VulnerabilityMap: An Open Framework for Mapping Vulnerability among Urban Disadvantaged Populations in the United States

Lin Chen[1] , **Yong Li**[2] , **Pan Hui**[1,3]
[1]Hong Kong University of Science and Technology
[2]Tsinghua University
[3]Hong Kong University of Science and Technology (Guangzhou)
lchencu@connect.ust.hk, liyong07@tsinghua.edu.cn, panhui@ust.hk

Abstract

Cities are crucibles of numerous opportunities, but also hotbeds of inequality. The plight of disadvantaged populations who are "left behind" within urban environments has been an increasingly pressing concern, which poses substantial threats to the realization of the UN SDG agenda. However, a comprehensive framework for studying this urban dilemma is currently absent, preventing researchers from developing AI models for social good prediction and intervention. To fill this gap, we construct *VulnerabilityMap*, a framework to meticulously dissect the challenges faced by urban disadvantaged populations, unraveling their vulnerability to a spectrum of shocks and stresses that are categorized through the prism of Maslow's hierarchy of needs. Specifically, we systematically collect large-scale multi-sourced census and web-based data covering more than 328 million people in the United States regarding demographic features, neighborhood environments, offline mobility behaviors, and online social connections. These features are further related to vulnerability outcomes from short-term shocks such as COVID-19 and long-term physiological, social, and self-actualization stresses. Leveraging our framework, we construct machine learning models that exhibit strong performance in predicting vulnerability outcomes from various disadvantage features, which shows the promising utility of our framework to support targeted AI models. Moreover, we provide model-based explainability analysis to interpret the reasons underlying model predictions, shedding light on intricate social factors that trap certain populations inside vulnerable situations. Our constructed dataset is publicly available at https://github.com/LinChen-65/VulnerabilityMap/.

1 Introduction

In the process of unprecedented global urbanization, cities have emerged as vibrant hubs teeming with opportunities and innovations, drawing millions of people seeking better lives with remarkably fast-paced development, intricate social networks, and diverse cultural experiences [Bettencourt *et al.*, 2007]. However, beneath the surface of urban dynamism lie marginalized and hence disadvantaged populations along multiple dimensions including but not limited to income, race, ethnicity, and education level [Nijman and Wei, 2020]. Such marginalization stems from long-standing structural inequalities that often go unaddressed, some of which even end up in a vicious circle of self-reinforcement. For instance, rapid population growth strains infrastructures, leading to housing shortages and increased cost of facility access. Limited access to quality education, healthcare, and employment opportunities creates barriers to upward mobility for many urban residents, further exacerbating inequalities [Kearney and Levine, 2014]. Moreover, urban life is riddled with challenges ranging from immediate shocks such as disease outbreaks and climate disturbance [Abedi *et al.*, 2021] to persistent stresses such as violent crimes and social segregation [Leitner *et al.*, 2018]. These adversities constitute a web of vulnerabilities, particularly for disadvantaged communities. Understanding the situation of these communities in the face of urban challenges is essential for fostering inclusive and sustainable urban development, as emphasized by the UN's Sustainable Development Goals (SDG), and ensuring the well-being of all residents, as anticipated by the UN's Leave No One Behind (LNOB) Principle.

Addressing such a crucial issue necessitates a comprehensive analytic framework empowered by multi-sourced data and targeted AI models. Regrettably, such a united framework is currently absent, with previous studies predominantly focusing on one aspect of vulnerability. To surmount this obstacle faced by the research community, we endeavor to dissect this urban complexity by meticulously curating the *VulnerabilityMap* framework. To comprehensively characterize urban disadvantages and vulnerabilities, our framework integrates the power of multiple types of data sources. Alongside census and survey statistics, we incorporate web-collected human mobility data, online social network characteristics, and other digital footprints of human activities. Fusing such data richness not only enables multidimensional profiling of social vulnerabilities with nuanced temporal and spatial details, but also provides up-to-date reflections of on-the-ground realities. For instance, mobility traces can un-

Please find our Appendix here: https://rb.gy/b5lgq2 cover the lived experiences of disadvantaged communities in their daily activities beyond their residential neighborhood [Moro *et al.*, 2021; Wang *et al.*, 2018], and online interest data facilitate understanding of diverse lifestyles that are not covered by any demographic census [Araujo *et al.*, 2017; Fatehkia *et al.*, 2020]. Through this comprehensive exploration, we aim to shed light on the nuanced experiences of urban disadvantaged populations, unraveling their vulnerability to a spectrum of shocks and stresses.

As depicted in Figure 1, our constructed framework consists of two key components: the *disadvantage features* and the *vulnerability outcomes*. The *disadvantage features* encompass data that reflect the unequal distribution of intrinsic demographic features and extrinsic living experiences, which are organized into four indices: *demographic disadvantage index*, *neighborhood disadvantage index*, *mobility disadvantage index*, and *social disadvantage index*. Combination of these four indices further yields a composite index that summarizes the experienced disadvantage of urban populations. The *vulnerability outcomes* are classified into two categories: *vulnerability to shocks*, which typically occur abruptly and exert most impacts in a relatively short time, and *vulnerability to stresses*, which are generally milder but accumulate gradually in a long term.

We conduct a series of experiments including geographical visualization, correlation analysis, prediction, and temporal analysis to showcase the usefulness of our constructed dataset in supporting various AI research on vulnerability faced by disadvantaged populations. Notably, our constructed machine learning models achieve strong prediction performances for all vulnerability outcome variables by jointly considering different disadvantage dimensions. Moreover, we can distinguish the importance of different disadvantage features for predicting each vulnerability outcome, so as to seek a better understanding of what contributes to certain dimensions of vulnerability.

To summarize, the contribution of this work is three-fold:

- We construct *VulnerabilityMap*, the first comprehensive framework for mapping vulnerability outcomes with a carefully categorized list of disadvantage features, which are extracted from multi-sourced web-collected data covering demographic characteristics, neighborhood environment, offline mobility patterns, and online social networks and interests.
- We conduct a series of experiments to validate the usefulness of our framework in supporting the development of targeted AI models. Specifically, we construct machine learning models based on the constructed dataset that achieve strong prediction performances, underscoring the effectiveness of selected feature dimensions.
- We provide explainability analysis to interpret the reasons underlying model predictions, shedding light on intricate social factors resulting in vulnerability traps for certain populations.

VulnerabilityMap seeks to pave the way for research on targeted interventions and policy initiatives, striving to create more equitable and resilient urban communities toward sustainable development.

2 Methods

2.1 Data Identification

We aim to create a comprehensive framework as a foundation for dissecting the root of inequalities and vulnerabilities in urban space. Therefore, we gather data from multiple sources, including official census data, tech-company-released data, and self-collected data published by researchers. We mainly consider three criteria when selecting data sources:

- **Accessibility criteria**. Our objective is to create an open framework accessible to everyone in the research community. Thus, we select data sources that either possess an open-source license or allow open-source distribution of certain derived features, if not all raw data.
- **Granularity criteria**. Previous research [Levy *et al.*, 2022; Hsu *et al.*, 2021] highlighted socio-economic heterogeneities among granular communities such as neighborhoods. To enable cross-comparison with these findings and facilitate new insights, we select data sources that can at least be decomposed to the county level, excluding those at the MSA level, state level, or even national level.
- **Timeliness criteria**. If the data are widely dispersed along the temporal dimension, the significance of their correlations is expected to decrease considerably. Thus, we opt for data collected as early as 2000, which not only ensures a substantial duration for analysis with an observational window of approximately 20 years but also maintains data cohesiveness.

With these criteria, our final set of data sources is summarized in Tables 1-4 in Appendix A.

2.2 Data Retrieval and Pre-Processing

As the collected data are heterogeneous in spatial resolution, time period, and data type, we design a pipeline for retrieving and processing the raw data into a unified format, as shown in Figure 2.

Processing demographic and neighborhood data. Our primary demographic data source is the American Community Survey (ACS), which is an annual official census capturing a wide range of socioeconomic features for every census block group (CBG) in the United States. Extracted features include: (1) *Non-Hispanic white rate*, reflecting persisting racial-ethnic disparities in many sections of urban life; (2) *Disability rate*, indicating households with at least one person with a disability, who are marginalized groups with highly-constrained capability in moving around to access various urban resources; (3) *Median household income*, a key factor shaping lifestyle choices and resilience capability; (4) *Bachelor rate*, representing the percentage of the population with at least a bachelor's degree, which is connected with socio-economic status and has a lasting impact on social mobility; and (5) *Creative job rate*, indicating the percentage of the working-age population in creative industries, as listed in the "Management, business, science, and arts occupations" category in the ACS [Credit *et al.*, 2021].

We also complement the dataset with researcher-collected data sources. For example, we extract the number of polluting

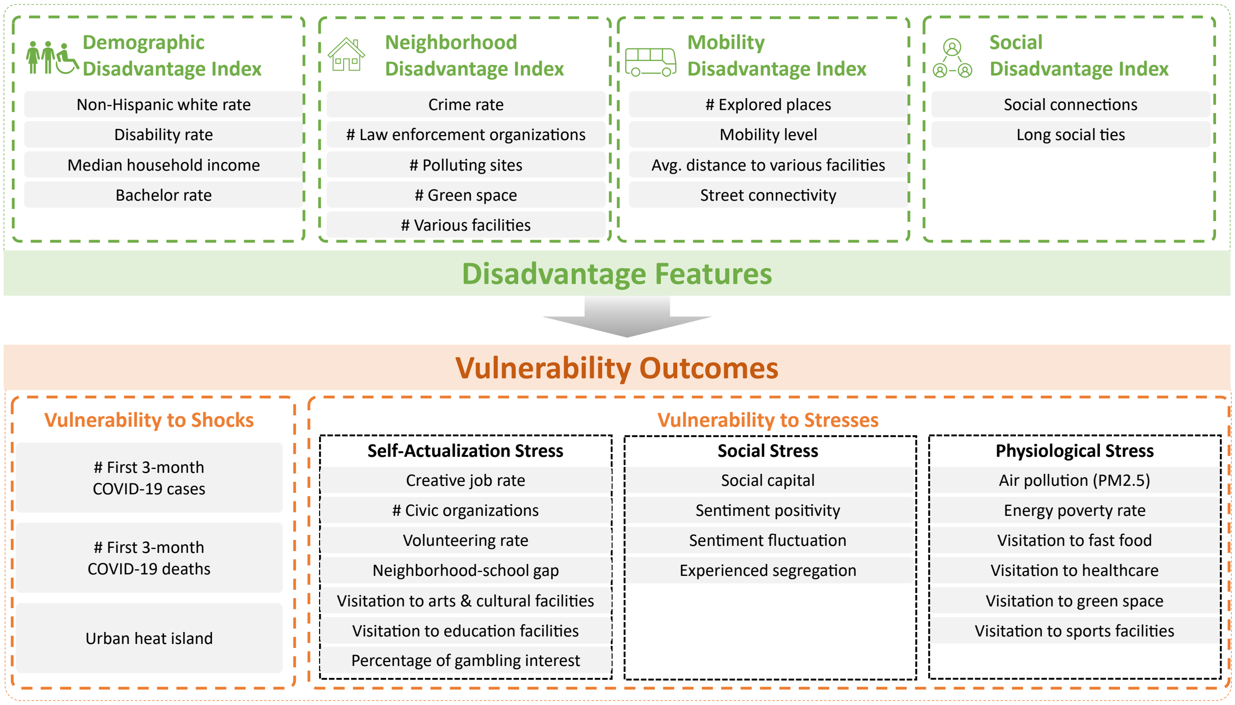

Figure 1: Schematic framework of VulnerabilityMap.

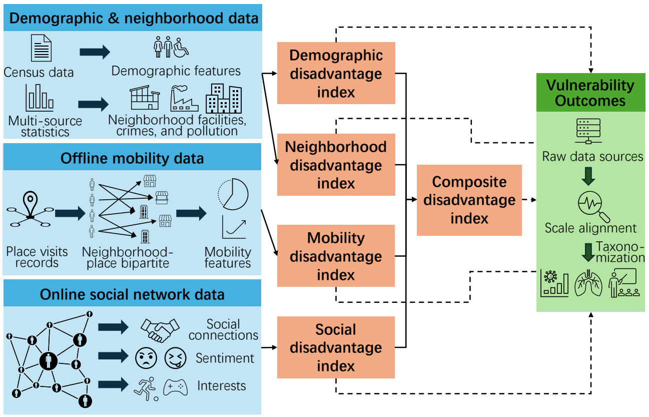
Demographic & neighborhood data
Census data
Demographic features
Multi-source statistics
Neighborhood facilities, crimes, and pollution
Offline mobility data
Place visits records
Neighborhood-place bipartite
Mobility features
Online social network data
Social connections
Sentiment
Interests
Demographic disadvantage index
Neighborhood disadvantage index
Mobility disadvantage index
Social disadvantage index
Composite disadvantage index
Vulnerability Outcomes
Raw data sources
Scale alignment
Taxonomization

Figure 2: Open workflow of the dataset construction.

sites [Finlay *et al.*, 2022], and the neighborhood-school gap which measures the discrepancy between the demographics of a public school and its surrounding community [Gomez-Lopez *et al.*, 2021].

Processing offline mobility data. Human mobility in urban space enables access to various facilities and spatial interaction, which is a critical element holding the social fabric. To include mobility features, we perform feature extraction from the Safegraph Patterns Data [1], which has been used to analyze inequality of place access [Fan *et al.*, 2022], reveal uneven behavioral changes during COVID-19 [Chen *et al.*, 2023], and design policy interventions [Chen *et al.*, 2022]. Safegraph collects data from online applications with location service about the visitations to each place of interest (POI), which can be traced back to the residential neighborhoods (CBGs) these visitors come from. To ensure privacy protection, the collected visitation data are aggregated spatially to the CBG level and temporally on a monthly basis, with differential privacy protection techniques further applied. From Safegraph Patterns, we extract the following features: (1) *Mobility level*, measured by the total visitation frequency of a neighborhood; (2) *# Explored places*, measured by the number of unique places visited by people from a neighborhood, which reflects people's exploration willingness in urban space; (3) *Visitation frequencies* to representative urban facilities that satisfy people's essential needs, including healthcare facilities, educational facilities, stores providing healthy food, green space, arts and cultural facilities, and sports facilities; and (4) *Experienced segregation* (ES), which quantifies the *de facto* segregation phenomenon beyond static residential perspectives that people from different socio-economic backgrounds are far from uniformly mixing in urban space, albeit seemingly moving freely from place to place [Moro *et al.*, 2021]. To calculate this feature, we divide CBGs into $N = 5$ groups with similar population sizes according to their median household income, and obtain the segregation at each POI by calculating the deviation of the observed mixing of different income groups from the perfect uniform mixing (with the greatest entropy), illustrated below:

[1] https://www.safegraph.com/

$$ES_p = \sum_{i=1}^{N} \left| p_i - \frac{1}{N} \right|, \qquad (1)$$

where p_i is the probability of Group i to visit a certain POI p.

Then, we obtain the segregation experienced by each CBG by taking the average of the segregation at all POIs that have been visited by this CBG, weighted by the visitation frequency to each POI.

Processing online social network data. We extract personal interest data from the Facebook Advertising Platform with a Python-based wrapper library named *pySocialWatcher* [2], which has been utilized in ”Digital Demography” [Cesare *et al.*, 2018] studies for socioeconomic indicator mapping [Fatehkia *et al.*, 2020] and chronic disease surveillance [Araujo *et al.*, 2017]. For each geographical region, we first query the expected number of all Monthly Active Users (MAU) in the Facebook user population (denoted as MAU$_{all}$), and then those with a pre-defined interest (denoted as MAU$_x$). Assuming a uniform distribution of personal interest between the sampled Facebook population MAU and the whole population N, we can divide MAU$_x$ by MAU$_{all}$ to obtain an estimation of the percentage of people with a specific interest in a geographical area, illustrated as follows:

$$\% \text{ people interested in } x = \frac{N_x}{N_{all}} = \frac{\text{MAU}_x}{\text{MAU}_{all}}. \qquad (2)$$

We also gather other online social network data to capture inequality in interpersonal connections and sentiments. First, we obtain features regarding online social interactions from the Facebook Data for Good Platform [3], including *social connections*, *social capital*, and number of *long social ties*. Second, we process the Twitter Sentiment Geographical Index Dataset [Chai *et al.*, 2023] to extract a *sentiment positivity* (measured as the average level of sentiment) and a *sentiment fluctuation* (measured as the standard deviation of sentiment).

2.3 Construction of Disadvantage Indices

Organizing data within our framework offers a degree of convenience; however, dealing with the raw values of diverse dimensions remains challenging due to two primary sources of complexity. First, distinct data ranges exist, where some features span from 0 to 1, while others theoretically range from 0 to infinity. Second, these dimensions exhibit different directions: in some cases (e.g., *disability rate*), a higher value signifies a greater disadvantage, whereas for others (e.g., *median household income*), the opposite holds true. To provide a clearer and more consistent perspective, we construct disadvantage indices using the following processing steps, inspired by the approach outlined in [Hale *et al.*, 2021]:

- **Step 1**: For each disadvantage dimension, rank all neighborhoods from the least disadvantaged to the most advantaged, which assigns an ordinal number to each neighborhood.
- **Step 2**: Apply z-score normalization to these rankings along each disadvantage dimension to obtain normal distributions.
- **Step 3**: Calculate the composite disadvantage index for each neighborhood, by summing up all the normalized rankings, which summarizes the corresponding set of disadvantage dimensions.

In the Results Section, we will further validate the effectiveness of the constructed disadvantage indices by correlating them with both the individual disadvantage features and the vulnerability outcomes.

2.4 Taxonomization of Vulnerability Outcomes

While living in cities enjoys considerable life convenience and thriving opportunities, it is also accompanied by a number of intertwined urban challenges. In a general sense, urban challenges can be divided into two categories, i.e., shocks and stresses, where shocks refer to those that typically occur abruptly and exert most impacts in a relatively short time, and stresses refer to those that are generally milder but accumulate gradually in a long term [Leitner *et al.*, 2018]. Correspondingly, we classify *vulnerability outcomes* into two categories, i.e., those to shocks and those to stresses. For *vulnerability to shocks*, we include data about the COVID-19 pandemic [The New York Times, 2021] and the Urban Heat Island (UHI) effect [Hsu *et al.*, 2021]. For *vulnerability to stresses*, we further taxonomize them along Maslow’s hierarchy of needs, as shown in Figure 3. From bottom to top, human needs can generally be categorized into three levels: basic needs (for food, water, safety, etc.), psychological needs (for friends, prestige, etc.), and self-fulfillment needs (for achieving one’s full potential, including creative activities). In accordance with this categorization, we divide urban stresses into physiological stresses (regarding healthcare, crime rate. etc.), social stresses (regarding social capital, experienced segregation, etc.), and self-actualization stresses (regarding political participation, volunteering, etc.). Note that this classification does not imply differentiated respect for different levels of needs, but rather serves as a slicing method to observe urban life.

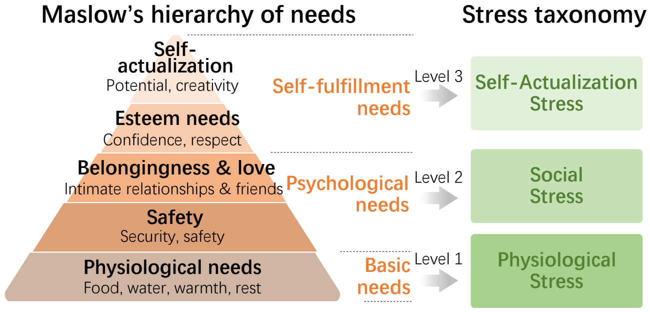

Figure 3: Taxonomy of urban stresses along Maslow’s hierarchy of needs.

[2] https://github.com/maraujo/pySocialWatcher

[3] https://dataforgood.facebook.com/dfg/tools

2.5 Machine Learning Model for Predicting Vulnerability

To further exemplify the utility of our meticulously constructed dataset for AI research, we train machine learning models to predict the values of vulnerability outcomes based on the comprehensive set of disadvantage features we collected. Our methodology involves partitioning the dataset into training and test sets, with 80% of CBGs randomly selected for training and the remaining 20% reserved for testing. For each vulnerability outcome variable, we opt for a random forest regression model, a choice driven by its ability to strike a balance between predictive performance and interpretability. The random forest comprises 100 decision trees, and we assess the model's predictive performance using the coefficient of determination (R^2).

The notable advantage of employing random forest models lies in their ability to provide feature importance scores as a natural outcome of the iterative data-splitting process based on features that yield the greatest information gain. This feature importance analysis is pivotal in unraveling the driving forces behind our predicted results. For each vulnerability outcome, we extract the top-3 important disadvantage features. This granular exploration aids in gaining a nuanced understanding of the factors contributing to specific dimensions of vulnerability, offering valuable insights into the intricacies of the urban challenges faced by disadvantaged populations.

3 Results

3.1 Checking Data Distributions

We discretize the constructed disadvantage index into 5 levels, where Group 0 represents the least disadvantaged and Group 4 represents the most disadvantaged, after which we visualize its distribution in the continental United States in Figure 4. We mainly observe two interesting characteristics: First, geographical adjacency does not ensure similar disadvantage levels, as some of the most disadvantaged neighborhoods are immediately adjacent to a neighborhood belonging to the least disadvantaged group. Second, there exists an unbalanced distribution of disadvantage at the national scale. Specifically, many neighborhoods from the most disadvantaged group cluster in the southwestern part intersecting with New Mexico and Arizona, both ranking relatively low in GDP per capita among all US states and territories. Meanwhile, many neighborhoods among the least disadvantaged cluster in the northeastern part where major cities like New York, Boston, and Washington, D.C. reside, which are hubs for finance, technology, education, and research. These observations validate that our index can capture the macroscopic heterogeneity in disadvantage distribution across geographical regions.

3.2 Effectiveness of Individual Disadvantage Dimensions

As our dataset contains multiple dimensions of potential disadvantages, we examine the correlation between single-dimensional disadvantage indices and representative vulnerability outcomes. As shown in Table 1, *demographic disadvantage index* shows reasonable results regarding the directions of correlations. For example, the negative correlation with *average sentiment* signals that neighborhoods with more disadvantaged demographic characteristics generate more negative feelings. Likewise, these disadvantaged neighborhoods are associated with fewer resources to receive help from other high-SES neighborhoods from social connections (reflecting more physiological stress), greater fluctuation and thus less stability in their sentiment (reflecting more social stress), and smaller participation rate in volunteering activities (reflecting more self-actualization stress). However, it is worth noting that we find a positive correlation between *demographic disadvantage index* and *support ratio*, suggesting that these neighborhoods with more disadvantaged demographic characteristics have a tighter (albeit maybe smaller) support network. The reason may be that socio-economically disadvantaged populations having fewer chances to explore the urban space are more confined in the family and close-friend networks. The concrete impact of such social network formation may be two-fold and needs further investigation.

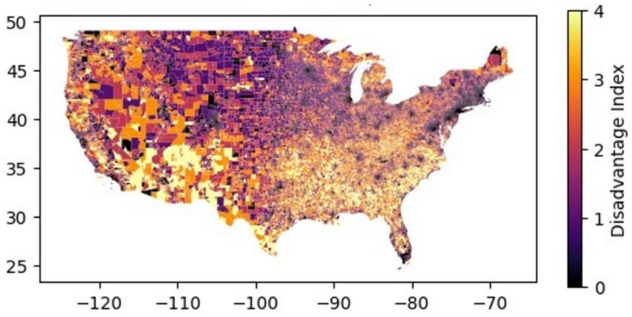

Figure 4: Geographical distribution of disadvantage index.

Vulnerability Outcome	Spearman Correlation
Sentiment positivity	-0.1384
Economic connectedness	-0.3810
Sentiment fluctuation	0.1294
Support ratio	0.2523
Volunteering rate	-0.1551

Table 1: Correlation between demographic disadvantage index and sampled vulnerability outcomes

3.3 Effectiveness of Composite Disadvantage Indices

After confirming the reasonability of the individual disadvantage indices, we take a further step to examine the effectiveness of our generated composite disadvantage indices. As shown in Figure 5, we identify the ”most disadvantaged neighborhoods” and the ”least disadvantaged neighborhoods” by sorting all neighborhoods based on the values of their *composite disadvantage index*, and compare their vulnerabilities when facing multiple types of challenges. The results are consistent with our expectations: the most disadvantaged neighborhoods are more negative and less stable in sentiment, exposed more to both the Urban Heat Island effect and air pollution, participating less in political and volunteering activities, and enjoying less economic connectedness but showing a stronger support ratio. These results show that our designed metrics are capable of accurately locating disadvantaged populations in urban spaces when taking into consideration the balance between multiple dimensions of disadvantage, providing convenience for designing intervention policies.

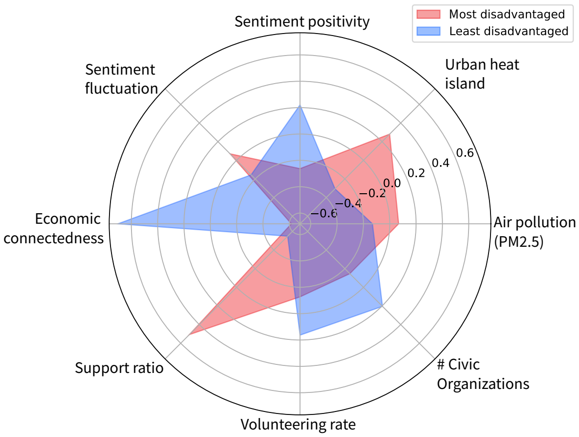

Figure 5: Comparison of most- and least-disadvantaged neighborhoods on vulnerability outcomes.

3.4 Prediction Performance and Interpretation

We present the prediction performances for all vulnerability outcomes in Figure 6, where the vulnerability outcomes are grouped into one type of *shock* and three types of *stresses*. Across these four categories, our predictive models demonstrate consistently strong performance. Specifically, all 20 outcomes are predicted with $R^2 > 0.4$, with 15 of them achieving an even higher $R^2 > 0.7$. These results substantiate the effectiveness of our selected variables as reliable indicators of vulnerability occurrences.

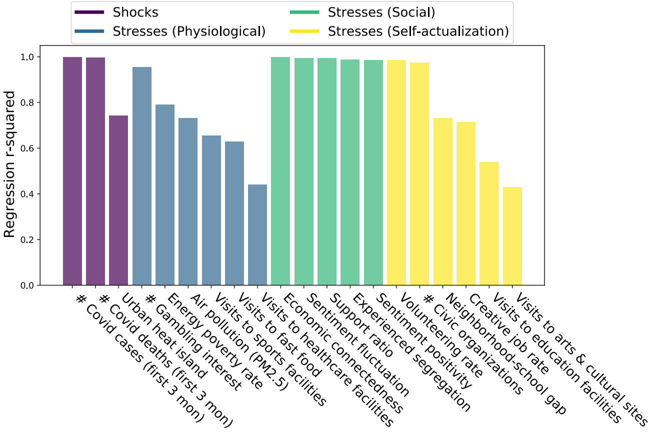

Figure 6: Prediction performance measured by R^2.

To gain insights into the reasons behind our prediction results, we visualize in Figure 7 the top-3 important features for each vulnerability outcome within the *self-actualization stresses* category, using it as a case study. We make the following three observations: First, the feature importance results expose more intricate interactions among social factors that trap certain populations inside vulnerability. Notably, *creative job rate* is strongly influenced by *bachelor rate*, underscoring that better-educated populations have a distinct advantage in securing jobs associated with exciting innovation processes. However, *median household income* and *non-Hispanic white rate* follow as the second and third most important feature, highlighting the non-negligible impact of poverty and racial-ethnic minority backgrounds on employment decisions. Second, the crucial disadvantage features influencing various vulnerability outcomes can exhibit considerable variation. For instance, *visitation to education facilities* is predominantly influenced by *per-capita social connections*, indicating that education resources tend to favor populations with robust social capital and thus are more informative through social network interactions. In contrast, *visitation to arts and cultural sites* is more influenced by *bachelor rate* than social connections. This discrepancy may be attributed to the idea that appreciating arts and specific cultures requires intentional aesthetic training and exposure, elements not easily transferrable through social network connections. Third, vulnerability outcomes with closer conceptual connections tend to be influenced by similar disadvantage features. Specifically, *volunteering rate* and *# civic organizations* describe the extent of civic life participation, thus relating to each other in a closed way. As a result, the top-3 important features for predicting both outcome variables are identical, albeit with different orders.

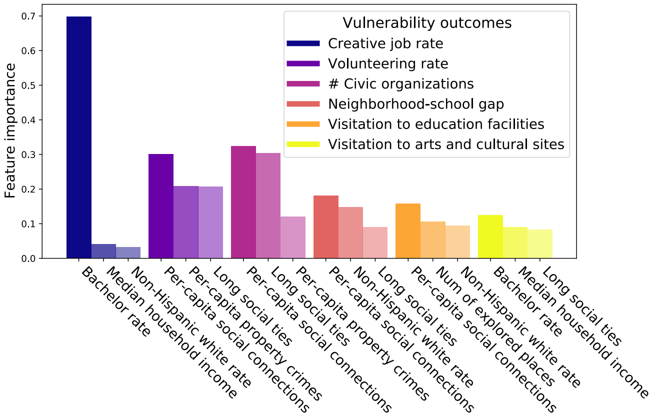

Figure 7: Top-3 important features for predicting various vulnerability outcomes.

3.5 Temporal Dynamics

As our dataset records feature dynamics with time passing by, we also analyze the temporal dynamics of the disadvantage indices. Here we take the offline mobility patterns across different periods of the COVID-19 pandemic as an example, as human mobility can change fast and correspond well to impacts brought by urban challenges, especially urban shocks. As shown in Figure 8, the correlation between *demographic disadvantage index* and *mobility level* is -0.2062 in Year 2019 before the pandemic, indicating that people with better socio-economic statuses move around more freely and actively in urban space, potentially getting in touch with more opportunities. However, in 2020 after the pandemic began, the magnitude of the correlation was significantly reduced to -0.1461. This corresponds to the fact that during the COVID-19, people with better socioeconomic statuses are more capable of cutting their mobility to conform to the ”stay-at-home” policies, while disadvantaged neighborhoods generally have more essential workers who had to stick to offline work [Weill *et al.*, 2020; Jay *et al.*, 2020]. Moreover, we observe a recovery of the correlation to -0.1542 in 2021 when the pandemic was gradually under control with vaccines and targeted drugs. The correlation between *demographic disadvantage index* and *num of explored places* sees a similar “reduce-rebound” dynamic.

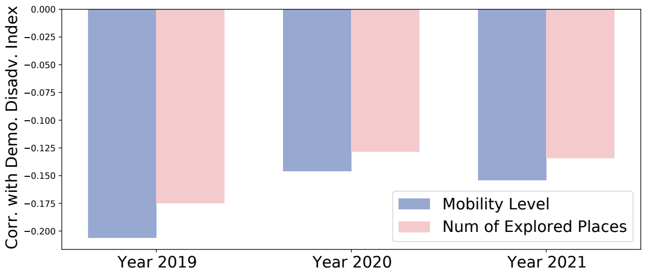

Figure 8: Temporal change in correlation between *demographic disadvantage index* and mobility outcomes.

4 Related Work

The interconnections between disadvantaged dimensions and vulnerability outcomes are highly complex, remaining an intriguing and inviting problem for the urban social network research communities. Some of the links have been studied in previous works, resulting in interesting findings. For example, [Williams *et al.*, 2020] reveals that neighborhoods with lower income and more racial/ethnic minorities not only enjoy less green space, but also suffer from more criminal threats. In the context of COVID-19, [Coleman *et al.*, 2022] finds that low-income and non-white neighborhoods faced larger risks of COVID-19 exposure. However, most of such studies focused their scope on one or several specific pairs of links, compromising on studying the bigger picture.

There are also some indices designed for summarizing certain dimensions of inequalities faced by urban populations. The Atlas of Inequality made by MIT Media Lab[4] maps the economic inequality encoded in people’s everyday movements onto communities. However, it does not proceed beyond the potential to study the concrete vulnerability outcomes experienced by communities. In contrast, our dataset provides both disadvantage dimensions and vulnerability outcomes, thus offering opportunities to study the processes of transformation in-depth. The Disaster Risk Index (DRI) proposed by [Peduzzi *et al.*, 2009] mainly reflects the risks of facing natural disasters such as droughts and earthquakes, and is only provided at the national level. The Social Vulnerability Index (SVI) developed by the US CDC and ATSDR[5] combines a set of demographic characteristics to evaluate the potential negative impacts on communities brought by either natural or human-caused disasters. However, it primarily concentrates on assessing the impact on human physical health, thereby overlooking other crucial aspects of human life. In contrast to these works, our framework goes beyond the scope of the SVI. Not only does it encompass both long-term stresses and short-term shocks experienced by communities, but it also takes a nuanced approach by taxonomizing stresses according to Maslow’s hierarchy of needs, providing a comprehensive landscape of vulnerabilities.

To summarize, given the absence of frameworks supporting comprehensive analyses of the disadvantage-vulnerability relationships, we collect multi-sourced data that reflects demographic disadvantages, neighborhood disadvantages, mobility disadvantages, and social disadvantages, covering most of the vital aspects of urban life. We also collect data regarding various types of urban shocks and stresses to measure vulnerability to urban challenges from a multi-dimensional perspective. More importantly, we organize these variables into an analytic framework, and validate its utility for AI research with well-performed machine learning models.

5 Discussion

We believe that our proposed framework, *VulnerabilityMap*, is promising in bringing several new opportunities for the web and urban research community. The most direct use is to analyze the relationship between dimensions of disadvantages and dimensions of vulnerabilities to various urban challenges. Based on these understandings, our framework can be further used to inform the design or generation of intervention strategies. Moreover, with the temporal dimension, our framework can support analyses of Granger causality and prediction/simulation tasks. Overall, our framework provides a nuanced understanding of the challenges faced by urban communities, serving as a foundation for targeted interventions and informed policy-making.

Although we made our best effort in constructing the framework, *VulnerabilityMap* is not without limitations. First, the varying updating frequencies of raw data sources introduce potential inconsistencies in estimation. Despite this, our experimental results establish a reliable lower bound for discovering inequality in vulnerability outcomes and their association with various disadvantage factors. Second, the complex web of entities and relationships in urban spaces raises the possibility of overlooking contributive factors to vulnerability outcomes. Third, our current approach to obtaining disadvantage indices involves a straightforward summation of normalized individual features, potentially disregarding more sophisticated methods like calibrated weighted sums. Thus, researchers using our framework should acknowledge these limitations to make grounded hypotheses and reasonable claims about the derived research findings.

For future work, we will continue improving the comprehensiveness of *VulnerabilityMap* to cover more potential determinants and vulnerabilities. We will also keep track of various indices along the temporal dimension, and develop a web interface to facilitate public use and citizen communication.

[4] https://inequality.media.mit.edu/

[5] https://www.atsdr.cdc.gov/placeandhealth/svi/index.html

References

[Abedi *et al.*, 2021] Vida Abedi, Oluwaseyi Olulana, Venkatesh Avula, Durgesh Chaudhary, Ayesha Khan, Shima Shahjouei, Jiang Li, and Ramin Zand. Racial, economic, and health inequality and covid-19 infection in the united states. *Journal of racial and ethnic health disparities*, 8:732–742, 2021.

[Araujo *et al.*, 2017] Matheus Araujo, Yelena Mejova, Ingmar Weber, and Fabricio Benevenuto. Using facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. In *Proceedings of the 2017 ACM on Web science conference*, pages 253–257, 2017.

[Bettencourt *et al.*, 2007] Luís MA Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences*, 104(17):7301–7306, 2007.

[Cesare *et al.*, 2018] Nina Cesare, Hedwig Lee, Tyler McCormick, Emma Spiro, and Emilio Zagheni. Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5):1979–1999, 2018.

[Chai *et al.*, 2023] Yuchen Chai, Devika Kakkar, Juan Palacios, and Siqi Zheng. Twitter sentiment geographical index dataset. *Scientific Data*, 10(1):684, 2023.

[Chen *et al.*, 2022] Lin Chen, Fengli Xu, Zhenyu Han, Kun Tang, Pan Hui, James Evans, and Yong Li. Strategic covid-19 vaccine distribution can simultaneously elevate social utility and equity. *Nature Human Behaviour*, 6(11):1503–1514, 2022.

[Chen *et al.*, 2023] Lin Chen, Fengli Xu, Qianyue Hao, Pan Hui, and Yong Li. Getting back on track: Understanding covid-19 impact on urban mobility and segregation with location service data. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 126–136, 2023.

[Coleman *et al.*, 2022] Natalie Coleman, Xinyu Gao, Jared DeLeon, and Ali Mostafavi. Human activity and mobility data reveal disparities in exposure risk reduction indicators among socially vulnerable populations during covid-19 for five us metropolitan cities. *Scientific Reports*, 12(1):15814, 2022.

[Credit *et al.*, 2021] Kevin Credit, Gustavo Dias, and Brenda Li. Exploring neighbourhood-level mobility inequity in chicago using dynamic transportation mode choice profiles. *Transportation research interdisciplinary perspectives*, 12:100489, 2021.

[Fan *et al.*, 2022] Chao Fan, Xiangqi Jiang, Ronald Lee, and Ali Mostafavi. Equality of access and resilience in urban population-facility networks. *npj Urban Sustainability*, 2(1):9, 2022.

[Fatehkia *et al.*, 2020] Masoomali Fatehkia, Isabelle Tingzon, Ardie Orden, Stephanie Sy, Vedran Sekara, Manuel Garcia-Herranz, and Ingmar Weber. Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):22, 2020.

[Finlay *et al.*, 2022] Jessica Finlay, Robert Melendez, Michael Esposito, Anam Khan, Mao Li, Iris Gomez-Lopez, and Megan Chenoweth. National neighborhood data archive (nanda): Polluting sites by census tract, united states, 2000-2018. https://doi.org/10.3886/E159961V1, 2022.

[Gomez-Lopez *et al.*, 2021] Iris Gomez-Lopez, Min Hee Kim, Mao Li, Dominique Sylvers, Michael Esposito, Philippa Clarke, and Megan Chenoweth. National neighborhood data archive (nanda): Neighborhood-school gap by census tract, united states, 2009-2010 and 2015-2016. https://doi.org/10.3886/E156043V1, 2021.

[Hale *et al.*, 2021] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar, et al. A global panel database of pandemic policies (oxford covid-19 government response tracker). *Nature human behaviour*, 5(4):529–538, 2021.

[Hsu *et al.*, 2021] Angel Hsu, Glenn Sheriff, Tirthankar Chakraborty, and Diego Manya. Disproportionate exposure to urban heat island intensity across major us cities. *Nature Communications*, 12(1):2721, 2021.

[Jay *et al.*, 2020] Jonathan Jay, Jacob Bor, Elaine O Nsoesie, Sarah K Lipson, David K Jones, Sandro Galea, and Julia Raifman. Neighbourhood income and physical distancing during the covid-19 pandemic in the united states. *Nature human behaviour*, 4(12):1294–1302, 2020.

[Kearney and Levine, 2014] Melissa S Kearney and Phillip B Levine. Income inequality, social mobility, and the decision to drop out of high school. Technical report, National Bureau of Economic Research, 2014.

[Leitner *et al.*, 2018] Helga Leitner, Eric Sheppard, Sophie Webber, and Emma Colven. Globalizing urban resilience. *Urban Geography*, 39(8):1276–1284, 2018.

[Levy *et al.*, 2022] Brian L. Levy, Karl Vachuska, S. V. Subramanian, and Robert J. Sampson. Neighborhood socioeconomic inequality based on everyday mobility predicts covid-19 infection in san francisco, seattle, and wisconsin. *Science Advances*, 8(7):eabl3825, 2022.

[Moro *et al.*, 2021] Esteban Moro, Dan Calacci, Xiaowen Dong, and Alex Pentland. Mobility patterns are associated with experienced income segregation in large us cities. *Nature communications*, 12(1):4633, 2021.

[Nijman and Wei, 2020] Jan Nijman and Yehua Dennis Wei. Urban inequalities in the 21st century economy. *Applied Geography*, 117:102188, 2020.

[Peduzzi *et al.*, 2009] Pascal Peduzzi, Hy Dao, Christian Herold, and Frederic Mouton. Assessing global exposure and vulnerability towards natural hazards: the disaster risk index. *Natural hazards and earth system sciences*, 9(4):1149–1159, 2009.

[The New York Times, 2021] The New York Times. Coronavirus (covid-19) data in the united states. https://github.com/nytimes/covid-19-data, 2021. Retrieved on Nov 7, 2023.

[Wang *et al.*, 2018] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences*, 115(30):7735–7740, 2018.

[Weill *et al.*, 2020] Joakim A Weill, Matthieu Stigler, Olivier Deschenes, and Michael R Springborn. Social distancing responses to covid-19 emergency declarations strongly differentiated by income. *Proceedings of the national academy of sciences*, 117(33):19658–19660, 2020.

[Williams *et al.*, 2020] Tim G Williams, Tom M Logan, Connie T Zuo, Kevin D Liberman, and Seth D Guikema. Parks and safety: a comparative study of green space access and inequity in five us cities. *Landscape and urban planning*, 201:103841, 2020.