Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Pandemic vulnerability index of US cities: A hybrid knowledge-based and data-driven approach

Md. Shahnoor Rahman a,*, Kamal Chandra Paul b, Md. Mokhlesur Rahman c,h, Jim Samuel d, Jean-Claude Thill e, f, Md. Amjad Hossain g, h, Kamal Chandra Paul i, G. G. Md. Nawaz Ali j

a Department of Earth and Environmental Sciences, New Jersey City University, Jersey City, NJ, 07305, USA
b Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC, 28223, USA
c The William States Lee College of Engineering, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC, 28223, USA
d E.J. Bloustein School of Planning & Public Policy, Rutgers University, NJ, 08901, USA
e Department of Geography and Earth Sciences, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC, 28223, USA
f Department of Accounting, Information Systems, and Finance, Emporia State University, Emporia, KS, 66801, USA
g Department of Computer Science and Information Systems, Bradley University, Peoria, IL, 61625, USA
h Department of Urban and Regional Planning, Khulna University of Engineering & Technology (KUET), Khulna, Khulna, 9203, Bangladesh
i School of Data Science, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC, 28223, USA

A B S T R A C T

Cities become mission-critical zones during pandemics and it is vital to develop a better understanding of the factors that are associated with infection levels. The COVID-19 pandemic has impacted many cities severely; however, there is significant variance in its impact across cities. Pandemic infection levels are associated with inherent features of cities (e.g., population size, density, mobility patterns, socioeconomic condition, and health & environment), which need to be better understood. Intuitively, the infection levels are expected to be higher in big urban agglomerations, but the measurable influence of a specific urban feature is unclear. The present study examines 41 variables and their potential influence on the incidence of COVID-19 infection cases. The study uses a multi-method approach to study the influence of variables, classified as demographic, socioeconomic, mobility and connectivity, urban form and density, and health and environment dimensions. This study develops an index dubbed the pandemic vulnerability index at city level (PVI-CI) for classifying the pandemic vulnerability levels of cities, grouping them into five vulnerability classes, from very high to very low. Furthermore, clustering and outlier analysis provides insights on the spatial clustering of cities with high and low vulnerability scores. This study provides strategic insights into levels of influence of key variables upon the spread of infections, along with an objective ranking for the vulnerability of cities. Thus, it provides critical wisdom needed for urban healthcare policy and resource management. The calculation method for the pandemic vulnerability index and the associated analytical process present a blueprint for the development of similar indices for cities in other countries, leading to a better understanding and improved pandemic management for urban areas, and more resilient planning for future pandemics in cities across the world.

A R T I C L E I N F O

Keywords:
Coronavirus
COVID-19
Pandemic vulnerability index
PVI
US cities
Resiliency

1. Introduction

The highly transmittable coronavirus has created a public health crisis that has disrupted many spheres of society across the globe, including health infrastructure, business, economy, education, transportation, and the emotional well-being of people (Ali et al., 2021; Codagnone et al., 2020; Salama, 2020; Samuel, Ali, Rahman, Esawi, & Samuel, 2020; Samuel et al., 2020). The impacts and the full consequences of the highly contagious COVID-19 infection are still evolving and presently unclear. While it is established that more than 250 million (Worldometers, 2021) people have been infected across the globe, the factors behind the rapid spread of the COVID-19 infection in urban settings have still not been addressed satisfactorily. At the beginning of this pandemic, many researchers hypothesized that climate and weather could be mitigating factors for the spread of COVID-19 (Bashir et al., 2020; Jamil, Alam, Gojobori, & Duarte, 2020; Rahman et al., 2021), but infections in every corner of the world, irrespective of climate and weather, proved the weakness of these and similar conjectures (Notari, 2021). It has also become part of the dominant.

* Corresponding author.
E-mail address: mrahman1@njcu.edu (Md. Shahnoor Rahman).

https://doi.org/10.1016/j.scs.2023.104570
Received 9 August 2022; Received in revised form 1 April 2023; Accepted 1 April 2023
Available online 11 April 2023
2210-6707/© 2023 Elsevier Ltd. All rights reserved.
narrative that global mega-cities and other metropolitan regions are the zones most severely affected by this pandemic (Shekhar et al., 2022; Stier, Berman, & Bettencourt, 2020). It is commonly understood that cities are the primary hot-spots for COVID-19 infection due to the higher concentration of population and economic activities. Cities in the United States (US) are no exception to the global trend (Patel, Cancel, Chatterjee, & Rahman, 2021; Rahman, Thill, & Paul, 2020; Smith & Judd, 2020). They are said to have been highly vulnerable to this pandemic because of their urban vibrancy, including high connectivity, density, socioeconomic makeup, high mobility, and diverse activities with high levels of direct human to human interactions such as personal and cultural services, and healthcare (Cordero & Castro, 2020; DeWeese et al., 2020; Fortaleza, Guimarães, de Almeida, Pronunciate, & Ferreira, 2020; Hanzl, 2020; Kashem, Baker, González, & Lee, 2021; Nikparvar, Rahman, Hatami, Thill, et al., 2021; Rahman et al., 2021; Sharifi & Khavarian-Garmsi, 2020). Since urban regions house more than 80% of the US population (US Census Bureau, 2021) and serve as centers of economic activities, it is important to understand the scale of their vulnerability and their underlying exposure to the pandemic.

Furthermore, a clearer understanding of the vulnerability of cities will help us rethink and reshape post-pandemic urban planning strategies and policies and craft robust policies shielding against future global pandemics. It is observed that infection rates are not uniform across cities. While the limiting factors of the spread of COVID-19 in urban areas continue to defy actionable comprehension, many studies have found relationships between the escalation of coronavirus infections and inherent features of cities (e.g., population size, density, mobility pattern, socioeconomic conditions, health infrastructure, and others). Overall, the infection rate is high in larger urban agglomerations, yet the influence of various urban features remains unsettled. Some researchers have found socioeconomic and demographic indicators to be the pivotal drivers of the pandemic, while others have pointed to urban forms (e.g., employment-household entropy, city size, land-use pattern), mobility patterns as well as transportation systems as significant factors of disease transmission. Fortaleza et al. explored the spatial and demographic indicators that affect the vulnerability of populations to the COVID-19 pandemic in 604 municipalities in Brazil (Fortaleza et al., 2020). They found that the metropolitan land area, population density, how populous a place is, its Human Development Index (HDI), and its Gini index (i.e., income inequality) have a positive association with COVID-19 prevalence (Fortaleza et al., 2020). Similarly, another study (Campos, Aratani, Cabral, Limongi, & Oliveira, 2021) from Brazil assessed vulnerabilities of 853 cities and concluded that high urbanization and population density, inadequate sanitation, high illiteracy, high Gini index, low HDI and household income, weak healthcare infrastructure, and a high number of people with underlying health conditions increased urban vulnerabilities to COVID-19 infection. Pourghasemi et al. concluded that distance from roads, mosques, hospitals and density of cities are the main drivers of the diffusion of COVID-19 in Iran (Pourghasemi et al., 2020). A study in Uganda found that vulnerability is highest in the areas with shopping malls, transport hubs, and transactional offices (Bamweyana et al., 2020). Urban areas with more low-income households, weak healthcare infrastructure, poor socioeconomic conditions, and underprivileged communities (e.g., lack of electricity, education, water etc.) are more afflicted by the COVID-19 pandemic (Coelho et al., 2020; Coven & Gupta, 2020; Imdad et al., 2020). Furthermore, a study across 276 Chinese cities observed that urban governance capacity, as measured by the city comprehensive development index, which is based on indicators of environment, society, and economy, is an important factor to control the COVID-19 pandemic, particularly in the later stages (Chu, Cheng, & Song, 2021). The study reported a 2.4% increase in COVID-19 case recovery for each 1% increase in the capacity of urban governance. US-based studies also indicate that COVID-19 susceptibility is directly related to underlying conditions of societies including socioeconomic vulnerability, deprivation index (i.e., more disadvantaged neighborhoods: racial minorities, economically deprived strata), percentage of the elderly population, health conditions, educational attainment, income level, and race (Aliiberti et al., 2021; Hatem, Cheng, Kitchen, Weiner, & Kharrazi, 2020; Karaye & Horney, 2020; Rhodes et al., 2021).

Vulnerability assessment has been commonly practiced in various contexts, particularly in relation to extreme events. When urban assessment is contemplated, vulnerability is the condition determined by physical, social, economic, and environmental factors or processes that increase the susceptibility of a community to the impacts of hazards (UN-ISDR, 2004). As a result, vulnerability is viewed as a multidimensional, differential, and dynamic phenomenon that encompasses physical, social, economic, environmental, and institutional features (Birkmann, 2006). The assessment of vulnerability to the spread of COVID-19 is important for examining the preconditions and context of cities to reduce health risks and enhance the resilience of humans and societies. The vulnerability score of a community can be determined using its various underlying characteristics as well as the impacts of past events. Both heuristic and data-driven approaches have been used widely to assess the vulnerability of communities (Amraim et al., 2020; Chan, Huang, Lam, Wong, & Zou, 2019; Ho et al., 2017; Marvel et al., 2021; Papathom-Köhle, Schlögl, & Fuchs, 2019). For instance, Ho et al. (2017) calculated the socio-environmental vulnerability to geriatric depression in Hong Kong using odds ratios estimated by a data-driven binomial logistic regression. Chan et al. (2019) used a two-stage dimension reduction statistical method to develop a health vulnerability index from nine indicators in the regions of China’s Belt and Road Initiative. Like for natural hazards, data-driven approaches have also been used to assess vulnerabilities of cities to the COVID-19 pandemic. Sangiorgio and Parisi (2020) assessed the vulnerability of 257 urban districts based on the relative importance of 56 variables, where they relied on a Generalized Reduced Gradient method for model optimization to produce a metric of COVID-19 risk. Snyder and Parks (2020) assessed county-level socio-ecological vulnerability by aggregating the percentile ranks of 18 indicators. Hamidi, Sabouri, and Ewing (2020) utilized regression-based structural equation modeling where the natural log of the COVID-19 infection rate serves as the outcome variable and urban form, health facilities, and socioeconomic characteristics as predictors of COVID-19 risk in 913 US metropolitan counties. Sarkar and Chouhan (2021) identified four fundamental dimensions of district level vulnerability in India by applying principal component analysis on 16 indicators. Various machine learning methods have been applied in a wide range of human response estimation studies (Rahman, Paul et al., 2021; Samwel, 2017). Many data-driven vulnerability assessment studies have utilized machine learning approaches and have statistically generated weights to estimate the importance of various indicators. Weights of different indicators of community vulnerability can also be assigned by experts. For instance, a recent study (Marvel et al., 2021) proceeded this way to develop a pandemic vulnerability index (PVI) of counties in the contiguous US states. Similarly, Smittenaar et al. (2021) developed a COVID-19 community vulnerability index based on relative weights of 40 indicators elicited from subject-matter experts. The Analytic Hierarchy Process (AHP) heuristic was used to demarcate vulnerability levels of urban districts in India (Mishra, Gayen, & Haque, 2020). Since the importance of underlying factors of cities to COVID-19 vulnerability is largely unknown and expertise on this subject matter is limited, we argue that a data-driven approach is advantageous for developing a vulnerability index of US cities. This study analyzes an extensive set of variables pertaining to demographic and socioeconomic features, urban form, mobility and connectivity, and health and environmental indicators to understand the compound risk factors inherent to different cities. We leverage a hybrid knowledge-based and data-driven approach to comprehend the associations of complex and interlinked relations among various indicators of urban areas to COVID-19 cases. Although many previous studies have focused
on specific US states and vulnerability indices have been developed for counties, as discussed above, cities are the places on the front line of the COVID-19 pandemic. Therefore, this study aims to develop a PVI at the city level – hereafter referred to as PVI-CI – by investigating 41 influential variables over 3069 cities in the contiguous US. The outcomes of this research will help us to understand the most influential variables driving the pandemic in urbanized areas. The vulnerability index can also be used to evaluate risks by studying the combination of hazard and exposure components. The findings will be helpful for city-specific policy development and implementation. They may also reveal the role of sustainability measures towards developing resilient cities. The lessons learned from this research may contribute to post-pandemic city planning.

2. Methods

This research is based on publicly available data only; therefore, neither human subjects nor animals are used as part of the design. All analyses were performed in accordance with relevant guidelines and regulations. The methodological steps of data collection and cleaning, weight generation, and vulnerability index calculation are illustrated in Fig. 1. The study follows a hybrid approach where weights are determined by a data-driven process and the association of COVID-19 cases with co-variates are determined by an extensive literature review and by expert judgements. At first, based on an extensive review of the literature, a set of 41 important characteristics (factors) of cities are identified. The detailed scope and rationale underlying these factors are discussed in the following subsection. The relative importance of vulnerability factors is then determined through three widely used feature ranking algorithms. Vulnerability scores of cities (PVI-CI) are calculated next on the basis of the weights (importance) of these factors. We end this section with a comparison of the results obtained with each of the three feature ranking algorithms.

2.1. Data scoping and variable selection

This study addresses fundamental questions of what conditions have lead cities to take the brunt of the COVID-19 pandemic and what drives the variability in their vulnerability to the pandemic across the contiguous US. In this study context, there are over 3000 cities that vary in size, density, socioeconomic makeup, air quality, healthcare infrastructure, and mobility and connectivity patterns. On this basis, this study has collected 41 variables in five dimensions for 3069 cities to develop a synthetic index of vulnerability to the pandemic. Table 7 in Appendix shows the descriptive statistics of 41 selected variables. The choice of these variables is primarily based on the findings of recent literature related to the COVID-19 infection rate and its association with city environment characteristics. Data on the outcome variable were collected on June 30, 2021, from the USA Facts web portal (US COVID-19 cases and deaths by state, 2021). Data on COVID-19 cases are determined by a data-driven process and the association of COVID-19 with no work or residency permits have limited access to healthcare services. If residential locations and workplaces are located in the same area, people have a higher propensity to walk or use bicycles to their workplaces, whereas having residence and workplace in different areas increases overall commuting distance (Ding, Wang, Liu, Zhang, & Yang, 2017; Manaugh, Miranda-Moreno, & El-Geneidy, 2010). Thus, we assumed that people who live in cities have shorter commuting distance to work compared to those who live outside of the cities but work within the city limits. Therefore, it is anticipated that the percentage of the working population working within the city and outside of the city is related to COVID-19 spreading because of the fact that people who live outside of the city area need more commuting by public transport and have a higher probability of being infected. Data for all demographic variables are sourced from the urban area census databases of the US Census Bureau. All demographic variables are 5-year estimated averages from the American Community Survey (ACS) for 2019, except for the non-citizen population data, which are drawn from the ACS’s 1-year estimates of 2019.

2.1.1. Demographic variables

Many studies have reported that demographic makeup is one of the prime factors of COVID-19 outbreaks in many locations (Azad & Ghandehari, 2021; Coelho et al., 2020; Coven & Gupta, 2020; Imdad et al., 2020; Kashem et al., 2021; Xu et al., 2021). A total of 10 variables are selected to apprehend the demographic characteristics of cities (Table 1). As a scale factor, the total population is an important parameter that describes city size, which has a strong influence on the intensity of contacts and interactions in the population, on resource allocation, on delivering healthcare facilities to a greater community and on a host of other considerations. The assumption is that the larger the population of a city, the higher the chance of COVID-19 spreading. Older (age > 65 years) and younger populations (age < 15 years) are identified as vulnerable age groups because of their underlying health condition (comorbidity) and exclusion from the vaccine scheme, respectively. Hence the shares of these age groups in the general city population are taken as potentially influential factors.

Group quarter living, average household size, and family size are proxies that represent the congested living condition of a city. Both the minority and non-citizen populations are vulnerable because of their limited access to health insurance coverage, formal jobs, and poor economic and living conditions (Nazroo & Williams, 2006; Tarraf, Vega, & González, 2014). Many undocumented people (i.e., illegal migrants) with no work or residency permits have limited access to healthcare services (Lee & Choi, 2009; Philipsborn, Sorscher, Sexton, & Evans, 2021; Seo, 2011). Additionally, they are prone to coronavirus infection due to their involvement in jobs that require direct contact with other people such as delivery persons, and retail or hospitality workers. Despite more inclusive healthcare policies and moral engagements to fundamental human rights irrespective of citizenship and residency status in some local communities, many governments and non-governmental organizations have had limited resources to support migrants and undocumented people during this global pandemic. Thus, considering this state of affairs, it is assumed that non-citizen populations are vulnerable to the pandemic compared to the citizen population due to barriers to access to healthcare services. If residential locations and workplaces are located in the same area, people have a higher propensity to walk or use bicycles to their workplaces, whereas having residence and workplace in different areas increases overall commuting distance (Ding, Wang, Liu, Zhang, & Yang, 2017; Manaugh, Miranda-Moreno, & El-Geneidy, 2010). Thus, we assumed that people who live in cities have shorter commuting distance to work compared to those who live outside of the cities but work within the city limits. Therefore, it is anticipated that the percentage of the working population working within the city and outside of the city is related to COVID-19 spreading because of the fact that people who live outside of the city area need more commuting by public transport and have a high probability of being infected. Data for all demographic variables are sourced from the urban area census databases of the US Census Bureau. All demographic variables are 5-year estimated averages from the American Community Survey (ACS) for 2019, except for the non-citizen population data, which are drawn from the ACS’s 1-year estimates of 2019.

2.1.2. Socioeconomic variables

Fragile socioeconomic conditions render people vulnerable to extreme conditions, and therefore also expose them more directly to the
COVID-19 pandemic. As presented in Table 2, a total of 10 social and economic variables are selected in this theme based on their relation to community vulnerability in the extant literature (Campos et al., 2021; Coven & Gupta, 2020; Fortaleza et al., 2020; Hossain, Chakma, Tasnim, & Zahra, 2021; Kashem et al., 2021; Snyder & Parks, 2020). In the US, people’s poverty level is measured based on household or family income. However, as a complement to this measure, we also account for population working for a low wage as the former encompasses individuals who are not working. Therefore, both low wage and poverty are considered important variables. Additionally, the inclusion of the low-wage workers also accounts for their working environment and low socioeconomic status. Population in weak financial condition is vulnerable to any extreme events because of their exposure to many informal service jobs. In addition, they cannot shelter at home during lockdown because they have been treated as essential workers and they may also not have enough financial backup to support their family during the pandemic. Finally, education level is also an important factor since educated people are more conscious of health guidelines from public health officials. The data on all 10 socioeconomic variables are sourced from the US Census Bureau; they are 5-year estimated averages by ACS in 2019.

2.1.3. Urban form and density variables

Many recent studies have found a relationship between the spread of COVID-19 and variables of urban form & density (Campos et al., 2021; Fortaleza et al., 2020; Pourghasemi et al., 2020; Xu et al., 2021). This study considers 10 variables to apprehend the important characteristics inherited in cities that may make them hot spots for the COVID-19 infection (Table 3). Some variables require further elaboration on how they are constructed. Employment and housing entropy expresses the evenness of the distribution across the types of employment and housing without consideration of the aggregated quantity of jobs or housing (Chapman et al., 2021). Employment density, residential density, and employment and housing entropy are extracted from the Smart Location Database (SLD) of the US Environmental Protection Agency (EPA) (Environmental Protection Agency, 2020). The values of these density indicators are calculated at the city level by taking the mean of block-group values. The total built-up area of each city is extracted from the GAP/LANDFIRE National Terrestrial Ecosystems database by aggregating developed & urban land cover classes (McKerrow, Davidson, Earnhardt, & Benson, 2014); the total built-up area is divided by the total city area to get the built-up ratio.

2.1.4. Mobility and connectivity variables

As indicated in Table 4, a total of six variables are taken into account to assess the mobility and connectivity of cities. Walkability is a component of the built environment (Chapman et al., 2021). The city-level index is extracted from the national index by taking the mean value of all pixels within the city boundary. Airport locations are derived from the US airports base layer available on the Mapcruzin web portal (MapCruzin, 2021). The distance from each city to the nearest large metropolitan (population 1M+) is measured by Euclidean distance between city centers.

2.1.5. Health and environment variables

Health care infrastructure is an important consideration to assess COVID-19 vulnerability because many events and actions are associated with health care infrastructure, including testing, vaccination, health awareness building, and treatments (Campos et al., 2021; Pourghasemi et al., 2020). Five variables in the health care infrastructure dimension are considered for this study (Table 5). People with underlying health conditions, especially respiratory illness (e.g., asthma), have a high risk of COVID-19 infection. Prolonged exposure to air pollution has direct impacts on public health, including asthma, diabetes, and high blood pressure. Many initial studies have shown that air quality may play a vital role in COVID-19 spreading as well as in the mortality rate (Bashir et al., 2020; Isphording & Pestel, 2021; Rahman, Paul et al., 2021). Therefore, this study considers the air quality of cities as an important

### Table 1

| Factor of pandemic vulnerability | Measurement | Conceptual relevance to pandemic vulnerability |
|---------------------------------|-------------|-----------------------------------------------|
| Total population                | Total population within the city limits | Total population is an indicator of city size. The larger the city, the greater the opportunities to spread the virus. |
| Size of younger population      | Percentage of total population that is under 15 years old | Young population is vulnerable to COVID-19 for two reasons: first, they are mostly not covered by vaccination program; second, they are socially active and less consciously aware of safety protocols. |
| Size of elderly population      | Percentage of total population over 65 years old | Elderly population is vulnerable to COVID-19 because of their compromised immune systems and co-morbidity conditions. |
| Group quarter living           | Percentage of total population living in a group living arrangement | People living in group quarter settings are highly exposed to this pandemic since many people live together in a single facility. Thus, the risk of community infection is increased. |
| Minority population            | Percentage of the minority population in total population | Minority population historically has limited access to education and healthcare services, and less secure income sources, which make them vulnerable to a spreading pandemic. |
| Non-citizen population         | Percentage of the non-citizen population in total population | Non-citizen population may lack health insurance and legal work status. They are often engaged in informal work, which increases their exposure to a viral infection and reduces their recourse to health care. |
| Population working in a designated principal city | Percentage of working population with job in a principal city | People working within the city may not travel by public transportation for a long duration. They have a lower risk of being infected while commuting. |
| Population working outside of a principal city | Percentage of working population whose job is outside of principal city | Population working outside the principal city may be at risk because of their longer exposure in public transport while commuting. |
| Household size                  | Average number of people in a household | Household size is an indicator of congested living where large households have raised exposure to COVID-19 infection. |
| Family size                     | Average number of people living in a family | Similar to household size, family size is also an indicator of congested living. Larger families have a larger contact network, increasing the risk of COVID-19 infection. |
variable for the COVID-19 vulnerability index. County-level annual air quality index (AQI) data for 2019 are collected from the Air Data Portal EPA since the year immediately before this pandemic (Environmental Protection Agency, 2021). The median annual air quality index value of the nearest available county is assigned to each city.

### Table 2: Description of socioeconomic variables.

| Factor of pandemic vulnerability | Measurement                                                                 | Conceptual relevance to pandemic vulnerability |
|----------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------|
| Mobile housing                   | Percentage of households in mobile homes                                   | People living in mobile housing may face greater economic precarity, which engages them more in informal and essential work activities, where contacts are more common and risk of infection higher. |
| Working population               | Total number of working people                                             | Working population requires to go outside the home daily in the business-as-usual situation. |
| Income inequality                | Gini index of a city                                                        | A higher Gini coefficient indicates social disparity, where disadvantaged people may not have adequate access to formal income and avail health care. |
| Congested living                 | Percentage of households with 4 or more members                             | Four or more people in a household indicates a congested living condition, which increases contact across a larger network and raises the risk of COVID-19 infection. |
| Per capita income                | Per capita income of workers in a city                                     | Low-income people are most engaged in work with a larger number of person-to-person contacts and have high social vulnerability in terms of access to formal work and healthcare services. |
| Median income                    | Median household income of a city                                           | Median income of a city indicates financial strength of the city. High median income means city dwellers face lower risk exposure and have higher ability to avoid dire consequences. |
| Poverty level                    | Percent of the population below poverty level                               | People below the poverty level are at high risk of pandemic exposure and lower ability to avoid dire consequences. |
| Unemployment rate                | Percentage of population 16 years and above with no employment             | Unemployed people are often engaged in informal works. hence higher risk exposure and more dire consequences. |
| Education level                  | Percentage of the population with high school education or above            | Educated people are usually health conscious, aware of COVID-19 related risks, experience less face-to-face contacts at work, make more use of formal health care, and follow health safety protocols. |
| Low wage population              | Percentage of the working population paid low wages                        | Similar to people below the poverty line, low-wage workers are more vulnerable to the COVID-19 pandemic. |

### Table 3: Description of urban form and density variables.

| Factor of urban form and density variables | Measurement                                                                 | Conceptual relevance to pandemic vulnerability |
|-------------------------------------------|-----------------------------------------------------------------------------|-----------------------------------------------|
| City size (Area)                          | Total land area                                                             | Large cities may have a large population and more economic activities, which increase the risk of disease spreading. |
| Housing units                             | Total number of housing units of a city                                     | Similar to the number of residents, the total number of housing units is also an indicator of city size. Many recent studies have pointed that large cities are more exposed to the pandemic. |
| Population density                        | Number of population per unit area                                          | Population density is an indicator of congested living of a city. Higher population density reduces the scope for maintaining social distancing and vice versa. Thus, densely populated areas raise the chance of spreading respiratory diseases. |
| Built-up ratio                            | Ratio of built-up area to the total area of a city                         | This is also a density indicator. A high built-up ratio indicates the high density of a city. |
| Number of dependent cities                | Number of cities located within 50 km of a city                            | This number is an indicator of the spatial setting of a city, it if is part of a large urban agglomeration versus an isolated population settlement. |
| Population of dependent cities            | Sum of the total population of cities located within 50 km of a city        | Similar to the number of dependent cities, people of dependent cities pose an extra risk of disease spreading. |
| Residential density                       | Gross residential density (housing units per acre) on unprotected land      | Residential density is a density factor directly related to congested living conditions. High residential density lower possibilities for social distancing and is a recognized indicator of higher risk. |
| Employment density                        | Gross employment density (jobs per acre) on unprotected land                | High employment density indicates the social interaction of more workers in an area. |
| Retail employment density                 | Retail jobs per acre on unprotected land                                   | Retail shops are some of the hot spots of COVID-19 spreading due to face-to-face contacts. Thus, their high density indicates a high risk of virus spreading. |
| Diversity of land uses                    | Employment and housing entropy, which refers to the mix of employment types and occupied housing | Diversity of land uses indicates that people live mostly near their workplaces. High land-use mix indicates high walkability, which implies short are possibly more frequent trips. Thus, high entropy of the land use mix reduces the COVID-19 infection rate. |

#### 2.2. Feature ranking

A city-level PVI can be calculated based on a series of factors of the city that contribute to the incidence of COVID-19 cases. However, not all variables have equal influence on the spread of the virus.
To strengthen the reliability of the feature ranking process, three alternative filter-based methods are used, namely F-Classification, Mutual Information Regression (MIR), and Select K Best (SKB). Each of these methods generates ranking scores of each variable based on different statistical measures. The selected feature ranking methods are presented hereafter.

1. **F-Classification method:** This method selects important features among the variables in a dataset by computing the ANOVA F-value of the sample points (Pedregosa et al., 2011; Polat & Güneş, 2009). Features are often ranked according to their order of importance (Costa, Llobodanin, Castro, & Barbosa, 2018; Güneş, Polat, & Yosunkaya, 2010; Liang et al., 2020; Polat & Güneş, 2009). Here, we use the f-classic method of scikit-learn, a Python package, to perform the feature ranking. From the feature ranking, we derive corresponding weights for each feature that serve to determine PVI-CI scores.

2. **Mutual Information Regression (MIR) method:** In this method, the mutual dependence between the target and an independent variable is estimated based on the mutual information between them (Chen, Yeo, Francis, & Lau, 2016; Gao & Wu, 2020; Zheng & Wang, 2018). The MIR method incorporates non-parametric algorithms where the entropy is estimated from distances of k-nearest neighbors (Doquire & Verleysen, 2013; Kraskov, Stögbauer, & Grassberger, 2004; Ross, 2014). It generates a non-negative value. The larger this value is, the greater the dependency. For two totally independent variables, the value will be zero. We use the “mutual_info_regression” function of scikit-learn for the implementation (Pedregosa et al., 2011). The feature weights are calculated from the values of the different features in the ranking.

3. **Select K Best Method:** We also use the Select K Best function of scikit-learn to find the ranking of our selected variables based on the k highest scores (Ma et al., 2015; Pedregosa et al., 2011). Regression is the scoring function. The derived feature rankings are taken as the weights used towards calculating the city-level PVI, as explained in the following sub-section.

2.2. Developing the pandemic vulnerability index

This section describes the methodological steps for determining the Pandemic Vulnerability Index (PVI-CI). After collecting data from secondary sources, we first clean our dataset by removing the cities that have missing values for one or more variables. The final dataset includes 3069 cities to assess their vulnerability to the COVID-19 pandemic. Then, we normalize each variable to achieve the same range of values on all features. Since the variables are measured on diverse scales and units (e.g., percentage, absolute number, ratio), normalization brings them to a consistent format which is an essential data-preprocessing step for model building. All 41 variables are transformed to the scale between 0 to 1 by max-min normalization.

The feature ranking methods mentioned above are applied to determine the relative importance of selected variables. These measures
serve as the weights \( (w) \) of corresponding variables in computing the composite vulnerability score of cities. COVID-19 cases per 1000 population are the target variables (dependent variables) against all other parameters (independent variables). We repeat the following mathematical procedures with each method and determine the pandemic vulnerability index of cities.

The variable importance scores derived by feature ranking are normalized among the variables of each dimension to determine their corresponding weights \( (w) \). For instance, the importance scores of the 10 demographic variables are normalized using the maximum and minimum values of those variables. In order to find the relative score, cities are sorted either in ascending \((+)\) or descending \((-)\) order for each variable based on their positive or negative association with COVID-19 spreading (sign indicated in Table 6). Then, the percentile ranks of all cities are calculated for each variable based on the ascending or descending orders as indicated by the aforementioned figure.

The vulnerability scores of each dimension are computed as the sum of products of the percentile ranks and of corresponding weights of variables of this dimension, divided by the sum of weights. The vulnerability scores of dimensions are calculated by combining percentile ranks and weights derived from the data-driven process using Eq. (1). The vulnerability score for the demographic dimension \( (S_D) \) is computed as:

\[
S_D = \frac{\sum_{d \in D} d \times w_d}{\sum_{d \in D} w_d}
\]  

(1)

where \( D = \{P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, P_9, P_{10}\} \) and \( w_d \) denotes the weight of the corresponding variable \( d \in D \), which is calculated by each feature selection method. In a similar fashion, we find the scores for the variables under the socioeconomic dimension \( (S_S) \), the urban form and density dimension \( (S_U) \), the mobility and connectivity dimension \( (S_M) \), and finally the health infrastructure and environment dimension \( (S_H) \).

The weight of each of the five dimensions is the mean value of feature importance scores derived with the feature ranking methods using Eq. (2):

\[
W_{avg}(\mathcal{D}) = \frac{\sum_{d \in \mathcal{D}} w_d}{|\mathcal{D}|}
\]  

(2)

where \( \mathcal{D} \in \{D, S, U, M, H\} \). Here, \( D, S, U, M, H \) represent demographic, socioeconomic, urban form and density, mobility and connectivity, and health and environment dimensions, respectively. The average weight \( W_{avg}(\mathcal{D}) \) of each dimension is normalized using Eq. (3):

\[
W_{norm}(\mathcal{D}) = \frac{W_{avg}(\mathcal{D})}{\sum_{d \in \{D, S, U, M, H\}} W_{avg}(d)}
\]  

(3)

where \( W_{norm}(\mathcal{D}) \) denotes the normalized weight of the corresponding dimension. Finally, a city’s PVI-CI score is derived from the percentile ranking and normalized weights of different dimensions using Eq. (4):

\[
PVI - CI = \sum_{d \in \{D, S, U, M, H\}} (S_d \times W_{norm}(d))
\]  

(4)

where \( S_d \) denotes a dimensional score for that city.

The PVI-CI scores are visualized in Fig. 3 using a choropleth mapping technique available in the ArcGIS Pro software. The figure aims to depict the geographic pattern in the variation of pandemic vulnerability across the US over its statistical range from very low vulnerability to very low vulnerability. For benchmarking purposes, we further compare the scores derived in our study with the vulnerability scores recently devised by the National Institute of Environmental Health Sciences (NIEHS) (Marvel et al., 2021) for a 10% random sample of cities. Although NIEHS calculated their pandemic vulnerability index at the county level rather than at the city level, we argue this is a suitable validation approach to compare the PVI-CI scores with them by matching locations and gauge the level of consistency. It should be stressed that no other studies has attempted to estimate the vulnerability of cities under the spread of the COVID-19 pandemic. Thus, PVI-CI scores from the most effective of the three feature ranking methods are compared with the PVI scores provided by NIEHS.

After determining the vulnerability scores of all 3069 cities using the most effective method, a cluster and outlier analysis is performed with the Anselin Local Moran’s I statistic to identify the spatial clusters of cities with high and low vulnerabilities, respectively. Cluster types for each statistically significant city are determined based on Local Moran’s I value and z-score at the 95% confidence level. A positive Moran’s I statistic indicates a city has neighboring cities with similarly high or low vulnerability scores. In contrast, a negative Moran’s I indicates that a city is surrounded by cities with dissimilar vulnerability levels (Sánchez-Martín, Rengifo-Gallego, & Blas-Morato, 2019).

2.4. Comparison of feature ranking methods

It is not possible to make a strict validation of the outcomes of this research because of the unavailability of ground truth for validation. Therefore, this study provides an indicative soft validation by comparing the PVI-CI score with the county-level vulnerability score developed by the NIEHS (Marvel et al., 2021). Thus, technically, this process can be called a relative comparison with an available and established vulnerability index instead of a strict validation. The main goal of this comparison is to select the feature ranking method that performs relatively well to calculate vulnerability scores for all cities in the US. In addition, this comparison process can tell us how close the vulnerability score of a city is related to their corresponding county.

A total of 300 cities, about 10% of the 3069 cities, are randomly selected for validation purposes. The PVI scores of 300 counties corresponding to the sampled cities were collected from the PVI dashboard of the NIEHS. Although the PVI scores by the NIEHS are at the county level, this can be utilized as proxy ground truth for the city score since PVI scores at the city level are not available. We have used county-level PVI data to represent scores of cities. In most cases, the representative county lies within the city. Only a few cities in the sampled cases spread over more than a single county. In that case, the PVI score of a county consisting of the city center is taken into account. At first, the absolute differences between the PVI scores of the NIEHS and our individual proposed methods are calculated for each of the 300 cities. Then, the sum of the squared differences is calculated and divided by the number of sampled cities (300). Finally, the root mean square error (RMSE) of the PVI scores for the selected cities are calculated by taking the square root of the above value. This procedure is repeated for all three feature ranking methods (FC, MIR, and SKB). From the deviations of the PVI-CI scores with respect to the NIEHS method, we have also calculated the number of cities that lie within the deviation range of 0.1. The purpose of this calculation is to see how many cities have PVI-CI scores closer to NIEHS’s PVI scores. The deviation in vulnerability scores of the sampled cities is presented in Figs. 2 and 6. Both PVI-CI and NIEHS vulnerability scores are normalized using Z-score normalization to calculate the deviation between corresponding scores. In addition, a geo-visualization is given in Fig. 6 that depicts the deviation between NIEHS county-level PVI scores and PVI-CI scores computed with the FC method; for the simplicity of the result interpretation and discussion, deviations for other scoring methods are not reported. Fig. 2 depicts the deviation of PVI scores of sampled cities calculated by alternative feature ranking methods from the respective county’s scores determined by the NIEHS. The results show that the FC method yields the lowest deviation in scores. A total of 31.7%, 24.7%, and 29.2% of sampled cities have PVI score deviation below 0.05 with the FC, MIR, and SKB methods, respectively. Cumulatively, 62.7%, 44.7%, and 55.4% of sampled cities have PVI score deviation less than 0.10 with these three methods, respectively. In contrast, more than one-third of sampled cities have a PVI score deviation greater than 0.15 with the MIR method. Based on this evaluation, it can be concluded that both
FC method and SKB generates vulnerability scores of sampled cities that are the closest to the PVI scores from the NIEHS. In addition, we have also calculated the RMSE to select the most effective method based on how close the PVI-CI score is to the PVI of their corresponding counties. The RMSEs of the FC, MIR and SKB methods are 0.097, 0.155, and 0.111, respectively. As a result, based on both evaluation methods, the FC method can be considered as the most effective at determining the vulnerability scores for cities. The rest of the study will focus on results of this scoring method.

3. Results

3.1. Weights and scores

The influence of variables retained as indicative of the prevalence of COVID-19 infections in US cities is studied with variable weights derived with the F-Classification (FC) feature selection algorithm. All five vulnerability dimensions demographics, socioeconomic features, urban form & density, mobility & connectivity, and health & environment have scores in close range, namely between 0.17 and 0.22. As indicated in Table 6, demographics (0.22) is identified as the most influential dimension for the spread of COVID-19, followed by mobility & connectivity (0.21) and health & environment (0.21), socioeconomic (0.19), and finally urban form & density (0.17).

Among the 10 demographic variables, the size of the younger population (0.15) is identified as the most important factor for COVID-19 infection. Since younger population cohorts have a greater tendency to socialize and interact with others and to do so without abiding by COVID-19 safety protocols, and since much of the lower aged population was excluded from early-stage vaccine coverage, they quite likely contributed to the spread of the virus. Some other demographic variables such as household size (0.14), family size (0.14), the share of the non-citizen population (0.12), the population working outside of a principal city (0.11), the size of elderly population (0.11), and the share of the minority population (0.09) also stand out by their high contributions to the spread of COVID-19, as articulated in Table 6. A higher household and family size indicates that people have more crowded living conditions under the same roof, which increases the risk of infection of family members from one infected member after exposure on large contact network. Moreover, a larger household and family size often indicates the precarious socioeconomic conditions that may compromise people’s ability to withstand this highly infectious disease via established health care services. The non-citizen and minority populations are often involved in informal job sectors such as food delivery and retail businesses which may also increase the chances of their exposure to COVID-19. Additionally, the non-citizen population has limited access to health care services, which makes them reluctant to undergo testing and to seek health care. The population working outside of a principal city contributes to the COVID-19 diffusion, probably indicating the degree of contacts between place-based communities living a certain distance away from each other. Elderly populations are among the critical victims of the COVID-19 pandemic due to their underlying health conditions. Thus, a higher number of elderly residents increases the probability of viral transmission and infection. Interestingly, the total population (0.03) and people living in group quarters (0.06) show a negligible influence on COVID-19 cases, given the other factors integrated in the model.

Among the socioeconomic variables, congested living is the most influential, which has the high weight of 0.17. Congested living conditions limit the ability of city dwellers to maintain adequate social distancing and heighten the risk of community infection. Therefore, it appears to be one of the most critical socioeconomic determinants of the spread of COVID-19 in cities. Likewise, education level (0.15), per capita income (0.13), median income (0.11), poverty level (0.09), and low wage population (0.09) also have a major influence on COVID-19 diffusion and infection. Educated people tend to be health-conscious, and therefore a higher share of educated people in a city makes it relatively less vulnerable to the pandemic; they are also less likely to be engaged in professional activities with a lot of face-to-face interactions with other people. A similar argument can be made for the other socioeconomic factors listed. However, working population (0.04), income inequality (0.07), mobile housing (0.08), and unemployment rate (0.08) have weak or no influence on the incidence of COVID-19 cases, as presented in Table 6.

Residential density and population density, which measure the state of compactness of a city, have high weights (0.18 and 0.13, respectively); they have a determining influence on the vulnerability of a city. The built-up ratio is also an important variable, which receives a relatively high weight (0.13). The number of dependent cities (0.13) around a major city is another influential variable among the ten variables related to urban form & density due to better transport connections and greater interaction opportunities for social, economic, and business activities. Additionally, the retail employment density (0.10) has a relatively robust influence on the incidence of COVID-19 infection, as presented in Table 6.

On the other hand, the land area of the city (0.04), the employment density (0.09), the population of dependent cities (0.07), the number of housing units (0.04), and the diversity of land uses (0.08) have had a limited effect on the incidence of COVID-19 cases across cities.

The results in Table 6 indicate that, among the variables of human mobility and transport connectivity, distance to the nearest metropolitan city (0.25) is the most influential. This is in line with the strong effect of dependent cities, population density, and congested living mentioned earlier. Metropolitan living and higher density have made social distancing quite challenging, which renders cities closer to a major metropolitan city more vulnerable to the COVID-19 pandemic.
Underscores the great impact of healthcare infrastructure on controlling hospital beds have higher weights (0.17 and 0.13, respectively), which air quality index (Table 6). Similarly, the number of hospitals and of mobility reduction has significant impacts on controlling the COVID-19 has indicated that air quality improvement due to lockdown-induced variables. Thus, this is another critical variable of viral diffusion which, turn, contributes to a higher PVI score in cities. Air quality is sured population (0.28) has the higher weight compared to other variables. This, thus, is another critical variable of viral diffusion which, in turn, contributes to a higher PVI score in cities. Air quality is observed to be another influential variable. Extant research (Bashir et al., 2020; Isphording & Pestel, 2021; Rahman, Paul et al., 2021) has indicated that air quality improvement due to lockdown-induced mobility reduction has significant impacts on controlling the COVID-19 diffusion. In particular, we observe here a higher weight (0.27) for the air quality index (Table 6). Similarly, the number of hospitals and of hospital beds have higher weights (0.17 and 0.13, respectively), which underscores the great impact of healthcare infrastructure on controlling the pandemic. As already established for to the mobility & connectivity dimension, all variables in the health & environment dimension have strong effects on COVID-19 transmission.

### Table 6: Variables and their corresponding weights estimated by the FC method.

| Dimensions                  | Scores | Variables                        | Symbols | Weights |
|-----------------------------|--------|----------------------------------|---------|---------|
| Demographic                 | 0.22   | Total population (+)             | 𝑃𝑡     | 0.03    |
|                             |        | Size of younger population (+)   | 𝑃𝑠yers  | 0.15    |
|                             |        | Size of elderly population (+)   | 𝑃elder  | 0.11    |
|                             |        | Group quarter living (+)         | 𝑃гран   | 0.06    |
|                             |        | Minority population size (+)     | 𝑃mgr    | 0.09    |
|                             |        | Non-citizen population size (+)  | 𝑃non    | 0.12    |
|                             |        | Population working in a designated principal city (−) | 𝑃pdp   | 0.05    |
|                             |        | Population working outside of a principal city (+) | 𝑃pdx   | 0.11    |
|                             |        | Household size (+)               | 𝐻𝑖     | 0.14    |
|                             |        | Family size (+)                  | 𝐹𝑖     | 0.14    |
| Socioeconomic               | 0.19   | Mobile housing (+)               | 𝑎ℎ     | 0.08    |
|                             |        | Working population (+)           | 𝑃𝑤p    | 0.04    |
|                             |        | Income inequality (+)            | 𝑔𝑖     | 0.07    |
|                             |        | Congested living (+)             | 𝑃𝑐𝑙    | 0.17    |
|                             |        | Per capita income (−)            | 𝐹𝑐𝑖    | 0.13    |
|                             |        | Median income (−)                | 𝐹𝑚𝑖    | 0.11    |
|                             |        | Poverty level (+)                | 𝑃𝑝𝑜    | 0.09    |
|                             |        | Unemployment rate (+)            | 𝑉𝑢     | 0.08    |
|                             |        | Education level (−)              | 𝑓𝑒     | 0.15    |
|                             |        | Low wage population (+)          | 𝑃𝑙𝑤    | 0.09    |
| Urban form & density        | 0.17   | City size (area) (+)             | 𝐴𝑐     | 0.04    |
|                             |        | Housing units (+)                | 𝐻𝑢     | 0.04    |
|                             |        | Population density (+)           | 𝐷𝑝     | 0.13    |
|                             |        | Built-up ratio (+)               | 𝐵𝑟     | 0.13    |
|                             |        | Number of dependent cities (+)   | 𝑁𝑑     | 0.13    |
|                             |        | Population of dependent cities (+) | 𝑃𝑑     | 0.07    |
|                             |        | Residential density (+)          | 𝑅𝑑     | 0.18    |
|                             |        | Employment density (+)           | 𝑊𝑑     | 0.09    |
|                             |        | Retail employment density (+)    | 𝐷تعليه | 0.10    |
|                             |        | Diversity of land uses (−)       | 𝐸Lv     | 0.08    |
| Mobility & connectivity     | 0.21   | Commute to work (+)              | 𝑉𝑐     | 0.13    |
|                             |        | Walkability (−)                  | 𝐴𝑤     | 0.14    |
|                             |        | Distance to nearest metropolitan city (−) | 𝐷𝑐.trace | 0.25    |
|                             |        | Travel duration in public transport (+) | 𝑃𝑡𝑟.trace | 0.19    |
|                             |        | Number of airports (+)           | 𝑁𝑎     | 0.18    |
|                             |        | Public transport usage (+)       | 𝑇𝑝     | 0.12    |
| Health & environment        | 0.21   | Uninsured population (+)         | 𝑃𝑢     | 0.28    |
|                             |        | Healthcare employment density (−) | 𝑃𝑐𝑒     | 0.14    |
|                             |        | Number of hospitals (−)          | 𝑁ℎ𝑖     | 0.17    |
|                             |        | Hospital beds (−)                | 𝑁𝑏     | 0.13    |
|                             |        | Air quality (+)                  | 𝐴𝑞     | 0.27    |

Compared to others. We also note that travel duration in public transport (0.19) and number of airports (0.18) are significant variables for viral propagation owing to limited opportunity for social distancing and fertile ground for exposure to populations from other regions and countries. As for travel duration, we also observe that the increased use of public transportation and higher travel duration to work exert a strong influence on the spread of the virus due to the possibility of infection from other passengers when social distancing is hardly feasible. On the other hand, walkability (0.14) has a comparatively greater role in reducing viral transmission, which indicates that neighborhoods with a higher walkability score have lower rate of COVID-19 infection, given the other factors accounted for in the model. Overall, all of the variables in the mobility & connectivity dimension have a sturdy impact on the COVID-19 pandemic.

Among the variables in the health & environment dimension, uninsured population (0.28) has the higher weight compared to other variables. Thus, this is another critical variable of viral diffusion which, in turn, contributes to a higher PVI score in cities. Air quality is observed to be another influential variable. Extant research (Bashir et al., 2020; Isphording & Pestel, 2021; Rahman, Paul et al., 2021) has indicated that air quality improvement due to lockdown-induced mobility reduction has significant impacts on controlling the COVID-19 diffusion. In particular, we observe here a higher weight (0.27) for the air quality index (Table 6). Similarly, the number of hospitals and of hospital beds have higher weights (0.17 and 0.13, respectively), which underscores the great impact of healthcare infrastructure on controlling the pandemic. As already established for to the mobility & connectivity dimension, all variables in the health & environment dimension have strong effects on COVID-19 transmission.

#### 3.2. Pandemic vulnerability index

Considering the number of COVID-19 cases per 1000 residents as the target variable, Fig. 3 portrays city-wise vulnerability categories based on the PVI-CI scores estimated by the FC method. The PVI-CI scores are categorized in five equal-range classes with the moderate class centered on the median PVI-CI value, and visualized for a comparative analysis. Overall, the number of cities with very high, high, moderate, low, and very low vulnerability are 152, 735, 1390, 682, and 110, respectively. We find that 29% of US cities have a high or very high level of vulnerability to the COVID-19 pandemic. At the other end of the range, about the same share of cities (26%) have low or very low vulnerability.

The map depicts that many cities in California, Oregon, Arizona, Texas, Florida, and North Carolina have a high or very high level of vulnerability to the COVID-19 pandemic. In contrast, cities located in the Rockies (e.g., Montana, Wyoming, New Mexico) and in Midwestern states (e.g., North Dakota, South Dakota, Iowa, Nebraska, Minnesota, Wisconsin) are mostly in the low and very low vulnerability categories. Similarly, many cities located in the Northeast (e.g., Maine, New Hampshire, Vermont, New York, and Pennsylvania) face low or very low vulnerability.
Among the 100 most vulnerable US cities, the vast majority are located in three states only: 60 are in California, 14 in Texas, and 9 in Florida. About 70% of California’s 201 cities have either high or very high vulnerability to COVID-19. However, this is not to say that all cities in California have high vulnerability. A number of them show low vulnerability to the COVID-19 pandemic, such as Portola which is in fact one of the least vulnerable cities in the country. The findings of this study also show that large cities are not always the most exposed to the pandemic, which is a departure from a widely held view. For example, despite being one of the most populous cities in the state, San Francisco is not rated as a city with very high vulnerability, like Los Angeles and Riverside, but only 26% of it belongs to the high vulnerability category of the PVI-CI. Thus, the size of a city by itself does not determine the level of vulnerability. However, a few large cities are very highly vulnerable. Such is the case with some of the most populous cities in TX, Houston, Dallas, and San Antonio.

Additionally, we study the distribution of PVI-CI scores by state with a series of box plots (Fig. 4). Each box plot shows the minimum, first quartile, median, third quartile, interquartile range, and a maximum of the PVI-CI scores in each state. The median PVI-CI scores in California, Oregon, Arizona, Texas, Florida, and North Carolina are above the highest score of many states. In contrast, the median PVI-CI scores of cities in Montana, Wyoming, New Mexico, North Dakota, South Dakota, Iowa, Nebraska, Minnesota, and Wisconsin are below the lowest scores of many other states. Thus, the pattern in PVI-CI scores of cities depicted in Fig. 4 is consistent with the findings presented in Fig. 3. Pandemic vulnerability shows great heterogeneity across cities, and some of this can be accounted for by differences between states, but not all of it. Some states are also much more homogeneous than others.

The mapping of the PVI-CI scores (Fig. 3) suggests that vulnerability to the pandemic is not geographically random and that different regions...
of the country stand out on this basis. Indeed, the global Moran’s I of 0.52 indicates that cities are geographically clustered according to their level of vulnerability to the COVID-19 pandemic overall. To dig further into the nature and properties of this clustering tendency, Fig. 5 depicts the results from a local analysis of spatial autocorrelation using Local Moran’s I on PVI-CI scores (Anselin, 1995). In this analysis, clusters highlight the spatial relationships between cities pertaining to their shared vulnerability characteristics (Sánchez-Martín et al., 2019), while outliers pinpoint the abnormality in the vulnerability behaviors of a city against nearby cities (Duan, Xu, Liu, & Lee, 2009). We find that 35% of cities are clustered in geographic groupings of similarly extreme vulnerability at the 95% confidence level. The pink dots in the figure indicate clusters of cities with a higher level of vulnerability (High-High), while light blue dots show clusters of cities of lower vulnerability (Low-Low). On the other hand, red dots denote cities with higher vulnerability surrounded by cities with a lower level of vulnerability (High-Low); dark blue dots represent the reverse situation (Low-High). Cities appearing as gray dots are not statistically significant for being in any cluster or outlier groups.

The geography of higher-vulnerability city clusters (Fig. 5) shows a distinctive pattern. We can delineate five large clusters of higher vulnerability. They encompass California and Southern Arizona, Eastern Texas, South and Central Florida, Northern and Central Georgia and large sections of North and South Carolina, and the Washington DC-Baltimore-Philadelphia-New York City Megalopolis. Some smaller clusters of vulnerability are also found in Northwestern Oregon, Northwestern Utah, the Front Range Urban Corridor of Colorado, and Western Mississippi. Geographic clusters of lower vulnerability mainly encompass cities at higher latitudes from the Rockies to Upper Michigan and as far South as Kansas and Central Illinois, as well as from New England to the mid-section of Appalachia in Kentucky.

The outlier analysis shows that a small number of cities behave differently compared to their nearby cities. According to Fig. 5, most of the cities classified as Low-High outliers are located in California, Texas, and North Carolina. For example, Portola in California is a low vulnerability outlier in a High-High cluster due to its small population and low residential density, higher number of children and elderly, low household and family size, lower rate of unemployment, lower number of airports, higher supply of infrastructure per person for health care facilities, high walkability index, lower use of public transport, and better air quality. As discussed earlier, these variables are also regarded as the key factors to contain the vulnerability of cities, despite being spatially located close to cities showing higher vulnerability to the COVID-19 pandemic. The study also shows that, of 18 cities that are High-Low outliers, a preponderance are in three Midwestern states (five in Minnesota, three in Iowa, and three in Illinois). The underlying
causes of high city vulnerability to the COVID-19 pandemic discussed earlier are the main reasons for such high-vulnerability outliers (red dots in Fig. 5).

The study also observes that the ten most vulnerable cities are typically characterized by high population density, large average household size, large average family size, high residential density, high unemployment rate, a large share of non-citizen population, long exposure in public transport, low air quality, low per capita income, and low educational attainment. These are the most influential characteristics that make cities vulnerable to the rapid spread of COVID-19 and therefore a high incidence of COVID-19 infection cases. On the other hand, a high share of elderly people, distance from large metropolitan cities, and a large share of working population are associated with low vulnerability. Thus, population and housing density, demographic and socioeconomic characteristics of the population, and the openness of the layout of the city are the salient variables signaling the level of vulnerability of a city to viral infections like COVID-19.

3.3. Relative agreement between PVI-CI and NIEHS county level vulnerability

We seek to establish the distinctiveness of the PVI-CI by comparing it to the PVI recently devised by the NIEHS. For the set of 300 randomly selected cities, the Pearson’s correlation between the two series is 0.1602, which indicates the two indicators of vulnerability are not equivalent overall, but account for different aspects of the vulnerability of place-based communities to the COVID-19 pandemic. The deviation of standardized scores estimated by NIEHS to their corresponding standardized PVI-CI is mapped in Fig. 6 in five classes around the mean to better contextualize the differences between the two indicators. The vast majority of cities have deviation between $-1.5$ and $1.5$ (75.33%). In this group, 82 cities have very low positive or negative deviation ($-0.5$ to $0.5$), while a large number of sampled cities (108) have moderately positive deviation between $0.5$ to $1.5$. Only 36 cities have deviation between $-1.5$ to $-0.5$. Specifically, we find that 8 cities have a very low deviation of their PVI-CI score compared to their county NIEHS score (from $-0.05$ to $0.05$), including Jerseyville (IL), Olympia (WA), Laurens (SC), Dumas (TX), Guymon (OK), Winnaboro (SC), Salem (OR), and Guthrie (OK). Also, there are only 10 sampled cities that show large negative deviation ($<-1.5$) between NIEHS PVI and PVI-CI scores, against 64 with large position deviations. The 10 cities that show the largest deviations (large positive or large negative) between scores are Tooele (UT), Harvard (IL), Roosevelt (UT), Santa Paula (CA), Longmont (CO), Fairfield Glade (TN), Newport (VT), Sutherlin (OR), Newton (MS), and McComb (MS). The first five cities in this list have higher PVI-CI scores compared to the NIEHS score (large negative), whereas the other five are in the reverse situation. A moderately positive skewness in deviations is also observed.

Overall, cities with a higher positive deviation in their vulnerability scores tend to be over-represented in the southern states, particularly in Tennessee, Mississippi, Georgia, North Carolina, Arkansas, and Alabama. In contrast, some clusters of cities in the western states (esp. California, Utah, and Colorado) have large negative deviations. However, the large number of cities whose PVI-CI and NIEHS scores are similar (very low deviation) are scattered across the US with no notable spatial trend. It is important to note that there is no discernible variable that is particularly relevant to the magnitude of the deviation between the vulnerability scores. There are some probable data-related reasons for high deviation between vulnerability scores. First, the cities with high deviation tend to be located in relatively large counties; thus, these counties are not a good representation of the PVI-CI score at the county level. Second, these cities are often located at the boundary of a county and therefore would be best represented by several nearby counties. Third, another city in the same county may overshadow the randomly selected city in that particular county. As a whole, we conclude the PVI-CI is an effective measure of the vulnerability of cities to COVID-19 and this is corroborated by the alignment with the PVI devised by the NIEHS.

4. Discussion

In this study, we focused on COVID-19 infection cases to characterize the magnitude, variability, and determinants of the pandemic vulnerability of cities from the perspective of the susceptibility to
contracting the disease. The analysis successfully identified meaningful trends across US cities as far as that significant, but limited, aspect of viral infection is concerned. A related interest calls for the evaluation of pandemic vulnerability from the perspective of mortality due to the disease. The goal here would be to see if the vulnerability status of cities is meaningfully different accordingly to the severity of the consequences, that is when we calculate the feature importance based on death rates instead of infection case rates while following the same methodological steps. Because a whole host of important considerations regulate the course of the disease once infection has happened, the same co-variates found to be important for the incidence of COVID-19 infection cannot be assumed to also be the critical determinants of a deadly outcome. Therefore, we are arguing that the study of vulnerability to death outcomes of COVID-19 infection should be conducted to guide public health policies in cities with full efficacy.

While the weights and vulnerability scores were calculated using three alternative methods, the FC method was selected based on the performance metric (i.e., RMSE) to report the results on city vulnerability and discuss the underlying factors of vulnerability. We observed some incongruities between the weights provided by the different methods, which cannot be easily reconciled and merits additional research. As an example of conflicting feature weighting, air quality was ranked as a variable of low influence with the K-Best method, while FC and MIR gave it a much higher weight. Similar large differences were observed for the number of hospitals and population of dependent cities. Further research is in order to sort out the reasons for these inconsistencies to alleviate lingering concerns about the proposed feature ranking algorithms and to advance robust policy recommendations.

According to the FC method, demographic variables have the greatest influence on city vulnerability, followed by the variables of mobility & connectivity and health & environment, urban form & density, and finally the socioeconomic features of residents. Population and housing density, demographic and socioeconomic characteristics of the population (e.g., household size, average family size, congested living, share of non-citizen population, per capita income, educational attainment), travel by public transportation, air quality, and the openness of the city are the key considerations to influence the vulnerability of a city. We observe several surprising factors, which include a weak association with the size of the elderly population which is opposite to the findings of several studies in the USA that considered hospital mortality (Mikami et al., 2021; Richardson et al., 2020; Wadhera et al., 2020). This could be due to a greater degree of caution and care among the elderly since the identification of their higher susceptibility levels. Similarly, the total population, which is intuitively anticipated to rank high, displays a relatively weak influence. This could be ascribed to the fact that a large total city population does not necessarily translate into proportionately high population density or proportionately large segments of vulnerable persons. Although it there is a common understanding that the percentage of people living in group quarter would have high influence On COVID-19 infection, our analysis shows low significance of this factor. Our conjecture in this respect is as follows: group quarter living includes nursing homes, cells, nursing homes, student dormitories, and homeless shelters in US census data; Since the variable importance is evaluated based on total COVID-19 cases in July 2021, many group living facilities were closed by then. Some income variables (e.g. inequality, low wage population, and poverty level) also show low significance in COVID-19 case rate, whereas other studies have identified them as determining factors of susceptibility to the COVID-19 pandemic (Cordes & Castro, 2020). This study found that race and education level are influential determinants for COVID-19 vulnerability, which is very similar to the findings of some other studies. Cordes and Castro (2020) and Karaye and Horney (2020) The Cluster and Outlier Analysis shows that cities are strongly clustered geographically based on their vulnerability score (i.e., hot spots where highly vulnerable cities are located in close proximity, versus cold spots with low vulnerability cities also clustered together). Cold spots are almost absent from the Sun Belt states, but large coalescences of hot spots of high vulnerability are defining traits, which is very
similar to the conclusion of another study based on social vulnerability (Snyder & Parks, 2020). However, some outlier cities are also found. This is the case of low-vulnerability outlier cities located geographically near highly vulnerable cities. The present study evaluated vulnerability scores of cities (PVI-CI) on an aggregate basis using data from a mature phase of the pandemic. It would be quite informative to the pandemic diffusion process to evaluate vulnerability temporally because it can be anticipated that PVI-CI ranked features vary across stages of the pandemic. Therefore, it will also be useful to analyze PVI-CI in a time-phased manner. The present study will provide an important basis for the understanding of the features associated with city vulnerability to pandemics, epidemic and the spread of contagious diseases more generally. This would help governmental agencies to identify vulnerable spots faster and help advance preparation to mitigate the spread and the mortality rates caused by future variations of COVID-19 or other deadly diseases. The present research has created a framework that can be applied as a valuable tool for ranking pandemic vulnerability of urban areas. However, this also opens possibilities for generalization of the tool with adaptive features and customization of locally relevant features. This can lead to artificial intelligence solutions which are

Fig. 7. Feature ranking using (a) Select K Best, (b) F-Classification, and (c) Mutual Info Regression methods.
adapt in nature to improve performance in complex information environments, and policies which support artificially intelligent applications of such tools for public good (Samuel, 2021; Samuel, Kashyap, Samuel, & Pelaez, 2022). This can serve as an outcome to catalyze future research on generalization and adaptation of the pandemic vulnerability index.

This study has a few limitations. Firstly, it largely identified vulnerable cities and outliers based on a large set of variables, however, it is hard to state unambiguously why certain cities are vulnerable and others are not. A detail study on causes of vulnerability would be helpful to answer why a particular city is vulnerable. Secondly, there could be some other variables of pandemic vulnerability that this study did not incorporate. Third, we only considered three feature selection algorithms but there are many other algorithms that could be useful and potentially of great efficacy. Fourth, many variables are reproduced at city scale from other data through data engineering and enrichment, which is prone to errors, since these data are not available through census. Data on many variables are gathered from secondary sources for which the level of accuracy is unknown.

5. Outcomes

Based on the analysis of this study, we find the following significant outcomes:

1. City size (population or area) has limited influence on pandemic vulnerability. Data analysis does not support that big cities consistently have a very high level of vulnerability to the pandemic. For instance, San Francisco, New York, Chicago, and Philadelphia do not lie in the very high vulnerability group, although they are considered as some the largest in the country.

2. Cities in California, Texas, Florida, Arizona, and North Carolina are mostly in high and very high vulnerable categories.

3. The most influential urban vulnerability demographic variables are the size of the younger population and family size. Congested living and educational attainment are influential factors along the socioeconomic dimension. Retail employment density is a highly weighted variable among all variables in the urban form and density dimension. Distance to the nearest big metropolitan city is the most significant mobility and connectivity variable. Both uninsured population and air quality are also determinant variables for making cities vulnerable to COVID-19 infection.

4. Although county-level vulnerability and city-level vulnerability are not consistently related, city-level vulnerability scores are close to the vulnerability of corresponding counties.

5. This research has broader impact implications as the tool itself and the methodology presented can both be adapted to global scenarios for the development of geography-specific indices, which will demonstrate greater sensitivity to locally critical variables, and hence function with greater accuracy.

6. Conclusions

This research makes significant contributions on three different levels: (a) it factors a broad range of variables to identify the pandemic vulnerability ranking of US cities; (b) it demonstrates the critical factors of pandemic vulnerability using a multi-method approach, which generates deeper insights in the form of variations in weights of the variables, with an evaluation of COVID-19 case rates as the target variable and detailed interpretation using case rates; and (c) it points out dimensions that require further attention and vulnerability modeling, namely the need to study temporal aspects of vulnerability, conflicting weights for a few important variables indicating possible interactions between variables, and the need for better integrated and more detailed data for future PVI modeling. It also underscore the need for the study of vulnerability to the deadly outcome of the viral infections to grow more resilient cities. This study is therefore expected to be valuable to governmental agencies in their effort to manage pandemic-associated variables and policy development, as well as for future PVI-generalization and PVI-adaptation. The PVI is also expected to inform urban vulnerability research and post-pandemic urban planning.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix

See Table 7 and Fig. 7.

References

Ali, G. G. Md. Nawaz, Rahman, Md. Mohhleus, Hossain, Md. Amjad, Rahman, Md. Shahinoor, Paul, Kamal Chandra, Thilli, Jean-Claude, et al. (2021). Public perceptions of COVID-19 vaccines: Policy implications from US spatiotemporal sentiment analytics. Healthcare, 9(9), Article 1110, URL: https://www.mdpi.com/2227-9032/9/9/1110.

Aliberti, Márton Juliano Romero, Covinsky, Kenneth E., Garces, Flavia Barreto, Smith, Alexander K., Cariati, Pedro Kallas, Lee, Sei J., et al. (2021). A fuller picture of COVID-19 prognostic: the added value of vulnerability measures to predict mortality in hospitalised older adults. Age and Ageing, 50(1), 32–39.

Amrnam, Ofer, Amiri, Solmaz, Thorn, Emily L., Mansfield, Jacob J., Melachterov, Pronav, Monsivais, Pablo, et al. (2020). Data-driven development of a small-area COVID-19 vulnerability index for the United States. MedRxiv.

Anselin, Luc (1995). Local indicators of spatial association—LISA. Geographical Analysis, 27(2), 93–115.

Aazad, S., & Ghandehari, M. (2021). Spatio-temporal study of the determinants of residential satisfaction in new york city during COVID-19 using crowdsourced data. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 8, 3–10.

Banweyana, Ivan, Okello, Daniel A., Scengendo, Ronald, Mazimwe, Allan, Ojiorut, Patrick, Mubiru, Fahad, et al. (2020). Socio-economic vulnerability to COVID-19: The spatial case of Greater Kampala Metropolitan Area (GKMA). Journal of Geographic Information System, 12(40), 302.

Bashir, Muhammad Farhan, Ma, Benjiang, Komal, Kushar, Bashir, Muhammad Adnan, Tan, Daojiao, Bashir, Madhia, et al. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of the Total Environment, 728, Article 138835.

Birkmann, Jörn (2006). Measuring vulnerability to promote disaster-resilient societies: Conceptual frameworks and definitions. In Measuring vulnerability to natural hazards: Towards disaster resilient societies, Vol. 1 (pp. 9–54).

Campos, Igor Silva, Aratani, Vinicius Ferreira, Cabral, Karina Bafior, Limongi, Jean Ezequiel, & Oliveira, Stefan Vilges de (2021). A vulnerability analysis for the management of and response to the COVID-19 epidemic in the second most populous state in Brazil. Frontiers in Public Health, 9, 331.

Chan, Emily Yang Ying, Huang, Zhe, Lam, Holm Ching Yu, Wong, Carol Ka Po, & Zou, Qiang (2019). Health vulnerability index for disaster risk reduction: application in belt and road initiative (BRI) region. International Journal of Environmental Research and Public Health, 16(3), 380.

Chapman, Jim, Fox, Eric H., Bachman, William, Frank, Lawrence D., Thomas, John, & Reyes, Alexis Rourk (2021). Smart location database technical documentation and user guide: Tech. rep. V3, United States Environmental Protection Agency.

Chen, Zhaomin, Yeo, Chai Kiat, Francis, Bu Sung Lee, & Lau, Chiew Tong (2016). Combining MIC feature selection and feature-based MPCA for network traffic anomaly detection. In 2016 third international conference on digital information processing, data mining, and wireless communications DIPDMWC, (pp. 176–181). IEEE.

Chu, Zhen, Cheng, Mingwang, & Song, Malin (2021). What determines urban resilience against COVID-19: City size or governance capacity? Sustainable Cities and Society, Article 103304.

Codagnone, Cristiano, Bogliacino, Francesco, Gómez, Camilo, Charris, Rafael, Montalegre, Felipe, Liva, Giovanni, et al. (2020). Assessing concerns for the economic consequences of the COVID-19 response and mental health problems associated with economic vulnerability and negative economic shock in Italy, Spain, and the United Kingdom. PloS One, 15(10), Article e0240876.
Samuel, Jim, Kashyap, Rajiv, Samuel, Yana, & Pelaez, Alexander (2022). Adaptive cognitive fit: Artificial intelligence augmented management of information facets and representations. *International Journal of Information Management, 65*, Article 102505.

Samuel, J., Rahman, M. M., Ali, G. G. M. N., Samuel, Y., Pelaez, A., Chong, P. H. J., et al. (2020). Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics. *IEEE Access, 8*, 142173–142190.

Sánchez-Martín, José-Manuel, Rengifo-Gallego, Juan-Ignacio, & Blas-Morato, Rocío (2019). Hot spot analysis versus cluster and outlier analysis: An enquiry into the grouping of rural accommodation in Extremadura (Spain). *ISPRS International Journal of Geo-Information, 8*(4), Article 176.

Sangiorgio, Valentino, & Parisi, Fabio (2020). A multicriteria approach for risk assessment of Covid-19 in urban district lockdown. *Safety Science, 130*, Article 104862.

Sarkar, Apurba, & Chouhan, Pradip (2021). COVID-19: District level vulnerability assessment in India. *Clinical Epidemiology and Global Health, 9*, 204–215.

Seo, Jennifer Y. (2011). Justice not for all: Challenges to obtaining equal access to health care for non-citizen immigrants in the United States. *Georgetown Journal of Law & Modern Critical Race Perspectives, 3*, 143.

Sharifi, Ayyoob, & Khavarian-Garmsir, Amir Reza (2020). The COVID-19 pandemic: Impacts on cities and major lessons for urban planning, design, and management. *Science of the Total Environment, Article 142391.*

Shekhar, Himanshu, Rautele, Malvika, Khan, Mehmooda, Paris, Ricardo, de León, Rafael Maximiliano Flóres, Romero-Aguirre, María Fernanda, et al. (2022). Are leading urban centers predisposed to global risks-a review of the global south from COVID-19 perspective. *Habitat International, 102517.*

Smith, James A., & Judd, Jenni (2020). COVID-19: vulnerability and the power of privilege in a pandemic. *Health Promotion Journal of Australia, 31*(2), 158.

Smittenaar, Peter, Stewart, Nicholas, Suterman, Staci, Coome, Lindsay, Dibner-Dunlap, Aaron, Jain, Mokshada, et al. (2021). A COVID-19 Community Vulnerability Index to drive precision policy in the US. *MedRxiv.*

Snyder, Brian F., & Parks, Vanessa (2020). Spatial variation in socio-ecological vulnerability to Covid-19 in the contiguous United States. *Health & Place, 66*, Article 102471.

Stier, Andrew J., Berman, Marc G., & Bettencourt, Luis (2020). COVID-19 attack rate increases with city size. *arXiv preprint arXiv:2003.10376.*

Tarraf, Wessim, Vega, William, & González, Hector M. (2014). Emergency department services use among immigrant and non-immigrant groups in the United States. *Journal of Immigrant and Minority Health, 16*, 595–606.

UN-ISDR (2004). Living with risk: A global review of disaster reduction initiatives. *International Strategy for Disaster Reduction.*

US Census Bureau (2021). 2010 census urban and rural classification and urban area criteria. https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html. accessed on 11-23-2021.

US COVID-19 cases and deaths by state. (2021). accessed on 2021-08-16. URL: https://usafacts.org/visualizations/coronavirus-covid-19-spread-map.

Wadhera, Rishi K., Wadhera, Priya, Gaba, Prakriti, Figueroa, Jose F., Maddox, Karen E. Joynt, Yeh, Robert W., et al. (2020). Variation in COVID-19 hospitalizations and deaths across New York City boroughs. *Jama, 323*(21), 2192–2195.

Worldometers (2021). COVID-19 coronavirus pandemic. USA: Worldometer, accessed on 11-23-2021. URL: https://tinyurl.com/fe3bajs9.

Xu, Gang, Jiang, Yuhan, Wang, Shuai, Qin, Kun, Ding, Jiechen, Liu, Yang, et al. (2021). Spatial disparities of self-reported COVID-19 cases and influencing factors in Wuhan, China. *Sustainable Cities and Society, Article 103485.*

Zheng, Kangfeng, & Wang, Xiujuan (2018). Feature selection method with joint maximal information entropy between features and class. *Pattern Recognition, 77*, 20–29.