social vulnerability index across eight countries. We first use the same data chosen to represent theoretical concepts driving social vulnerability based on consensus within the social science community about major factors that influence social vulnerability (referred to as "Level 1" analysis) Cutter et al. (2003). We compare findings with a data-driven analysis, which expands the included variables to all possible variables covering the same social vulnerability factors (referred to as "Level 2" analysis). Both approaches were implemented for seven countries, for which all variables were available through IPUMS. Analyses were also performed on United States data using the American Community Survey, as a benchmark with previous social vulnerability index construction. Further, for the United States data, we also examined a deep learning method and resulting variable contributions and social vulnerability concepts. Finally, although no standardized pre-existing multi-dimensional measures of vulnerability are available across such a group of countries, as an external validation, we examined how the constructed social vulnerability index relates to childhood mortality, which is known to relate to social vulnerability across multiple contexts.

RESULTS

Impact of Data Driven Approach on Social Vulnerability

As discussed in creation of the original social vulnerability index, the theoretical concepts which underpin social vulnerability are agreed upon within the social science community. Yet, this same work also contends that the specific data and variables chosen to represent the concepts do not have the same level of consensus Cutter et al. (2003). While some efforts for individual countries outside the United States have captured such concepts in their own country-specific datasets (often census data such as in Nepal or Bangladesh Aksha et al. (2019); Rabby et al. (2019)), we were able to capture these for an international context by leveraging the IPUMS data resource. While there are subtle differences between variables gathered from IPUMS versus that of the American Community Survey (a derivative of the dataset used in the initial index construction Cutter et al. (2003)), following other international focused efforts (e.g., Aksha et al. (2019)), we included relevant proxies and overlapping data that led to similar variables across each context. We began with those concepts that influence social vulnerability most often found in the literature, and used the same benchmark method as in Cutter et al. (2003). These include socioeconomic status, gender, race and ethnicity, age, commercial and industrial development, employment loss, rural/urban, residential property, infrastructure and lifelines, renters, occupation, family structure, education, population growth, medical services, social dependence, and special needs populations Cutter (2002); Perry et al. (2001); Wolshon et al. (2005). We then selected variables in line with those provided in this previous work (as best as could be matched based on the IPUMS and current American Community Survey (ACS) datasets, for the United States) and merged each with one more data source (Open Street Map) to cover all of the concepts. The Open Street Map data is used to fill in the gap associated with the concept of medical services - such variables are not initially included in the ACS nor IPUMS. Countries for which all possible domains are available are: Cambodia, Costa Rica, Dominican Republic, Morocco, Nepal, Panama, and Senegal. This selection resulted in 61 variables and is referred to as the "Level 1" analysis. Full details are discussed in the Materials and Methods section.

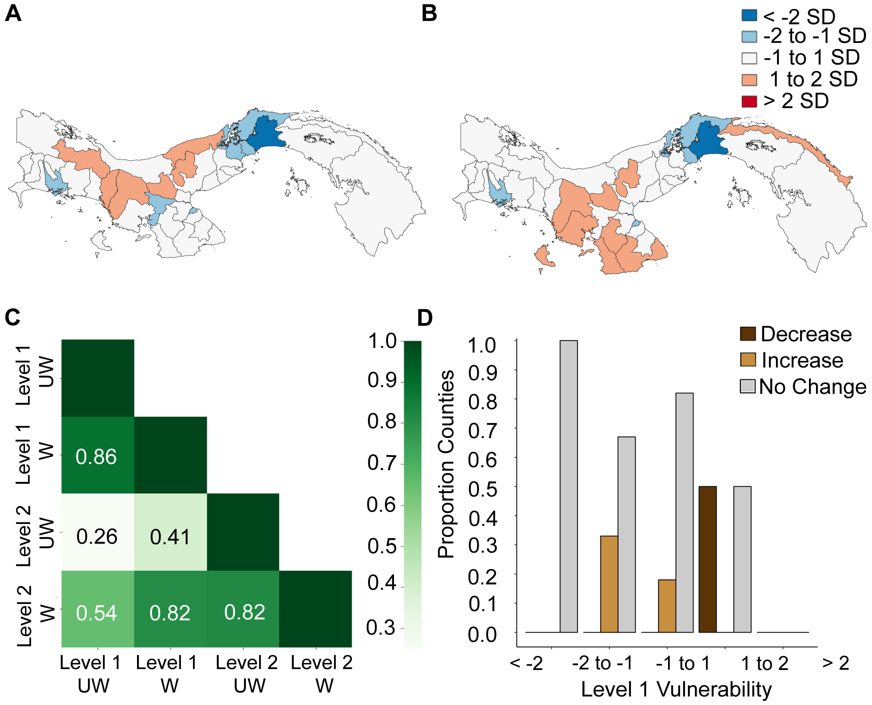

Figure 1. Examination of the implications of Level 1 (*A*) versus Level 2 (*B*) analyses, example of Panama. Correlation of the final social vulnerability index based on the Level 1 or 2 weighted (W) or unweighted (UW) approach; all correlations are significant to a $p < 0.05$ level (*C*). Across level switches from Level 1 to Level 2 analyses, the direction of vulnerability changes in terms of standard deviations (*D*).

Building on this approach, we leveraged a large amount of standardized data to also explore a data-driven method. Indeed, as reported in previous work, while the major concepts composing social vulnerability are agreed upon, disagreements arise in the selection of specific variables to represent these broader concepts. Expanding the list of variables resulted in a range of variables from 164 from the ACS representing the United States to 304 from IPUMS representing Nepal. This data-driven approach is referred to as the ”Level 2” analysis. Expansion of the number of variables in the data-driven approach was primarily the result of an increase in variables associated with age categories, race/ethnicity, family structure, socio-economic status, and residential domains. For example, Nepal has 130 ethnicity categories. In terms of ethnicity/race, the Level 2 analysis includes the full 130 ethnicity categories for Nepal, while for the Level 1 analysis, these were re-coded to two overarching categories - major ethnicity (the most populated ethnicity) and minor. Another example is the category of household characteristics, such as ownership of kitchens, toilets, refrigerators and computers, including 35 variables total mostly with binary responses of “yes” and “no” in Level 1. In the Level 2 analysis, the number of items owned is also included, expanding the category to 58 variables. Similarly, in the Cambodia case, the Level 1 analysis differentiates whether a household has a single family or multiple families, and the Level 2 analysis includes “one family”, “two families”, all the way to “8 families” and “9 and more families”. Supplementary tables S1-S8 describe the number of variables used per level for each country in the analysis.

Despite the large variability in specific variables used, two methods showed strong social vulnerability index correlation for resulting vulnerability levels (Figure 1C). As previous work calls for more attention to how components are weighted in the social vulnerability index construction Oulahen et al. (2015), we also tested the effect of weighting each component by the variance explained. Considering both Level-1 and Level-2 unweighted and weighted indices, the weighted Level 1 and 2 indices had the highest correlation across six of the eight countries. The vulnerability levels for results for Panama via the Level 1 and 2, as well as unweighted and weighted PCA methods are illustrated and compared in Figure 1. Comparisons between Level 1 and Level 2 unweighted and weighted methods for all countries are summarized in Supplementary Table S10. Results show that for each country, expanding the set of data included in computing the social vulnerability index (Level 1 to Level 2) shows consistency of the index with a more expansive data set, based on largely no change in the categorization of vulnerability levels by considered spatial unit.

For all countries except Nepal, movement in vulnerability levels was largely a decrease for those in the vulnerability ranges originally greater than 1 standard deviation, and largely an increase for those with vulnerability originally lower than -1 SD, suggesting that extremes were brought to the middle with more data. Shifts for each country are detailed in Supplementary Table S11 and total shifts are 0.71 – 0.93 proportion of units no change, 0.03 – 0.18 decrease in vulnerability, and 0.02 – 0.17 increase in vulnerability. In sum, the expansion in data does not show changes in vulnerability. Figure 2 illustrates an example of the vulnerability levels mapped by administrative unit for the Level 1 and 2 analyses. 26 (0.74%) counties stay with in -1 to 1 SD of vulnerability, while 6 (0.17%) increase and 3 (0.09%) decrease.

Data Reduction - Variables that explain the most variance and have the highest importance

First, to reduce the data we used the statistical procedure, principal components analysis (PCA), which has been the standard approach in social vulnerability index creation to define composite factors that differentiate places according to their relative level of social vulnerability Cutter et al. (2003). Using the same number of variables selected in previous analyses shows, at the second administrative level, 2 to 8 components (Level 1) that differentiated each unit, while the data-driven approach results in 3 to 10 (Level 2). The United States shows 13 principal components (Level 1) and 22 (Level 2). In both cases, the lowest number of components is in Panama, and the highest in Cambodia. A summary of the total number of principal components and total percent variation explained by the dominant principal component is summarized in Table 1. The total percent variation explained based on all components ranges from 62.9 to 74.4 % (in the Level 2 analysis). The first component explained from 21.6 to 42.1% of the variance. Each of the components is described in SI Appendix Tables S1 to S8, and the component type, determined through the same approach as in previous work Cutter et al. (2003) and fully described in the section explaining the most variation.

Household assets were the most frequent dominant component of vulnerability (explained the highest amount of variance) (Table 1). As described in survey methodology in international contexts, an asset-based measure of wealth is common in international contexts such as the Demographic Health Survey Rustein and Johnson (2004). In the United States data, in both Level 1 and 2, income measures are highly dominant. The ACS lacks questions about households’ wealth Chenevert et al. (2017), but it should be noted that education level as well as home ownership, which is also indicative of wealth in the United States Turner and Luea (2009) were also present in the PCA component explaining the highest variance.

Building upon the standard PCA, we used an autoencoder, a type of artificial neural network, to learn an efficient representation of the data Rumelhart et al. (1985). The autoencoder learns a representation (encoding) for a set of data, typically for dimensionality reduction, allowing for non-linear relationships and more flexibility than the PCA. In order to interpret the learned representations, supervised learning is often used to assess feature importance in relation to them.

Accordingly, we use Shapley values Lundberg and Lee (2017), combined with agglomerative hierarchical clustering to interpret the clusters by learning how they predicted different variables. SHapley Additive exPlanations (SHAP) methodology is a common method for ascertaining the importance of features in machine learning models and is used here to highlight which variables are most important in defining vulnerability Lundberg and Lee (2017). While not directly comparable to PCA, the idea behind autoencoder is similar to defining the principal component (and associated variables) that explains the most variance in the dataset.

Based on the best fitting model (chosen through minimizing the reconstruction loss), four resulting clusters resulted from agglomerative clustering with each cluster including 30.2%, 24.8%, 22.9%, and 22.0% of counties, respectively. Though this approach is not directly comparable to the PCA approach, the same themes dominate each outcome. Specifically, the clusters that include the largest number of counties show factors such as high median income, the proportion of the population with professional/graduate education, and the cost of rent having importance, broadly grouping wealthy, well-educated counties. Other clusters are those with a high percentage of American Indians and agricultural workers (both groups which have demonstrated increased vulnerability to natural and other hazards Lanjwani et al. (2012); Hathaway (2021)), middle-income ($50,000-$74,999), manufacturing employment having importance and finally low-income ($10,000-$14,999), the proportion of mobile homes, and no high-school degrees (Figure S3).

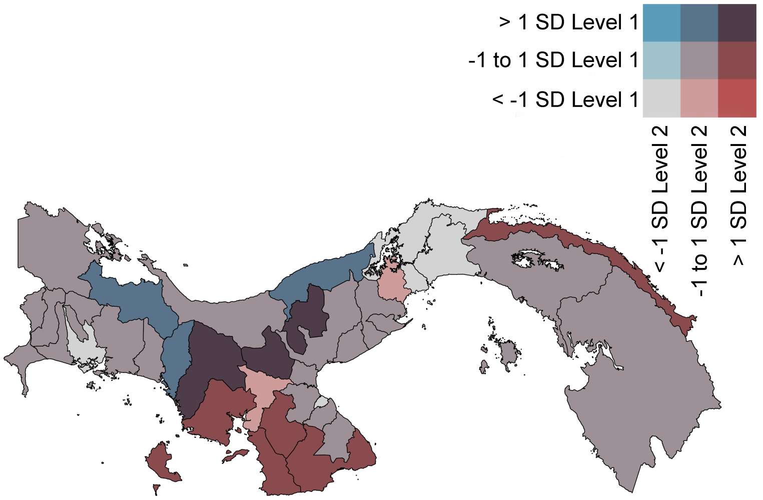

Figure 2. Panama social vulnerability by district (second administrative level). Moving from concept-driven (Level 1) to data-driven variable (Level 2) inclusion results in most districts remaining at the same vulnerability level. Some southern districts, such as Macaracas, Pedasí, Pocrí, Tonosí in the Province of Los Santos, as well as a northern district Comarca Kuna Yala in San Blas increased in vulnerability in the Level 2 analysis compared to the Level-1 analysis. While Chiriquí Grande, Tolé, Müna and Chagres and Donoso districts are more vulnerable in the Level 1 analysis.

In summary, methodological techniques that do not impose strict linear assumptions upon the data, such as an autoencoder, show consistent patterns in social vulnerability as compared to results that arise from traditional procedures that do impose such assumptions including PCA. Further, by weighting resulting PCA components by their variance we have shown that similar outcomes arise from using more (Level 2) or fewer data (Level 1) - an outcome that may impact how we explore social vulnerability in areas where data may be sparse or difficult to ascertain from traditional sources. Using these findings, social vulnerability indices were computed for each country using all available data (Level 2) (visualized in Figure 3) and are interpreted in the following section.

Geography of most and least vulnerable areas

To assess how the resulting indices capture social vulnerability across locations, we qualitatively examine geographies of the resulting most and least vulnerable areas. While gold-standard social vulnerability indices for comparison are not available, we assess how the multi-dimensional measures relate to existing understanding of economic and poverty related indicators in the included countries, at the same geographic resolution (second administrative level).

In Cambodia (Figure 3A), areas identified to have least social vulnerability overlap with districts such as Chamkar Mon and Tuol Kouk, part of central Phnom Penh which has generally lower household poverty rates Agency (2010). Banlung Municipality, which surrounds the capital of Ratanakiri Province, shows a lower level of vulnerability. This is

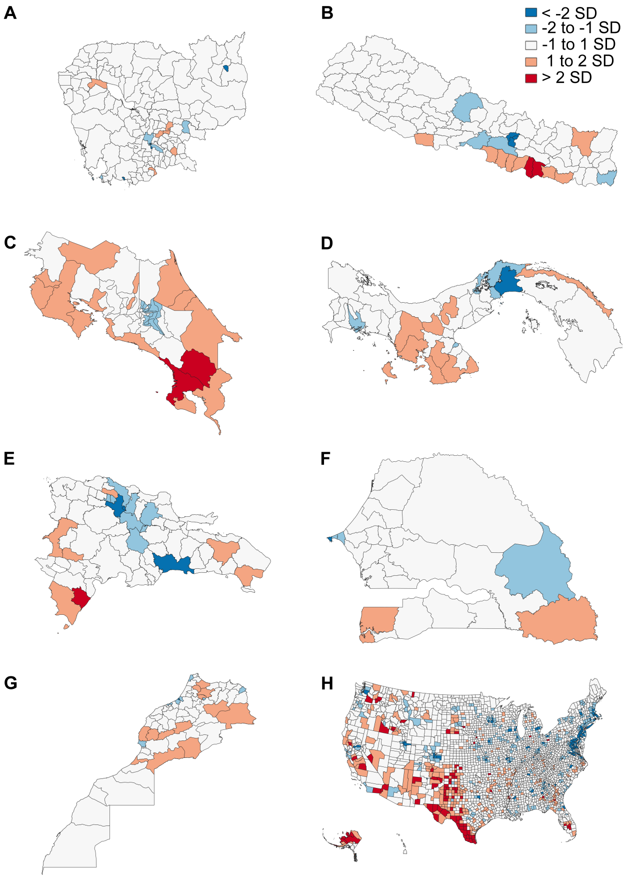

Figure 3. Maps of SoVI scores on each administrative unit. Each administrative unit is visualized by vulnerability using a standard deviation representation similar to Cutter et al. Cutter et al. (2003). Places with SoVI values between -1 and 1 standard deviation from the mean are shown in gray and indicate neutral vulnerability. Scores greater than 1 standard deviation above the mean are shown in orange and red indicating higher vulnerability. Scores less than -1 standard deviation below the mean are shown in light and dark blue indicating lower vulnerability. All SoVI scores were computed using the same harmonized IPUMS variables, besides the United States, and the scale is standardized across all countries.

understood as Ban Lung is a lively commercial area with considerably more wealth spread throughout the population - leading to additional elements to explore such as the rural/urban divide. Further, we compare our results in Nepal to previous work using the Cutter framework Aksha et al. (2019). There are certainly subtle differences in our results due to differing spatial scales and overall methodological criteria in how components were selected. However, in general, our results corroborate those of previous work including our work (Figure 3B) highlighting similar areas of poverty and poor infrastructure that are also highlighted in Aksha et al. (2019) (Figure 3) Aksha et al. (2019). Vulnerability in Costa Rica aligns with poverty maps highlighting several areas including Osa and Buenos Aires Cantons in Puntarenas Province, and richer areas in the capital San Josè Cavatassi et al. (2004) (Figure 3C). In Panama, low areas of vulnerability include the Panamá district in Panamá Province, while there are areas of higher vulnerability in Guna Yala Comarca, and Montijo, Las Palmas, Soná Mariato Districts (Veraguas Province) Assessment (2021), Figure 3D. Studies of poverty and the Human Development Index (HDI) in the Dominican Republic highlight areas of increased vulnerability including El Seybo Province, Pedernales, La Estrelleta, Bahoruco Provinces 3E. Less vulnerable areas include parts of Duarte, Monseñor Mouel and Santo Domingo DRP (2019). Previous work using census data from Senegal showed some overlapping areas of vulnerability in K/'edougou Region, Goudiry Départment in Tambacounda Region. A key difference between our work and the previous work is the characterization of Dakar as more or less vulnerable, Figure 3F. However, it should be noted that the compared work has limited details on the vulnerability index construction, the exact variables used, and how they may relate to those from IPUMS Schwarz et al. (2018). Reports from Morocco show economic vulnerability based on job loss during Covid-19 were centered in areas around Tanger-Tetouan-Al Hoceima (Chefchaouen Province, Province d Ouezzane) and Marrakech-Safi (Essaouira, Chichaoua and Al Haouz Provinces) which are also represented in Figure 3G Haddad et al. (2020). Lastly, Figure 3H identifies high vulnerability areas in parts of Southern Texas, areas in mid-California, South-west Florida, and Alaska Cutter et al. (2003) which have also been cited in the latest social vulnerability map from the United States created from 2010 census data Cutter and Finch (2008). Combined, our results here demonstrate a remarkable degree of overlap with previous country-specific analyses, further highlighting the validity of the approach highlighted here.

Social Vulnerability and Child Mortality

In addition to robustness and consistency checks for internal validation assessments (Level 1 vs Level 2) Schmidtlein et al. (2008), we also aim to examine the constructed social vulnerability measures to assess if they are measuring what we intend. While no pre-existing multi-dimensional measures of social vulnerability exist across the same set of countries for precise external construct validation, here we find that measures of social vulnerability are correlated with child mortality - a possible measure of vulnerability. Indeed, child mortality is known to be a proxy for the social, economic, environmental, and healthcare systems into which children are born Macharia and Beňová (2022). We find that increased social vulnerability is significantly positively correlated with child mortality in all countries except Nepal (correlations reported in Supplementary Table S9). It should be noted that the dominant component in the Level-2 model for Nepal is different than the rest of the countries based on the first component being race and ethnicity, instead of the household asset component. It is possible that the high number of race/ethnicity categories created by the Level 2 approach could be driving this and skewing results for Nepal.

Table 1. Summary of Dimensions of Social Vulnerability Across Countries, Dominant Component and Percentage of Variation it Explains.

Country	Level 1			Level 2		
	Total Principal Components	Percent Variation Explained	Dominating Principal Component	Total Principal Components	Percent Variation Explained	Dominating Principal Component
Cambodia	8	71.8	Household assets	10	63.1	Household assets
Costa Rica	4	73.4	Household assets, education and employment	10	66.2	Household assets, education and employment
Dominican Republic	5	74.4	Household assets, education and occupation	6	64.5	Household assets, education and occupation
Morocco	4	73.0	Household assets and education	8	70.0	Household assets and education
Nepal	4	77.2	Household assets, education and occupation	7	61.9	Ethnicity and religion
Panama	2	71.2	Household assets, occupation and education	3	63.9	Household assets, occupation and education
Senegal	3	71.3	Household assets, occupation and education	4	62.9	Dwelling characteristics, household assets, age and occupation
US	13	71.6	Wealth/income	22	73.4	Wealth/income

DISCUSSION

Our results indicate that even when considering larger types of data to represent social vulnerability, as well as allowing for more flexible deep learning algorithms, similar concepts related to wealth are most important in defining social vulnerability. Furthermore, we find that across eight countries of varied contexts (in North, Central and South America, Asia and Africa), the same concepts are important. Though previous studies have been focused on specific geographies and types of modelling approaches, our findings are significant in that given these methodological improvements, the findings still resonate those of several studies which show the importance or correlation of poverty with social vulnerability Goodman et al. (2021); Wisner et al. (2014); Fatemi et al. (2017).

Our work could have a range of implications for both research and policy. Given the increasing relevance of social vulnerability based on natural, anthropogenic and socio-natural hazards, our findings can inform data collection and development of indices for other places. Although IPUMS provides an important harmonized data resource, the base (Level-1) data needed to compose the social vulnerability index was only available in 7 countries. Accordingly, understanding of the components that capture most variance in social vulnerability can be used to prioritize data collection in new places or estimating social vulnerability in places where data covering all base concepts are not available.

Our findings also reinforce knowledge regarding global wealth trends and rise of wealth inequality which have been strongly increasing since the mid-1970s Dabla-Norris et al. (2015). Wealth is known to be driven by of a number of interrelated economic, social, and political channels, and wealth inequality, even larger than income inequality, makes it further difficult for middle- and lower-income individuals to set aside money for saving Dabla-Norris et al. (2015). This understanding of wealth enforces both the social costs and highlights the positive feedback mechanisms that will further exacerbate wealth inequities without any positive actions.

There are a additional avenues for future work to improve how researchers and policymakers define and measure social vulnerability. First, this work considers data at one time-point. Previous work tracking social vulnerability across four decades (1960-2000) in the United States has shown that while similar components consistently increased social vulnerability, there was considerable regional changes over this time period Cutter and Finch (2008), suggesting that making consistent data available over time in resources such as IPUMS would be useful for understanding changes and results of interventions. While our analysis is global in scale with eight countries represented, data availability ultimately led the decision to *only* include the selected countries. Therefore, social vulnerability indices in countries not included such as Slovenia and the Czech Republic who contain varying economic systems might distill other aspects of social vulnerability relevant in settings where wealth inequality is decreased. It is possible that data on further aspects of socio-ecological experiences that are not currently captured in census and IPUMS data resources could be used to improve social vulnerability index creation. For example, recent research highlights how discrimination affects vulnerability Carter (2021). This would include, for example, data at the individual-level, on the experience of people with diverse sexual orientations, gender identities, and also at the structural level based on policies and population-level characteristics such as segregation. Otherwise, with increasing flexibility in variable selection and categorization from existing sources (the data driven approach in Level 2), as well as flexibility in component aggregation (including using an autoencoder allowing for more than linear relationships in clustering variables), methods here still show consistency in the type of variables that matter most when measuring social vulnerability.

MATERIALS AND METHODS

Data Source Description

The Integrated Public Use Microdata Series (IPUMS)-International data which contains harmonized and analogous data (census micro-data) on a broad range of population characteristics, was leveraged to create and facilitate comparison of social vulnerability indices across multiple countries Ruggles et al. (2015).

Countries included (Cambodia, Costa Rica, Dominican Republic, Morocco, Nepal, Senegal, and Panama) based on availability of all variables needed from the vulnerability framework Cutter et al. (2003). United States was also included as a benchmark as it was the country in which the original social vulnerabilty index was produced. American Community Survey (ACS) 5-year data profile data from 2015-2019 was used U.S. Census Bureau (2020), allowing for consistency with, though more recency compared to selection based on previous work Cutter et al. (2003); Cutter and Finch (2008). To account for data on medical services, a core component of the vulnerability framework which is not available from both IPUMS and ACS 2015-2019 data (previous work had augmented United States census data with City and County Data Books from 1994 and 1998 Cutter et al. (2003)), medical service Point of Interest data from OpenStreetMap (OSM) was used. The OSM data was filtered to relevant medical facilities based on metadata of the POI tags OpenStreetMap contributors (2017).

While IPUMS provides a generalized framework to compare similar variables between global contexts, it should be noted that there is still an element of country-specific information to capture, as also implemented in a previous country-specific reproduction of the SoVI Aksha et al. (2019). The most relevant category not captured across the harmonized data within IPUMS (however is captured in the country-specific ACS) is that of race/ethnicity. To operationalize this factor within the IPUMS data we used five elements from each country-specific survey as a proxy for categories associated with race and ethnicity. These elements derive from questions related to ethnicity, religion, race, indigenous status and languages spoken. The most common identifier is religion and is present in surveys from Cambodia, Nepal and Senegal. On the other hand, race is most prevalent in the Costa Rica survey, while language is most prevalent in Morocco. The variable ”indigenous status” is only present in Panama, while the Dominican Republic does not have any variable relating to any of these elements. The Level 2 analyses incorporated all possible values of these variables, and they were recoded to include aggregated information for the level-one analysis. For example, there are 130 ethnicity categories for Nepal, which were all treated as binary variables for Level 2, the variable was recoded to major ethnicities and minor ethnicities for Level 1. In terms of the other Level 1 variables, the religion variable was recoded to Buddhist, Hindu, Jewish, Muslim, Christian, No religion, and Other religions. The race variable was recoded to White, Black and

Other races that include all other categories. The indigenous variables were consistent across both levels as one binary variable. Note, the ACS data included a a social race category and is included in both sets of analyses. All analyses were conducted on second-level administrative units (similar to counties in the United States or equivalent units such as districts or municipalities) for all countries. All geographic data comes from the GIS (geographic information system) boundary files in the IPUMS repository.

Data Selection

For consistency with the initial social vulnerability index Cutter et al. (2003), variables listed as close to those described in previous work were selected, resulting in a dataset size of 67 variables. Starting with the broad concepts enumerated within previous work using both ACS Cutter et al. (2003) and IPUMS Aksha et al. (2019) that influence social vulnerability: socioeconomic status, gender, race and ethnicity, age, commercial and industrial development, employment loss, rural/urban, residential property, infrastructure and lifelines, renters, occupation, family structure, education, population growth, medical services, social dependence, and special needs population, we came up with a broad list of variables that define social vulnerability across each country in the study. We present two levels of analysis to determine if the construction of the index is sensitive to the number of variables used to explicate each concept. The use of all variables selected represents what we define as a “Level 2” analysis (164 total variables). The inclusion of all available variables may result in collinearity between variables, but it eliminates the subjective process of selecting only certain variables as in previous works Cutter et al. (2003); Aksha et al. (2019). See supplementary material for a complete list of variables including within each level.

SOVI Construction

For both the Level-1 and Level-2 datasets, for each country, following the Cutter et al. (2003) method, principal components analysis (PCA) was used to construct an index of social vulnerability, reducing the selected list of variables into a lower-dimensional set of “components”. Once data is selected (Level-1 and Level-2), we leveraged findings from recent work examining the influence of options applied in the steps in construction of the index; PCA rotation, PCA component selection, and the weighting scheme used to combine the components to create the index Schmidtlein et al. (2008).

First, in the PCA process, data is linearly transformed into a new coordinate system wherein the variation can be described in fewer dimensions than the initial data. This involves first normalizing and centering the variables to have mean zero and then rotation of the axes. Rotation is performed so that the first axis contains as much variation as possible, the second axis contains as much of the remaining variation and so on. The rotation process can involve a change of coordinates, and Varimax rotation is one such method which maximizes the sum of the variances of the squared loadings as all the coefficients will be either large or near zero, with few intermediate values. Previous work shows that different rotation methods (no rotation, Proxmax, Varimax and Quartimax) showed fairly similar results Schmidtlein et al. (2008). Accordingly, the Varimax method, which typically leads to easier component interpretation due to loading of each variable highly on just one component, was used.

Following the PCA implementation, and following the procedure used initially Cutter et al. (2003), the most significant variables with a factor loading of more than 0.7 or 0.5 if none of the variables has a loading of more than 0.7 were assumed as drivers of each component and provided the rationale for the labels and corresponding cardinality according to their influence on social vulnerability (e.g., median household income loads on component 1 in the United States, and since higher income decreases social vulnerability, the sign of this component becomes negative because it reduces overall social vulnerability).

Next, component selection (i.e. which resulting components are used in the social vulnerability index construction), can be performed by a range of methods from expert choice to selection on the eigenvalues of components based on a threshold. We utilized Horn’s parallel analysis, which uses simulated data sets to compare the eigenvalues to expected eigenvalues for each component to determine which to retain, providing a rigorous threshold for selection Dinno (2009). For combining the selected and interpreted components, each were weighted by the proportion of total variation that particular component explains. As qualitative examination of a social vulnerability index with practitioners in Canada reported, weighting the variables the same was identified as a major source of improvement over existing methods opposed to using the raw components without weighting by variance Oulahen et al. (2015). Once each components is signed, they are weighted by their total variance and summed to create a social vulnerability score for each spatial unit. The social vulnerability score is a unitless measure whose interpretation is dependent upon geographic context.

Social vulnerability was stratified into five groups based on standard deviations from the mean, for visualization and interpretation for each country. We then examined the impact of variable set size changes on index construction, sensitivity to weighting variables in the PCA construction, and examine sensitivity across geographic contexts using the same approach as in previous work which focused on specifics of the PCA algorithm Schmidtlein et al. (2008). This includes a Pearson's correlation matrix across spatial units for each country, for each of Level-1 and Level-2 weighted and unweighted indices. Further, rank changes of vulnerability levels, stratified into the five groups are also computed and visualized.

Social Vulnerability and Child Mortality

In order to assess construct validation we assess Pearson correlation of the created social vulnerability index with another measure of vulnerability Rufat et al. (2019). For this test, we use child mortality which is known to be is a proxy for the social, economic, environmental, and health-care systems into which children are born Macharia and Beňová (2022). This proxy also can be generated from the IPUMS data at the same geographic level of the social vulnerability index. Children ever born (CHBORN in IPUMS) was subtracted from children surviving (CHSURV in IPUMS) for each record, and averaged by administrative spatial regions used. The child mortality data and weighted level 2 social vulnerability scores were compared, for each country, using a Pearson's correlation coefficient test.

Deep Learning for Vulnerability Clustering

In recent years, new deep learning techniques have been found powerful for finding structure in data. Autoencoders are a type of deep learning which have performed well to learn latent feature representations in a variety of applications such as image recognition Peng et al. (2017), pattern matching Dehghan et al. (2014), speech recognition Lee et al. (2009), and social determinants Rosati et al. (2020); Luo et al. (2021). A deep learning approach allows for nonlinear dimensionality reduction, and good generalization properties due to the inclusion of regularization methods Goodfellow et al. (2016). These aspects are of particular relevance to social factors considered here due to their complex pathways of action Mhasawade et al. (2021).

The architecture of an autoencoder consists of two elements: (i) an encoder that converts input features into a lower dimension representation called latent representation, and (ii) a decoder that reconverts the latent representations into the output corresponding to the reconstructed input. The structure of an autoencoder is similar to a Multi-Layer Perceptron with the number of neurons in the output layer equal to the number of neurons in the input layer.

We built the autoencoder using the Keras library with Tensorflow. The model was trained with ADAM Kingma and Ba (2014) defining batches of data resampled with repetition over the empirical distribution to ensure convergence. A Tanh activation function was used to allow for negative values and preserve the distribution of the data around zero. We split the full dataset into two-thirds for training and one-third for testing. With the train set, we trained a model using K-fold cross-validation (K = 10) to obtain hyperparameters (e.g., the best number of latent nodes in the latent layer). To optimize the number of hidden layers, we repeated this process varying the number of hidden layers from 1 to 32. After that, we selected the model with the lowest reconstruction loss on the test set. The estimated model has 7 hidden layers and 10 latent dimensions. Additionally, to interpret the latent layer from the autoencoder, we applied agglomerative hierarchical clustering with group-average as the inter-cluster similarity measure to categorize counties into similar clusters. The number of clusters was determined by the Davies Boulden (DB) score, which gives a measure of how similar clusters are to themselves compared to other clusters. Lower values of the DB index means that clusters are dense and well separated. Based on the DB score, the number of clusters was set to four. The SHapley Additive exPlanations (SHAP) methodology, a common method for ascertaining the importance of features in machine learning models, to a gradient boosting classification model (for predicting each of the four clusters), to identify the 20 most important variables for each cluster Lundberg and Lee (2017). The SHAP method is based on game theory to evaluate the contribution of each feature by calculating its Shapley value, the difference between the actual prediction and the mean prediction of machine model output given the current set of feature values Shapley (2016). The larger the mean SHAP value of a feature, the more important that feature is to the model prediction.

REFERENCES

(2019). Ranking: These are the poorest places in the dominican republic. Accessed: 2022-12-01.

Agency, J. I. C. (2010). Kingdom of cambodia study for poverty profiles in the asian region.

Aksha, S. K., Juran, L., Resler, L. M., and 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.

Assessment, N. D. P. B. (2021). Panama disaster risk profiles. Accessed: 2022-12-01.

Carter, B. (2021). Impact of social inequalities and discrimination on vulnerability to crises.

Cavatassi, R., Davis, B., and Lipper, L. (2004). Estimating poverty over time and space: construction of a time-variant poverty index for costa rica.

Chenevert, R., Gottschalck, A., Klee, M., and Zhang, X. (2017). Where the wealth is: The geographic distribution of wealth in the united states. *US Census Bureau*.

Cutter, S. L. (2002). *American hazardscapes: The regionalization of hazards and disasters*. Joseph Henry Press.

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

Cutter, S. L. and Emrich, C. T. (2006). Moral hazard, social catastrophe: The changing face of vulnerability along the hurricane coasts. *The Annals of the American Academy of Political and Social Science*, 604(1):102–112.

Cutter, S. L. and Finch, C. (2008). Temporal and spatial changes in social vulnerability to natural hazards. *Proceedings of the national academy of sciences*, 105(7):2301–2306.

Dabla-Norris, M. E., Kochhar, M. K., Suphaphiphat, M. N., Ricka, M. F., and Tsounta, M. E. (2015). *Causes and consequences of income inequality: A global perspective*. International Monetary Fund.

Dehghan, A., Ortiz, E. G., Villegas, R., and Shah, M. (2014). Who do i look like? determining parent-offspring resemblance via gated autoencoders. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1757–1764.

Dinno, A. (2009). Implementing horn's parallel analysis for principal component analysis and factor analysis. *The Stata Journal*, 9(2):291–298.

Fatemi, F., Ardalan, A., Aguirre, B., Mansouri, N., and Mohammadfam, I. (2017). Social vulnerability indicators in disasters: Findings from a systematic review. *International journal of disaster risk reduction*, 22:219–227.

Fekete, A. (2009). Validation of a social vulnerability index in context to river-floods in germany. *Natural Hazards and Earth System Sciences*, 9(2):393–403.

Flanagan, B. E., Hallisey, E. J., Adams, E., and Lavery, A. (2018). Measuring community vulnerability to natural and anthropogenic hazards: the centers for disease control and prevention's social vulnerability index. *Journal of environmental health*, 80(10):34.

for Economic Information, N. C. (2023). Billion-dollar weather and climate disasters. `https://www.ncei.noaa.gov/access/billions/`. Accessed: 2022-01-01.

Fothergill, A. and Peek, L. A. (2004). Poverty and disasters in the united states: A review of recent sociological findings. *Natural hazards*, 32(1):89–110.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Goodman, Z. T., Stamatis, C. A., Stoler, J., Emrich, C. T., and Llabre, M. M. (2021). Methodological challenges to confirmatory latent variable models of social vulnerability. *Natural Hazards*, 106(3):2731–2749.

Haddad, E. A., El Aynaoui, K., Ali, A. A., Arbouch, M., and Araújo, I. F. (2020). The impact of covid-19 in morocco: Macroeconomic, sectoral and regional effects.

Hamidi, A. R., Jing, L., Shahab, M., Azam, K., Atiq Ur Rehman Tariq, M., and Ng, A. W. (2022). Flood exposure and social vulnerability analysis in rural areas of developing countries: An empirical study of charsadda district, pakistan. *Water*, 14(7):1176.

Hathaway, E. D. (2021). American indian and alaska native people: Social vulnerability and covid-19. *The Journal of rural health*.

Karaye, I. M. and Horney, J. A. (2020). The impact of social vulnerability on covid-19 in the us: an analysis of spatially varying relationships. *American journal of preventive medicine*, 59(3):317–325.

Keim, M. E. (2008). Building human resilience: the role of public health preparedness and response as an adaptation to climate change. *American journal of preventive medicine*, 35(5):508–516.

Khan, S. U., Javed, Z., Lone, A. N., Dani, S. S., Amin, Z., Al-Kindi, S. G., Virani, S. S., Sharma, G., Blankstein, R., Blaha, M. J., et al. (2021). Social vulnerability and premature cardiovascular mortality among us counties, 2014 to 2018. *Circulation*, 144(16):1272–1279.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lanjwani, B. A., Gaho, G. M., et al. (2012). Debt bondage of agriculture workers in the wake of floods, 2011 sindh. *The Government-Annual Research Journal of Political Science.*, 1(01).

Lee, H., Grosse, R., Ranganath, R., and Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural* *information processing systems*, 30.

Luo, D., Caldas, M. M., and Goodin, D. G. (2021). Estimating environmental vulnerability in the cerrado with machine learning and twitter data. *Journal of Environmental Management*, 289:112502.

Macharia, P. M. and Beňová, L. (2022). Double burden of under-5 mortality in lmics. *The Lancet Global Health*, 10(11):e1535–e1536.

Mhasawade, V., Zhao, Y., and Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8):659–666.

OpenStreetMap contributors (2017). Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`.

Oulahen, G., Mortsch, L., Tang, K., and Harford, D. (2015). Unequal vulnerability to flood hazards:"ground truthing" a social vulnerability index of five municipalities in metro vancouver, canada. *Annals of the Association of American Geographers*, 105(3):473–495.

Peng, X., Li, Y., Wei, X., Luo, J., and Murphey, Y. L. (2017). Traffic sign recognition with transfer learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.

Perry, R. W., Lindell, M. K., and Tierney, K. J. (2001). *Facing the unexpected: Disaster preparedness and response in the United States*. Joseph Henry Press.

Rabby, Y. W., Hossain, M. B., and Hasan, M. U. (2019). Social vulnerability in the coastal region of bangladesh: An investigation of social vulnerability index and scalar change effects. *International Journal of Disaster Risk Reduction*, 41:101329.

Rosati, G. F., Olego, T. A., and Vazquez Brust, H. A. (2020). Building a sanitary vulnerability map from open source data in argentina (2010-2018). *International Journal for Equity in Health*, 19(1):1–16.

Rufat, S., Tate, E., Emrich, C. T., and Antolini, F. (2019). How valid are social vulnerability models? *Annals of the American Association of Geographers*, 109(4):1131–1153.

Ruggles, S., McCaa, R., Sobek, M., and Cleveland, L. (2015). The ipums collaboration: integrating and disseminating the world's population microdata. *Journal of demographic economics*, 81(2):203–216.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1985). Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Rustein, S. and Johnson, K. (2004). The dhs wealth index. `https://dhsprogram.com/pubs/pdf/cr6/cr6.pdf`. Accessed: 2022-12-01.

Schmidtlein, M. C., Deutsch, R. C., Piegorsch, W. W., and Cutter, S. L. (2008). A sensitivity analysis of the social vulnerability index. *Risk Analysis: An International Journal*, 28(4):1099–1114.

Schwarz, B., Pestre, G., Tellman, B., Sullivan, J., Kuhn, C., Mahtta, R., Pandey, B., and Hammett, L. (2018). Mapping floods and assessing flood vulnerability for disaster decision-making: A case study remote sensing application in senegal. In *Earth observation open science and innovation*, pages 293–300. Springer, Cham.

Shapley, L. S. (2016). 17. a value for n-person games. In *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–318. Princeton University Press.

Sobek, M. and Ruggles, S. (1999). The ipums project: An update. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 32(3):102–110.

Turner, T. M. and Luea, H. (2009). Homeownership, wealth accumulation and income status. *Journal of Housing Economics*, 18(2):104–114.

U.S. Census Bureau (2020). 2015 - 2019 American Community Survey 5-year Public Use Microdata Samples .

Wallace, L. M., Theou, O., Pena, F., Rockwood, K., and Andrew, M. K. (2015). Social vulnerability as a predictor of mortality and disability: cross-country differences in the survey of health, aging, and retirement in europe (share). *Aging Clinical and Experimental Research*, 27(3):365–372.

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

Wolshon, B., Urbina, E., Wilmot, C., and Levitan, M. (2005). Review of policies and practices for hurricane evacuation. i: Transportation planning, preparedness, and response. *Natural hazards review*, 6(3):129–142.

Zhang, W., Xu, X., and Chen, X. (2017). Social vulnerability assessment of earthquake disaster based on the catastrophe progression method: A sichuan province case study. *International journal of disaster risk reduction*, 24:361–372.

Zhao, Y., Wood, E. P., Mirin, N., Cook, S. H., and Chunara, R. (2021). Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *American journal of preventive medicine*, 61(4):596–605.