THE INFLUENCE OF DATA UNCERTAINTY ON GEOGRAPHIC SOCIAL VULNERABILITY MODEL ACCURACY

by

Carlos Alejandro Soto, B.S.

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Geography
May 2023

Committee Members:

Ron Hagelman III, Chair

T. Edwin Chow

Sarah Blue

DEDICATION

This thesis is dedicated to Amy and Madeline. It is also dedicated to all the people who may be inadequately counted or represented in ACS data, to the detriment of their individual, family, neighborhood, and community's quality of life.

ACKNOWLEDGEMENTS

I would like to thank several key people who were instrumental in the development of this research. My advisor Dr. Ron Hagelman III was patient, generous, and wise throughout this process. Dr. Ed Chow helped me learn and develop new skills that enabled me to take this research to a higher level and inspired me to achieve a deeper understanding of the data. I am also grateful for Dr. Sarah Blue's experiences and perspective, which helped me stay grounded with the human aspects of this research. Allison Glass-Smith's attention to detail and experience expedited the process, while Nathaniel Dede-Bamfo's broad skillset and diverse background made consultations with him uniquely insightful. Finally, I would like to express my appreciation to Mark Carter for the enthusiasm with which he introduced me and all his other introductory statistics students to the wonderful world of quantitative methods. Without the encouragement, support, and guidance I received from these faculty and staff, and the broader Texas State University community, this research would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

Abbreviation	Description
ACS	American Community Survey
ATSDR	Agency for Toxic Substances and Disease Registry
CDC	Centers for Disease Control and Prevention
CV	Coefficient of Variation
GRASP	Geospatial Research, Analysis, and Services Program
LISA	Local Indicator of Spatial Association
MOE	Margin of Error
MSA	Metropolitan Statistical Area
SQL	Signal Query Language
SVI	Social Vulnerability Index
USCB	United States Census Bureau

I. INTRODUCTION

By the time COVID-19 became a global pandemic in 2020, it was well established that the level and types of risk posed by hazards and disasters are different for different groups of people, and that some groups are more vulnerable than others. As the COVID-19 pandemic unfolded, certain populations were found to have increased risk along social, economic, political, and institutional divisions in much the same way that natural hazards and other emergent situations have done so before (Karaye and Horney, 2020, Kim and Bostwick, 2020). To more effectively plan and prepare for hazards and disasters, stakeholders must therefore have a clear understanding of the social vulnerabilities that may exist in their region of interest. Historically, the United States Census Bureau (USCB) has been an important source of socioeconomic, demographic, and other data, and it continues to serve in that role. During the 20th century, the USCB introduced the decennial census long form to gather numerous key socioeconomic data points from a subset of the population. However, the 10-year gap between censuses was a significant limitation to the use of these data in short-term planning applications. Later, the United States Census Bureau developed the American Community Survey (ACS) in part to address the temporal resolution issue resulting from the decennial nature of the traditional census count.

After having replaced the "long form" in 2010, the USCB's American Community Survey (ACS) has been the premier source for demographic data in the USA. Since the ACS is carried out on a sample of the United States (US) population, there is a degree of statistical uncertainty associated with each ACS estimate. As a result, the margins of error (MOE) inherent in ACS data negatively impact the reliability of some ACS estimates (United States Census Bureau, 2020a). For example, Table 1 shows the ACS unemployment rate for a contiguous group of five census tracts in Boston, Massachusetts. The estimated unemployment rates range from 1.4% to 29.4% (Table DP03, 2014-2018 ACS 5-Year Estimates, USCB). However, due to the MOE, it is unclear which census tract has the lowest (or highest) unemployment rate. Still, these data are currently used in the US Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry Social Vulnerability Index (CDC/ATSDR SVI), an important tool for people working in fields such as urban planning, disaster management, and public health (ATSDR, 2021). Quite often, social vulnerability indexes such as the CDC/ATSDR SVI are utilized to identify places with high levels of vulnerability. This task is facilitated by tools such as the CDC/ATSDR's interactive SVI map (Figure 1). However, a previous analysis of an early version of the CDC SVI shows that, because of the margins of error (MOE) associated with the underlying data, the index's precision tends to decrease as vulnerability increases (Tate, 2013).

Estimates with a high MOE run the risk of over counting or under counting the indicator being measured, which will have an impact on the relative vulnerability derived from these estimates. For an example of how adding or subtracting the MOE may maps made with ACS data, see Figures 2-4. Figure 2 shows a map displaying the percentage of the population below the poverty level for census tracts in the Austin Metropolitan Statistical Area (MSA). Subtracting the MOE from the estimates and mapping the result produces a "lower estimate", illustrated in Figure 3. Adding the MOE to the estimates produces an "upper estimate", illustrated in Figure 4. An analysis of the reliability of the CDC SVI will provide an example with which to demonstrate how the margin of errors in the sample data can affect the reliability of tools derived from the data. A spatial analysis of the SVI's ACS source data and the margins of errors inherent in those data uncovers the geographic extent of over counting or under counting and illustrates the ways in which the CDC SVI may, or may not, accurately reflect community conditions.

Table 1. ACS unemployment rate for a contiguous group of five census tracts in Boston, Massachusetts.

Census Tract	Percent Unemployed	Margin of Error
102.03	1.4%	1.1%
103	15.7%	3.6%
104.05	22.4%	4.1%
104.08	3.7%	2.8%
9818	29.4%	30.9%

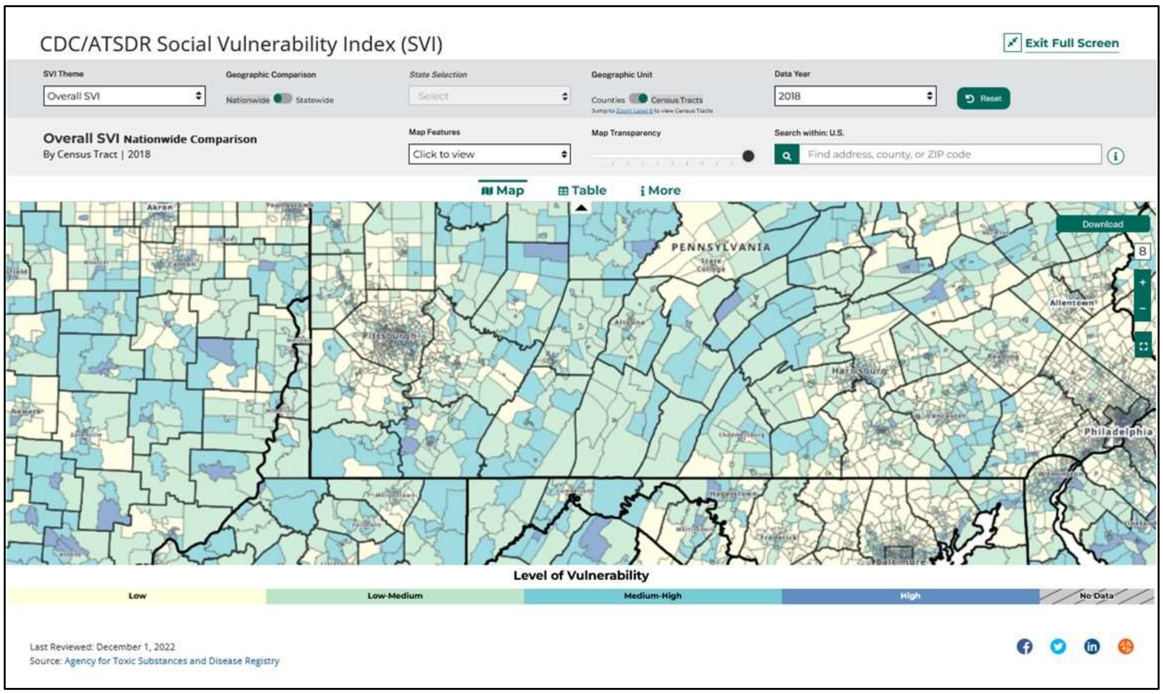


Figure 1. The 2018 CDC/ATSDR SVI interactive map.

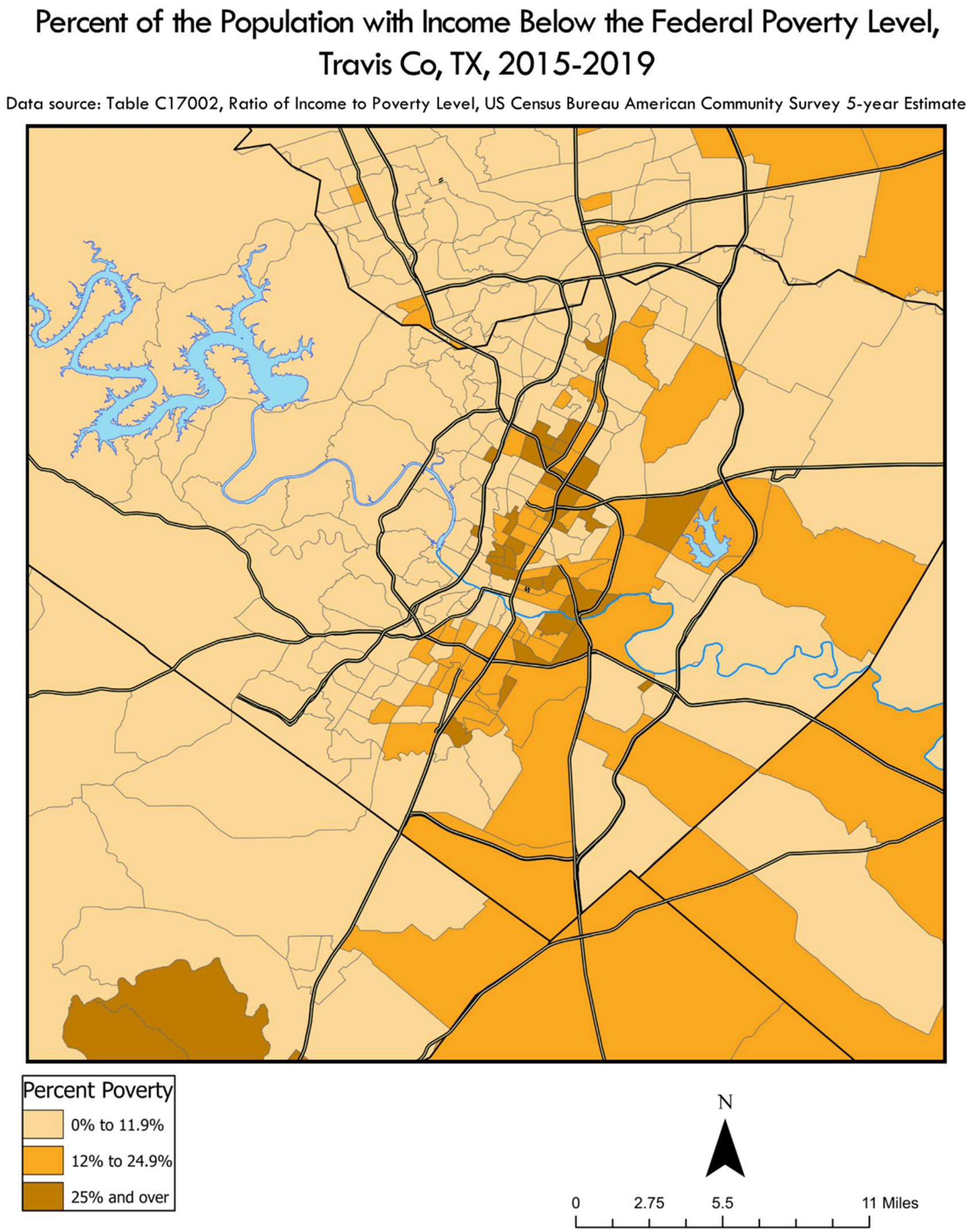


Figure 2. Percent of the Population with Income Below the Federal Poverty Level, Travis Co, TX, 2015-2019.

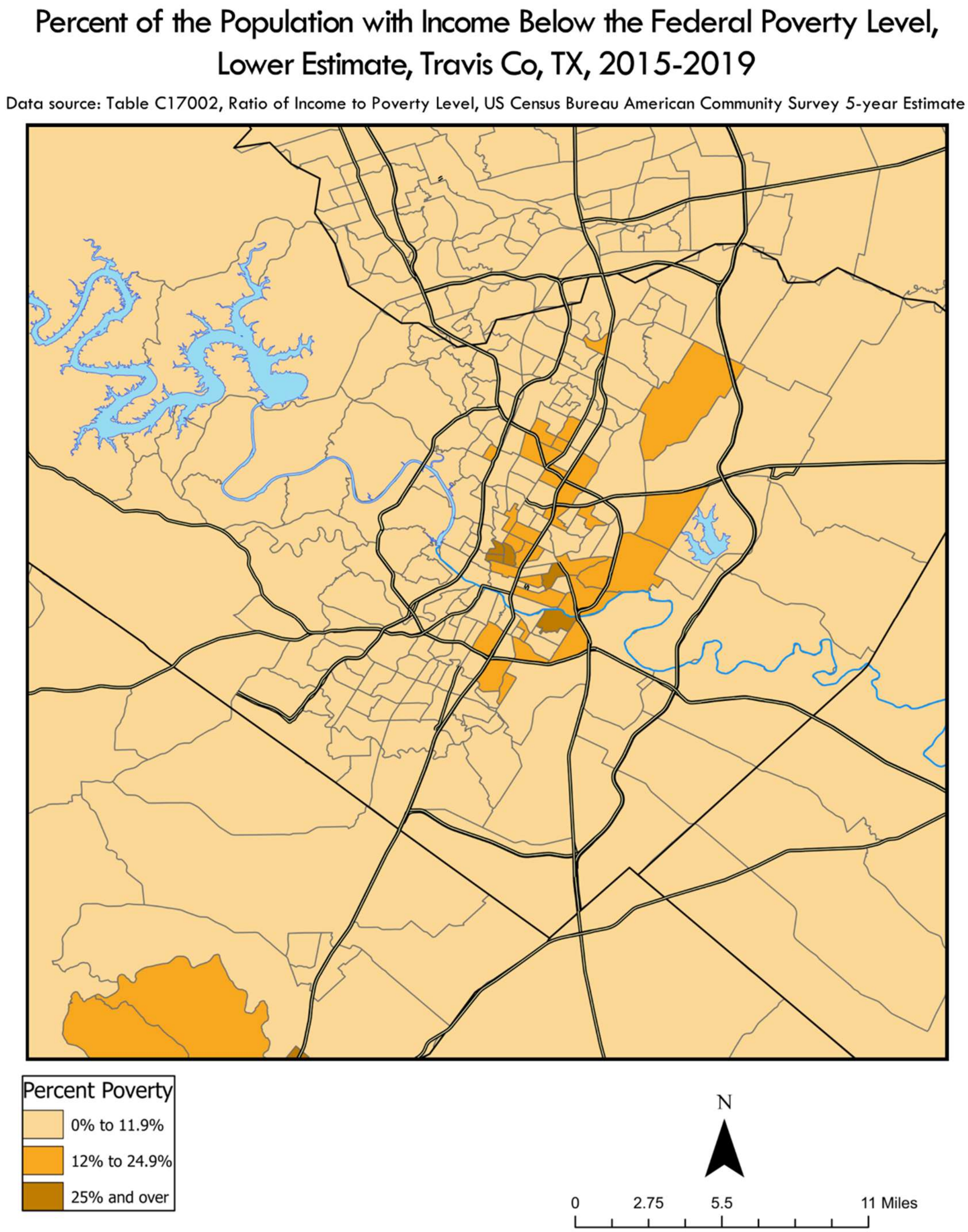


Figure 3. Percent of the Population with Income Below the Federal Poverty Level, Lower Estimate, Travis Co, TX, 2015-2019.

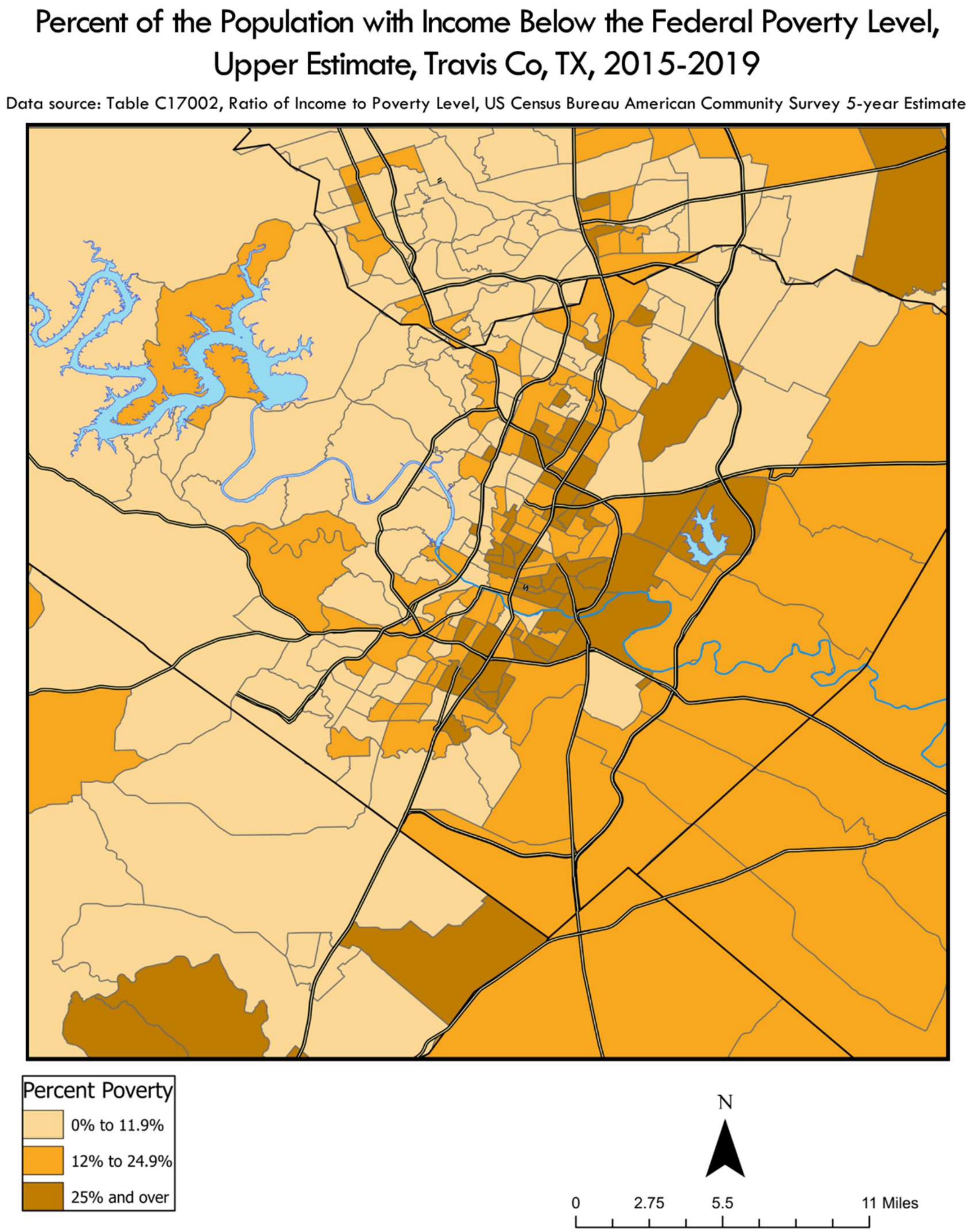


Figure 4. Percent of the Population with Income Below the Federal Poverty Level, Upper Estimate, Travis Co, TX, 2015-2019.

The idea of social vulnerability became integrated into the field of disaster management during the second half of the 20th century after researchers recognized that socioeconomic factors also shape community resilience (Cutter et al. 2003, Flannagan et al. 2011, Morrow 1999). As a result of the efforts of many who have contributed to a growing body of scientific research carried out over the past several decades, today we have a clearer understanding of the variety of factors that may influence social vulnerability. An individual's age, economic level, connections, and surrounding community can influence their vulnerability to different kinds of hazards (Cutter et al. 2003, Wolkin et al. 2015). Although local, national, and state government entities are tasked with mitigating, planning for, and responding to a variety of hazards and emergencies, in many cases the needs of the most vulnerable are not sufficiently met (Tate 2013, Kim and Bostwick 2020). For this and other reasons, it is of utmost importance to identify the locations of socially vulnerable populations and which groups are most vulnerable. Social vulnerability indexes, which rank geographies from most vulnerable to least vulnerable according to different models, are typically utilized for this purpose. But an uncertainty analysis of an earlier version of the CDC SVI found that it was precisely in the locations that were identified as most vulnerable "where the precision of the model is at its nadir." (Tate 2013). The margins of error in the ACS estimates (which are not included with the CDC SVI's derived vulnerability rankings), can lead to an over count or under count of vulnerable populations and an inaccurate representation of social vulnerability by the SVI model.

Over counting or under counting vulnerable populations can have social implications. That vulnerability exists within a context of different hazard types and potentially available options to reduce risk is contended by many (Spielman at al. 2020). Measurable aspects of social vulnerability may be context dependent and lead to unexpected results (Spielman et al. 2020). For instance, the percent of the population that is African American within a census tract is unlikely to have the same meaning for social vulnerability in a place like Atlanta, GA, where an estimated 52% of the population identified as African American, that it might somewhere like Helena, Montana, where an estimated 0.509% of the population identified as African American (Table B02001, 2014-2018 ACS 5-year Estimates). Census tract level data is often used for determining where to allocate resources at the local level. Since these kinds of resources are often the ones whose funds are first cut from local budgets in times of austerity, a clearer understanding of the strengths and weaknesses of tools such as the CDC/ATSDR SVI could help better target the use of what limited funds are available, potentially leading to a more enduring impact.

The USCB warns about the limitations and proper use of ACS data in guidance materials provided to ACS data users (USCB 2020a). These uncertainties are greater for small geographies, such as census tracts, or subsets of the general population, such as ethnic minorities or persons with disabilities. Previous research indicates that overlooking the MOE in ACS data can render problematic results (Bazuin and Fraser 2013, Spielman et al. 2020, Sun and Wong 2010). Other research indicates that planners who use ACS data rarely communicated its uncertainty (Jurjevich et al. 2018). The few studies that highlight issues related to SVI reliability only do so at the local or regional level (Bakkensen et al. 2017, Tate 2013). Therefore, the need continues to exist for improved understanding of the uncertainty of the measures that make up the CDC SVI at the census tract level across the USA, and the potential error(s) in the vulnerability rankings that make up the index.

II. PURPOSE STATEMENT

The purpose of this research is twofold. The first goal is to improve the next generation of social vulnerability indexes by assessing the uncertainty of an existing, widely used social vulnerability index. A quantitative characterization of the statistical uncertainty inherent in the CDC's SVI estimates can provide its users with a better idea of the strengths and weaknesses of the index as it relates to the reliability of its source data. It can also provide researchers who design and produce social vulnerability indexes with important data about the geographic and social location of potential ACS data uncertainties at the census tract level. An uncertainty analysis of the CDC SVI's source data will contribute to a better understanding of which social groups are potentially most affected by the variability in ACS source data and the CDC Social Vulnerability Index. The second goal of this research is to illustrate the importance of how vulnerability modeling can misidentify populations as a consequence of the uncertainties in the underlying data. In the case of the CDC SVI, since the margins of error in the ACS source data sometimes make the estimates uncertain to the point of being misleading, it is essential to have a clear understanding of the geographic and social distribution of the uncertainties of the source data and how these may impact the calculated overall vulnerability metric.

III. RESEARCH QUESTIONS

This research is intended to help improve our understanding of the kinds of limitations that arise from inherent uncertainties in the data used for the development of a social vulnerability index. To achieve this goal, this research will seek answers to the following questions:

1. What are the quantitative characteristics and geographic patterns of the statistical uncertainties in the CDC SVI census data measures?
2. What statistically significant relationships exist, if any, among tracts most likely to be over counted or under counted in an SVI such as the CDC's? Are certain communities systematically misrepresented?
3. Considering the quantitative characteristics and geographic patterns of the uncertainties, in what places and among which social groups do the statistical uncertainties cluster?

To respond to these questions, the measures in the 2018 CDC SVI at the census tract scale for the 50 states and the District of Columbia were subjected to an uncertainty analysis. This uncertainty analysis included a focus on quantifying the reliability of the ACS variable estimates from which the CDC SVI rankings are derived, and examining what the MOEs of the underlying data may mean for tools such as the interactive map on the ATSDR/CDC website. To measure the reliability of the 15 ACS variable estimate data, the coefficient of variation (CV) was calculated. The MOE was used in combination with the 15 ACS variable estimates to calculate social vulnerability rankings corresponding to the upper and lower boundaries of the 90% confidence intervals of the SVI variables. The CVs were analyzed to determine the existence of geographic or social patterns in the distribution of uncertainty in the CDC SVI's underlying data, and to determine the reliability of the CDC SVI vulnerability rankings.

IV. CONCEPTUAL FRAMEWORK

This investigation of the CDC's Social Vulnerability Index is based on the premise that uncertainties in the underlying data may lead to uncertainties in the estimated level of vulnerability for some areas or groups of individuals. Previous research shows that, in many cases, the margin of error is overlooked, even though this error can result in significantly unreliable representations (such as maps, charts, etc.) made with the data (Jurjevich et al., 2018; MacEachren et al., 1998, Sun and Wong, 2010). In a list of guidelines for planners who use ACS data, the first recommendation is to "report the corresponding MOEs of ACS estimates" (Jurjevich et al. 2018). Although the CDC SVI includes MOEs for the ACS estimates in its data set, the model does not consider the margin of error when the percentile rank is calculated for each vulnerability estimate (the percentile rank is used to determine the relative vulnerability of each census tract). A diagram visualizing the conceptual framework underlying this research is available in Figure 5. As currently implemented, the CDV SVI rankings (like many estimates derived from survey data), do not make use of the MOE data when calculating percentile rankings to determine relative vulnerability.

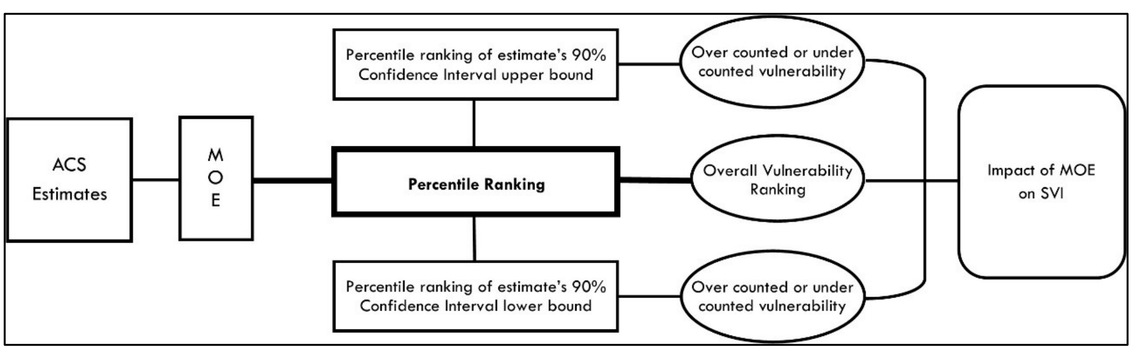


Figure 5. Conceptual Framework.

V. THEMES OF RELATED RESEARCH

Identifying Social Vulnerability

The Hazards-of-Place model of vulnerability includes social vulnerability within an interacting set of components that include social and place-based inequalities (Figure 6, Cutter at al. 2003). In recent decades, the idea of social vulnerability gained importance within the disaster management field as researchers developed a deeper understanding of the types of risks, emerging from socioeconomic and demographic factors, that affect resilience at the local level (Flanagan et al. 2011). Major factors impacting social vulnerability can include insufficient access to information, knowledge, and technology; barriers to political power and effective representation; demographics; and undesirable housing conditions (Cutter et al. 2003). Previous research shows that specific groups of people, such as individuals with incomes below the poverty level, elderly individuals, households led by women, and people who have recently established residence, are subject to higher levels of risk over the course of the disaster response process (Morrow 1999). Using community indicator data from widely accepted sources, researchers developed social vulnerability indices to help support disaster management work and research (Cutter at al. 2003, Flannagan et al. 2011). These indices continue to be popular resources in a variety of fields today.

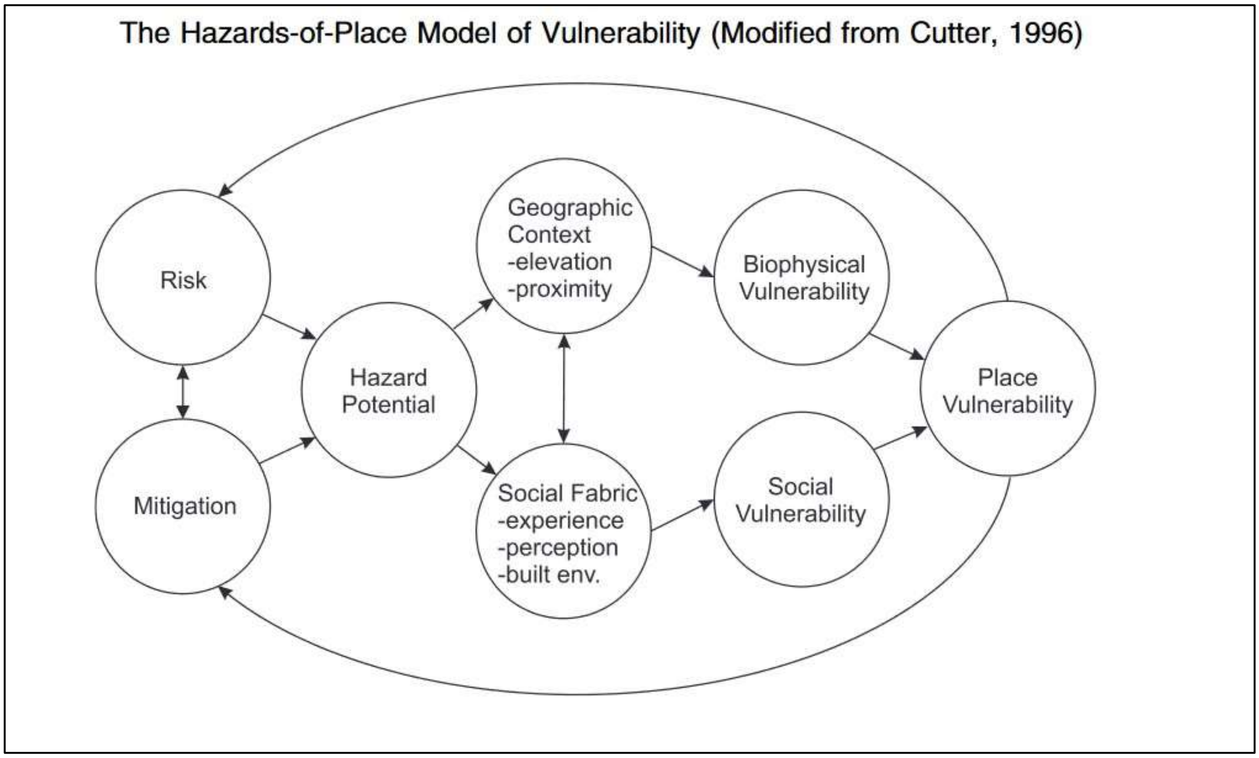


Figure 6. The-Hazards-of-Place model of vulnerability (modified from Cutter, 1996). Source: Cutter at al. 2003.

The CDC SVI is one of the most prominently recognized and applied social vulnerability configurations (Rufat et al. 2019). The Geospatial Research, Analysis, and Services Program (GRASP) created the first version of the US Centers for Disease Control and Prevention/Agency for Toxic Substances and Disease Registry Social Vulnerability Index (CDC/ATSDR SVI) in 2011, with the goal of providing the public health community a tool they could use for the quick and precise identification and planning of assistance for socially vulnerable populations throughout the duration of hazardous events (ATSDR 2021). At its introduction the index consisted of a series of composite measures derived from data taken from the US decennial census Long Form. Currently, CDC SVI data are sourced from the United States Census Bureau American Community Survey (ACS). Due to differences in the methodology between the decennial census long form and the ACS, ACS data is less reliable than the decennial long form (Jurjevich et al. 2018, Spielman et al. 2014). What follows is a brief overview of the history of the ACS and major differences between it and the decennial long form.

From the first census of the US population in 1790 to the present day, the USCB continues to stand out as the main source of data regarding the US population. With the goal of a complete enumeration of the population within the US, the decennial census constitutes the most precise tool available to researchers. However, the data collected by the decennial census was limited in scope and the need for a richer set of data prompted the development of a more detailed questionnaire to be sent to a sample of the US population. The decennial census long form was used from 1970 to 2000 to collect detailed household data from the US population (USCB 2021). Despite the variety of data offered by the long form, one major drawback was the 10-year gap between data points. To overcome this limitation, a collaborative effort was put in place to develop a survey that would gather the same data as the long form on a continuous basis (USCB 2006). After a long period of design and development, the American Community Survey began full implementation in 2005. The ACS replaced the long form in 2010 (USCB 2021).

The development and introduction of the ACS was prompted by the major temporal resolution limitation of US decennial census long form and the need for a data source based on continuous measurement (USCB 2006). The complete enumeration philosophy of the decennial census also influenced how the sampling for the long form was carried out, which led to high costs of surveying each household sampled (Spielman et al. 2014). The ACS solved some of these problems, but at the same time it brought a new set of problems for data users to contend with. The creators of the CDC SVI recognize the limitations of the data that make up the index and CDC SVI documentation warns users that some of the MOEs are high (ATSMR 2022).

An important difference between the decennial long form and the ACS is in the number of households sampled for each survey. The most recent long form was administered in 2000. It sampled 21,107,353 households, representing 18.2% of households enumerated in the 2000 decennial census (Gbur and Hefter 2002, USCB 2000). Because of the COVID-19 pandemic on ACS activities, the 2020 sample size and final interview number were much lower than in a normal year. In 2019, the most recent year the ACS was carried out to its full extent, the sample consisted of 3,544,301 addresses (USCB 2021). This represents a rate of 2.5% compared to the number of households enumerated in the 2020 decennial census (140,498,736) (USCB 2020b). Previous research has found that the statistical uncertainty of ACS data was found to be close to 75% higher than the decennial census long form (Jurjevich et al. 2018). The uncertainty inherent in ACS data will underly any products derived from this data, such as the CDC SVI.

Indexes such as the CDC SVI fall short in other ways in addition to the errors in the underlying data. An important aspect to consider about social vulnerability is that, although it may exist in a person or place, it cannot be directly observed. Instead, social vulnerability must be statistically derived. This makes social vulnerability an example of a latent variable, and the correctness of the quantities provided by the SVI a challenge to verify (Spielman et al. 2020). Analysis of outcomes resulting from Hurricane Sandy found that the explanatory power of the SVI was poor, and the construct validity was weak (Rufat et al. 2019).

The principal focus of this research is in developing a better understanding of the quantitative uncertainties that underly ACS data and how these uncertainties manifest in tools developed from ACS estimates, using the CDC SVI as an example**.** To develop a better understanding of one of the potential limitations present in tools made from ACS data, this project utilizes an approach developed from methods employed in previous uncertainty analyses.

The research proposed in this project will be the first time an analysis this type will be carried out at a national scale and has the potential to uncover new geographic and social patterns in the uncertainty that underlies the CDC SVI.

GIS and Social Vulnerability Modeling

Scalar characterizations of vulnerability (such as the quartile scale used in the CDCSVI) are inevitably impacted by the uncertainties in the underlying data. The MOEs of the underlying data can impact ACS estimates and, as a result, the vulnerability measurement produced with an index such as the CDC SVI. The MOEs are provided within the original ACS data as well as with the CDC SVI data documentation at a 90% confidence interval (USCB 2020a, ATSDR 2022), which is a rough measurement. The effect of the margin of error on the estimate may affect the percentile ranking upon which the vulnerability level of a census tract is based, which may lead to an over count or under count of the relative vulnerability of that census tract. Considering the MOE when calculating the relative vulnerability ranking can provide measures of relative vulnerability corresponding to the upper and lower bounds of the 90% confidence interval. Another way of assessing the reliability of the CDC SVI involves measuring the variability of the individual variable estimates using a measure such as the coefficient of variation.

The coefficient of variation (CV) is a measure calculated by dividing the standard deviation of an estimate by the mean for that estimate. It measures the relative amount of sampling error associated with a specific sample estimate that helps to distinguish between cases where sampled responses lie close to the median (representing low uncertainty) from cases where the sampled responses are dispersed throughout the confidence interval (representing high uncertainty). The CV can be used to measure the reliability of an estimate and is a standard statistical quality measure of the United States Census Bureau (Heimel 2014, USCB 2020a). A lower CV means that the relative reliability of the estimate is higher. Furthermore, the CV is a practical measure in that it represents the uncertainty in the form of a relative percent of the estimate. A previous analysis of the patterns and causes of uncertainty in the ACS found that census tracts with similar CVs tend to cluster around each other, and that census tracts in the center of cities tend to have a higher CV for income than suburban census tracts (Spielman et al. 2014). The CV was also used in a previous analysis of the CDC SVI for Sarasota County, Florida (Tate 2013). This project builds on this research by providing a census tract-level analysis of the reliability of the CDC SVI for the 50 states and the District of Columbia.

An important shortcoming to consider about the social vulnerability approach is the challenge for nonexperts to comprehend the construction of the index or how to best utilize it, particularly since an independent variable that the vulnerability index can be calibrated against is unavailable (Zhou et al. 2014). In terms of the statistical uncertainty of ACS data, practitioners who understand the uncertainty indicated that they do not report or otherwise address MOE values because they are unable to convince others of its significance (Jurjevich et al., 2018), perhaps because general audiences likely lack knowledge about the meaning of MOEs or how to best utilize them (Bazuin and Fraser, 2013). The result is that many of the stakeholders who recur to tools such as the CDC/ATSDR SVI may not realize that the vulnerability estimates they are interested in may be quite unreliable. The ease with which maps and GIS allow complex issues to be visualized has a disadvantage in that an audience without an understanding of mapping effects or without clear descriptions and documentation pertaining to the map may be lead into false interpretations (Fekete, 2012). Furthermore, the growing focus on social vulnerability may be drawing attention away from institutional vulnerability. Institutional vulnerability has been shown to have the potential to significantly increase the vulnerability of people to the hazards presented by disasters such as Hurricane Katrina (Birkmann, Wisner 2006). Developing an improved understanding about how the uncertainties underlying the CDC/ATSDR SVI may affect the reliability of its estimates is an important step in improving a key tool that national, state, and local institutions depend on for making decisions that impact the lives and livelihoods of vulnerable populations.

VI. RESEARCH METHODS

Site & Situation

The full set of estimates reported for the 73,057 census tracts on the 2018 CDC SVI was analyzed to examine the uncertainty of the SVI variable estimates and the vulnerability rankings. The area under investigation encompassed the totality of census tracts within the 50 US states and the District of Columbia. This research was carried out at the census tract level because this is the highest-resolution data available from the CDC SVI website. Census tracts are a statistical division of a county or other entity. Census tracts typically have a population size between 1,200 and 8,000 people. Their spatial extent can vary depending on population density, with urban census tracts typically covering a much smaller area than rural census tracts. Census tract boundaries are updated in response to local population growth or decline every ten years as determined through the decennial census. Access to census tract level data is important for community planners, agencies, and other stakeholders because it provides neighborhood level data invaluable for localized decision making.

The population residing within the area of study is heterogeneously dispersed and socially diverse. According to USCB data, 81% of the US population lived in urban areas in 2010 (USCB 2021). These urban areas are distributed throughout the 50 states and DC. States in the western USA have lower population densities than states in the eastern USA. Of the 3,006 counties in the USA, 769 are considered coastal watershed counties, where a substantial portion of their land area intersect coastal watersheds, and 452 are considered coastal shoreline counties, directly adjacent to the open ocean, major estuaries, and the Great Lakes. In 2010, 52% of the US population lived in coastal watershed counties, and 39% of the US population resided in coastal shoreline counties (USCB 2021). Predicted sea level rise may displace millions of people residing in coastal communities, which will consequently impact the inland communities where the displaced people would eventually settle (Hauer 2017).

The US population is increasing in racial/ethnic diversity. According to the 2020 decennial census, the most prevalent racial or ethnic group in the USA was the White alone non-Hispanic population with a 57.8% share, down from a 63.7% share in 2010 (Jensen et al. 2021). Appendix A shows a table provided by the USCB with race and ethnicity prevalence by state for 2020. Although the White alone non-Hispanic population represents the largest racial or ethnic group in most states when comparing the 50 US states, the District of Columbia, and Puerto Rico, five entries stand out as having other racial or ethnic groups represent the largest share of the population (California, the District of Columbia, Hawaii, New Mexico, and Puerto Rico). Within the group of states where White alone non-Hispanic people represent the largest racial or ethnic group, prevalence varies. While 90.2% of the population of Maine identifies as white alone non-Hispanic, this group made up only 39.7% of the population of Texas. For a map showing the most prevalent race or ethnic group by county for 2020, see Appendix B.

Data/Information

This investigation utilizes census tract level data from the 2018 CDC SVI available from the Data & Documentation Download section of the CDC SVR website. The ATSDR CDC platform offers users the option to download data according to two different geographic ranking methods. A national comparison method ranks all United States Census tracts against one another. A state level comparison method ranks tracts or counties against other tracts or counties in that state. An interactive SVI map is also available on the ATSDR CDC website, which utilizes the national comparison method. The ATSDR CDC SVI map presents the relative vulnerability percentile data in a graduated color map with 4 classes in quartile order, with categories ranging from low vulnerability (lowest quartile) to high vulnerability (highest quartile). Therefore, this analysis focuses on tracts ranked according to the national comparison method.

Census tract level data were downloaded for the 50 states and District of Columbia. The data set as provided by the CDC SVI consists of primary data from the United States Census Bureau American Community Survey and rankings derived by the ATSDR/CDC SVI program from the ACS data. The ACS data consists of estimates and percentages for each of the 15 ACS variables. The CDC SVI data includes the corresponding margin of error (at the 90% confidence interval) for each variable estimate. Because the CDC SVI is developed using Structured Query Language (SQL), which uses a different level of precision than Excel and other software (ATSDR 2022), the methods used in this analysis may yield marginally different results from calculations carried out using SQL.

The primary source of data for the ATSDR CDC SVI analyzed in this investigation are ACS 2014-2018 5-year estimates. The ACS estimates in the CDC SVI are collected from the following 2018 ACS 5-Year estimate tables: S0601, DP04, B17001, DP03, B9301, B06009, S0101, B09001, DP02, B01001H, B16005, DP04, and B26001. Estimates gathered from these tables are provided at the census tract level in the data and documentation section for the CDC SVI. In the SVI, tract level estimates are ranked on 15 social factors such as poverty, disability, education level, housing characteristics, and other factors which are divided into four corresponding themes. The four themes (socioeconomic status, household composition & disability, minority status & language, housing type & transportation) are further ranked to determine rankings specific to each theme. The ATSDR and CDC calculate a tract's overall vulnerability ranking by aggregating the sums for each theme, ordering the tracts, and calculating the overall percentile ranking for each census tract (ATSDR 2022). This overall vulnerability ranking is visualized in the default view of the interactive ATSDR CDC SVI map in the CDC website by classifying the value of the overall vulnerability into 5 categories according to the following scheme: 0.0-0.25 (least vulnerable 25%), 0.2501-0.50, 0.5001-0.75, 0.7501-1.0 (most vulnerable 25%), and No Data.

The ATSDR CDC SVI data and documentation section provides reference information that includes specific names for the variables and themes that make up the index. Throughout this study, the variables and estimates in the text, charts, graphs, tables, or other components utilize the nomenclature provided on the SVI website. For reference, Figure 7 shows the 15 ACS variables organized into the four themes for which vulnerability is ranked. The ATSDR provides data files that include the census tract estimates for: the 15 variables tracked in the CDC SVI, percentile estimates for each variable, percentile estimates for each theme, sum of series themes, overall percentile ranking, flags for each variable, flags for each theme, sum of flags for each theme, and four adjunct variables. The focus of this study is on the variables that are ranked to determine relative vulnerability, the corresponding MOEs, and the overall vulnerability ranking.

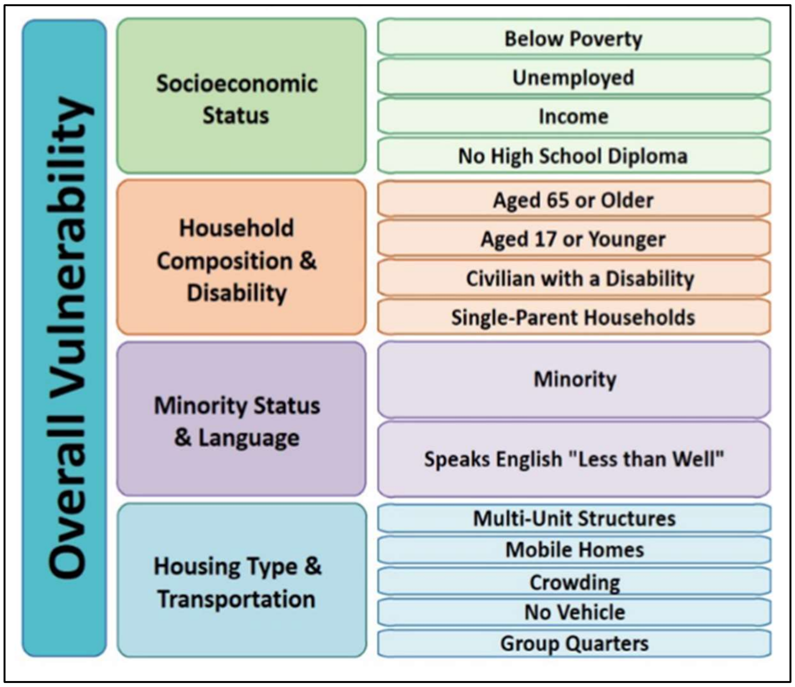


Figure 7. Fifteen ACS variables divided among four vulnerability themes make up the overall vulnerability ranking in the CDC SVI. Image source: ATSDR 2022.

Analysis/Techniques

This study makes use of the coefficient of variation (CV) between the ranked variable estimates and their MOEs to measure the statistical uncertainties in the CDC SVI data measures. The CV, calculated as ratio, quantifies the relative sampling error associated with each sample estimate. The USCB provide a method for calculating the CV with the estimate and its MOE (USCB 2020). This method serves as the basis for the formula utilized in this analysis. The USCB mentions that the magnitude of the CV can serve as an indicator of the reliability of the estimate, with lower magnitudes indicating greater reliability. Other research suggests that estimates with CV magnitudes below 12% are reliable and estimates with CV magnitudes above 40% are unreliable (Jurjevich et al. 2018, Tate 2013). The MOE and the estimate are provided in the ATSDR CDC SVI data and documentation for the 2018 SVI. The CV was calculated using Microsoft Excel, according to the following formula:

$$\mathrm{CV} = \frac{\frac{\mathrm{MOE}}{1.645}}{\mathrm{Estimate}}$$

Microsoft Excel was also used to analyze and visualize the quantitative characteristics and geographic patterns of the CVs for the 15 SVI variables.

To determine instances where there may be an overcount or undercount of the overall vulnerability ranking of a particular census tract, the overall vulnerability was recalculated twice for each census tract. These recalculations employed the MOE included with the SVI variable estimates and were carried out on Microsoft® Excel® for Microsoft 365 MSO (Version 2203 Build 16.0.15028.20152) 64-bit. In the first recalculation, the MOE was added to the SVI variable estimates and the result ranked according to the percentile ranking and aggregation methods specified in the 2018 CDC SVI data and documentation (ATSDR 2022), which make use of Microsoft Excel and its PERCENTRANK.INC function. The second recalculation involved subtracting the margin of error before continuing with the previously described SVI ranking and aggregation.

The values of the overall vulnerability rankings were then compared to the corresponding values in the two recalculations. Instances where the value of the overall vulnerability ranking are lower than the value of the recalculations can be understood as a potential undercount of the relative vulnerability for that census tract. Instances where the value of the overall vulnerability ranking are greater than the value of the recalculations can be understood as a potential overcount of the relative vulnerability for that census tract. When these discrepancies were enough to shift the category within which the census tract is classified on the SVI map, (for example, from lowest vulnerability to highest vulnerability or vice-versa), it was categorized as an over count or under count. In cases where the overall vulnerability was higher when compared to one of the recalculated vulnerabilities and lower when compared to another, the tract was categorized as “both” meaning both under counted and over counted. Census tracts which remained in the same category as the original overall vulnerability were categorized as U for unchanged. To test whether it would be possible to predict the categorical outcome (over counted, under counted, both, or unchanged) as a function of the CV of the 15 SVI variable estimates, a discriminant analysis was carried out utilizing JMP Pro 15 for Windows.

Spatial associations (clusters) among the uncertainties that were detected using univariate local Moran’s I were subsequently classified according to their relatedness utilizing local indicator of spatial association (LISA) statistics (Anselin 1995). Geovisualization of LISA statistics was accomplished utilizing GeoDa spatial modeling software (version 1.20.0.10) to generate four classes of clusters for the SVI variable CVs at the census tract level. These classes are: (1) high-high (HH), tracts with high standardized CV values (furthest from the mean) surrounded by tracts with high standardized CV values; (2) low-low (LL), tracts with low standardized CV values (closest to the mean) surrounded by tracts with low standardized CV values; (3) high-low (HL), tracts with high standardized CV values surrounded by tracts with low standardized CV values; and (4) low-high (LH), tracts with low standardized CV values surrounded by tracts with high standardized CV values.

The spatial clustering of similar values is indicated by the positive spatial autocorrelations which manifest as hot spots (HH) and cold spots (LL), while the spatial clustering of dissimilar values is indicated by the negative spatial autocorrelations which manifest as spatial outliers (LH and HL) (Anselin 2005). Statistically significant clusters were mapped using LISA scores with a significance of $p \geq 0.05$. An observation *i*’s local Moran statistic is defined in the following way:

$$I_i = z_i \sum_{i \neq j}^{n} w_{ij} z_j$$

Here, I_i is the local Moran's I indicating spatial autocorrelation; z_i and z_j are observations deviating from the mean; while the summation over j is carried out in a way that results in only the neighboring values $j \in J_i$ being included. The weight matrix (w_{ij}) defining the structure of the neighborhood for this analysis utilizes first-order queen contiguity, and $w_{ij} = 1$ if tracts i and j share a border (otherwise $w_{ij} = 0$). Thematic maps were created with ArcGIS Pro 2.8.0.

VII. RESULTS

CV Analysis

The reliability of the SVI variables varied across variables. Table 2 shows the 2018 CDC/ATSDR SVI variables with the highest percentage of census tracts with reliable CVs in metro, and non-metro areas. Table 3 shows the 2018 CDC/ATSDR SVI variables with the highest percentage of census tracts with unreliable CVs in metro, and non-metro areas. A chart showing the relationship between the CV and the SVI variable estimate was created for each of the 15 variables in the CDC/ATSDR SVI (Figures 10 – 24). At one end of the reliability spectrum were variables such as Per Capita Income, where most census tracts (88%) were found to have CV indicating their estimates were reliable (Figure 12). At the center of the reliability spectrum was Percent Minority (Figure 18), where 30% of census tracts had a CV indicating their estimates were reliable, 39% of tracts had a CV indicating their estimates were moderately reliable, and 31% of census tracts had a CV that indicated their estimates were unreliable. On the other end of the reliability spectrum were variables such as Crowding (Figure 22), where only 18% of tracts had a CV that indicated their estimates were reliable and 65% of census tracts were found to have a CV that indicated their estimates were unreliable.

Table 2. 2018 SVI variables with the highest percentage of reliable CVs, metro, and non-metro areas.

	Variables with the Highest Percentage of Reliable CVs in Metro Areas	Variables with the Highest Percentage of Reliable CVs in Non-Metro Areas
SVI Variables	Per Capita Income (88%), Aged 17 or Younger (57%), Mobile Homes (54%), Aged 65 or Older (52%), Group Quarters (41%)	Per Capita Income (89%), Aged 65 or Older (74%), Aged 17 or Younger (64%), Group Quarters (42%), Multi-Unit Structures (36%)

Table 3. 2018 SVI variables with the highest percentage of reliable CVs, metro, and non-metro areas.

	Variables with the Highest Percentage of Unreliable CVs in Metro Areas	Variables with the Highest Percentage of Unreliable CVs in Non-Metro Areas
SVI Variables	Crowding (63%), Speaks English "Less than Well" (56%), No Vehicle (43%), Unemployment (39%), Group Quarters (31.8%)	Crowding (72%), Speaks English "Less than Well" (60%), Minority (59%), No Vehicle (47%), Unemployment (40%)

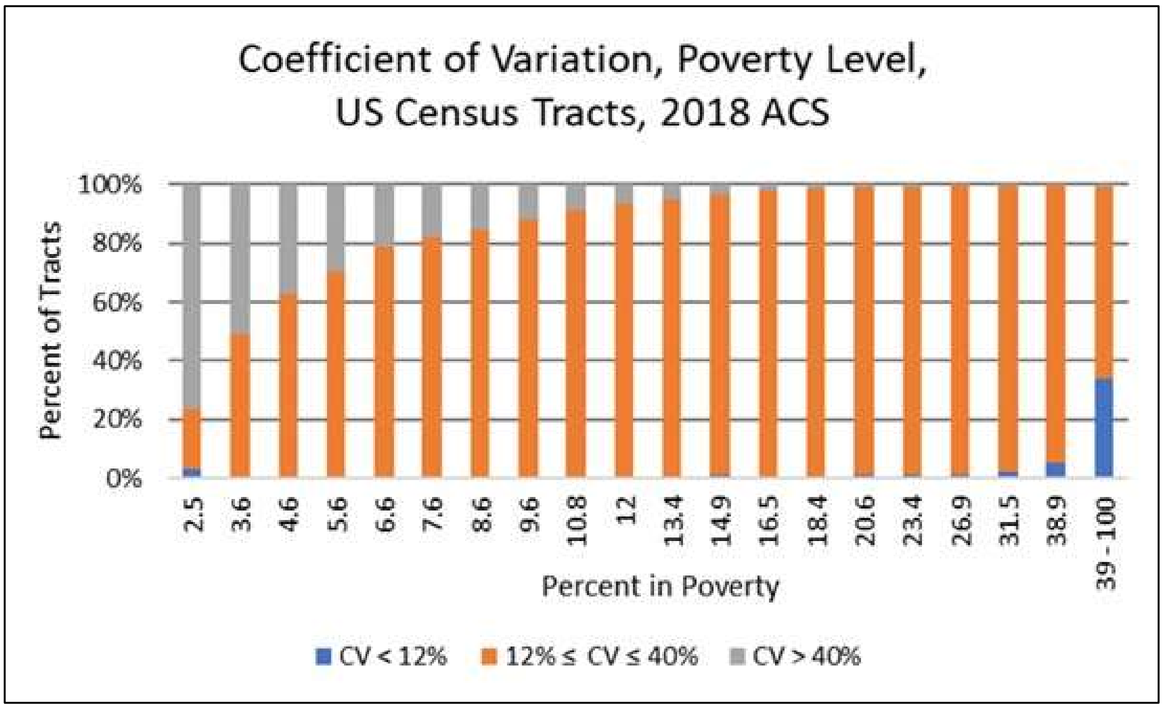


Figure 8. CV chart for 2018 poverty level.

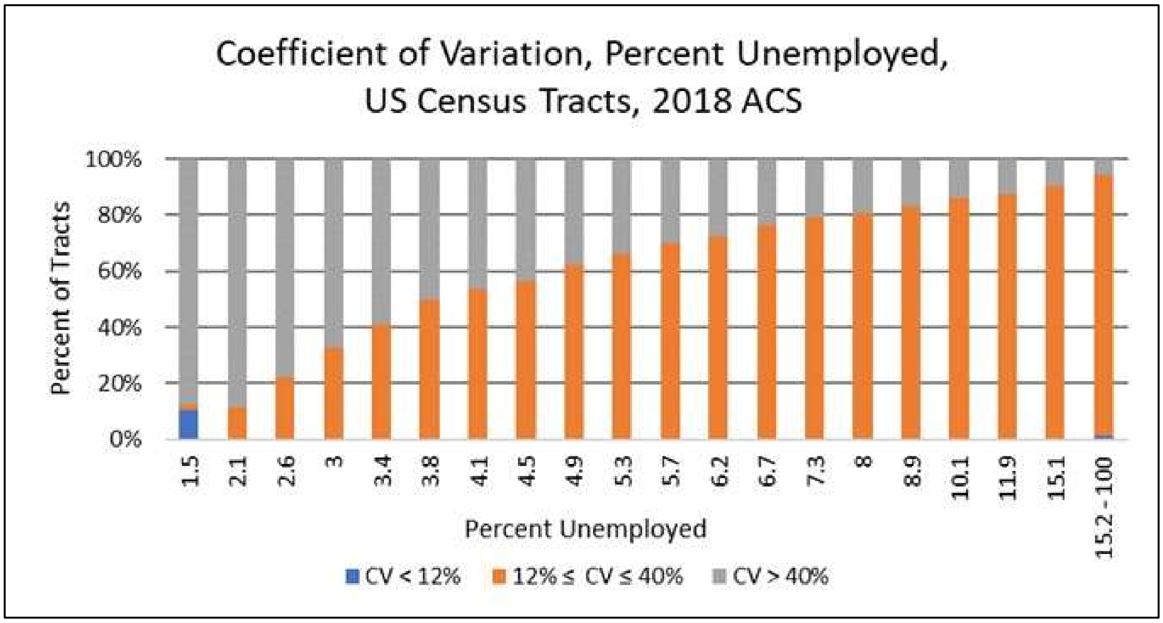


Figure 9. CV chart for 2018 percent unemployed.

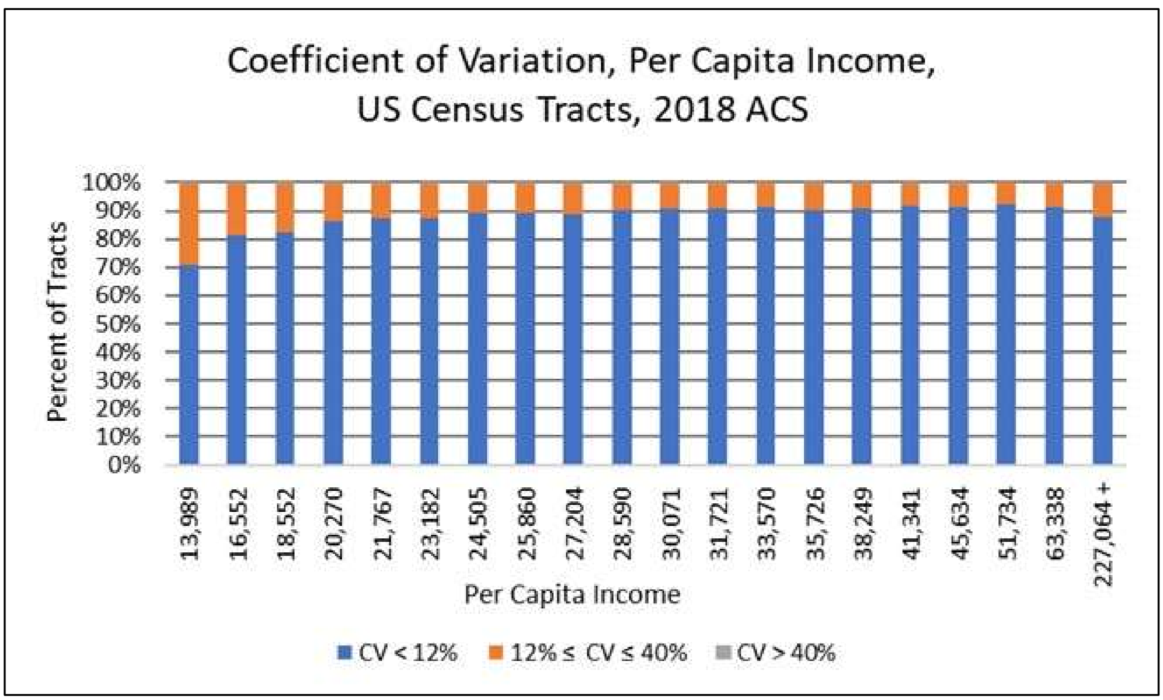


Figure 10. CV chart for 2018 per capita income.

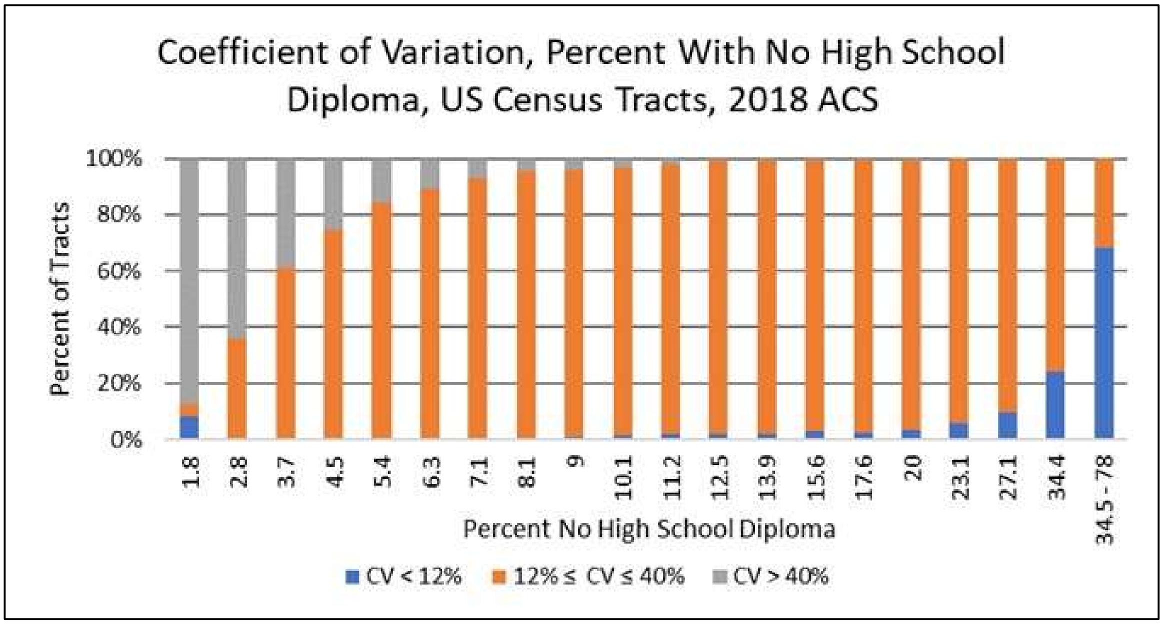


Figure 11. CV chart for 2018 no high school diploma.

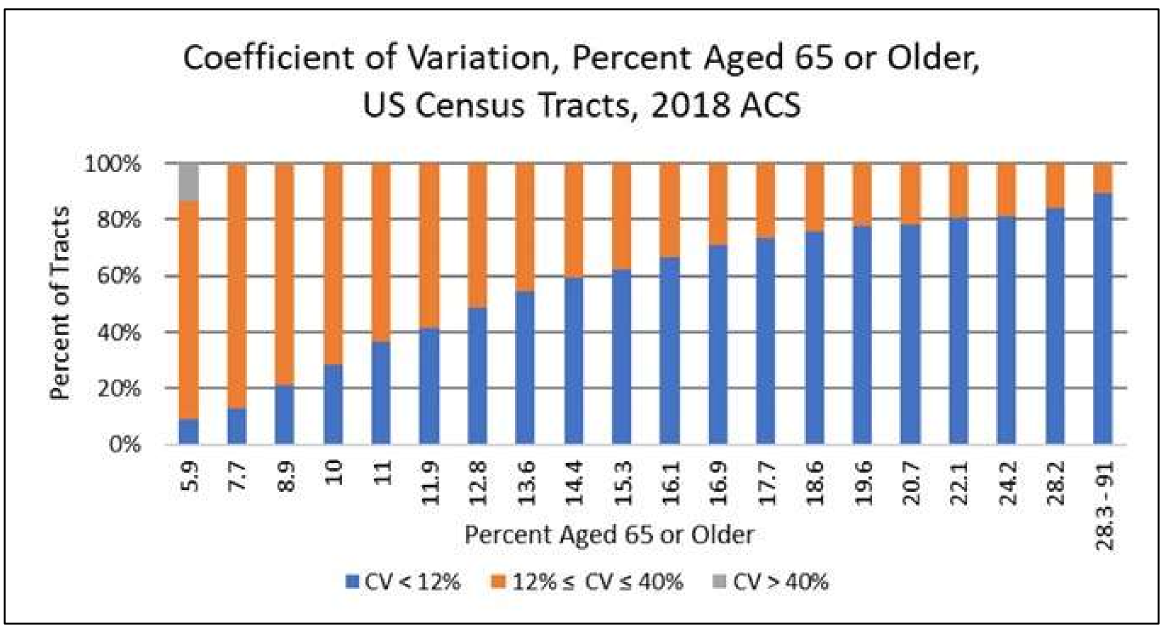


Figure 12. CV chart for 2018 percent aged 65 or older.

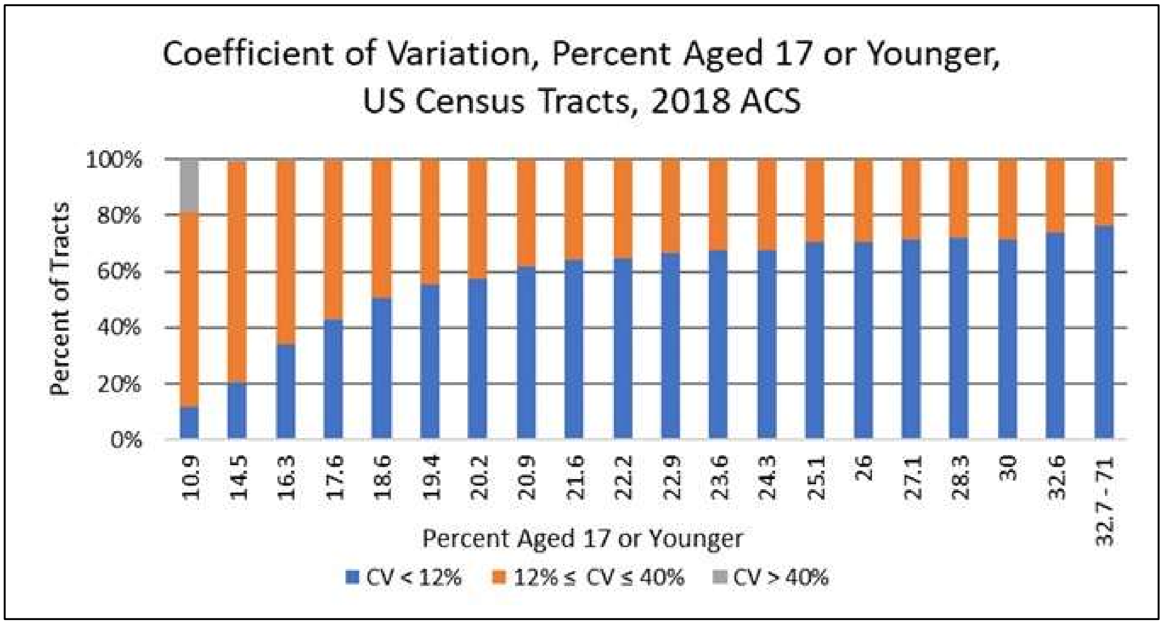


Figure 13. CV chart for 2018 percent aged 17 or younger.

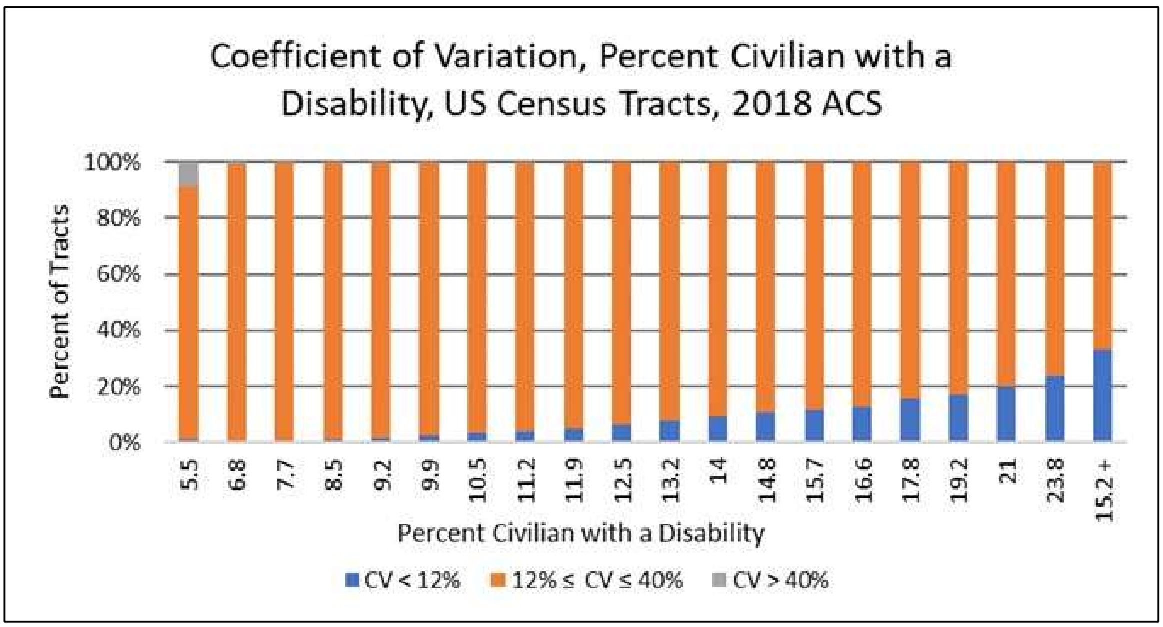


Figure 14. CV chart for 2018 percent civilian with a disability.

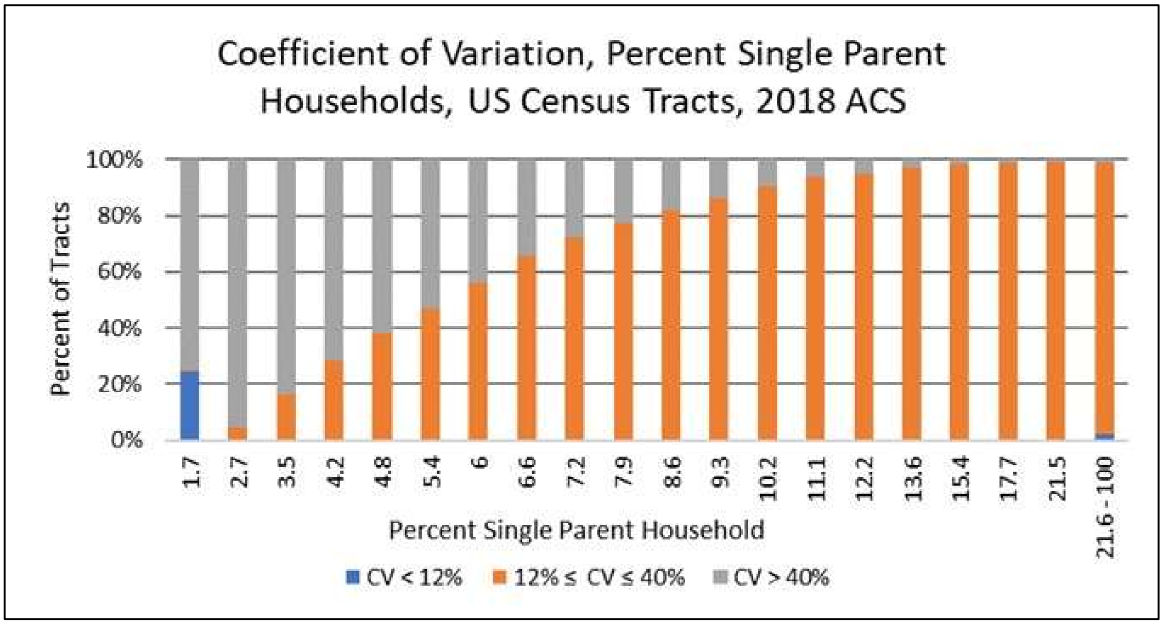


Figure 15. CV chart for 2018 percent single parent households.

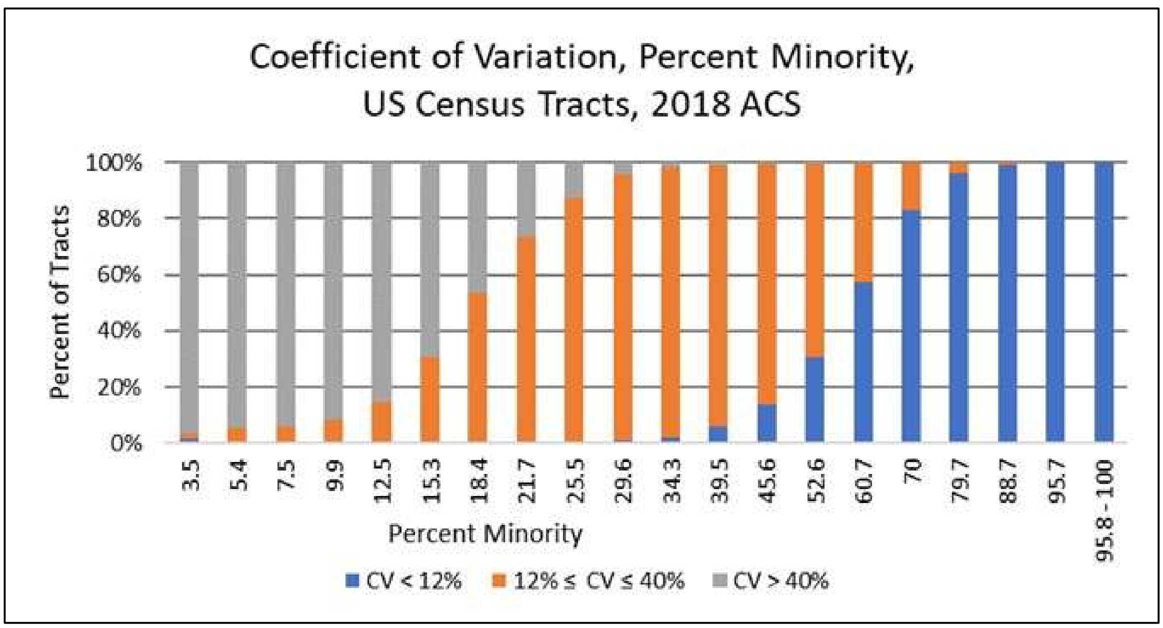


Figure 16. CV chart for 2018 percent minority.

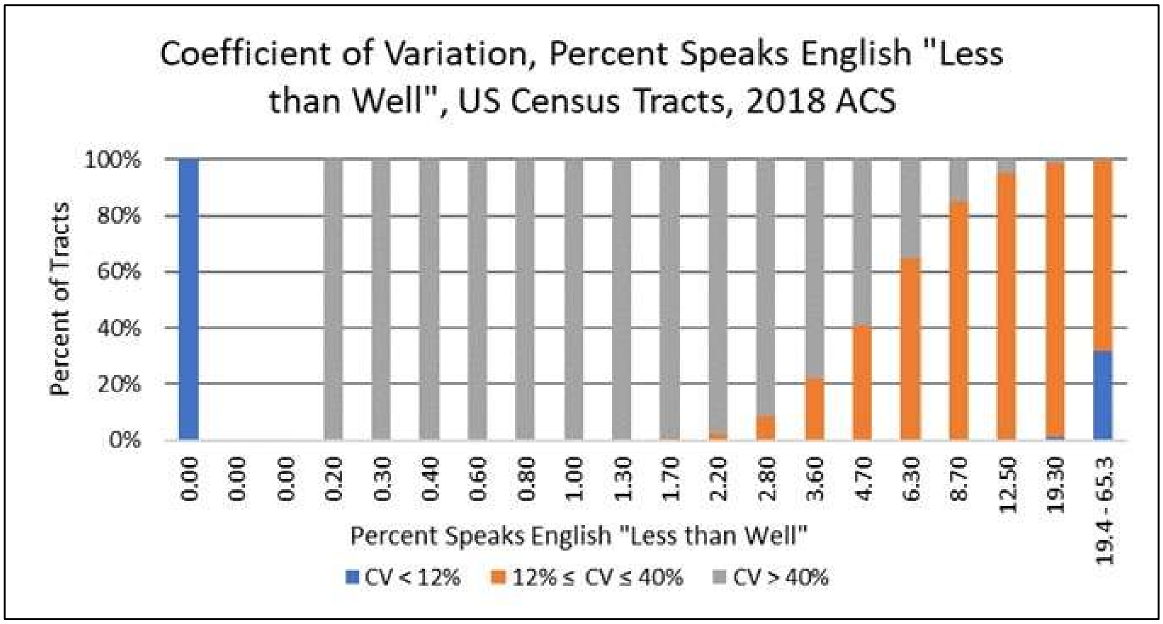


Figure 17. CV chart for 2018 percent speaks English "less than well".

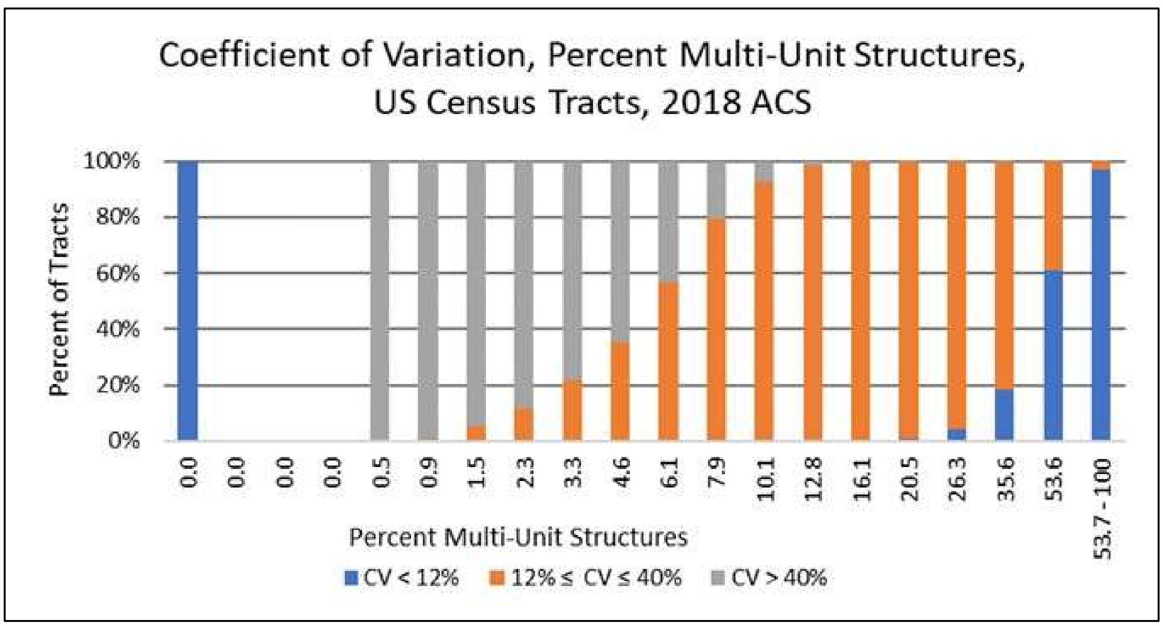


Figure 18. CV chart for 2018 percent multi-unit structures.

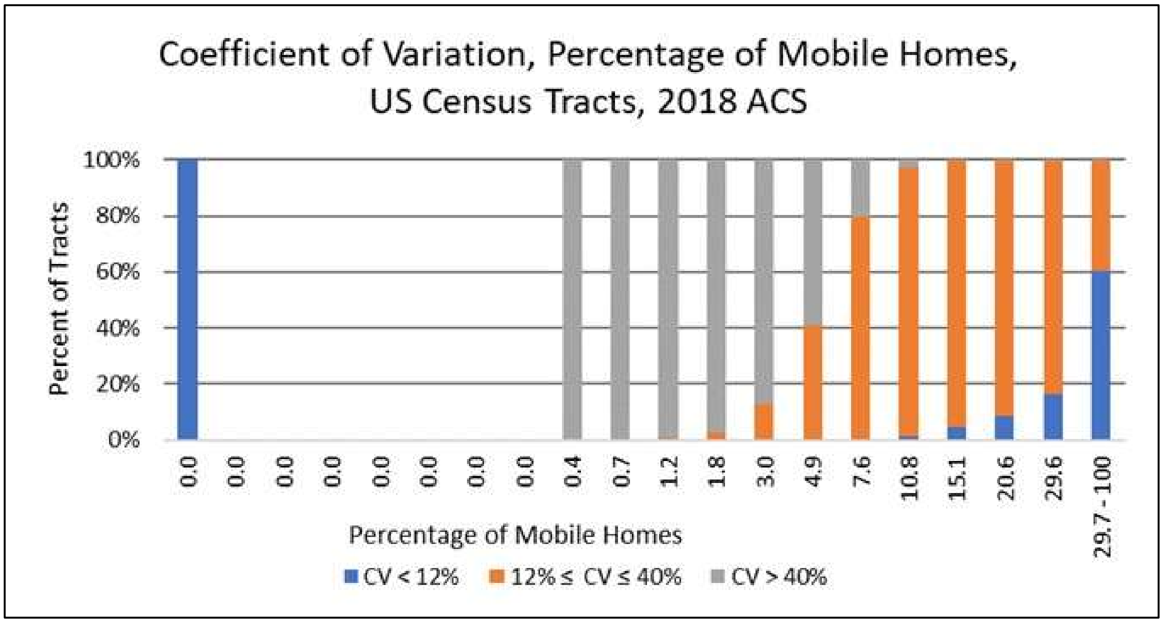


Figure 19. CV chart for 2018 percentage of mobile homes.

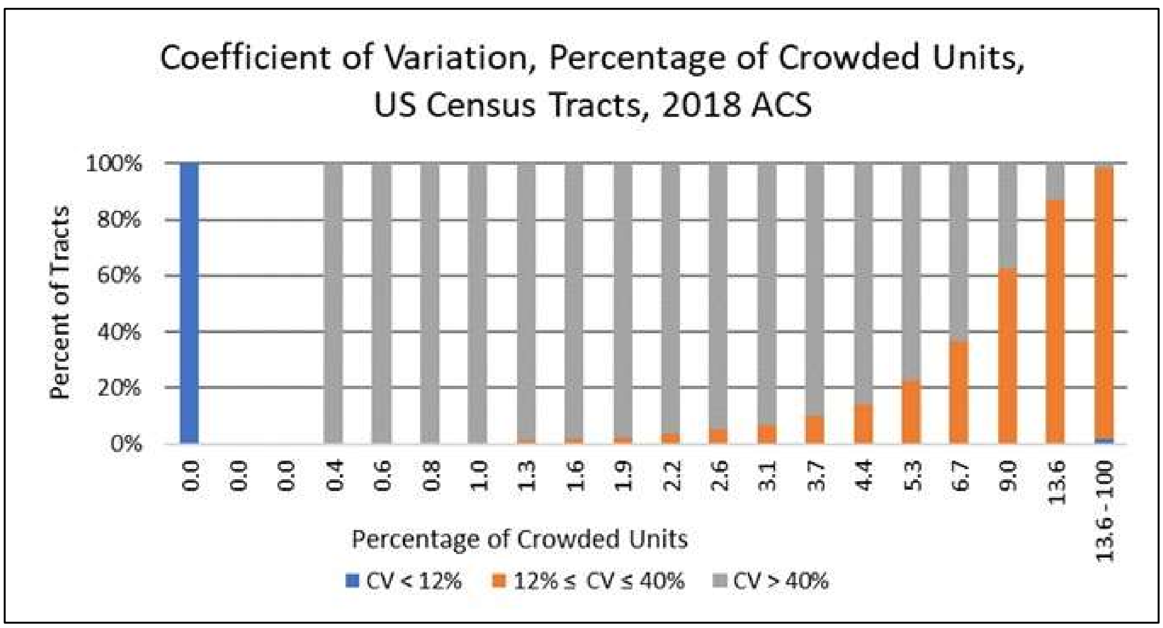


Figure 20. CV chart for 2018 percentage of crowded units.

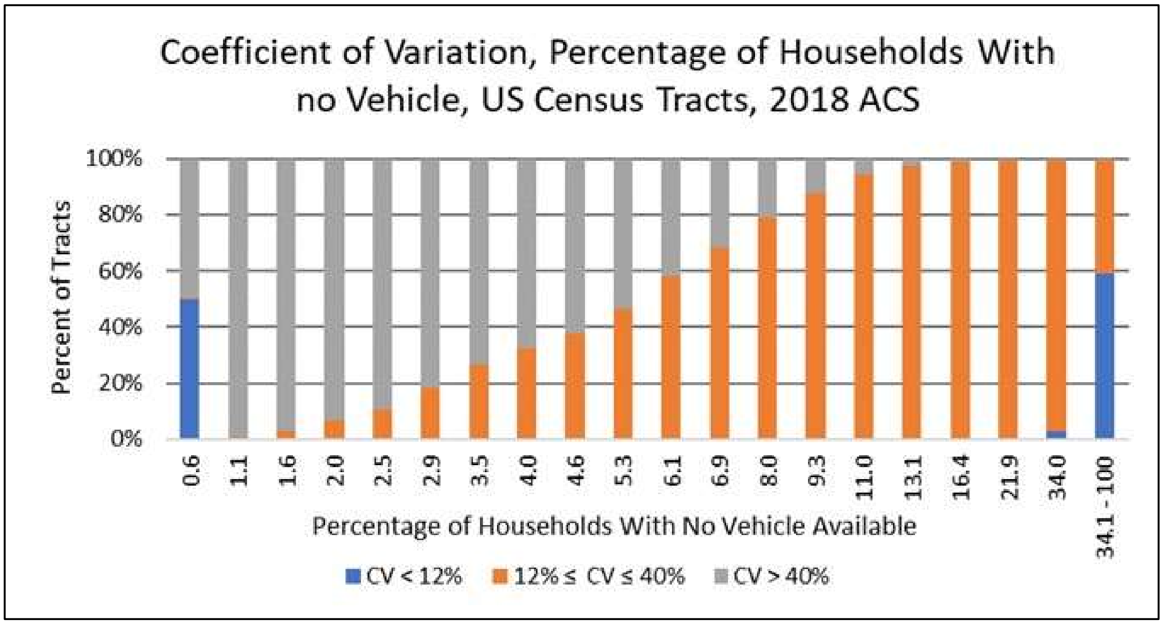


Figure 21. CV chart for 2018 percentage of households with no vehicle.

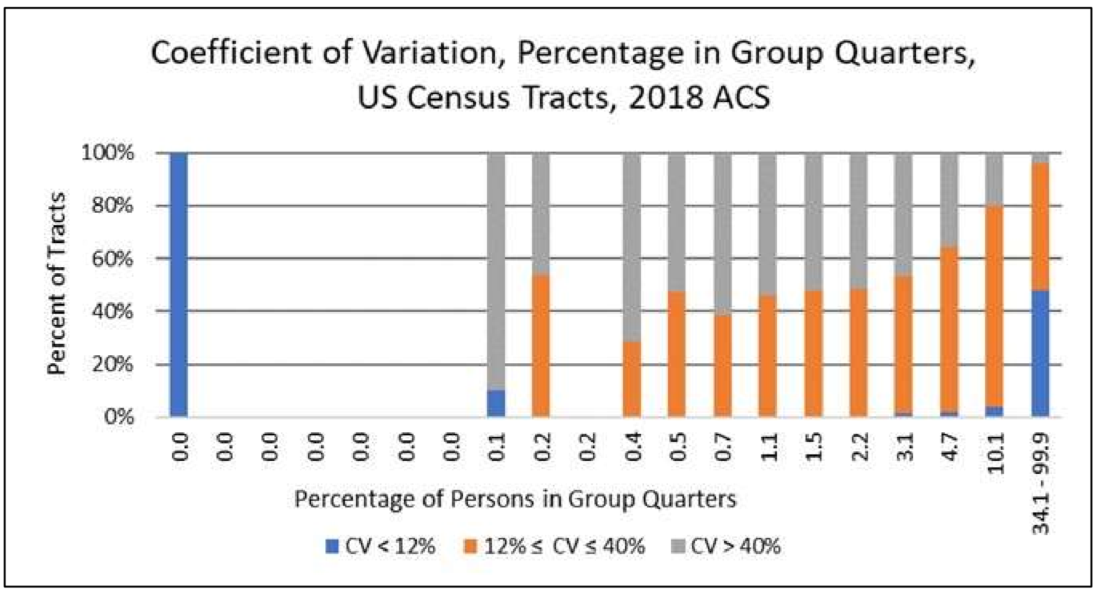


Figure 22. CV chart for 2018 percentage in group quarters.

The CVs for all 15 SVI variables were counted and grouped according to their reliability and whether they were within MSAs or not. The CV breakdown for all census tracts in all 15 variables (Table 4) shows similar proportions of reliable, moderately reliable, and unreliable census tracts across metro and non-metro areas, while the number of census tracts located within MSAs is much greater (Tables 5 and 6). When the proportions of each reliability category are charted for each variable (Figures 20 and 21), some variables show similarities between metro and non-metro areas (poverty, unemployed, per capita income, no high school diploma, aged 17 or younger, single parent households, no vehicle, and group quarters), while others did not (aged 65 or older, civilian with a disability, minority, speaks English "less than well", multi-unit structures, mobile homes, crowding).

Table 4. Overall CV reliability of metro and non-metro areas.

	Metro	Non-Metro
CV < .12	27%	28%
.12 ≤ CV ≤ .4	47%	44%
CV > .4	25%	28%
Total	100%	100%

Table 5. CV reliability of SVI variables in metro and non-metro areas, continued in Figure 22.

		Below Poverty	Unemployed	Income	No High School Diploma	Aged 65 or Older	Aged 17 or Younger	Civilian with a Disability	Single-Parent Households
Metro	CV < .12	1,386	301	48,585	3,520	28,914	31,380	2,512	854
	.12 ≤ CV ≤ .4	44,745	33,568	6,588	43,110	25,800	23,190	52,295	36,837
	CV > .4	9,089	21,351	47	8,590	506	650	413	17,529
Non-Metro	CV < .12	450	173	15,007	1,318	12,511	10,776	4,293	228
	.12 ≤ CV ≤ .4	14,992	9,937	1,933	14,392	4,368	6,024	12,574	11,161
	CV > .4	1,511	6,843	13	1,243	74	153	86	5,564

Table 6. CV reliability of SVI variables in metro and non-metro areas.

		Minority	Speaks English "Less than Well"	Multi-Unit Structures	Mobile Homes	Crowding	No Vehicle	Group Quarters
Metro	CV < .12	19,559	8,679	15,481	29,781	10,048	3,798	22,675
	.12 ≤ CV ≤ .4	23,046	15,870	25,289	9,278	10,365	27,410	15,001
	CV > .4	12,615	30,671	14,450	16,161	34,807	24,012	17,544
Non-Metro	CV < .12	1,815	5,291	6,169	4,005	2,622	577	7,160
	.12 ≤ CV ≤ .4	5,081	1,463	4,100	10,177	2,103	8,379	4,720
	CV > .4	10,057	10,199	6,684	2,771	12,228	7,997	5,073

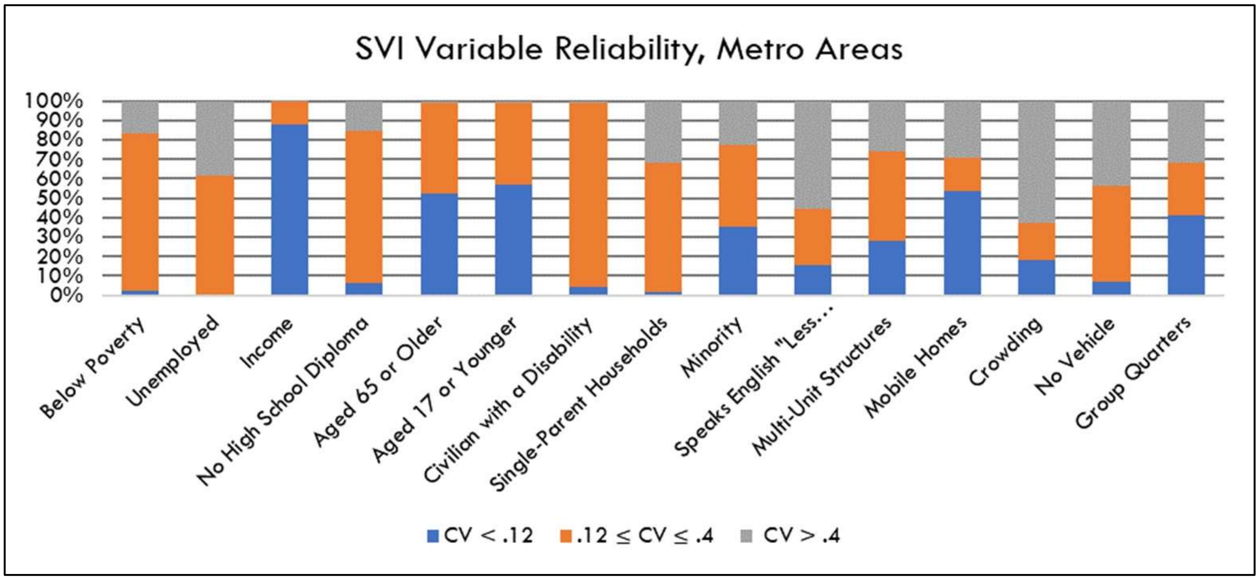


Figure 23. SVI variable reliability, metro areas

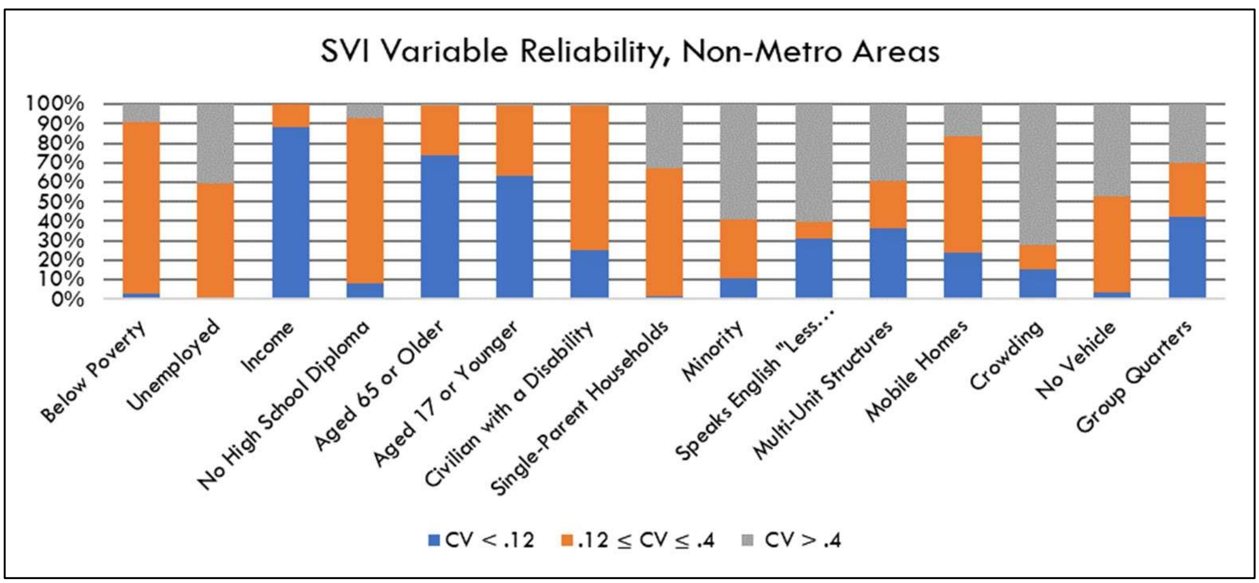


Figure 24. SVI variable reliability, non-metro areas

Overall Vulnerability Classification

Discriminant analysis provides a boundary-based statistical model for categorical outcome prediction as a function of a series of (continuous) predictor variables. The canonical plot (Figure 25) is a biplot where the axes are the first two canonical variables, which define the two dimensions providing maximum separation among the groups. The Score Summaries table (Figure 26) shows an overview of the discriminant scores, including the Entropy R^2 value, a measure of fit. The canonical plot visualizes the results in the score summary. A better fit is indicated by larger values, with an Entropy R^2 value of 1 indicating that the classifications were perfectly predicted. With an Entropy R^2 value of 0.02867, the results indicate little to no capacity for the CVs of the variable estimates to predict whether the overall vulnerability of a census tract may be over counted, undercounted, or remain the same. A discriminant analysis was also carried out on the variable estimates, the estimate percentages, and the CV of the estimates. All produced similar results. The canonical plot and score summaries tables for the results of these discriminant analyses are available in Appendix C.

Besides the incapacity of the uncertainties in the underlying data to predict the reliability of a census tract's vulnerability classification, the results of this analysis raise further questions regarding the reliability of the vulnerability classification of many census tracts in the first place. Figure 27 shows a map illustrating how the MOE may impact the CDC/ATSDR vulnerability categorization of several census tracts in the Austin, TX Metropolitan Statistical Area. Census tracts where the vulnerability classification may be overcounted can be found along the I-35 corridor, on both sides of the interstate. Census tracts where the vulnerability classification may be undercounted can be found more frequently in the western portions of Hays and Travis County.

With the vulnerability classification of 31,148 of 72,173 census tracts remaining unchanged at the national level, only a potential 43.2% of all census tracts in the CDC/ATSDR SVI may have reliable vulnerability classifications. This means that 56.8% of census tracts in this index may possibly have unreliable vulnerability classifications. This is a serious issue for the vulnerable people who miss out on resources because they were under counted, or because some other place was over counted, and that location received the resources instead. Areas with diverse populations may also contain disparities in wealth, which complicates the per-capita income variable. In similar ways, other variables may be unable to capture the nuances within the phenomenon they measure. The results of this analysis suggest that tools (such as the CDC/ATSDR SVI) derived from potentially unreliable data (such as the ACS) may be more sensitive to the uncertainties in the underlying data than many of their users may be aware of or able to comprehend.

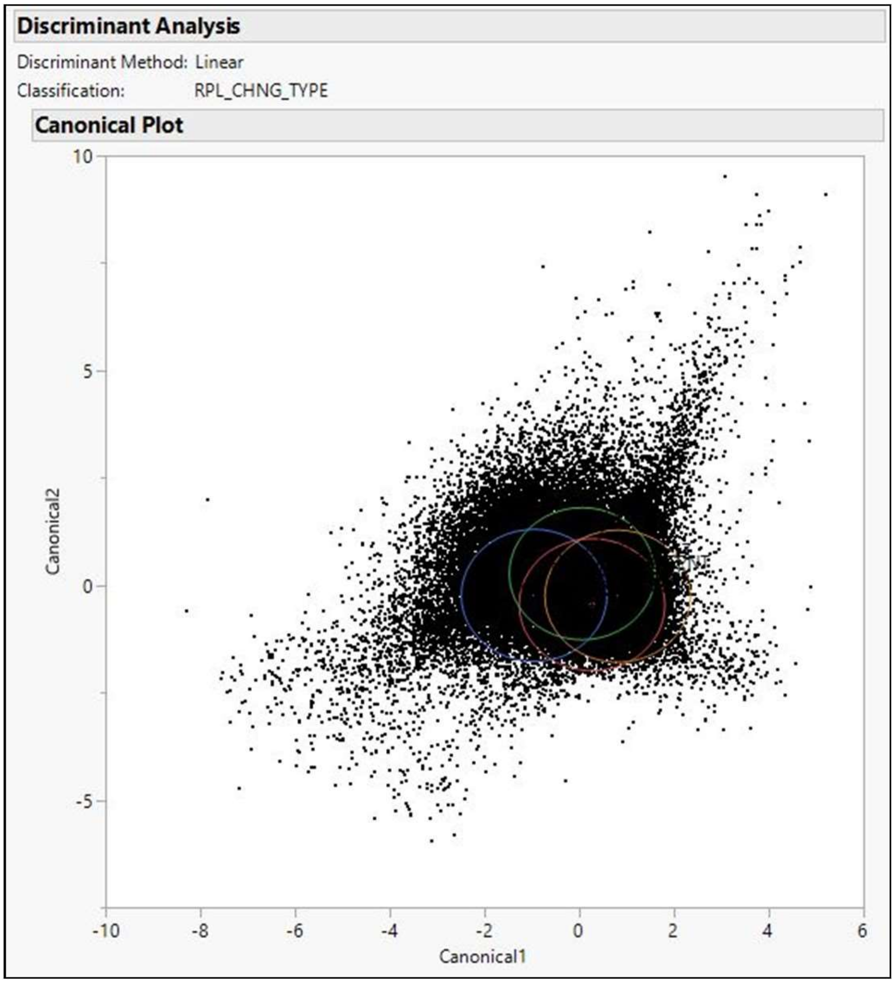


Figure 25. Discriminant analysis canonical plot.

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	72173	37983	52.6277	0.02867	168974

Training

Actual	Predicted Count			
RPL_CHNG_TYPE	B	N	O	U
B	1294	409	514	1130
N	5641	10529	6230	8748
O	3256	2655	10337	1373
U	4403	2974	650	12030

Groups

RPL_CHNG_TYPE	Count
B	3347
N	31148
O	17621
U	20057

Figure 26. Discriminant analysis score summaries

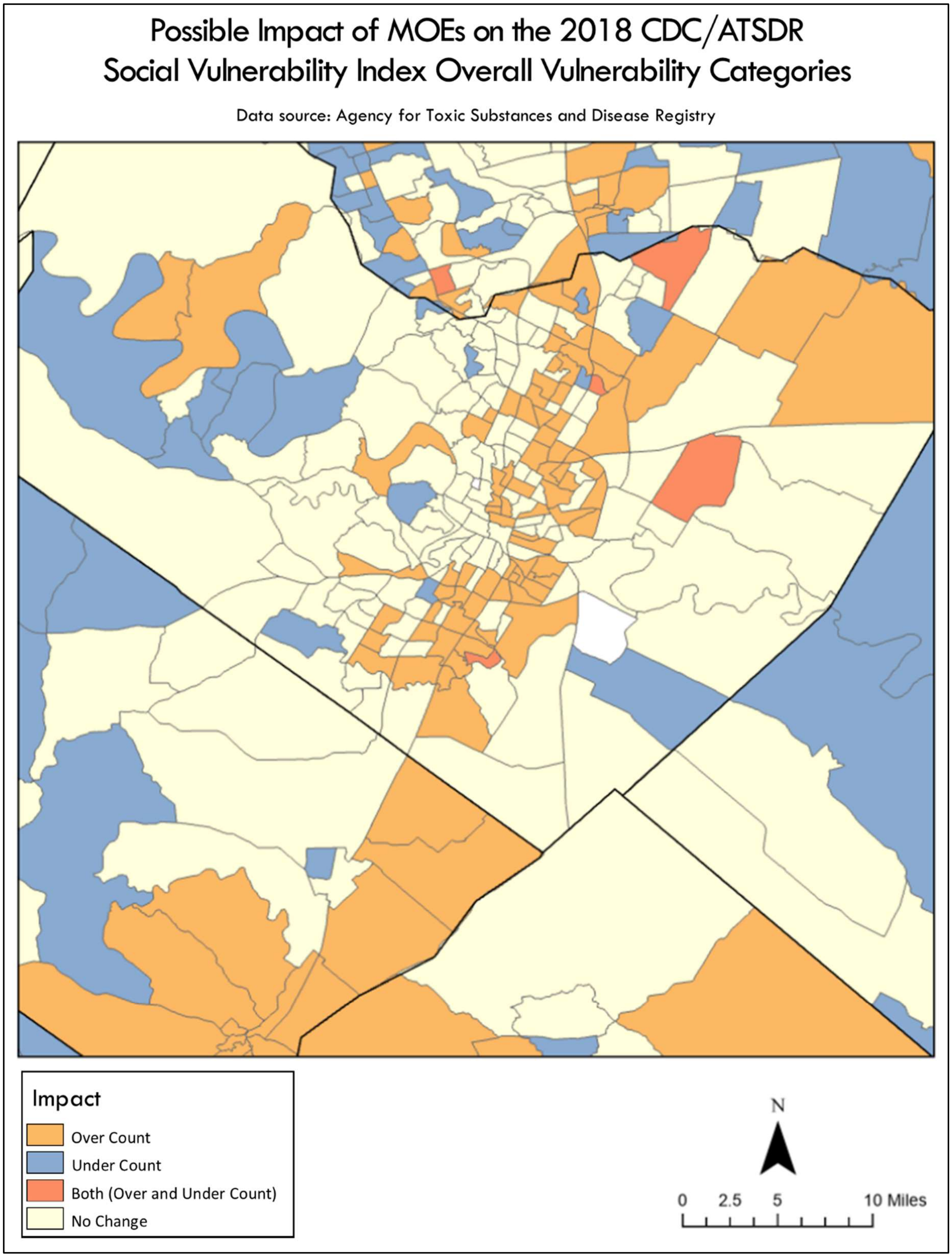


Figure 27. Possible Impact of MOEs on the 2018 CDC/ATSDR Social Vulnerability Index Overall Vulnerability Categories.

Cluster Analysis

A univariate local Moran's I analysis of each SVI variable's CV produced a Moran scatterplot, a LISA cluster map, and a LISA significance map. The Moran scatterplot visualizes the local Moran statistic, illustrating the relationship between the value of the CV at each census tract and the average value of the CV in neighboring census tracts. The LISA cluster maps show tracts with high CV values whose neighbors also have high CV (high-high) values in red, tracts with low CV values whose neighbors also have low CV values (low-low) in dark blue, tracts with low CV values whose neighbors have high CV values in light blue (low-high), tracts with high CV values whose neighbors have low CV values (high-low) in pink, and tracts that do not significantly cluster in light grey. Positive spatial autocorrelation is indicated by points in the high-high and low-low clusters, which correspond to quadrants 1 and 3 in the Moran scatterplot, while negative spatial autocorrelation is indicated by points in the low-high and high-low clusters, which correspond to quadrants 2 and 4 of the Moran scatterplot. Significance is assessed using a pseudo p-value which is visualized in the LISA significance maps. In these maps, darker shades of green indicate higher significance.

By far, most tracts for each variable were found to have pseudo p-values greater than $p = 0.05$, which are considered to have no significance. Crowding had the highest number of tracts with no significance (61,934 of 72,173). Near the middle of the range of the number of tracts with no significance was per capita income, with 54,649 tracts. Percent minority had the lowest number of tracts with no significance, with 44,757. Of the census tracts with pseudo p-values greater than $p = 0.05$, Moran's I range from 0.508 to -0.98. Within this subset, only one of the high-high categories (Aged 17 or Younger with a Moran's I of 0.208, compared to a Moran's I of 0.146 for the rest of the tracts) was found to have a Moran's I within the top quartile of the range. Meanwhile, 8 of the low-low categories had a Moran's I within the top quartile of the range. However, one of these low-lows (No High School Diploma with a Moran's I of 0.282) was found to have a lower Moran's I value when compared to the rest of the census tracts for that variable (which had a Moran's I of 0.318). When comparing the Moran's I of the tracts within the selected category to the Moran's I of the rest of the tracts, the low-low category (consisting of 11 of the 15 SVI variables) had greater value, compared to 10 of 15 SVI variables for the low-lows. These measures indicate that the strongest spatial autocorrelation tends to exist among census tracts within the low-low categories, that is census tracts with low CV values whose neighbors also have low CV values.

This section includes a map of the CV (Figures 27, 30, and 33), a LISA cluster map (Figures 28, 31, and 34), and a LISA significance map (Figures 29, 32, and 35) for each of the three variables highlighted previously (per capita income, percent minority, and crowding). These variables were selected to illustrate the kinds of spatial autocorrelation that may occur in variables with different proportions of reliable or unreliable census tracts. They were also selected to illustrate certain issues with the CV that will be discussed in the following section. Reference maps for each of the three variables, in addition to Moran scatterplots, are available in Appendix D.

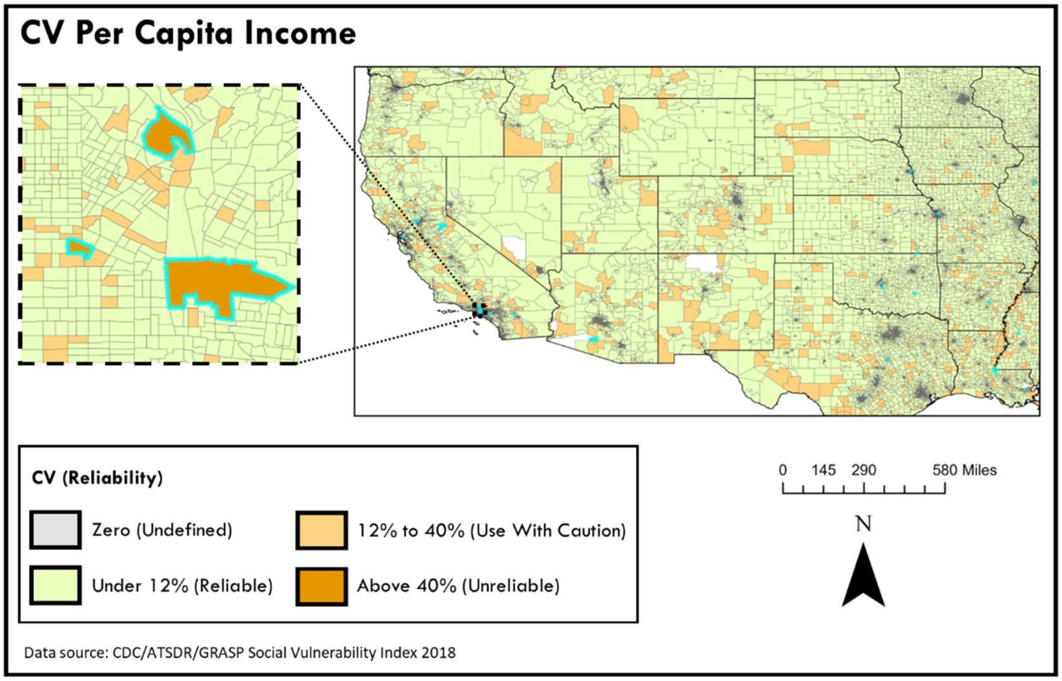


Figure 28. CV Per Capita Income. Census tracts with unreliable CVs highlighted.

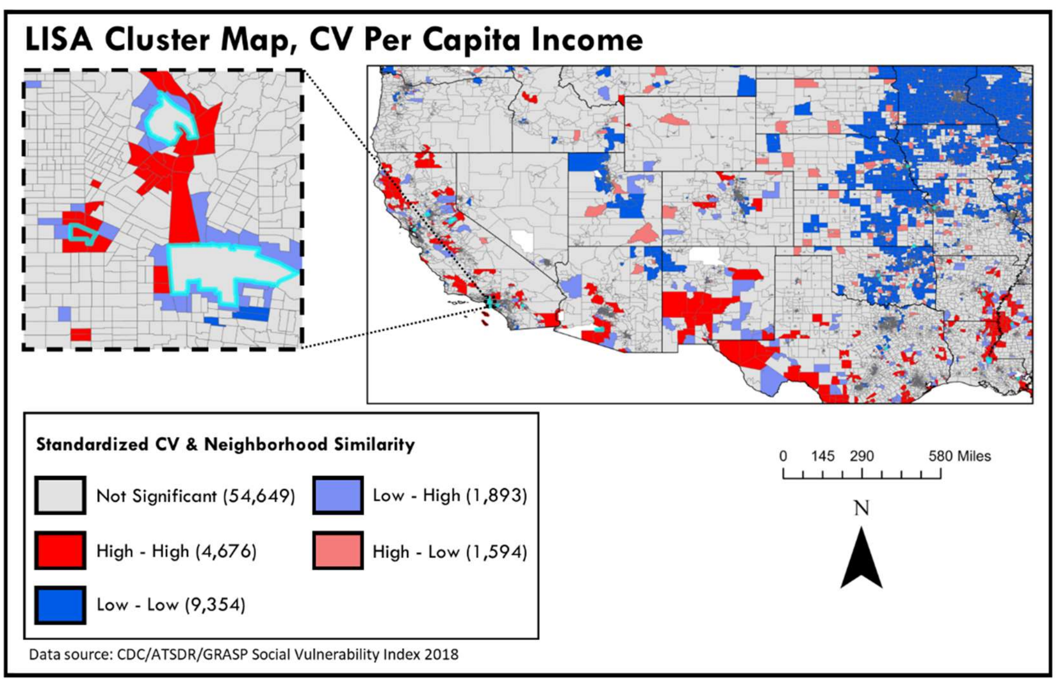


Figure 29. LISA Cluster Map, CV Per Capita Income. Census tracts with unreliable CVs highlighted.

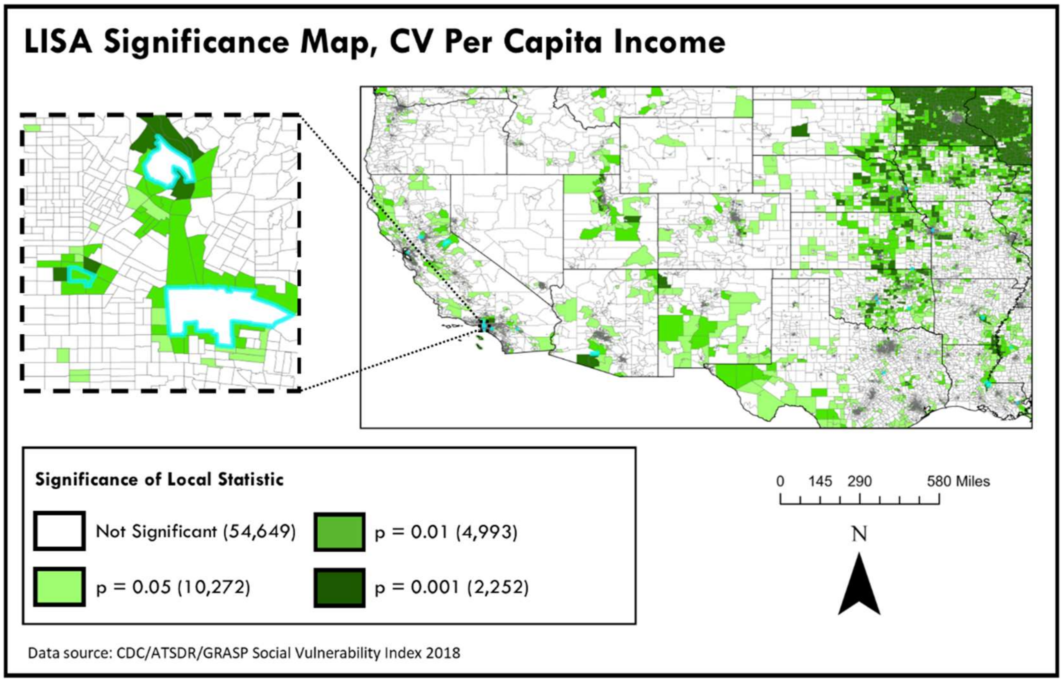


Figure 30. LISA Significance Map, CV Per Capita Income. Census tracts with unreliable CVs highlighted.

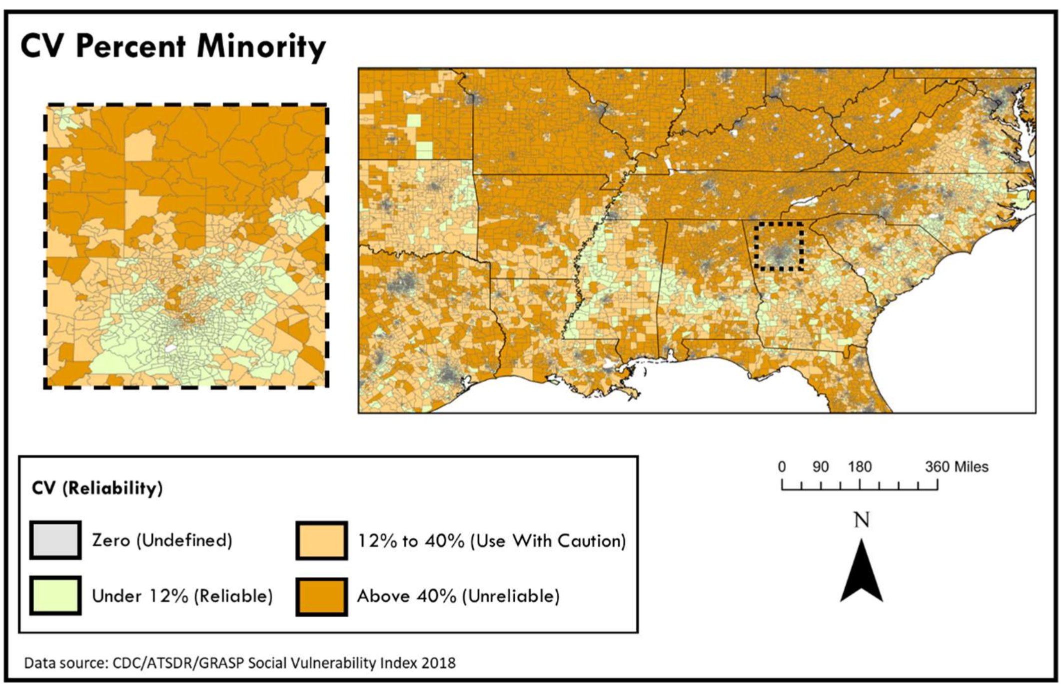


Figure 31. CV Percent Minority.

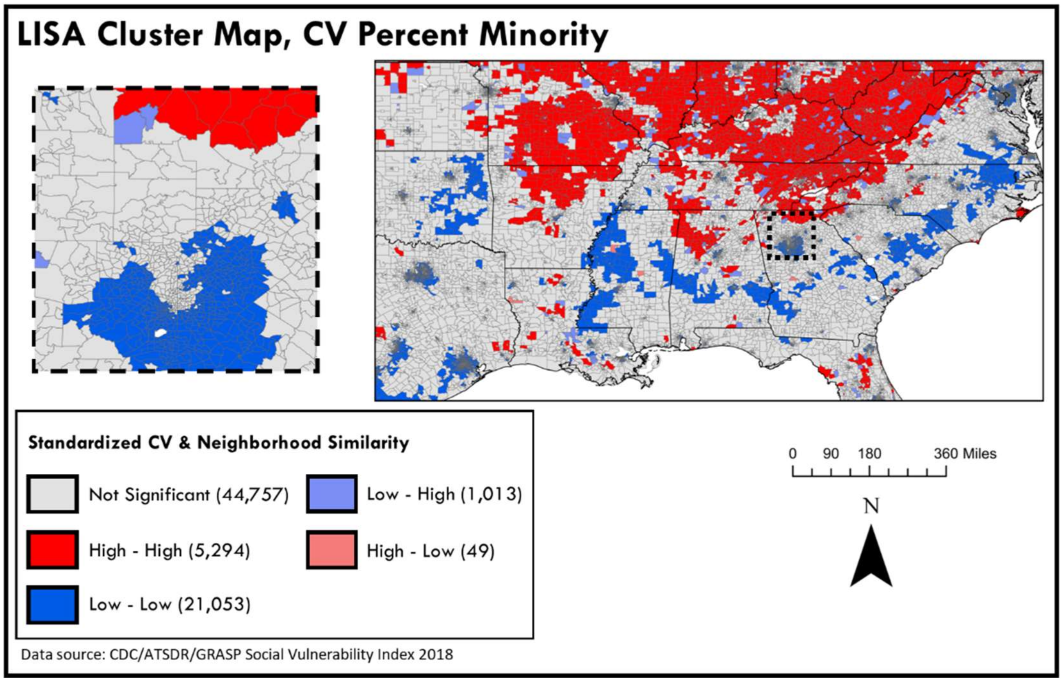


Figure 32. LISA Cluster Map, CV Percent Minority.

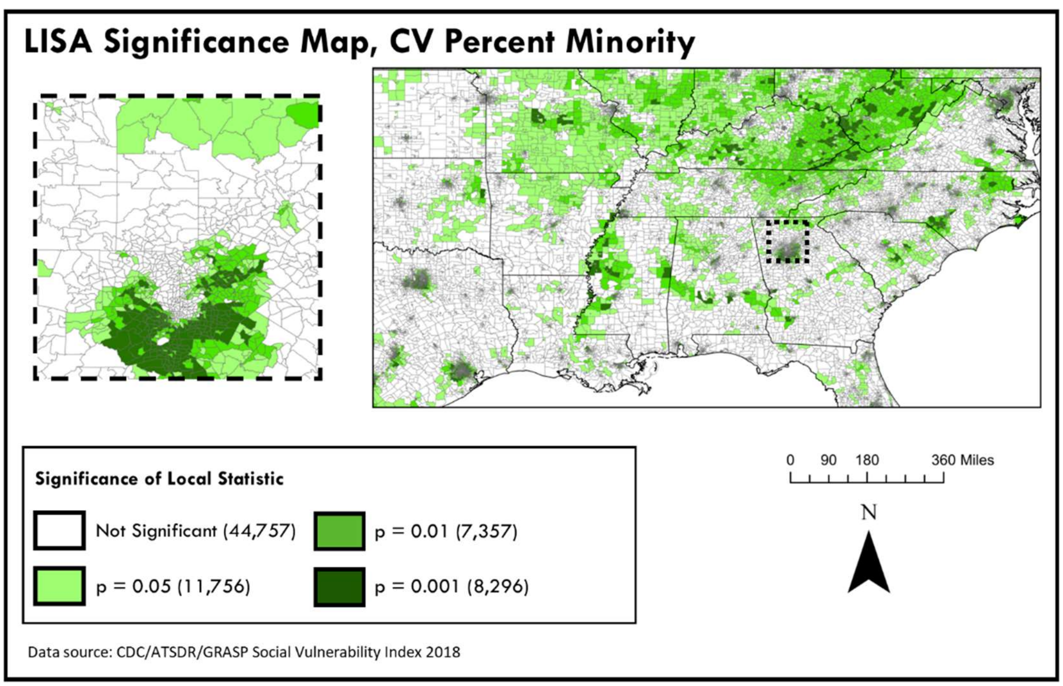


Figure 33. LISA Significance Map, CV Percent Minority

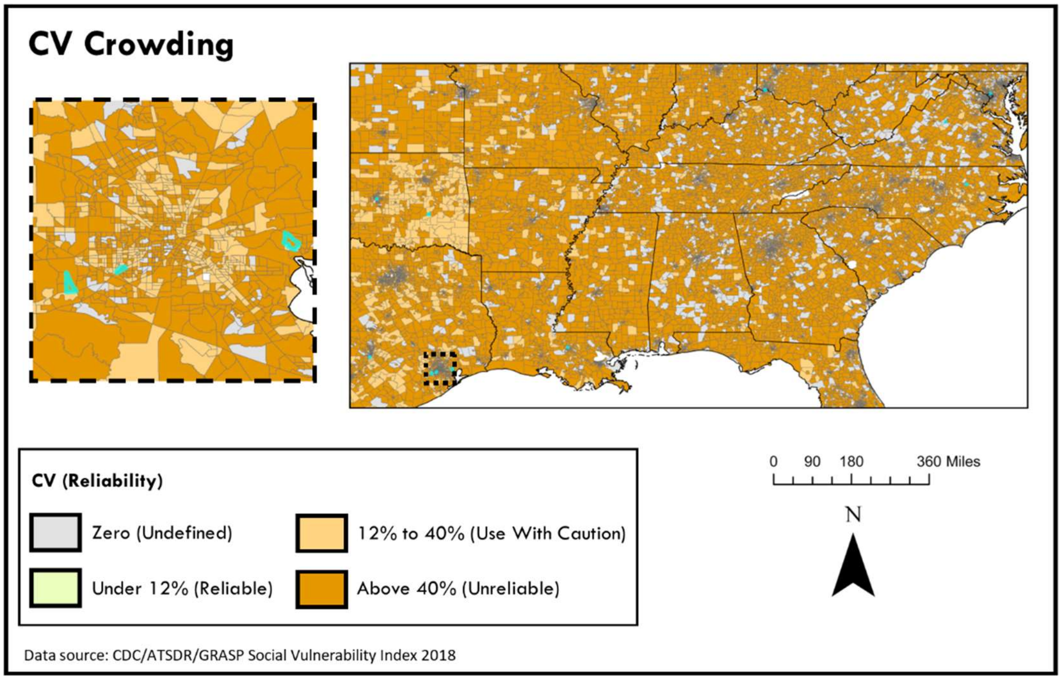


Figure 34. CV Crowding. Census tracts with unreliable CVs highlighted.

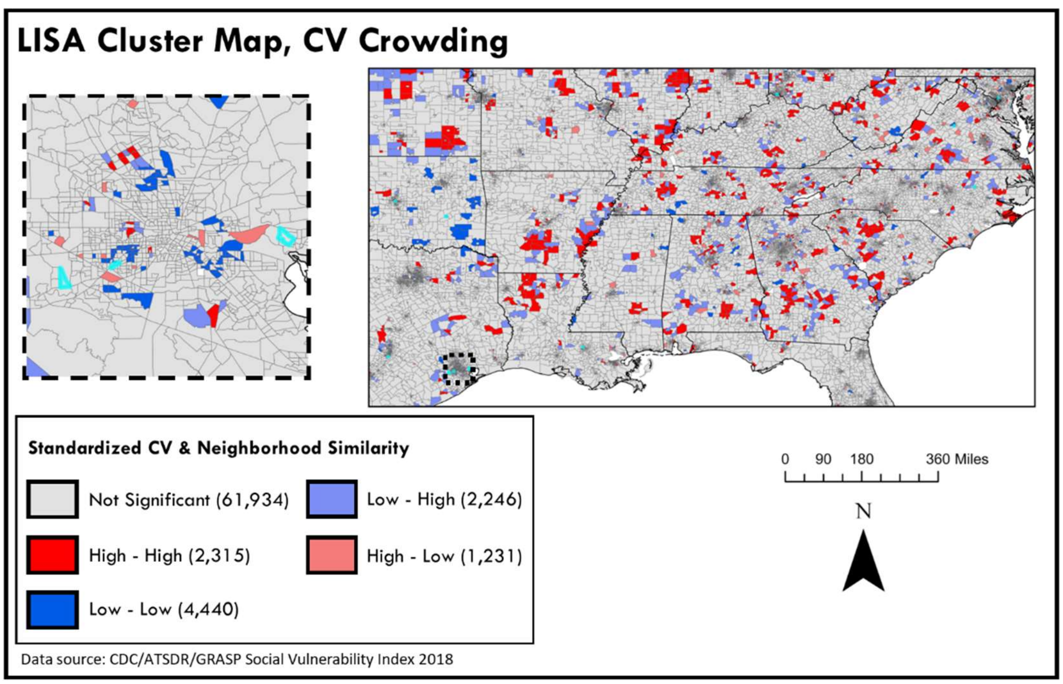


Figure 35. LISA Cluster Map, CV Crowding. Census tracts with unreliable CVs highlighted.

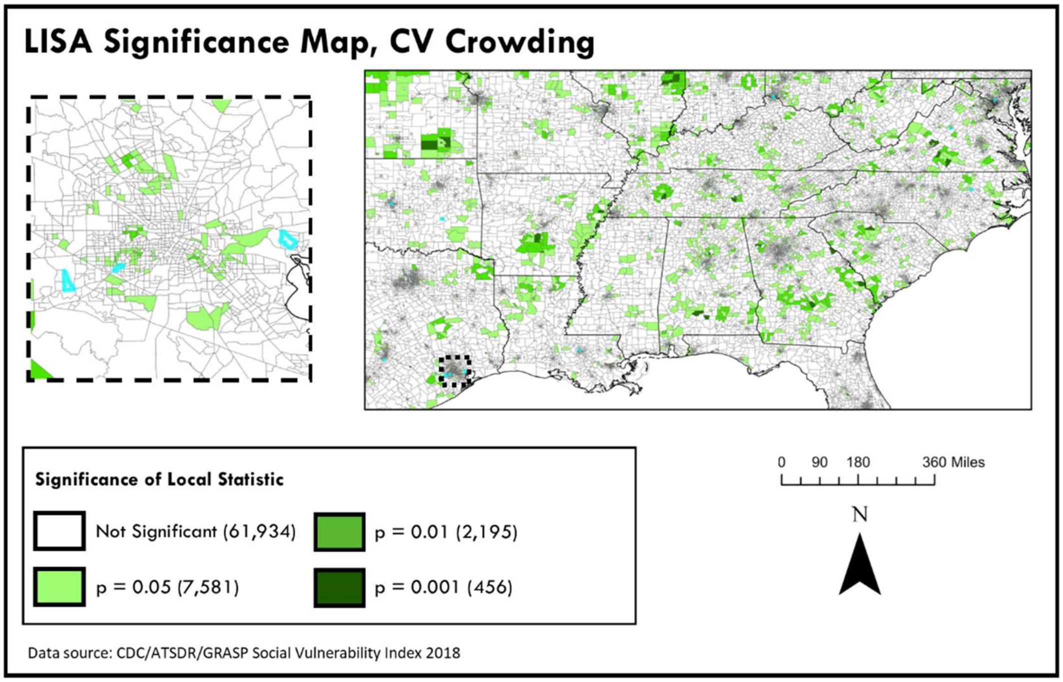


Figure 36. LISA Significance Map, CV Crowding. Census tracts with unreliable CVs highlighted.

VIII. DISCUSSION

Validating vulnerability indices has been a focus of previous research (Bakkensen et al. 2017, Rufat et al. 2019), as has evaluating indicator selection criteria (Spielman et al. 2020). Evaluating the uncertainties in the data underlying the CDC SVI has received attention (Tate 2013), but not at the national scale. A major issue with statistically derived measurements for topics so tightly bound to human lives and the social issues woven throughout is that they reduce a complex circumstantial outcome to a number on a spreadsheet or a color on a map. Although Fekete (2012) succinctly stated that "a map represents space and observations; it does not aim at reproducing reality," some social vulnerability maps may even go so far as to distort the observations within the space they represent. For example, some social vulnerability rankings have been found to show counterintuitive effects such as decreasing vulnerability scores when the unemployment percentage increased (Spielman et al. 2020). Complicating matters is the fact that aggregation and categorization reduce differences and hide outliers, a major issue in dense, diverse locales such as gentrifying neighborhoods in urban environments. Social vulnerability is also an important factor in the designation of Federal Opportunity Zones, where tax breaks meant to revitalize poor communities sometimes end up supporting the development of luxury facilities whose benefits to poor or working-class communities are questionable. An example of this is the Rybovich superyacht marina in West Palm Beach (Edwards et al. 2019), a resort built to accommodate members of the global 0.1% while their vessels (which can cost over $100 million) are being serviced.

For these and other reasons users must exercise caution when depending on tools such as the CDC/ATSDR SVI while planning and implementing policy or interventions intended to reduce vulnerability. There may also be local data, not based on surveys, that can help provide a richer context to the ACS or SVI data. For example, the Community Advancement Network's (CAN) We CAN! ATX dashboard provides data regarding the 5 principal service call categories to the local 211 service line from 2020 to 2022 (CAN 2023). Local stakeholders need to develop a better understanding of the types of uncertainties in the SVI data for their locality. To support local stakeholders in this task and to address the social risks inherent in tools such as the CDC/ATSDR SVI, further research can explore what local data may best help fill the gap(s) that remain in these tools because of the uncertainties in the underlying ACS data.

CV Analysis

Although the CV is a useful measure of the reliability of estimates originating from ACS data, there are certain cases where it may fall short. An important circumstance to consider are census tracts which have an SVI variable estimate value of zero. Any time a variable estimate is zero, it can be assumed that it may be an instance of an undercount of that variable. When the CV was calculated, the output for census tracts where the estimate was zero was a CV of zero, regardless of the magnitude of the MOE. This means that a hypothetical tract with a zero estimate and MOE of 1 would be grouped in the same class (reliable) as a hypothetical tract with a zero estimate and MOE of 1,000. A future analysis must take additional steps to sort the tracts with a CV of zero to distinguish between tracts with low magnitude MOES and tracts with high magnitude MOEs. One possible fix would be to assign a negligibly small value to census tracts where the estimate is zero, which would eliminate the issue of having a zero on the denominator and may produce more meaningful results. Minding the census tracts where the estimate is zero is important because these are tracts where the estimate may be potentially undercounted. Additionally, incorporating too many variables where too many estimates are zero runs the risk of making an index a zero-inflated model, that is one with frequent zero-valued observations.

When the CVs were mapped as part of the cluster analysis process, other issues were observed. In most cases, as the quantity of the phenomenon being measured increases, the CV tends to decrease. An example of this can be seen in Figure 31, mapping the CV of the percentage of the population who is a minority. Census tracts in the Deep South which have a greater percentage of the population belonging to minority groups will also tend to have lower CV values (and thus more reliable estimates). However, exceptions exist. For example, as observed with certain instances in the Crowding variable. Here, in some cases, census tracts where crowding is high (Appendix D11), also have high (unreliable) CVs (Figure 34). Similarly, just because the CV is high, one cannot assume that the estimate for that census tract is low. For example, Figure 28 shows three census tracts in the Long Beach-Anaheim MSA that have an unreliable per capita income estimate. When these census tracts are referenced in Appendix D1, we can see that each tract belongs in different income brackets with one of the tracts being from the second-to-highest income bracket. The census tract adjacent to this high-income, low reliability census tract that is also high-income (located on the west side of its northern boundary), in contrast, is in the reliable category.

A final observation worth mentioning here is how the map of the CV of some variables would sometimes show visible differences between states. For example, Oklahoma has lower CVs (and thus, generally, more reliable estimates) than neighboring states, to the extent that its outline is clearly visible in Figures 30 and 33. Perhaps this may be the result of how the ACS is administered or implemented within the state, as the starkness of the contrast in the reliability suggests there may be a statistical cause to the effect, although such inquiry is beyond the scope of this research.

Overall Vulnerability Classification

Classifying the overall vulnerability ranking into the four classes in the interactive SVI map on the CDC/ATSDR website may also potentially misrepresent the level of vulnerability in those census tracts if the MOES are considered. Each “variable” is a quantitative representation of a factor that plays a role in the daily life of people living in their various communities. Quantitative imprecision manifests as an over count or under count of the intensity of the factor being measured. This has an impact on the relative ranking for that variable, such that a census tract in the medium high overall vulnerability category may end up in the low-medium vulnerability category when the MOE is added and the result ranked, as well as in the high vulnerability category when the MOE is subtracted and the result ranked. Since community conditions are different and populations are not homogeneously distributed, the places where the SVI is unreliable are different in different places. The map in Figure 27 shows the locations where social vulnerability may be overcounted, undercounted, both undercounted and overcounted, or remain unchanged in parts of the Austin MSA. As mentioned, census tracts where the vulnerability classification may be overcounted can be found along the I-35 corridor, on both sides of the interstate, while census tracts where the vulnerability classification may be undercounted can be found more frequently in the western portions of Hays and Travis County.

Additionally, variables where many census tracts have an estimated value of zero (such as percent who speak English “less than well,” multi-unit structures, mobile homes, crowding, and group quarters) may risk misrepresenting how few vulnerable people may live in those census tracts, regardless of the CV value. These complexities may have played a role in the discriminant analysis showing that the CVs of the SVI variable estimates do not have much capacity to predict the type of uncertainty that may exist when classifying the overall vulnerability ranking into one of the four classes on the interactive SVI map on the CDC/ATSDR website. Further analysis at different geographic scale(s) may yield different results.

Cluster Analysis

It is important to consider that SVI variables with an estimate of zero and a nonzero MOE will be assigned a CV of zero when calculated with Excel. Relying on the CV as a measure of reliability is insufficient because tracts with a zero estimate end up with a CV of zero, so they get lumped into the "reliable" end of the CV range. This makes CV potentially unreliable, zeros must be accounted for and grouped with the unreliable tracts. It also affects the utility of GeoDa because the tracts with a CV of zero, which is only possible if the estimate is zero, were grouped into the low-low category, regardless of the size of their MOE.

It must be mentioned that the CDC/ATSDR SVI mitigates tracts with an estimated population of zero by removing these tracts from the ranking process. Tracts with unavailable values or where the value could not be calculated because of unavailable census data are not used for further calculations (ATSDR 2022). The results presented in this study suggest that mitigation steps should also be taken when census tracts with SVI variable estimates of zero are subjected to cluster analysis.

Limitations

This study focuses on a comparison of the relative reliabilities of the estimates with which the CDC SVI rankings are calculated, for all census tracts in the 50 US states and the District of Columbia. The principal limitation of this study is that the results and insights it will produce will only be comparable to the national-level CDC SVI rankings. This is because the CDC SVI rankings used in this study are the national-level comparisons provided by the ATSDR CDC. As such, this study is limited to a national-level comparison of relative vulnerability. The state-level index ranking provided by the ATSDR CDC is incompatible for comparison with the measures, maps, and analyses produced with this study. Therefore, caution must be exercised when implementing or applying any of the methods used in this study.

Other limitations to this study are related to the software that was used to analyze the ATSDR CDC SVI data. The cluster analysis was carried out using GeoDa, which eliminates isolates (polygons which lack neighbors). Although this means that certain types of census tracts, such as those located on islands with only one census tract, were not considered in the cluster analysis, the number of isolates was negligible (only 7 tracts were found to be isolates). These differences do not invalidate the results of this study, but caution must be exercised when comparing the results of this study to the results of others. In addition, this analysis was carried out using the version of Microsoft Excel available on the Microsoft 365 for Enterprise software suite. The CDC SVI Documentation notes: "When replicating SVI using Microsoft Excel or similar software, results may differ slightly from databases on the SVI website or ArcGIS Online. This is due to variation in the number of decimal places used by the different software programs. For purposes of automation, we developed SVI using SQL programming language. Because the SQL programming language uses a different level of precision compared to Excel and similar software, reproducing SVI in Excel may marginally differ from the SVI databases downloaded from the SVI website. For future iterations of SVI, beginning with SVI 2018, we plan to modify the SQL automation process for constructing SVI to align with that of Microsoft Excel." (ATSDR 2022).

IX. CONCLUSION

The susceptibility of social groups to the adverse impacts of natural hazards, which can include disproportionate death, injury, loss, or disruption of livelihood, is known as social vulnerability. Several tools, such as the CDC/ATSDR SVI have been developed that statistically derive measures of social vulnerability from demographic survey data. Statistically derived measurements for topics so tightly bound to human lives and the social issues woven throughout reduce complex circumstantial outcomes to numbers on a spreadsheet or colors on a map. This research consisted of an accuracy assessment of the CDC/ATSDR's Social Vulnerability Index, with the goal of improving our understanding of the types of limitations arising from the uncertainties of the underlying data.

The analysis made use of census tract level data where the CVs were calculated as a measure of uncertainty. The CVs were used to examine the quantitative characteristics and geographic patterns of the uncertainties. Different SVI variables have different proportions of census tracts with reliable or unreliable CVs. Some SVI variables showed similar proportions of the CV reliability categories for census tracts located within metropolitan areas and those located within non-metropolitan areas (poverty, unemployed, per capita income, no high school diploma, aged 17 or younger, single parent households, no vehicle, and group quarters), while others (aged 65 or older, civilian with a disability, minority, speaks English "less than well," multi-unit structures, mobile homes, crowding) showed different proportions in the CV breakdowns for metro and non-metro census tracts. The most reliable variables in metro and non-metro areas included income, aged 17 or younger, aged 65 or older, and group quarters. The most unreliable variables in metro and non-metro areas included crowding, speaks English "less than well", no vehicle, and unemployment. Stakeholders working with SVI or ACS data related to these topics may benefit from consulting alternative data sources to cross-reference with or enrich their analysis.

Alternative overall vulnerability rankings were calculated using the SVI variable estimates and the MOEs provided by the CDC/ATSDR to examine the relationships that may exist among tracts most likely to be over counted or under counted. The MOE was added to the SVI variable estimates to calculate a vulnerability ranking derived from the variable's value at the upper bound of the 90% CI. The MOE was also subtracted from the MOE to calculate its ranking at the lower bound of its 90% CI. These alternative rankings were compared to the original SVI vulnerability rankings to determine whether a census tract's SVI vulnerability classification may be overcounted, undercounted, or remain unchanged. A discriminant analysis was carried out to examine the capacity of the uncertainties in the underlying data of the CVs of the variable estimates to predict whether the SVI vulnerability ranking may be overcounted, undercounted, or remain the same. The result was an Entropy R^2 value of 0.02867, indicating that the uncertainties in the SVI variables have little to no capacity to predict whether the overall vulnerability of a census tract may be over counted, under counted, or remain the same. Approximately 56.8% of census tracts in this index may possibly have unreliable vulnerability classifications. Because so many census tracts may have unreliable vulnerability classifications, stakeholders who depend on SVI data may benefit from data sharing partnerships or agreements with other local stakeholders such as government agencies, nonprofits, or other service providers.

The SVI variable estimate CVs were also subjected to a cluster analysis using a univariate local Moran's I statistic as a measure local spatial autocorrelation. Most census tracts in all variables were found to be not significantly similar to their neighbors (pseudo p-value was greater than p = 0.05) in terms of their CV value, with numbers ranging from 44,757 (crowding) to 61,934 (percent minority) out of a total of 72,173 census tracts, or 62% and 86% respectively. Since the statistical uncertainty is a function of the estimate and the MOE, and populations are not homogeneously distributed, the places in which the uncertainties cluster (at the national scale) vary among SVI variables, as do the social groups in which the uncertainties cluster at each place.

X. FUTURE RESEARCH AND POLICY IMPLICATIONS

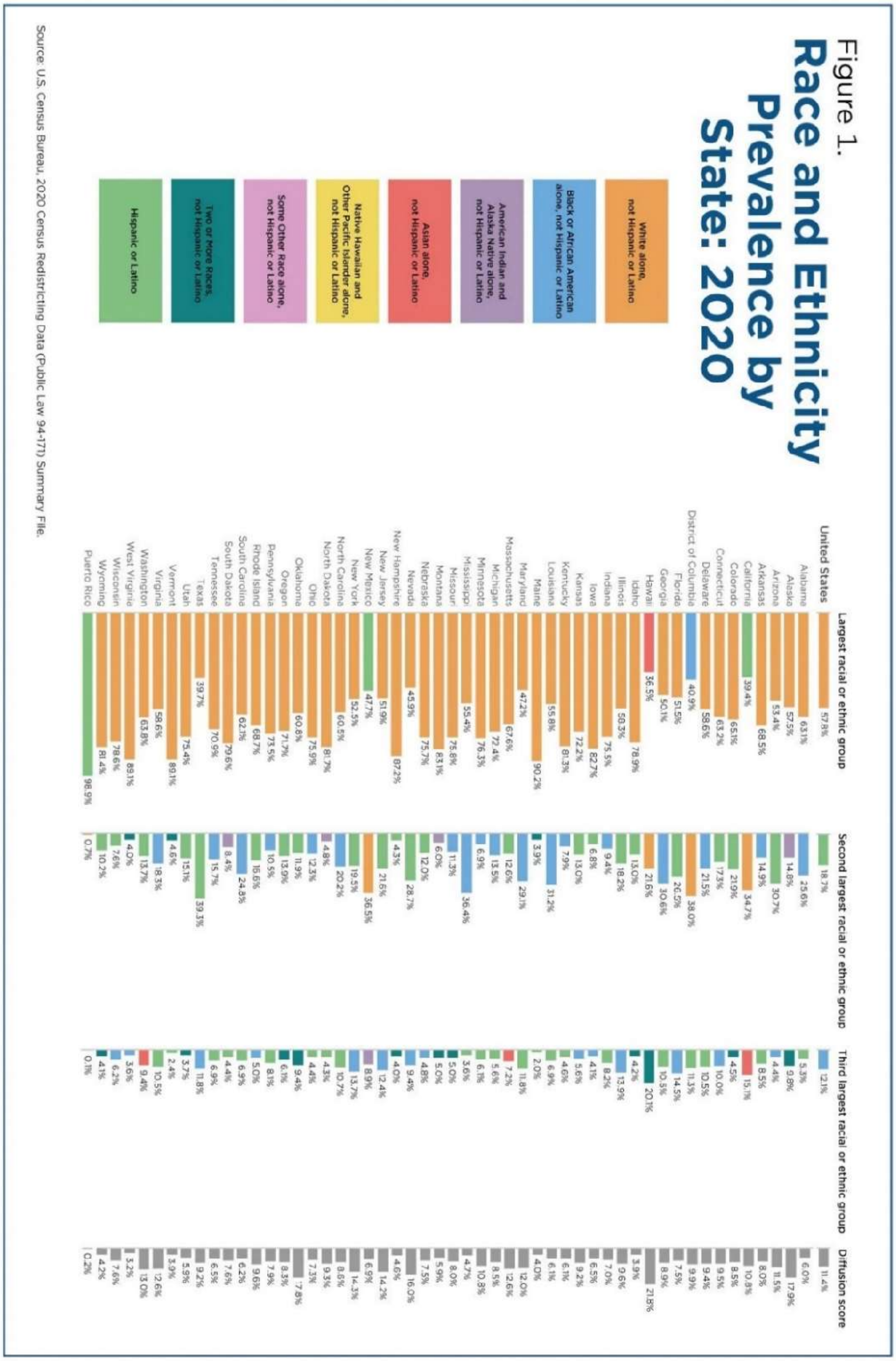
What can be done regarding the uncertainties in the CDC/ATSDR? There are steps that can be taken at different levels to achieve different goals related to the issue of uncertainty in this index and other tools that make use of ACS data. To make the CDC/ATSDR SVI more useful, its publishers in the federal government can recalculate the overall vulnerability classification of each census tract in a way that incorporates the MOE to determine whether the overall vulnerability classification of the census tracts may be over counted, under counted, both (over counted and under counted) or remain unchanged and include the results with the data and documentation. The dataset can also include CV's for variable data originating from the ACS. The research in this thesis shows this is a simple, straightforward process that shouldn't take many resources to implement, but may be of great value to stakeholders who lack the capacity to carry out the calculations independently. Taking these steps would also improve the quality of the data included with the CDC/ATSDR SVI and help align it with Jurjevich et al.'s (2018) guidelines for the ethical use of ACS data.

Since recalculating the overall vulnerability of the census tracts and calculating the CVs of the estimates is not a complex process, it can also be carried out at the state, regional, or community level by interested stakeholders with the capacity to do so. Those who are unable to carry out the analysis can partner with local universities, school districts, or other agencies or nonprofits who may have analysts on staff. Once the uncertainties are identified, alternative data sources may be consulted to enrich the information available regarding the overall vulnerability classification of census tracts with potentially unreliable vulnerability classifications. Some of these data sources may be publicly available, such as academic performance reports or public health authority reports, while others may require closer collaboration with local stakeholders, such as georeferenced aggregated service output data from local nonprofits.

Finally, policy or legislation may be necessary to improve the integrity of decisions where the CDC/ATSDR SVI plays a pivotal role, since so many public funds and resources may be at stake. This is part of why it may be most reasonable for the CDC/ATSDR to be responsible for identifying the census tracts where the overall vulnerability classification may be uncertain. That way, local stakeholders can focus their (more limited) resources on finding local data to correct for the errors in the CDC/ATSDR SVI data or developing partnerships that would gain them access to those data. Future research can help determine what types of data may best fill the gaps in the CDC/ATSDR SVI, and what kinds of partnerships may help make those data more accessible. It is also important to develop a better understanding of who uses the CDC/ATSDR SVI, their capacity to understand the uncertainties, the context in which it is used, the types of decisions made with it, and the outcome for the vulnerable populations subject to those decisions.

APPENDIX SECTION

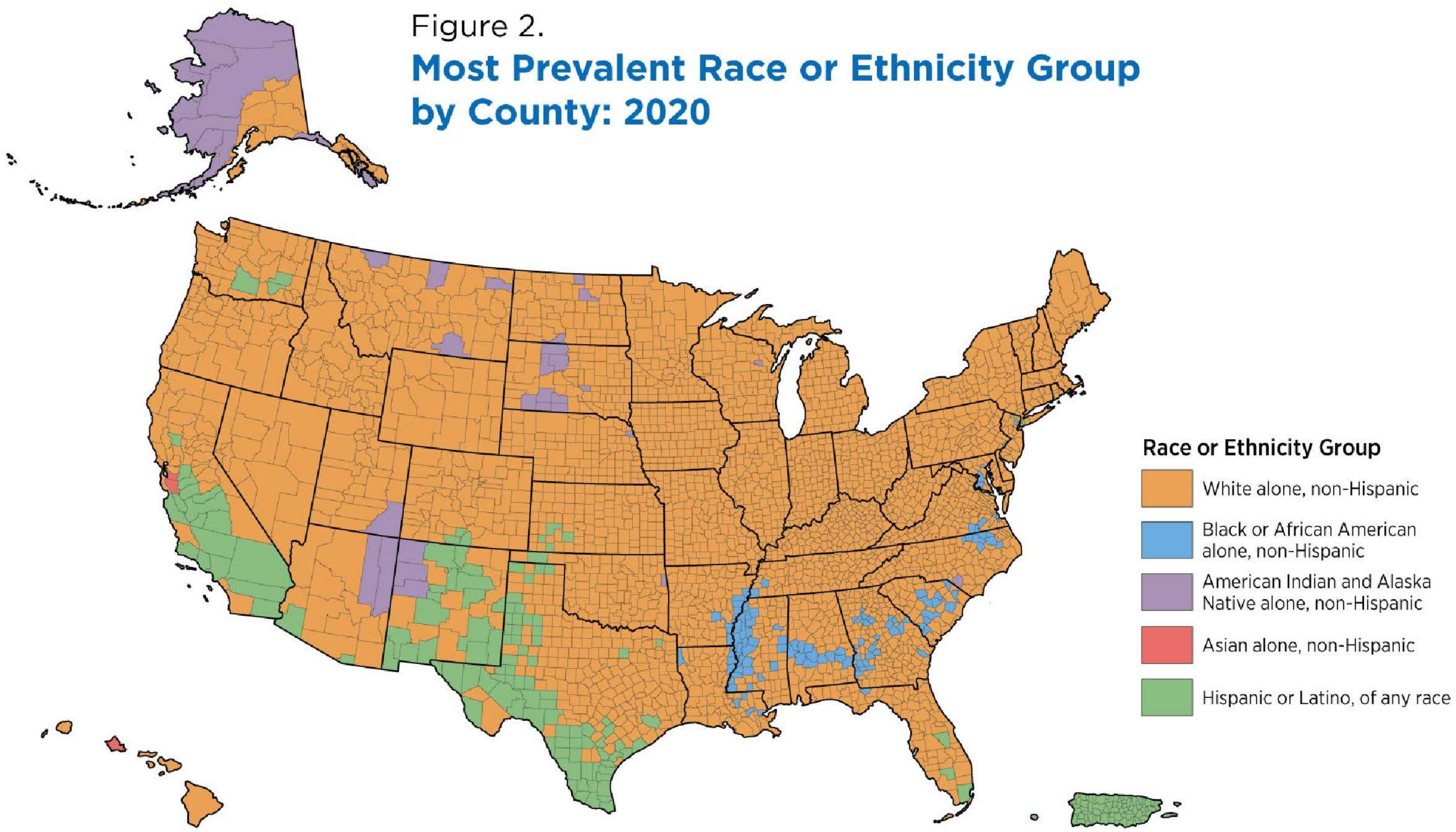
APPENDIX A



APPENDIX B

Figure 2.

Most Prevalent Race or Ethnicity Group by County: 2020



Note: Native Hawaiian and Other Pacific Islander alone, non-Hispanic; Some Other Race alone, non-Hispanic; and Two or More Races, non-Hispanic were not the most prevalent groups in any county.

Information on confidentiality protection, nonsampling error, and definitions is available at <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/>.

Source: U.S. Census Bureau, 2020 Census Redistricting Data (Public Law 94-171) Summary File.

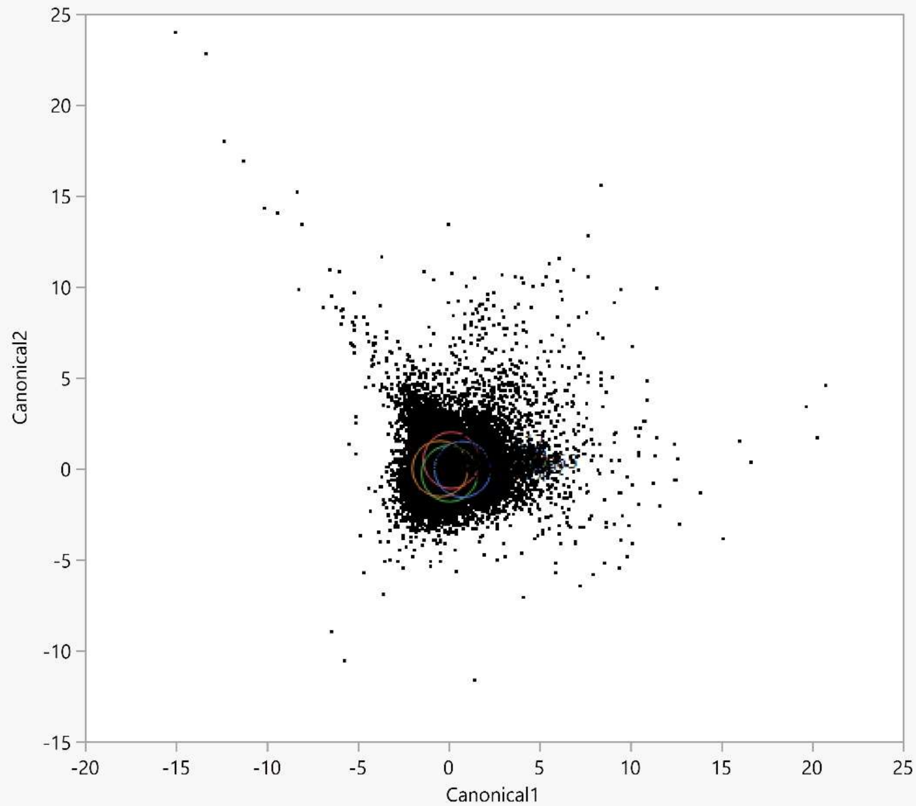
APPENDIX C

Discriminant Analysis

Discriminant Method: Linear

Classification: RPL_CHNG_TYPE

Canonical Plot



Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	72173	41276	57.1904	-0.0413	181153

Training

Actual	Predicted Count			
RPL_CHNG_TYPE	B	N	O	U
B	1082	373	655	1237
N	4129	9833	8425	8761
O	2988	3944	8890	1799
U	3090	4688	1187	11092

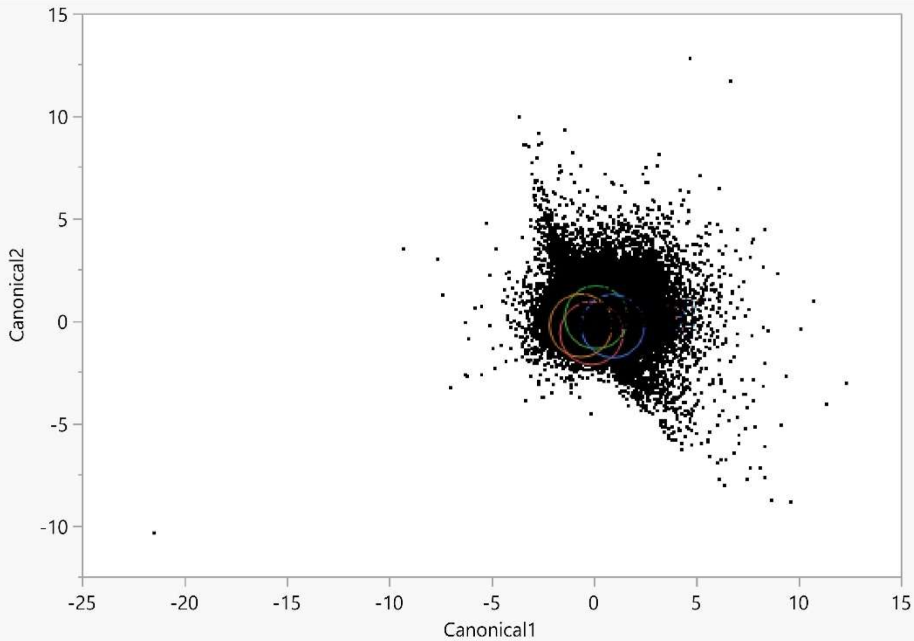
Appendix C1. Canonical plot (top) and Score Summaries table (bottom) for SVI variable percent estimates.

Discriminant Analysis

Discriminant Method: Linear

Classification: RPL_CHNG_TYPE

Canonical Plot



Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	72173	42017	58.2171	0.00338	173373

Training

Actual	Predicted Count			
RPL_CHNG_TYPE	B	N	O	U
B	2001	294	260	792
N	7405	8811	6750	8182
O	4196	2802	9102	1521
U	6655	2558	602	10242

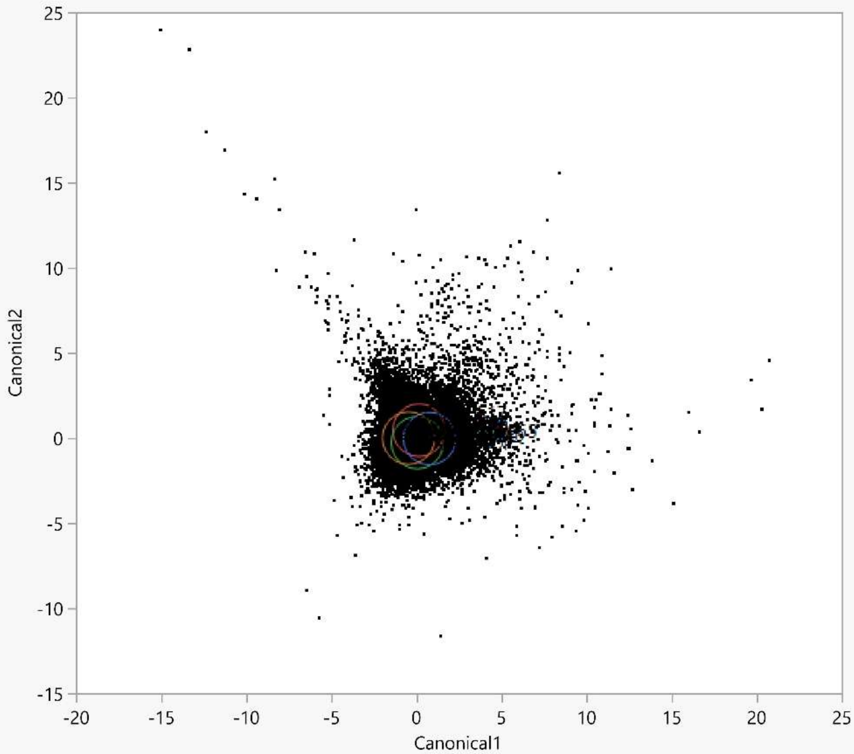
Appendix C2. Canonical plot (top) and Score Summaries table (bottom) for SVI variable estimates.

Discriminant Analysis

Discriminant Method: Linear

Classification: RPL_CHNG_TYPE

Canonical Plot



Score Summaries

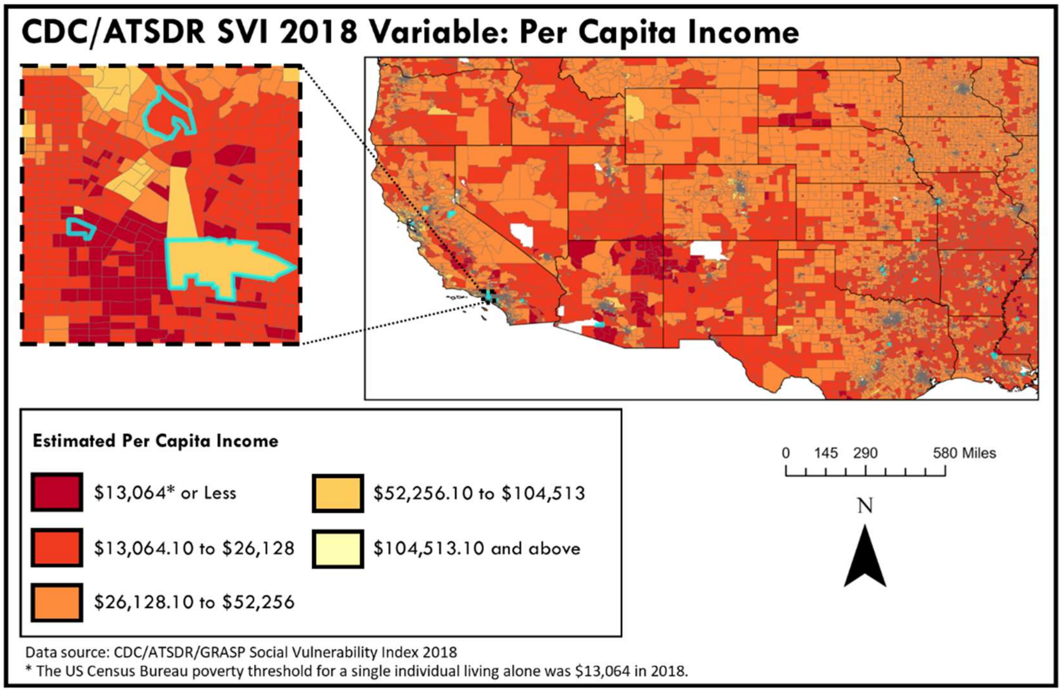
Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	72173	41891	58.0425	-0.0506	182757

Training

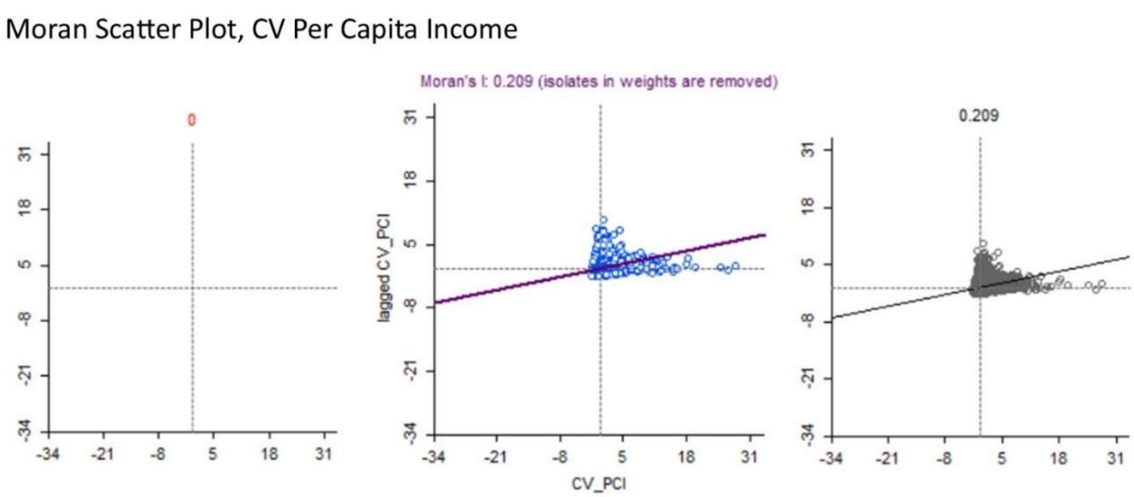
Actual RPL_CHNG_TYPE	Predicted Count B	N	O	U
B	1111	333	614	1289
N	4399	9407	8187	9155
O	3343	3751	8471	2056
U	2995	4546	1223	11293

Appendix C3. Canonical plot (top) and Score Summaries table (bottom) for SVI variable estimate CVs.

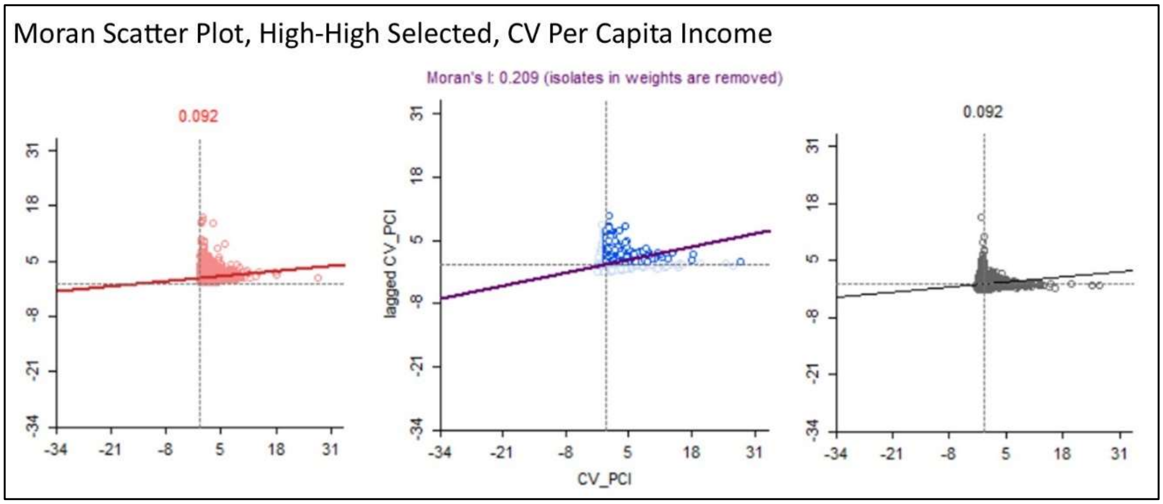
APPENDIX D



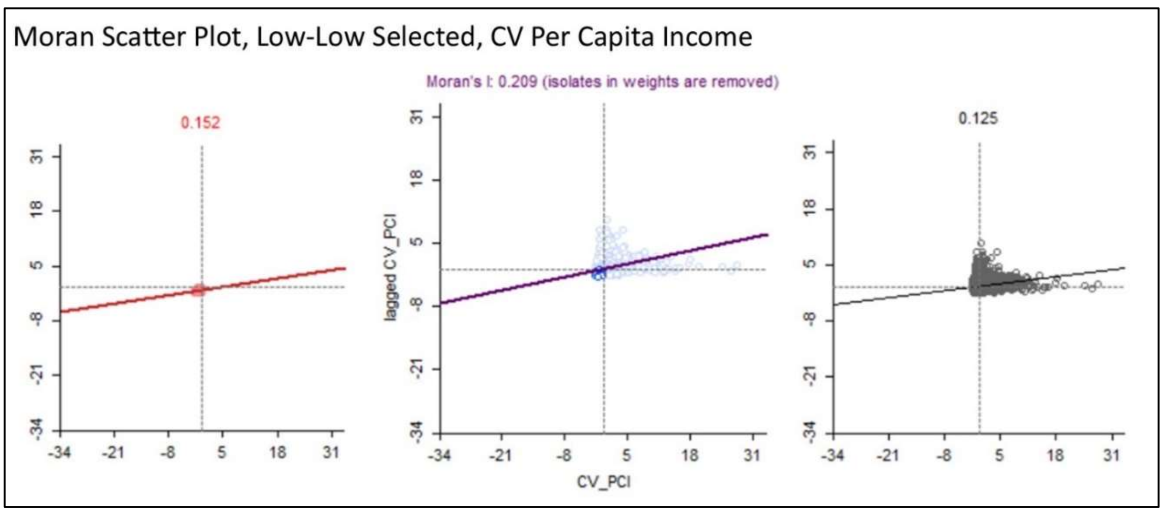
Appendix D1. Estimated Per Capita Income. Census tracts with unreliable CVs highlighted.



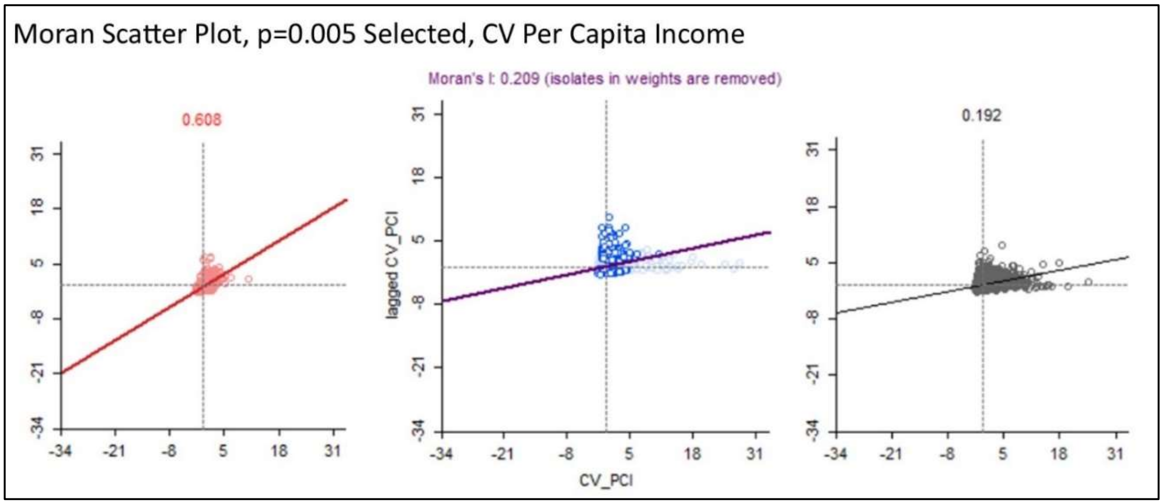
Appendix D2. Moran Scatter Plot, CV Per Capita Income, showing a local Moran statistic value of 0.209 for the dataset.



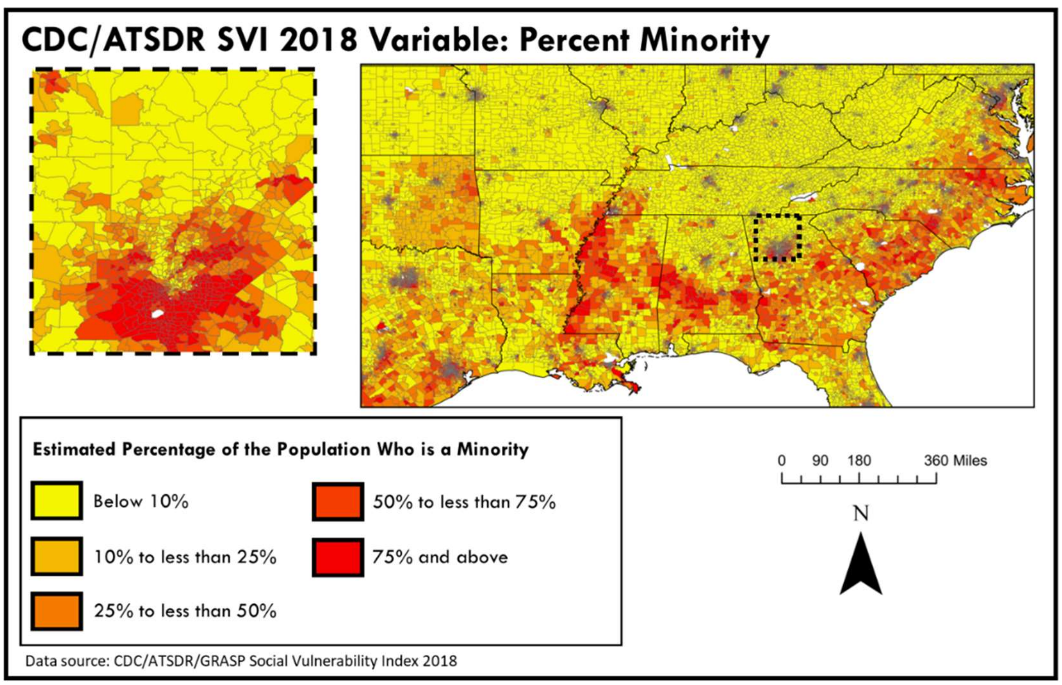
Appendix D3. Moran Scatter Plot, High-High selected, CV Per Capita Income, showing a local Moran statistic value of 0.092 for the High-High tracts, with the rest of the tracts also showing a value of 0.092.



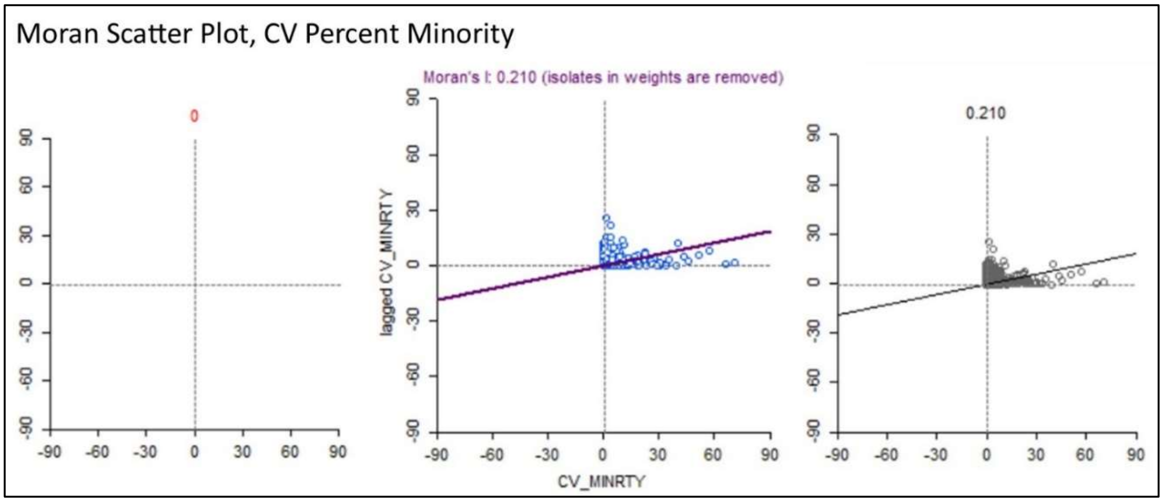
Appendix D4. Moran Scatter Plot, Low-Low selected, CV Per Capita Income, showing a local Moran statistic value of 0.152 for the Low-Low tracts, with the rest of the tracts showing a value of 0.125.



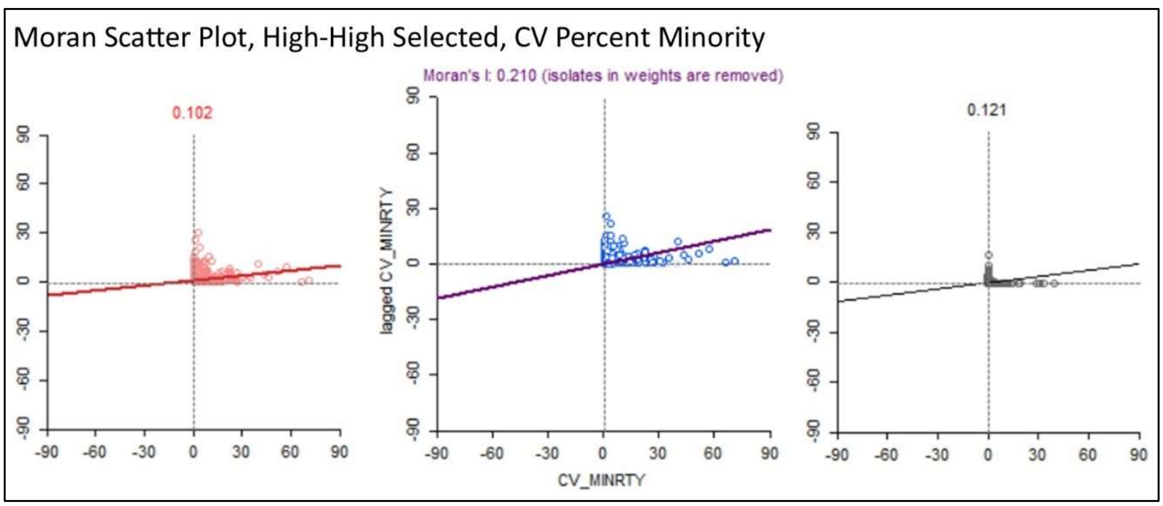
Appendix D5. Moran Scatter Plot, p = 0.005 Selected, CV Per Capita Income, showing a local Moran statistic value of 0.608 for the p = 0.005 tracts, with the rest of the tracts showing a value of 0.192.



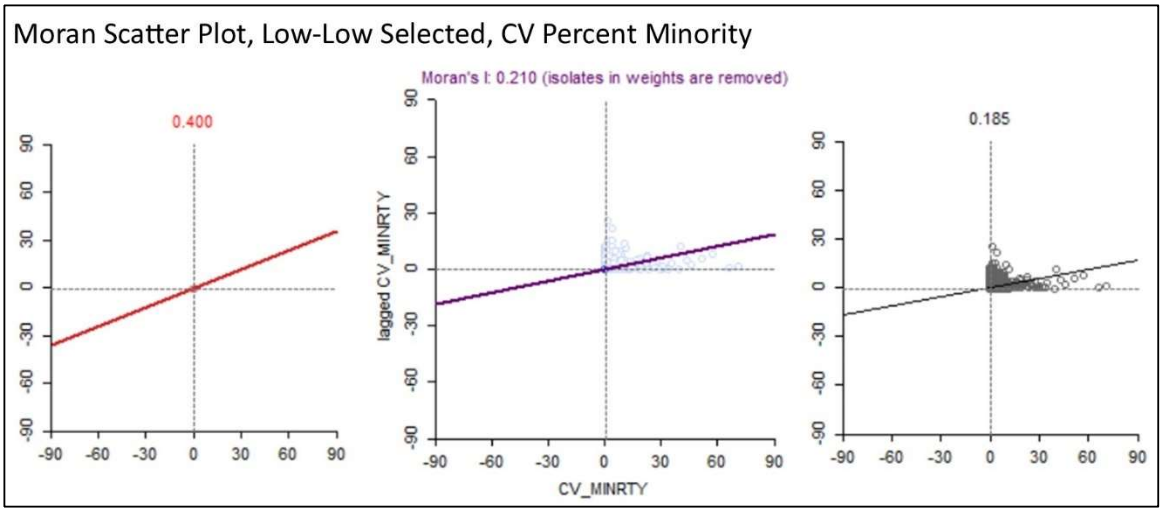
Appendix D6. Estimated percentage of the population who is a minority.



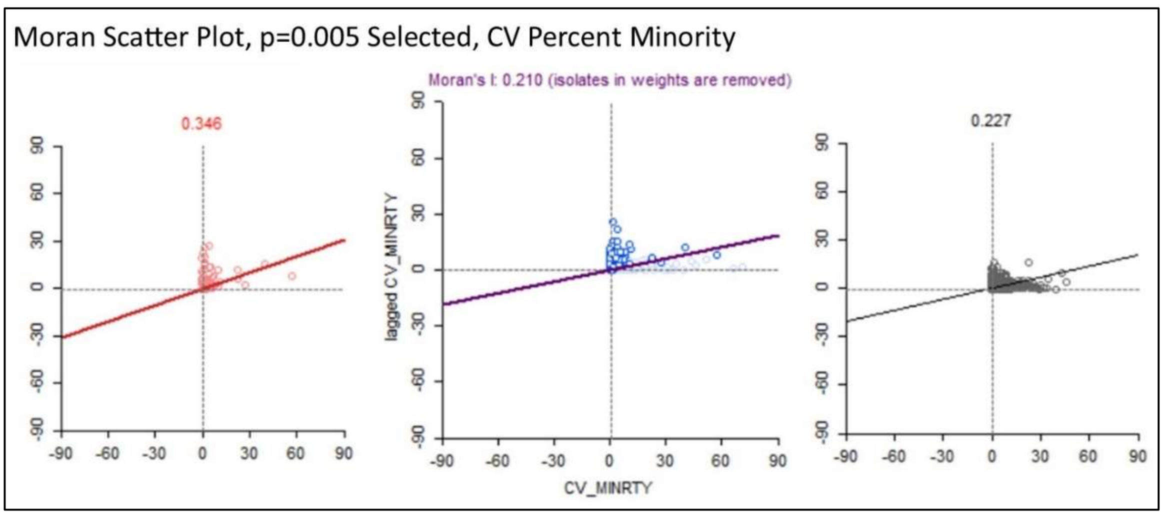
Appendix D7. Moran Scatter Plot, CV Percent Minority, showing a local Moran statistic value of 0.210 for the dataset.



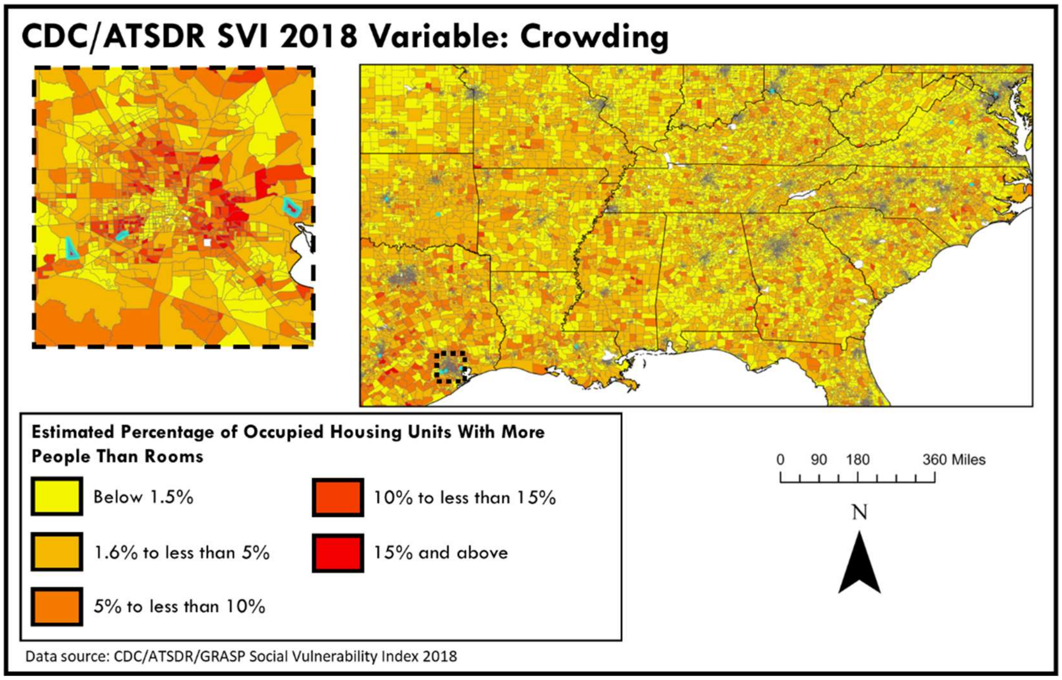
Appendix D8. Moran Scatter Plot, High-High selected, CV Percent Minority, showing a local Moran statistic value of 0.102 for the High-High tracts, with the rest of the tracts showing a value of 0.121.



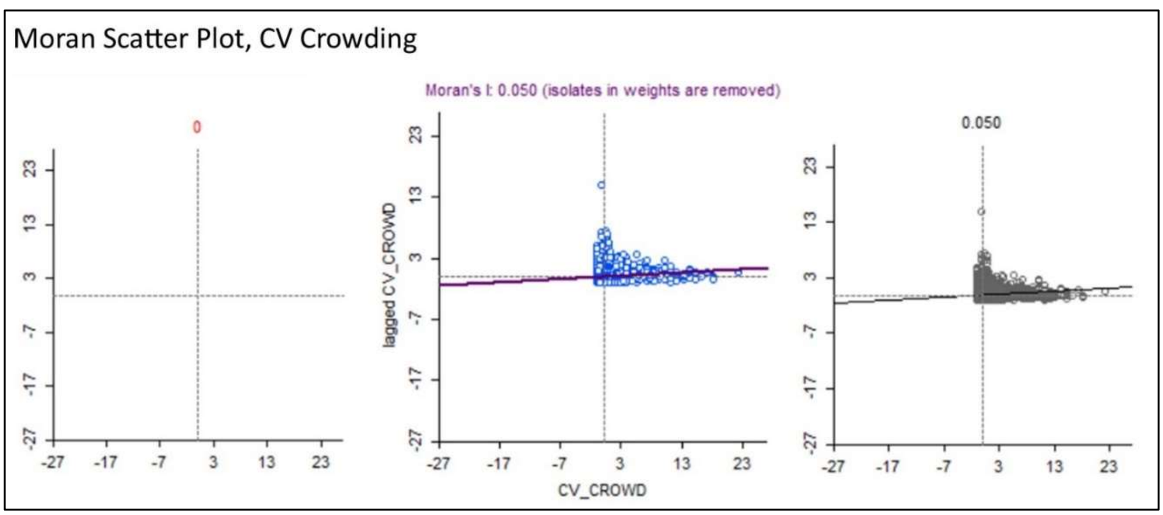
Appendix D9. Moran Scatter Plot, Low-Low selected, CV Percent Minority, showing a local Moran statistic value of 0.400 for the Low-Low tracts, with the rest of the tracts showing a value of 0.185.



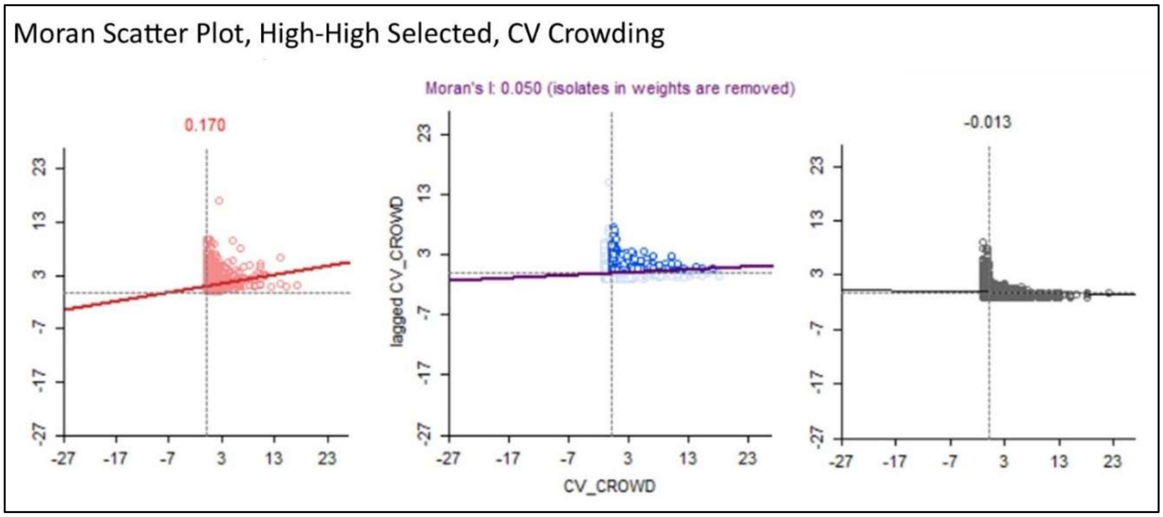
Appendix D10. Moran Scatter Plot, p = 0.005 Selected, CV Percent Minority, showing a local Moran statistic value of 0.346 for the p = 0.005 tracts, with the rest of the tracts showing a value of 0.227.



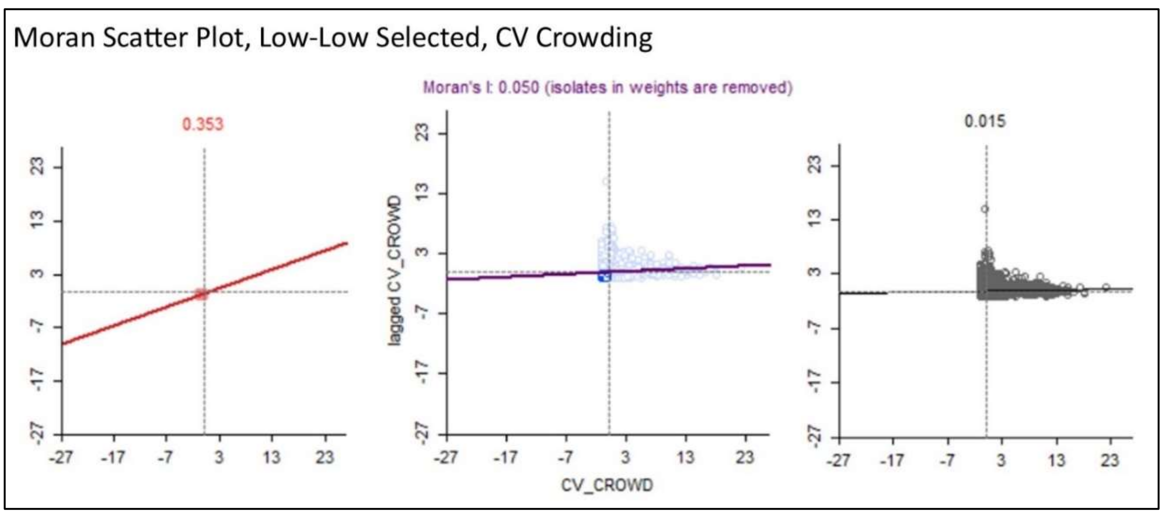
Appendix D11. Estimated percentage of occupied housing units with more people than rooms. Census tracts with unreliable CVs highlighted.



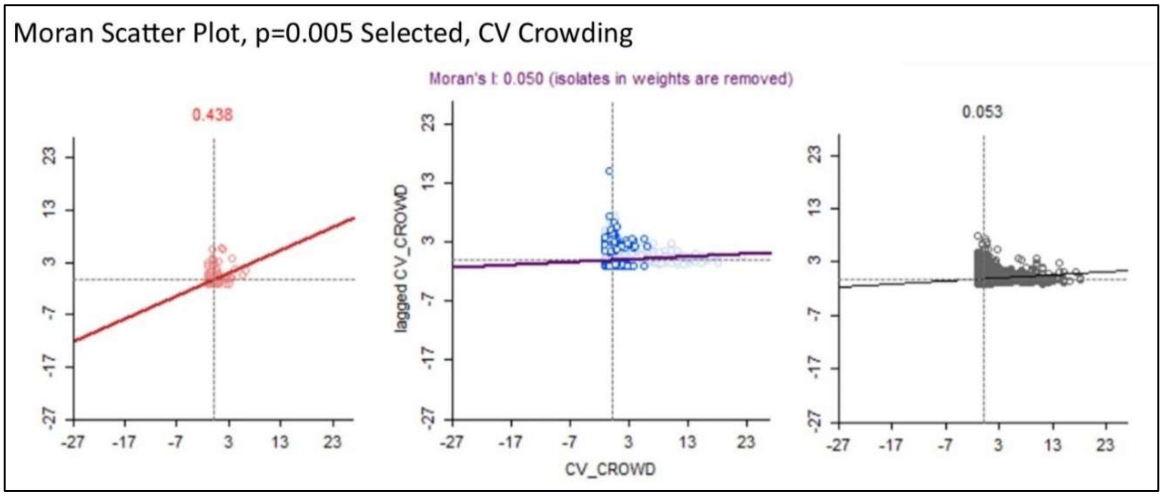
Appendix D12. Moran Scatter Plot, CV Percent Minority Income showing a local Moran statistic value of 0.050 for the dataset.



Appendix D13. Moran Scatter Plot, High-High selected, CV Percent Minority, showing a local Moran statistic value of 0.170 for the High-High tracts, with the rest of the tracts showing a value of -0.013.



Appendix D14. Moran Scatter Plot, Low-Low selected, CV Crowding, showing a local Moran statistic value of 0.353 for the Low-Low tracts, with the rest of the tracts showing a value of 0.015.



Appendix D15. Moran Scatter Plot, p = 0.005 Selected, CV Crowding, showing a local Moran statistic value of 0.438 for the p = 0.005 tracts, with the rest of the tracts showing a value of 0.053.

REFERENCES

Agency for Toxic Substances and Disease Registry. 2022. CDC SVI Documentation 2018. https://www.atsdr.cdc.gov/placeandhealth/svi/documentation/SVI_documentation_2018.html (last accessed 13 March 2022).

Agency for Toxic Substances and Disease Registry. 2021. At A Glance: CDC/ATSDR Social Vulnerability Index | Place and Health | ATSDR. https://www.atsdr.cdc.gov/placeandhealth/svi/at-a-glance_svi.html (last accessed 27 March 2022).

Anselin, L. 1995. Local Indicators of Spatial Association—LISA. *Geographical Analysis* 27 (2):93–115.

Anselin, L. 2005. Exploring Spatial Data with GeoDaTM: A Workbook. Spatial Analysis Laboratory. Urbana, IL: Department of Geography, University of Illinois at Urbana-Champaign.

Anselin, L. 2020. Local Spatial Autocorrelation (1) LISA and Local Moran. https://geodacenter.github.io/workbook/6a_local_auto/lab6a.html#fn1 (last accessed 24 February 2023).

Bakkensen, L. A., C. Fox-Lent, L. K. Read, and I. Linkov. 2017. Validating Resilience and Vulnerability Indices in the Context of Natural Disasters. *Risk Analysis* 37 (5):982–1004.

Bazuin, J. T., and J. C. Fraser. 2013. How the ACS gets it wrong: The story of the American Community Survey and a small, inner city neighborhood. *Applied Geography* 45:292–302.

Bolin, B. 2007. Race, Class, Ethnicity, and Disaster Vulnerability. In *Handbook of Disaster Research*, eds. H. Rodríguez, E. L. Quarantelli, and R. R. Dynes, 113–129. New York, NY: Springer https://doi.org/10.1007/978-0-387-32353-4_7 (last accessed 29 March 2022).

Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. "CDC/ATSDR Social Vulnerability Index," 2018. Accessed March 29, 2022. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html.

Community Advancement Network (CAN). "Assistance With Basic Needs - Calls to 2-1-1 Austin Round Rock 5-County MSA." Get Informed – We CAN! ATX. Accessed March 30, 2023. https://wecanatx.org/get-informed/.

Cutter, S. L., B. J. Boruff, and W. L. Shirley. 2003. Social Vulnerability to Environmental Hazards. *Social Science Quarterly* 84 (2):242–261.

Edwards, J., Elliott, J., and Ernsthausen, K. “A Trump Tax Break To Help The Poor Went To a Rich GOP Donor’s Superyacht Marina.” ProPublica. Last modified November 14, 2019. Accessed April 16, 2023. https://www.propublica.org/article/superyacht-marina-west-palm-beach-opportunity-zone-trump-tax-break-to-help-the-poor-went-to-a-rich-gop-donor.

Flanagan, B. E., E. W. Gregory, E. J. Hallisey, J. L. Heitgerd, and B. Lewis. 2011. A Social Vulnerability Index for Disaster Management. *Journal of Homeland Security and Emergency Management* 8 (1). https://www.atsdr.cdc.gov/placeandhealth/svi/img/pdf/Flanagan_2011_SVIforDisasterManagement-508.pdf (last accessed 25 February 2022).

Gbur, P. M., and S. P. Hefter. 2002. Overview of the U.S. Census 2000 Long Form Weighting. *2002 American Statistical Association Proceedings* :6.

Hauer, Mathew E. “Migration Induced by Sea-Level Rise Could Reshape the US Population Landscape.” Nature Climate Change 7, no. 5 (2017). Accessed April 1, 2022. https://www.nature.com/articles/doi:10.1038/nclimate3271.epdf.

Heimel, S. 2014. American Community Survey Fiscal Year 2014 Content Review - Median County-Level Coefficients of Variation of Key Estimates from ACS Data. American Community Survey Office: United States Census Bureau. https://www2.census.gov/programs-surveys/acs/operations_admin/2014_content_review/methods_results_report/Coefficients_of_Variation_Report.pdf (last accessed 20 March 2022).

Jensen, Eric, Nicholas Jones, Megan Rabe, Beverly Pratt, Lauren Medina, Kimberly Orozco, and Lindsay Spell. The Chance That Two People Chosen at Random Are of Different Race or Ethnicity Groups Has Increased Since 2010. US Census Bureau, 2021. Accessed April 2, 2022. https://www.census.gov/library/stories/2021/08/2020-united-states-population-more-racially-ethnically-diverse-than-2010.html.

Jung, P. H., J.-C. Thill, and M. Issel. 2019. Spatial autocorrelation and data uncertainty in the American Community Survey: a critique. *International Journal of Geographical Information Science* 33 (6):1155–1175.

Jurjevich, J. R., A. L. Griffin, S. E. Spielman, D. C. Folch, M. Merrick, and Nagle. 2018. Navigating Statistical Uncertainty: How Urban and Regional Planners Understand and Work With American Community Survey (ACS) Data for Guiding Policy. *Journal of the American Planning Association* 84 (2):112–126.

Karaye, I. M., and J. A. Horney. 2020. The Impact of Social Vulnerability on COVID-19 in the U.S.: An Analysis of Spatially Varying Relationships. *American Journal of Preventive Medicine* 59 (3):317–325.

Kim, S. J., and W. Bostwick. 2020. Social Vulnerability and Racial Inequality in COVID-19 Deaths in Chicago. *Health Education & Behavior* 47 (4):509–513.

MacEachren, A. M., C. A. Brewer, and L. W. Pickle. 1998. Visualizing Georeferenced Data: Representing Reliability of Health Statistics. *Environment and Planning A: Economy and Space* 30 (9):1547–1561.

Morrow, B. H. 1999. Identifying and Mapping Community Vulnerability. Disasters 23 (1):1.

National Research Council. 2015. *Realizing the Potential of the American Community Survey: Challenges, Tradeoffs, and Opportunities*. Washington, DC: The National Academies Press. https://www.nap.edu/catalog/21653/realizing-the-potential-of-the-american-community-survey-challenges-tradeoffs (last accessed 10 March 2022).

Palaiologou, P., A. A. Ager, M. Nielsen-Pincus, C. R. Evers, and M. A. Day. 2019. Social vulnerability to large wildfires in the western USA. *Landscape and Urban Planning* 189:99–116.

Rufat, S., E. Tate, C. T. Emrich, and F. Antolini. 2019. How Valid Are Social Vulnerability Models? Annals of the American Association of Geographers 109 (4):1131–1153.

Spielman, S. E., D. Folch, and N. Nagle. 2014. Patterns and causes of uncertainty in the American Community Survey. *Applied geography (Sevenoaks, England)* 46:147–157.

Spielman, S. E., J. Tuccillo, D. C. Folch, A. Schweikert, R. Davies, N. Wood, and E. Tate. 2020. Evaluating social vulnerability indicators: criteria and their application to the Social Vulnerability Index. *Natural Hazards* 100 (1):417–436.

Sun, M., and D. W. S. Wong. 2010. Incorporating Data Quality Information in Mapping American Community Survey Data. *Cartography and Geographic Information Science* 37 (4):285–299.

Tate, E. 2013. Uncertainty Analysis for a Social Vulnerability Index. *Annals of the Association of American Geographers* 103 (3):526–543.

United States Census Bureau. 2006. Design and Methodology, American Community Survey. https://www.census.gov/history/pdf/ACSHistory.pdf (last accessed 25 March 2022).

United States Census Bureau. 2020a. Understanding and Using American Community Survey Data: What All Data Users Need to Know. https://www.census.gov/programs-surveys/acs/library/handbooks/general.html (last accessed 13 March 2022).

United States Census Bureau. Table H1, 2020b Census Redistricting Data. https://data.census.gov/cedsci/table?g=0100000US&y=2020&d=DEC%20Redistricting%20Data%20%28PL%2094-171%29&tid=DECENNIALPL2020.H1 (last accessed 27 March 2022).

United States Census Bureau. Summary File 3, 2000 Census. https://data.census.gov/cedsci/table?g=0100000US&y=2000&tid=DECENNIALSF32000.H006 (last accessed 27 March 2022).

United States Census Bureau. 2021. Questionnaires - History - U.S. Census Bureau. https://www.census.gov/history/www/through_the_decades/questionnaires/ (last accessed 27 March 2022).

Wolkin, Amy, Jennifer Rees Patterson, Shelly Harris, Elena Soler, Sherry Burrer, Michael McGeehin, and Sandra Greene. "Reducing Public Health Risk During Disasters: Identifying Social Vulnerabilities." *Journal of homeland security and emergency management* 12, no. 4 (December 2015): 809–822.

Zhou, Yang, Ning Li, Wenxiang Wu, Jidong Wu, and Peijun Shi. "Local Spatial and Temporal Factors Influencing Population and Societal Vulnerability to Natural Disasters: Population and Societal Vulnerability to Natural Disasters." *Risk Analysis* 34, no. 4 (April 2014): 614–639.