ECONSTOR Make Your Publications Visible.

A Service of

ZBW

Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

Zhao, Yuan et al.

Article — Published Version

Constructing Social Vulnerability Indexes with Increased Data and Machine Learning Highlight the Importance of Wealth Across Global Contexts

Social Indicators Research

Suggested Citation: Zhao, Yuan et al. (2024) : Constructing Social Vulnerability Indexes with Increased Data and Machine Learning Highlight the Importance of Wealth Across Global Contexts, Social Indicators Research, ISSN 1573-0921, Springer Netherlands, Dordrecht, Vol. 175, Iss. 2, pp. 639-657, https://doi.org/10.1007/s11205-024-03386-9

This Version is available at: https://hdl.handle.net/10419/315619

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.



http://creativecommons.org/licenses/by/4.0/

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



WWW.ECONSTOR.EU

ORIGINAL RESEARCH



Constructing Social Vulnerability Indexes with Increased Data and Machine Learning Highlight the Importance of Wealth Across Global Contexts

Yuan Zhao¹ • Ronak Paul² • Sean Reid³ • Carolina Coimbra Vieira⁴ • Chris Wolfe⁵ • Yan Zhang⁶ • Rumi Chunara^{7,8} D

Accepted: 23 June 2024 / Published online: 3 July 2024 © The Author(s) 2024, corrected publication 2024

Abstract

We consider the availability of new harmonized data sources and novel machine learning methodologies in the construction of a social vulnerability index (SoVI), a multidimensional measure that defines how individuals' and communities may respond to hazards including natural disasters, economic changes, and global health crises. The factors underpinning social vulnerability-namely, economic status, age, disability, language, ethnicity, and location—are well understood from a theoretical perspective, and existing indices are generally constructed based on specific data chosen to represent these factors. Further, the indices' construction methods generally assume structured, linear relationships among input variables and may not capture subtle nonlinear patterns more reflective of the multidimensionality of social vulnerability. We compare a procedure which considers an increased number of variables to describe the SoVI factors with existing approaches that choose specific variables based on consensus within the social science community. Reproducing the analysis across eight countries, as well as leveraging deep learning methods which in recent years have been found to be powerful for finding structure in data, demonstrate that wealth-related factors consistently explain the largest variance and are the most common element in social vulnerability.

Keywords Social vulnerability · Principal component analysis · Autoencoder

1 Introduction

Social vulnerability, an important aspect of risk assessments, refers to the characteristics and circumstances of social groups that make them susceptible to the damaging effects of hazards (Fothergill & Peek, 2004; Cutter & Emrich, 2006). An understanding of social vulnerability is increasingly relevant; for example, the costs to the United States (US) government and private insurers to support socially vulnerable communities affected by weather and climate hazards in the US since 1980 are estimated at more than \$2.2 trillion dollars

Ronak Paul, Sean Reid, Carolina Coimbra Vieira, Chris Wolfe and Yan Zhang are co-second authors.

Extended author information available on the last page of the article

(National Centers for Economic Information, 2023). US 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). Indeed, social vulnerability indexes (SoVIs) are integral to assessing a variety of natural, anthropogenic, and socionatural hazards, identifying communities for targeted response and disaster planning, formulating public health policies, addressing environmental injustices, and informing policy to prioritize resources and interventions; such indices also offer a research tool for understanding the relationship between social vulnerability and many health and environmental outcomes (Cutter, 2002; Cutter & Emrich, 2006; Bevan et al., 2022; Karaye & Horney, 2020).

Originally developed in the American context, the social vulnerability framework has been applied in varied places such as Pakistan, Germany, Nepal, and China, with each application usually using data from a country's census or, in some cases, surveys (Aksha et al., 2019; Fekete, 2009; Hamidi et al., 2022; Zhang et al., 2017). While each local context may have specific data requirements that should be fine-tuned to represent appropriate social vulnerability concepts, the proliferation of intertwined environmental, social, health, and other hazards such as climate-driven disease pandemics or health effects give rise to a need to improve the holistic understanding of social vulnerability and to identify and compare the components of social vulnerability to strengthen societal resilience to these compounding shocks (Keim, 2008).

The new harmonized data sources enables us to assess a wider set of factors and their contribution to social vulnerability, compare social vulnerability across contexts based on comparable types of data, and identify consistent drivers of social vulnerability; relatedly, this approach can inform new data collection efforts in places without existing data resources by enhancing our understanding of the most important components of the social vulnerability construct (Oulahen et al., 2015). An understanding of social vulnerability across places is also especially relevant given the proliferation and compound character of global hazards that affect multiple countries (Keim, 2008). Alongside the potential of augmenting data, novel analytic techniques such as deep learning have been shown to capture nonlinear and potentially subtle patterns, in contrast to the standard statistical methods commonly used for grouping multidimensional social vulnerability measures, e.g., linear combinations through principal component analysis (Cutter et al., 2003; Cutter & Finch, 2008). Indeed, such methods show promise for text and vision based social science analyses (Bernasco et al., 2021; Wankmüller, 2019). These new data and analytic methods such as deep learning thus can be used to potentially improve measures and assess whether explicating the theoretical concept of vulnerability in different ways (through more data or better pattern recognition) allows identification of different important factors in vulnerability assessment.

Along these lines, existing work on SoVI construction has largely focused on using structured statistical methods (i.e., parametric models) to create and evaluate social vulnerability models (Cutter et al., 2003; Cutter & 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 the data resources used are often limited. The included variables are selected manually (including, albeit not always, by means of expert opinion) to represent specific concepts rather than through allowing the specific important variables representing each concept to be selected from a larger set. To date, the data are largely selected from one location's census, which may not be available to reproduce in another location.

Here, we use Integrated Public Use Microdata Series (IPUMS) project (Sobek & Ruggles, 1999) data to create and test consistent SoVIs across contexts. As census samples are not designed with compatibility in mind, here we produce a standardized SoVI for eight countries and test different data amounts and index construction methods, enabled by IPUMS's standardization and harmonization of census data, to address challenges of different sampling methods, record layouts, variable coding, and uneven documentation. We first use the data chosen to represent the theoretical factors driving social vulnerability as identified by consensus within the social science community (a step that we refer to as our "Level 1" analysis) (Cutter et al., 2003). Then, leveraging a wider data approach (Crocetta et al., 2021) we compare the findings to a procedure that expands the included variables to all possible variables related to social vulnerability (which we refer to as our "Level 2" analysis). As a benchmark, we perform both analyses on US data using the American Community Survey as it has previously been used for SoVI construction. The same approach was also implemented for seven countries for which all variables are available through IPUMS. Further, for the US data, we also examine a deep learning method and the resulting social vulnerability factors. Finally, although no standardized preexisting multidimensional measures of vulnerability are available across our group of countries, as an external validation, we examine how the constructed SoVI relates to childhood mortality, which is known to be correlated with social vulnerability across multiple contexts (Macharia & Beňnová, 2022).

2 Data Sources for Assessment of Social Vulnerability Across Contexts

We consider the US first as a benchmark, as it is the country for which the original SoVI was produced. We use American Community Survey (ACS) 5-year data profiles from 2015 to 2019 (Census Bureau, 2020), allowing our sample to be consistent with—although more recent than—the samples selected in previous work (Cutter et al., 2003; Cutter & Finch, 2008). Looking to other countries, we utilize IPUMS International, which contains harmonized and analogous (census micro) data on a broad range of population characteristics, to create SoVIs and facilitate comparison (Ruggles et al., 2015). The included countries (Cambodia, Costa Rica, Dominican Republic, Morocco, Nepal, Senegal, and Panama) are selected based on the availability of all needed variables from the theoretical vulnerability framework, while all other 96 countries in IPUMS are missing variables for at least one of the domains, and thus are excluded. (Cutter et al., 2003). To account for medical services, a core component of the vulnerability framework not available from either the IPUMS or ACS 2015–2019 data (previous work augmented the US census data with City and County Data Books from 1994 and 1998 (Cutter et al., 2003)), we use medical service point of interest data from OpenStreetMap (OSM). The OSM data are filtered to relevant medical facilities based on the metadata of the POI tags (OpenStreetMap contributors, 2017).

While IPUMS provides a generalized framework to compare similar variables across global contexts, it should be noted that there is still an element of country-specific information to capture (this was also done in a previous country-specific reproduction of the SoVI (Aksha et al., 2019) because ethnicity data is not collected in some countries or is classified differently in different countries). The most relevant category not captured across the harmonized data within IPUMS (albeit captured in the country-specific ACS) is that of race/ ethnicity. To operationalize this factor within the IPUMS data, we use five elements from each country-specific survey as a proxy for categories associated with race and ethnicity,

including: ethnicity, religion, race, indigenous status and languages spoken. The most common identifier present is religion, which is in the surveys from Cambodia, Nepal and Senegal. On the other hand, race is the most prevalent identifier in the Costa Rica survey and language the most prevalent in the Morocco survey. The indigenous status variable is present only in the Panama survey, while the Dominican Republic does not have a variable relating to any of these elements. Given the varying availability of these types of variables across the countries, the Level 2 analyses incorporate all possible values of these variables, and they are recoded as an aggregate indicator for the Level 1 analysis. For example, there are 130 ethnicity categories for Nepal, all treated as binary variables for Level 2, while the ethnicity variable is recoded as major and minor ethnicities for Level 1. Note that the ACS data include a social race category that is included in both sets of analyses. All analyses are conducted on second-level administrative units (similar to counties in the US or equivalent units such as districts or municipalities) for all countries. All geographic data come from the geographic information system (GIS) boundary files in the IPUMS repository.

2.1 Data Selection

To ensure consistency with the initial SoVI (Cutter et al., 2003), we select variables listed as close to those described in previous work, resulting in a dataset size of 67 variables. We start with the broad factors cited in previous work using both the ACS (Cutter et al., 2003) and IPUMS data (Aksha et al., 2019) as influential in 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—and come 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 whether the construction of the index is sensitive to the number of variables used to explicate each concept. In our Level 2 analysis, we use all the variables selected (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 authors have done in previous works (Cutter et al., 2003; Aksha et al., 2019). See the supplementary material for a complete list of the variables included within each level.

3 Index Construction via Best Practices

For both the Level 1 and Level 2 datasets, for each country, following the Cutter et al. (2003) method, we use principal components analysis (PCA) to construct an index of social vulnerability. For both the Level 1 and Level 2 datasets, for each country, following the method from Cutter et al. Cutter et al. (2003), we use principal components analysis (PCA) to construct an index of social vulnerability. PCA was chosen as it is typically used in the SoVI literature (Aksha et al., 2019; Tate, 2012; Cutter & Finch, 2008). Indeed, PCA is a popular technique in statistical analyses which involve spatial components, to bring together multiple components into a lower-dimensional set of components while preserving variation of original variable with the least possible loss of information and facilitate the interpretation of the original concepts (Libório et al., 2022). Once the data are selected (Level 1 and Level 2), we constructed the vulnerability index following standard steps;

PCA rotation, PCA component selection, and a weighting scheme (Schmidtlein et al., 2008).

First, in the PCA process, the data are linearly transformed into a new coordinate system to reduce the dimensions: the variables are first normalized and centered to have mean zero and then axes are rotated using the Varimax method, which maximizes the sum of the variances of the squared loadings such that 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) offer fairly similar results (Schmidtlein et al., 2008). Accordingly, we choose the Varimax method, which typically leads to easier component interpretation due to the loading of each variable highly on just one component.

After the PCA implementation, and following the procedure used for the initial SoVI (Cutter et al., 2003), we assume that 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) (Hair et al., 2010; Comrey & Lee, 2013) are drivers of each component and define the labels and their corresponding cardinality according to the variables' influence on social vulnerability (e.g., median household income loads on component 1 in the US, and since higher income decreases social vulnerability, the sign of this component becomes negative because it reduces overall social vulnerability).

Next, we utilized Horn's parallel analysis for components selection, 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). To combine the selected and interpreted components, we weight each by the proportion of total variation that particular component explains. As a qualitative examination of a SoVI with practitioners in Canada reported, weighting the variables in this way—as opposed to using the raw components without weighting by variance—is identified as a major source of improvement over existing methods (Oulahen et al., 2015). Once each component is signed, the components 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 is stratified into five groups based on standard deviations (SD) from the mean, for visualization and interpretation for each country. We then examine the impact of variable set size changes on index construction, the sensitivity to variable weightings in the PCA construction, and the sensitivity across geographic contexts using the same approach as in previous work focused on specifics of the PCA algorithm (Schmidtlein et al., 2008). This approach includes a Pearson's correlation matrix across spatial units for each country for each of the Level 1 and Level 2 weighted and unweighted indices. Further, rank changes in the vulnerability levels stratified into the five groups are also computed and visualized.

3.1 External Validity Assessment with Child Mortality

To validate our constructed index, we assess the Pearson correlation of the created SoVI with another measure of vulnerability (Rufat et al., 2019). For this test, we use child mortality, which is known to be a proxy for the social, economic, environmental, and health care systems into which children are born (Macharia & Beňnová, 2022). This proxy also can be generated from the IPUMS data at the same geographic level as the SoVI. Children ever born (CHBORN in IPUMS) is subtracted from children surviving (CHSURV in IPUMS) for each record and averaged by the administrative

spatial region used. The child mortality data and weighted Level 2 social vulnerability scores are compared for each country by means of a Pearson correlation coefficient test.

4 Deep Learning for Vulnerability Clustering

In recent years, new deep learning techniques have been found to be powerful for finding structure in data. Autoencoders are a type of deep learning that have performed well in learning 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 nonlinear dimensionality reduction and has good generalization properties due to the inclusion of regularization methods (Goodfellow et al., 2016). These aspects are of particular relevance to the social factors considered here due to their complex pathways of action (Mhasawade et al., 2021).

The architecture of an autoencoder consists of two elements: (1) an encoder that converts input features into a lower-dimensional representation called a latent representation and (2) a decoder that reconverts the latent representations into the output corresponding to the reconstructed input. The structure of an autoencoder is similar to that of a multilayer perceptron, with the number of neurons in the output layer equal to the number of neurons in the input layer.

We build the autoencoder using the Keras library with TensorFlow. We train the model with ADAM (Kingma & Ba, 2014), defining batches of data resampled with repetition over the empirical distribution to ensure convergence. A Tanh activation function is 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 train 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 repeat this process while varying the number of hidden layers from 1 to 32. After that, we select 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 apply agglomerative hierarchical clustering with group average as the intercluster similarity measure to categorize counties into similar clusters. The number of clusters is determined by the Davies-Bouldin (DB) score, which gives a measure of how similar clusters are to themselves compared to other clusters. Lower values of the DB index mean that clusters are dense and well separated. Based on the DB score, the number of clusters is set to four. The SHapley Additive exPlanations (SHAP) methodology, a common method for ascertaining the importance of features in machine learning models, is used with a gradient boosting classification model (for predicting each of the four clusters) to identify the 20 most important variables for each cluster (Lundberg & Lee, 2017). The SHAP method is based on game theory and evaluates the contribution of each feature by calculating its Shapley value, the difference between the actual prediction and the mean prediction of the 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.

5 Dominant Variables

First, to reduce the data, we use the statistical procedure PCA, which has been the standard approach in SoVI creation for defining 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 differentiate each unit, while the wider data approach results in 3 to 10 (Level 2). The US shows 13 (Level 1) and 22 (Level 2) principal components. In both cases, the lowest number of components corresponds to Panama and the highest to 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 explains from 21.6 to 42.1% of the variance. All components determined through the Level 1 and Level 2 data selection approaches and their level of variation explained are listed in Supplementary Tables S1 to S8.

Household assets are 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 & Johnson, 2004). In the US data in both Levels 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 and home ownership, which is also indicative of wealth in the US (Turner & Luea, 2009), are also present in the PCA component explaining the highest variance. Other common components across all included countries that explained less variance are labeled by topics such as dwelling characteristics, family composition (informed by variables such as no mother or father), employment and age characteristics.

Building upon the standard PCA, as described above, we use 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 nonlinear relationships and more flexibility than the PCA. To interpret the learned representations, supervised learning is often used to assess feature importance in relation to them. Accordingly, we use Shapley values (Lundberg & Lee, 2017), combined with agglomerative hierarchical clustering, to interpret the clusters by learning how they predicted different variables. The 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 & Lee, 2017). Though autoencoder is not directly comparable to PCA, the idea behind it is similar to defining the principal component (and associated variables) explaining the most variance in the dataset.

Based on the best-fitting model (chosen through minimization of the reconstruction loss), four resulting clusters result from agglomerative clustering, with the four clusters including 30.2%, 24.8%, 22.9%, and 22.0% of the counties, respectively. Though this approach is not directly comparable to the PCA approach (it shows common features in county clusters instead of features that cluster together), the same themes dominate the results yielded by each approach. 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

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	9	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	ε	71.3	Household assets, occupation and education	4	62.9	Dwelling characteristics, household assets, age and occupation
United States	13	71.6	Wealth/income	22	73.4	Wealth/income

American Indians and agricultural workers (both groups demonstrated to have an increased vulnerability to natural and other hazards (Lanjwani et al., 2012; Hathaway, 2021)), mid-dleincome (\$50,000–\$74,999) or low income (\$10,000-\$14,999), or with high manufacturing employment, proportions of mobile homes, and shares with no high-school degree (Figure S3).

In summary, a methodological technique that does not impose strict linear assumptions upon the data (autoencoder) shows patterns in social vulnerability consistent with the results that arise from the traditional PCA procedure that imposes these assumptions. Further, by weighting the resulting PCA components by their variance, we show that similar outcomes arise from using more (Level 2) or less (Level 1) data—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, we compute SoVIs for each country using all available data (Level 2) (visualized in Fig. 2) and interpret them in the following section.

6 Geography of Most and Least Vulnerable Areas

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

In Cambodia (Fig. 2A), the areas identified to have the 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 (Japan International Cooperation Agency, 2010). Banlung Municipality, which surrounds the capital of Ratanakiri Province, shows a lower level of vulnerability. This is understandable, as Banlung is a lively commercial area with considerable wealth spread throughout the population, which means that, for this area, elements in addition to poverty such as the rural/urban divide are relevant for vulnerability. Further, we compare our results for Nepal to those from previous work using the Cutter framework (Aksha et al., 2019). In brief, there are certainly subtle differences in our results are selected. However, in general, our results corroborate those from previous work (for example, Fig. 2B), highlighting similar areas of poverty and poor infrastructure also highlighted in other SoVI construction efforts for Nepal (Aksha et al., 2019).

Vulnerability in Costa Rica aligns with poverty maps highlighting several areas including the Osa and Buenos Aires Cantons in Puntarenas Province and richer areas in the capital San José (Cavatassi et al., 2004) (Fig. 2C). In Panama, areas of low vulnerability include the Panamá district in Panamá Province, while there are areas of higher vulnerability in Guna Yala Comarca and the Montijo, Las Palmas, and Soná Mariato Districts (Veraguas Province) (Assessment, 2021) (Fig. 2D). Studies of poverty and the Human Development Index (HDI) in the Dominican Republic highlight areas of increased vulnerability including the El Seibo, Pedernales, La Estrelleta, and Baoruco Provinces (Fig. 2E). Less vulnerable areas include parts of Duarte, Monseñor Nouel and Santo Domingo (Ranking: These are the poorest places in the Dominican Republic, 2019). Previous work using census data from Senegal showed some overlapping areas of vulnerability in the Kédougou Region and Goudiry Département in the Tambacounda

Region. A key difference between our work and the previous work is the characterization of Dakar as more or less vulnerable (Fig. 2F). 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 that economic vulnerability based on job loss during COVID-19 was centered in areas around Tanger–Tetouan–Al Hoceima (Chefchaouen and Ouezzane Provinces) and Marrakech–Safi (Essaouira, Chichaoua and Al Haouz Provinces), which are also represented in Fig. 2G (Haddad et al., 2020). Last, Fig. 2H identifies high-vulnerability areas in parts of southern Texas, areas in mid-California, southwest Florida, and Alaska (Cutter et al., 2003), which have also been cited in the latest social vulnerability map from the US created from 2010 census data (Cutter & Finch, 2008). Combined, our results here demonstrate strong overlap with previous country-specific analyses, further highlighting validity of the approach used.

7 Impact of Considering a Wider Set of Data on the SoVI

As discussed with respect to the creation of the original SoVI, the theoretical concepts that underpin social vulnerability are agreed upon within the social science community. However, this same work also contends that the specific data and variables chosen to represent the concepts do not enjoy the same level of consensus (Cutter et al., 2003). While some attempts at building SoVIs for individual countries other than the US have captured the necessary concepts in their own country-specific data sets (often census data, as in Nepal and Bangladesh (Aksha et al., 2019; Rabby et al., 2019)), we can capture these for an international context by leveraging the IPUMS data resource. Though there are subtle differences between variables gathered from IPUMS and those from 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 include relevant proxies and overlapping data that lead to similar variables across each context. We begin with those concepts identified most often in the literature as influencing social vulnerability and use 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 select variables in line with those provided in previous work (as best as we can match them between the IPUMS and current ACS datasets for the US) and merge each with one more data source (OpenStreetMap) to cover all of the concepts. The OpenStreet-Map data are used to fill in the gap associated with the concept of medical services—such variables were not initially included in the ACS nor IPUMS. The countries for which all possible domains are available are Cambodia, Costa Rica, Dominican Republic, Morocco, Nepal, Panama, and Senegal. This selection results in 61 variables and is referred to as the Level 1 analysis.

Building on this approach, we consider a method that selects from a wider set of data (Level 2 analysis). 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 to include all relevant variables from IPUM and ACS results in a number of variables ranging from the

164 from the ACS, representing the US, to the 304 from the IPUMS representing Nepal. This approach is referred to as the Level 2 analysis. The expansion of the number of variables is primarily the result of an increase in variables associated with age categories, race/ ethnicity, family structure, socioeconomic 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 are recoded into two overarching categories: major ethnicity (the most populated ethnicity) and minor ethnicities. Another example is the category of household characteristics, such as ownership of kitchens, toilets, refrigerators and computers, which includes 35 variables total, most of them 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", etc., 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.

The vulnerability levels for Panama in Levels 1 and 2 and under the unweighted and weighted PCA methods are illustrated and compared in Fig. 3. Comparisons of the Levels 1 and 2 unweighted and weighted methods for all countries are summarized in Supplementary Table S10. The results show that, for each country, expanding the set of data included in computing the SoVI (going from Level 1 to Level 2) yields index results consistent with those based on a more expansive data set, with largely no changes in the categorizations of vulnerability level in the considered spatial units. Despite the large variability in the number of variables used, the two methods show a strong correlation in the resulting social vulnerability levels (Fig. 3C). As previous work calls for more attention to how components are weighted in SoVI construction (Oulahen et al., 2015), we also test the effect of weighting each component by the variance explained. Considering both the Level 1 and Level 2 unweighted and weighted indices, the weighted indices had the highest correlation across six of the eight countries.

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

8 Social Vulnerability and Child Mortality

In addition to robustness and consistency checks for internal validation (Level 1 vs Level 2) (Schmidtlein et al., 2008), we examine the constructed SoVIs to assess whether they are measuring what they are intended to measure. While no preexisting multidimensional



Fig. 1 Panama social vulnerability by district (second administrative level). Moving from concept-driven (Level 1) to a wider data approach (Level 2) 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, and the northern district Comarca Kuna Yala in San Blas, increase in vulnerability in the Level 2 analysis compared to the Level 1 analysis, while the Chiriquí Grande, Tolé, Müna and Chagres and Donoso districts are more vulnerable in the Level 1 analysis

measures of social vulnerability exist across the same set of countries for precise external construct validation, we find here that our measures of social vulnerability are correlated with child mortality as an indicator of vulnerability. Child mortality is known to be a proxy for the social, economic, environmental, and health care systems into which children are born and are regarded as indicators of socio-economic development in a community (Macharia & Beňnová, 2022; Mishra et al., 2023). 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 from that for 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 the results for Nepal.



Fig. 2 Maps of SoVI scores for each administrative unit. The vulnerability of each administrative unit is visualized by means of a standard deviation (SD) representation similar to that of Cutter et al. (2003). Places with SoVI values between -1 and 1 SD are shown in gray and indicate neutral vulnerability. Scores greater than 1 SD are shown in orange and red, indicating higher vulnerability. Scores less than -1 SD are shown in light and dark blue, indicating lower vulnerability. All SoVI scores are computed from the same harmonized IPUMS variables, except for the US, and the scale is standardized across all countries



Fig.3 Comparison of Level 1 (**A**) and Level 2 (**B**) analyses, example of Panama. Correlation of the final SoVI based on the Level 1 or 2 weighted (W) or unweighted (UW) approach; all correlations are significant at a p < 0.05 level (**C**). Across the switches from Level 1 to Level 2 analyses, the direction of vulnerability changes in terms of the standard deviation (**D**)

9 Conclusions

Our results indicate that across eight countries in varied contexts (in North, Central and South America, Asia and Africa), when we consider an increased data size in generating our SoVI and allowing for more flexible algorithms to capture the common components, concepts related to wealth are consistently the most important in defining social vulnerability. Though previous studies have been focused on specific geographies and types of modeling approaches, our findings are significant in that, given these methodological improvements, the findings still resonate with those of several studies showing the importance of poverty to or its correlation with social vulnerability (Wisner et al., 2014; Fatemi et al., 2017; Goodman et al., 2021).

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 socionatural hazards, our findings can inform data collection and development of indices for new nations and regions. Although IPUMS provides an important harmonized data resource, the base (Level 1) data needed to compose the SoVI are available only for 7 countries. Accordingly, an understanding of the components that capture the most variance in social vulnerability can be used to prioritize data collection in new places or estimate social vulnerability in places where data covering all the base concepts are not available. Our findings also reinforce knowledge regarding global wealth trends and the rise of wealth inequality, which has 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, to an even greater extent than income inequality, makes it more difficult for middle- and lower-income individuals to set aside money for saving (van Krevel, 2023). This understanding of wealth also highlights the positive feedback that will occur via further exacerbation of wealth inequities due to the immense resource and social costs of hazards in the absence of any interventions to mitigate these inequities.

There are additional avenues for future work to improve how researchers and policymakers define and measure social vulnerability. First, this work considers data at one timepoint. Previous work tracking social vulnerability across four decades (1960–2000) in the US has shown that while similar components consistently increased social vulnerability, there were considerable regional changes over this period (Cutter & Finch, 2008), suggesting that making available data that are consistent over time in resources such as IPUMS would be useful for our understanding of changes and results of interventions. While our analysis is global in scale, with eight countries represented, data availability ultimately led us to decide to include *only* the selected countries. Therefore, SoVIs in countries not included such as Slovenia and the Czech Republic, which have 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 socioecological experiences not currently captured in census and IPUMS data resources could be used to improve SoVI creation. For example, recent research highlights how discrimination affects vulnerability (Carter, 2021). Extensions could consider, for example, data at the individual level on the experience of people with diverse sexual orientations and gender identities and at the structural level based on policies and population-level characteristics such as segregation. Otherwise, even with the greater flexibility in selecting and categorizing variables from existing sources (through the Level 2 approach) and in aggregating components (including by means of an autoencoder, which allows for more than linear relationships in clustering variables), the methods here still show consistency in the type of variables that matter most in measuring social vulnerability.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11205-024-03386-9.

Author Contributions YZ: Conceptualization, Data curation, Software, Methodology, Writing - Original Draft, Writing - Review & Editing, Visualization. RP, SR, CCV, CW, YZ: Conceptualization, Data curation, Software, Methodology, Writing - Original Draft, Visualization. RC: Conceptualization, Methodology, Funding Acquisition, Writing - Original Draft, Writing - Review & Editing, Supervision.

Funding We acknowledge funding from the Max Planck Society via a Sabbatical Award, and United States National Science Foundation grant 1845487 to RC.

Consent, Data, Materials and/or Code availability Previously published data were used for this work. The data underlying this article are all publicly available (Ruggles et al., 2003; Census Bureau, 2022).

Declarations

Conflict of interest None.

Ethical approval This study utilised secondary, de-identified data and was exempt from ethical review.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

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.
- Assessment, N. D. P. B. (2021). Panama disaster risk profiles. Accessed 02 December 2022.
- Bernasco, W., Hoeben, E., Koelma, D., Liebst, L. S., Thomas, J., Appelman, J., & Lindegaard, M. R. (2021). Promise into practice: Application of computer vision in empirical research on social distancing. *Sociological Methods & Research*, 52, 1239–1287.
- Bevan, G., Pandey, A., Griggs, S., Dalton, J. E., Zidar, D., Patel, S., & Al-Kindi, S. (2022). Neighborhood-level social vulnerability and prevalence of cardiovascular risk factors and coronary heart disease. *Current Problems in Cardiology*, 48, 101182.
- Carter, B. (2021). Impact of social inequalities and discrimination on vulnerability to crises. Institute of Development Studies.
- Cavatassi, R., Davis, B., & Lipper, L. (2004). Estimating poverty over time and space: construction of a time-variant poverty index for costa rica. ESA working paper.
- Chenevert, R., Gottschalck, A., Klee, M., & Zhang, X. (2017). Where the wealth is: The geographic distribution of wealth in the united states. US Census Bureau.
- Comrey, A. L., & Lee, H. B. (2013). A first course in factor analysis. Psychology Press.
- Crocetta, C., Carpita, M., & Perchinunno, P. (2021). Data science and its applications to social research. Social Indicators Research, 156(2–3), 339–340.
- Cutter, S. L. (2002). American hazardscapes: The regionalization of hazards and disasters. Joseph Henry Press.
- Cutter, S. L., Boruff, B. J., & Shirley, W. L. (2003). Social vulnerability to environmental hazards. Social Science Quarterly, 84(2), 242–261.
- Cutter, S. L., & 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., & 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., & Tsounta, M. E. (2015). Causes and consequences of income inequality: A global perspective. International Monetary Fund.
- Dehghan, A., Ortiz, E. G., Villegas, R., & Shah, M. (2014). Who do i look like? Determining parentoffspring resemblance via gated autoencoders. *Proceedings of the IEEE conference on computer* vision and pattern recognition (pp. 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., & 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.
- Fothergill, A., & 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., & Courville, A. (2016). Deep learning. MIT Press.
- Goodman, Z. T., Stamatis, C. A., Stoler, J., Emrich, C. T., & 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., &Araújo, I. F. (2020). The impact of covid-19 in Morocco: Macroeconomic, sectoral and regional effects.
- Hair, J. F., Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis*. Pearson Prentice Hall.
- Hamidi, A. R., Jing, L., Shahab, M., Azam, K., Atiq Ur Rehman, M., & 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*, 37, 256.
- Japan International Cooperation Agency. (2010). Kingdom of Cambodia study for poverty profiles in the Asian region. OPMAC Corporation.
- Karaye, I. M., & 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., & Nasir, K. (2021). Social vulnerability and premature cardiovascular mortality among us counties, 2014 to 2018. *Circulation*, 144(16), 1272–1279.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412. 6980
- Lanjwani, B. A., & Gaho, G. M. (2012). Debt bondage of agriculture workers in the wake of floods, 2011 Sindh. *The Government-Annual Research Journal of Political Science*, 1, 1.
- Lee, H., Grosse, R., Ranganath, R., & 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 (pp. 609–616).
- Libório, M. P., da Silva Martinuci, O., Machado, A. M. C., Machado-Coelho, T. M., Laudares, S., & Bernardes, P. (2022). Principal component analysis applied to multidimensional social indicators longitudinal studies: Limitations and possibilities. *GeoJournal*, 87(3), 1453–1468.
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 22.
- Luo, D., Caldas, M. M., & 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., & Beňová, L. (2022). Double burden of under-5 mortality in lmics. *The Lancet Global Health*, 10(11), e1535–e1536.
- Mhasawade, V., Zhao, Y., & Chunara, R. (2021). Machine learning and algorithmic fairness in public and population health. *Nature Machine Intelligence*, 3(8), 659–666.
- Mishra, P. S., Sinha, D., Kumar, P., Srivastava, S., & Syamala, T. (2023). Linkages of multi-dimensional vulnerabilities with infant and child mortality rates in India and its specific regions: Are social determinants of health still relevant? OMEGA-Journal of Death and Dying, 86(3), 1002–1018.
- National Centers for Economic Information. (2023). Billion-dollar weather and climate disasters. https:// www.ncei.noaa.gov/access/billions/. Accessed 01 January 2022.
- OpenStreetMap Contributors. (2017). Planet dump retrieved from.https://planet.osm.org, https://www.opens treetmap.org
- Oulahen, G., Mortsch, L., Tang, K., & 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., & Murphey, Y. L. (2017). Traffic sign recognition with transfer learning. 2017 IEEE symposium series on computational intelligence (SSCI) (pp. 1–7).
- Perry, R. W., Lindell, M. K., & Tierney, K. J. (2001). Facing the unexpected: Disaster preparedness and response in the United States. Joseph Henry Press.
- Rabby, Y. W., Hossain, M. B., & 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.
- Ranking: These are the Poorest Places in the Dominican Republic. (2019). https://dominicantoday.com/dr/ economy/2019/09/06/ranking-these-are-the-poorest-places-in-the-dominican-republic/. Dominican Today. Accessed 01 December 2022.
- Rosati, G. F., Olego, T. A., & 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., & Antolini, F. (2019). How valid are social vulnerability models? Annals of the American Association of Geographers, 109(4), 1131–1153.

- Ruggles, S., King, M. L., Levison, D., McCaa, R., & Sobek, M. (2003). Ipums-international. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 36(2), 60–65.
- Ruggles, S., McCaa, R., Sobek, M., & Cleveland, L. (2015). The ipums collaboration: Integrating and disseminating the worldâ€TMs population microdata. *Journal of Demographic Economics*, 81(2), 203–216.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (Tech. Rep.). California Univ San Diego La Jolla Inst for Cognitive Science.
- Rustein, S., & Johnson, K. (2004). The dhs wealth index. https://dhsprogram.com/pubs/pdf/cr6/cr6.pdf. Accessed 01 December 2022.
- Schmidtlein, M. C., Deutsch, R. C., Piegorsch, W. W., & 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., & 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* (pp. 293–300). Springer.
- Shapley, L. S. (2016). 17. A value for n-person games. Contributions to the theory of games (am-28) (Vol. 2, pp. 307–318). Princeton University Press.
- Sobek, M., & Ruggles, S. (1999). The ipums project: An update. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 32(3), 102–110.
- Tate, E. (2012). Social vulnerability indices: A comparative assessment using uncertainty and sensitivity analysis. *Natural Hazards*, 63(2), 325–347.
- Turner, T. M., & 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.
- U.S. Census Bureau. (2022). American Community Survey, 2010 American Community Survey 5-Year Estimates.
- van Krevel, C. (2023). Why cross-country convergence of income is unsustainable: Evidence from inclusive wealth in 140 countries. *Social Indicators Research* (pp. 1–29).
- Wankmüller, S. (2019). Introduction to neural transfer learning with transformers for social science text analysis. Sociological Methods & Research, 00491241221134527.
- Wisner, B., Blaikie, P., Cannon, T., & Davis, I. (2014). At risk: Natural hazards, people's vulnerability and disasters. Routledge.
- Wolshon, B., Urbina, E., Wilmot, C., & 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., & 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., & Chunara, R. (2021). Social determinants in machine learning cardiovascular disease prediction models: A systematic review. *American Journal of Preventive Medicine*, 61(4), 596–605.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yuan Zhao¹ • Ronak Paul² • Sean Reid³ • Carolina Coimbra Vieira⁴ • Chris Wolfe⁵ • Yan Zhang⁶ • Rumi Chunara^{7,8}©

Rumi Chunara rumi.chunara@nyu.edu

> Yuan Zhao yuan.zhao@nyu.edu

Ronak Paul greenophenn@gmail.com Sean Reid sean.reid@geog.ucsb.edu

Carolina Coimbra Vieira coimbravieira@demogr.mpg.de

Chris Wolfe wolfec23@ecu.edu

Yan Zhang yan.zhang@sociology.ox.ac.uk

- ¹ Department of Epidemiology, New York University, New York, USA
- ² Department of Public Health and Mortality Studies, International Institute for Population Sciences, Mumbai, Maharastra, India
- ³ Department of Geography, University of California, Santa Barbara, USA
- ⁴ Laboratory of Migration and Mobility, Max Planck Institute for Demographic Research, Rostock, Germany
- ⁵ Department of Anthropology, East Carolina University, Greenville, NC, USA
- ⁶ Department of Sociology, University of Oxford, Oxford, UK
- ⁷ Department of Biostatistics, New York University, New York, USA
- ⁸ Department of Computer Science and Engineering, New York University, New York, USA