VTRC
Virginia Transportation Research Council

We Bring Innovation to Transportation

Estimating Disadvantaged Populations Based on Public Data

https://vtrc.virginia.gov/media/vtrc/vtrc-pdf/vtrc-pdf/25-R23.pdf

YIQING XU, Ph.D.
Research Scientist

LANCE E. DOUGALD
Associate Principal Research Scientist

Final Report VTRC 25-R23

VIRGINIA TRANSPORTATION RESEARCH COUNCIL
530 Edgemont Road, Charlottesville, VA 22903-2454
vtrc.virginia.gov

Standard Title Page—Report on State Project

Report No.: VTRC 25-R23	Report Date: June 2025	No. Pages: 50	Type Report: Final	Project No.: 124695
			Period Covered: January 2024–February 2025	Contract No.:

Title:
Estimating Disadvantaged Populations Based on Public Data

Key Words:
Disadvantaged Populations, Estimation Models, Transportation Equity, Census Tract, Census Block Group

Author(s): Yiqing Xu, Ph.D., Lance E. Dougald

Performing Organization Name and Address:
Virginia Transportation Research Council
530 Edgemont Road
Charlottesville, VA 22903

Sponsoring Agencies' Name and Address:
Virginia Department of Transportation
1401 E. Broad Street
Richmond, VA 23219

Supplementary Notes:

Abstract:

This report develops robust estimation models to address data challenges for identifying disadvantaged populations, defined by low-income, racial minority, and limited English proficiency (LEP) status, which aligns with the Virginia Department of Transportation's (VDOT) SMART SCALE definition. The current reliance on American Community Survey data presents granularity limitations, double-counting risks, and large margins of error, particularly at finer geographic levels.

A custom tabulation from the U.S. Census Bureau, adjusted for noise errors and suppressed data, served as the foundation. The researchers developed three regression models at the census tract level: Basic, Option 1 (including LEP households), and Option 2 (including LEP population). All models achieved high reliability, with adjusted R^2 values exceeding 0.95 and mean absolute errors of 11% (169 persons) for the Basic model. Option 2 demonstrated superior accuracy, reducing errors to 8% (118 persons) and ensuring estimates for 95% of tracts were within the census margin of error. For Option 2, only 97 out of 2,188 tracts exhibited errors exceeding the census margin of error, averaging 599 persons.

At the block group level, a single regression model incorporated racial minority, poverty, and LEP populations. This model achieved an adjusted R^2 of 0.97, with a mean absolute error of 15.6% (88 persons). Approximately 90% of block groups met census margin-of-error thresholds, with average discrepancies for outliers at 227 persons.

The findings demonstrate that reliable estimates can be produced using public data, with models reducing error rates significantly. The Option 2 model at the census tract level and the block group model are recommended for statewide application to enhance transportation equity planning. VDOT's Transportation and Mobility Planning Division can leverage these tools to prioritize SMART SCALE projects effectively, ensuring equitable access for disadvantaged populations.

FINAL REPORT

ESTIMATING DISADVANTAGED POPULATIONS BASED ON PUBLIC DATA

Yiqing Xu, Ph.D.
Research Scientist

Lance E. Dougald
Associate Principal Research Scientist

Virginia Transportation Research Council
(A partnership of the Virginia Department of Transportation
and the University of Virginia since 1948)

Charlottesville, Virginia

June 2025
VTRC 25-R23

DISCLAIMER

The contents of this report reflect the views of the author(s), who is responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the Virginia Department of Transportation, the Commonwealth Transportation Board, or the Federal Highway Administration. This report does not constitute a standard, specification, or regulation. Any inclusion of manufacturer names, trade names, or trademarks is for identification purposes only and is not to be considered an endorsement.

ABSTRACT

This report develops robust estimation models to address data challenges for identifying disadvantaged populations, defined by low-income, racial minority, and limited English proficiency (LEP) status, which aligns with the Virginia Department of Transportation's (VDOT) SMART SCALE definition. The current reliance on American Community Survey data presents granularity limitations, double-counting risks, and large margins of error, particularly at finer geographic levels.

A custom tabulation from the U.S. Census Bureau, adjusted for noise errors and suppressed data, served as the foundation. The researchers developed three regression models at the census tract level: Basic, Option 1 (including LEP households), and Option 2 (including LEP population). All models achieved high reliability, with adjusted R^2 values exceeding 0.95 and mean absolute errors of 11% (169 persons) for the Basic model. Option 2 demonstrated superior accuracy, reducing errors to 8% (118 persons) and ensuring estimates for 95% of tracts were within the census margin of error. For Option 2, only 97 out of 2,188 tracts exhibited errors exceeding the census margin of error, averaging 599 persons.

At the block group level, a single regression model incorporated racial minority, poverty, and LEP populations. This model achieved an adjusted R^2 of 0.97, with a mean absolute error of 15.6% (88 persons). Approximately 90% of block groups met census margin-of-error thresholds, with average discrepancies for outliers at 227 persons.

The findings demonstrate that reliable estimates can be produced using public data, with models reducing error rates significantly. The Option 2 model at the census tract level and the block group model are recommended for statewide application to enhance transportation equity planning. VDOT's Transportation and Mobility Planning Division can leverage these tools to prioritize SMART SCALE projects effectively, ensuring equitable access for disadvantaged populations.

EXECUTIVE SUMMARY

Introduction: Lack of Accurate Disadvantaged Population Data

This project was directly driven by a research need identified by VDOT's Transportation Mobility Planning Division and was submitted to the Transportation Planning Research Advisory Committee (Ohlms, 2024). The project's goal is to develop a tool to estimate the percentage of households that are considered disadvantaged, as defined by meeting one or more of the following criteria: low income, racial minority, or limited English proficiency (LEP). The Commonwealth Transportation Board (CTB) established these three criteria (CTB, 2021). The low-income criterion refers to individuals whose household income in the past 12 months was below the poverty level. The racial minority criterion refers to "Black, American Indian, Asian, Pacific Islander, Other, and/or Hispanic individuals". Individuals who indicated two or more races were automatically included. The LEP criterion refers to persons for whom English is not their primary language and who have a limited ability to read, write, speak, or understand English. This criterion includes people who reported to the U.S. Census Bureau that they speak English at a level less than very well, not well, or not at all (Rogoff, 2012; U.S. Department of Health and Human Services, 2024; U.S. Department of Justice, 2020). The project goal was not to redefine equity but to address the best way to collect and refine data in alignment with CTB's existing criteria.

The process of VDOT's SMART SCALE project prioritization incorporates equity by evaluating projects on how well they improve accessibility for disadvantaged populations. This approach aligns with CTB's commitment to fostering inclusive growth and promoting equal opportunities, ensuring that transportation planning supports economic and social participation for all residents (CTB, 2021).

VDOT has sourced data on disadvantaged populations from the American Community Survey (ACS) (U.S. Census Bureau, 2024) since 2018, but challenges with data accuracy, granularity, and multicollinearity persist. For instance, LEP data are only available at the census tract level, which limits precision at the block group level. In addition, overlapping criteria (e.g., poverty and LEP) can lead to potential double-counting. A prior VTRC technical assistance project (Li et al., 2023) attempted to estimate disadvantaged populations across Virginia but relied on data from 56 Public Use Microdata Areas, which may lack sufficient granularity. Given these limitations, VDOT's Transportation and Mobility Planning Division (TMPD) is exploring alternative approaches to improve data accuracy for defining equity areas and better support SMART SCALE priorities for VDOT's Office of Intermodal Planning and Investment.

Problem: Current Disadvantaged Population Data Have Limited Granularity, Risks of Double-Counting, and Large Margins of Error

VDOT's current data on disadvantaged populations, derived from ACS, present several challenges that affect transportation equity efforts. First, data granularity is limited because LEP data are only available at the census tract level, thus reducing accuracy at smaller geographic levels, such as block groups. In addition, ACS provides separate tabulations for each criterion

(low income, racial minority, and LEP) without accounting for overlaps, which can lead to double-counting when combined. High multicollinearity among variables further complicates analyses because correlations between criteria hinder the ability to isolate each variable's effectiveness accurately. Lastly, ACS data for small geographic units often have large margins of error, making these estimates less reliable.

Purpose and Scope

In alignment with the need to estimate disadvantaged populations defined in VDOT's SMART SCALE, as well as the Justice40 Initiative under the Infrastructure Investment and Jobs Act (Government Finance Officers Association, 2023) and the refinements to the Office of Intermodal Planning and Investment's equity areas for Virginia's Transportation Plan (Virginia OIPI, 2022), this project aimed to devise a methodology for extracting data on disadvantaged populations from publicly accessible sources. The scope of this endeavor was bound by the currently available data, encompassing open sources and a special tabulation provided by the Census Bureau at the census tract level. The research team expanded the analysis to the census block group level as the data became available through requests for additional Census Bureau tabulations.

Methods

To meet the objectives of this project, the following tasks were undertaken:

1. Review relevant literature.
2. Obtain disadvantaged population data from a Census Bureau custom tabulation.
3. Cleanse the custom tabulation.
4. Adjust the population universe.
5. Develop the estimation model at the census tract level.
6. Develop the estimation model at the census block group level.

The methods in this study were organized into six key tasks aimed at developing a reliable model for estimating disadvantaged populations in Virginia. The first task involved a literature review analyzing the link between disadvantaged populations and socioeconomic, environmental, and health factors. However, the reviewed studies predominantly used broad indicators, such as income and education, and lacked finer geographic details, such as census blocks.

In the second and third tasks, the research team worked with the Census Bureau during an 18-month period to obtain custom tabulations that provided counts of disadvantaged individuals by census tract and block group based on three criteria: LEP, racial minority, and low income. However, cleansing these data revealed discrepancies because of limited population coverage, privacy-protection mechanisms (e.g., discrete Gaussian noise), and large margins of error. Specifically, the limited population universe excluded institutionalized individuals and certain suppressed census tracts, whereas the noise mechanism introduced rounding errors that

sometimes resulted in disadvantaged counts exceeding total population counts. The researchers made adjustments to align the disadvantaged population with total counts in such cases, improving the dataset's accuracy for modeling.

The final three tasks involved developing and evaluating estimation models at both census tract and block group levels. At the tract level, the research team created three models by using least squares regression, including a Basic model with racial minority and poverty variables and two optional models incorporating LEP data. At the block group level, the research team accounted for overlapping populations among variables using separate coefficients for each to reduce multicollinearity. The research team assessed all models for fit and predictive accuracy using adjusted R^2, mean testing error metrics, residual plots, and the inaccurate rate for tracts or block groups exceeding the margin of error to ensure reliable estimates for targeted transportation planning activities and investments.

Key Results

Census Tract Level Estimation

The three models developed (Basic, Option 1, and Option 2) are displayed in Table ES1.

Table ES1. Developed Models

Models	Formula
Basic	DP[a] = 19.932 + 1.105 * (S0601_Variable 1[b] + S1701_Variable 2[c])
Option 1[f]	DP =0.997 * (S0601_Variable 1 + S1701_Variable 2) + 0.745 * C16002_Variable 3[d]
Option 2[f]	DP =0.999 * (S0601_Variable 1 + S1701_Variable 2) + 0.330 * B16009_Variable 3[e]

[a] Disadvantaged population.
[b] Total racial minority population from Census Table S0601 (U.S. Census Bureau, n.d.b.).
[c] Total population of White population below the poverty level from Census Table S1701 (U.S. Census Bureau, n.d.c.).
[d] Number of limited English proficiency (LEP) households from Census Table C16002 (U.S. Census Bureau, n.d.a.).
[e] Total LEP population from Census Table B16009 (U.S. Census Bureau, n.d.d.).
[f] Regression without intercept.

Each model uses different independent variables to estimate disadvantaged populations, with coefficients logically reflecting the population definitions used. For instance, in the Option 2 model, the coefficient of 0.999 for the independent variable aligns well with the intended estimation accuracy, indicating a high degree of representation of racial minority and White populations below the poverty level in the disadvantaged group.

All three models exhibit strong performance, as shown by high adjusted R^2 values greater than 0.95, which indicates they explain more than 95% of the variance in disadvantaged population estimates. The mean absolute error is low across models, with the Basic model having 11% (169 people per tract/1,534 people per tract) of the mean size of the disadvantaged population. Adding LEP variables in the Option models improves accuracy slightly, reducing the mean absolute error by 44 persons per census tract compared with the Basic model. This improvement reflects the supplemental role of the LEP variable in representing disadvantaged individuals not captured by racial minority and poverty variables alone, especially in tracts with

significant LEP populations. Overall, this finding suggests that all models provide a reliable estimate, and if detailed LEP data are limited, the Basic model can still be effective for estimating disadvantaged populations.

The Basic model has a higher inaccurate estimation rate of 171 tracts (i.e., the rate of census tracts with an estimation exceeding the census margin of error), whereas the Option models, which include an LEP variable, perform slightly better. Applying the Option 2 model, the researchers found that only 97 out of 2,188 tracts had errors beyond the margin. For these tracts, the average estimation error was 599 people.

Among the three models, Option 2 provides the best accuracy with the lowest standard error (193) and mean absolute error (118). Introducing variables like LEP population makes the Option 2 model unbiased and homoscedastic.

Census Block Group-Level Estimation

The researchers developed the following model (Equation ES1) for estimating the disadvantaged population at the census block group level using three independent variables (racial minority, poverty, and LEP populations):

DP = 0.886 * total racial minority population + 0.422 * total poverty population
+ 0.125 * total LEP population (Eq. ES1)

Where:
DP = disadvantaged population.
LEP = limited English proficiency.

This model achieved a high adjusted R^2 of 0.97, which explained most of the variance. The standard error of estimate and mean absolute error are 138 persons and 88 persons, respectively, corresponding to 24.5% and 15.6% of the mean value (564 persons) of the disadvantaged population for all 5,943 census block groups involved in the analysis. Approximately 90% of the block groups had estimation errors within the census margin of error, with only 593 of 5,943 groups exceeding this threshold. For these 593 groups, the average error was 227 people. The residual plot for the block group model is unbiased and slightly heteroscedastic (based on visual inspection).

Key Conclusions

- *The preferred estimation models demonstrate high reliability, as indicated by a high adjusted R^2 (no less than 0.97).* When these models were tested on data not used to build them, they showed mean absolute error rates of 8% at the tract level and 16% at the block group level. For instance, the Option 2 model had an error of 118 disadvantaged persons per tract, which is 8% of the mean value of 1,534 disadvantaged individuals per tract.

- *Detailed, publicly available data are necessary to accurately estimate disadvantaged populations at a statistically significant level (i.e., $p < 0.05$).* At the census tract level, a

model that did not account for the missing LEP population, described previously as the Basic model, would have yielded a mean absolute error of 11% at the census travel level. In practical terms, accounting for overlapping variables is more important than producing the results at a finer level of geographic detail.

- *More than 95% of the census tracts and more than 90% of the census block groups will have accurate estimations, considering the marginal error of the census data, if the best model is applied.* That is, the testing error resulting from these models is less than the specified census marginal error for nearly all tracts and block groups.

- *Cross-verification demonstrates consistent estimation across time and geographic areas, indicating that the block group model's reliability is not significantly affected by special events (e.g., the COVID-19 pandemic) and produces reliable estimates when errors are aggregated at the tract level.* Applying the block group model (developed from 2020 data) to the 2021 and 2022 data showed that the testing error exceeded the specified census marginal error for 9.4 and 7.2% of block groups, respectively, which is an improvement over the 10% figure for 2020. By applying a block group model and aggregating the estimation error at the tract level, the block group model has an accuracy rate of 84.6%, which is approximately 10% less than the direct application of the Option 2 model at the tract level.

- *The overall yearly growth trend of the average disadvantaged population is statistically significant, as confirmed by the p-value less than 0.05.* These observed changes are meaningful and not random, meaning that the approach offers a useful tool for rough estimations of future populations.

Recommendations

1. *VDOT's TMPD should estimate statewide disadvantaged populations using publicly available data at the block group level to support SMART SCALE project prioritization.* The block group model has been validated for multiyear use with ACS data from 2013 to 2022, demonstrating more than 90% accuracy and reliability for long-term applications. VDOT's TMPD should apply the block group model to the most recent 5-year ACS data, including racial minority populations, poverty levels, and individuals with LEP, to estimate disadvantaged populations at the block group level for the entire state of Virginia. This estimated disadvantaged population should then be used in VDOT's SMART SCALE project prioritization.

2. *VDOT's TMPD should continue using the block group model in the coming years while closely monitoring changes in population over time.* Long-term forecasting methods and models have been proposed for projecting disadvantaged populations at the block group level. To ensure their accuracy and relevance, TMPD should compare the yearly changes in disadvantaged population estimates produced by these methods with the trends observed using the block group model. Based on these comparisons, appropriate adjustments to the estimation formulas of the proposed models should be made to enhance the accuracy of future forecasts.

3. *VDOT's TMPD should consider using the proposed disadvantaged population model to provide more accurate data in the Pathways for Planning (VDOT, n.d.).* The proposed block group model provides more accurate estimates of disadvantaged populations while reducing double-counting compared with the existing data in Pathways for Planning. VDOT's TMPD should apply the block group model to the most recent ACS 5-year data to provide users with the most up-to-date disadvantaged population information in Pathways for Planning.

TABLE OF CONTENTS

FINAL REPORT

ESTIMATING DISADVANTAGED POPULATIONS BASED ON PUBLIC DATA

Yiqing Xu
Research Scientist

Lance E. Dougald
Associate Principal Research Scientist

INTRODUCTION

For project prioritization purposes, the Virginia Department of Transportation's (VDOT) SMART SCALE process defines the disadvantaged population as individuals who meet one or more of three criteria: low income, racial minority, or limited English proficiency (LEP; Commonwealth Transportation Board [CTB], 2021). The SMART SCALE process incorporates equity by evaluating projects on how well they improve accessibility for disadvantaged populations. This approach aligns with CTB's commitment to fostering inclusive growth and promoting equal opportunities, ensuring that transportation planning supports economic and social participation for all residents (CTB, 2021).

Identifying disadvantaged populations is essential for understanding equity in transportation. Transportation equity can be defined as the goal of providing the same access to affordable and reliable transportation to everyone (U.S. Department of Transportation [USDOT], 2022). In transportation planning, disadvantaged populations often face unique mobility challenges due to limited financial resources, disabilities, or lack of access to personal vehicles or reliable public transportation. A disadvantaged population refers to individuals who have experienced racial, ethnic, or cultural biases in society due to their association with certain groups without considering their personal attributes (13 C.F.R. §124.103; Code of Federal Regulations, 2011).

The primary purpose of estimating disadvantaged populations is to promote equitable transportation infrastructure development. The Justice40 Initiative reinforces this need within the Infrastructure Investment and Jobs Act (IIJA; Government Finance Officers Association [GFOA], 2023). The Justice40 Initiative allocates a significant portion of Federal investments to marginalized, underserved, and pollution-burdened communities, mandating that at least 40% of Federal funds benefit these disadvantaged areas. These initiatives underscore the importance of equity and inclusivity in transportation planning.

Following the signing of the IIJA in November 2021, the Federal Highway Administration and the Federal Transit Administration jointly issued an updated set of planning emphasis areas (Federal Transit Administration, 2021), which are crucial considerations for planning agencies. One of the eight planning emphasis areas, "Equity and Justice40 in Transportation Planning," stresses the need for agencies to collaborate on strategies that advance racial equity and support underserved and disadvantaged communities. Key strategies encompass

enhancing the following: (1) providing public transport and infrastructure for pedestrian and nonmotorized travels in underserved regions; (2) prioritizing safety for all users; (3) minimizing single-vehicle travel on busy corridors to decrease air pollution; (4) offering discounted public transport fares when feasible; (5) focusing demand-response services on areas with older residents or with limited access to essential services; and (6) integrating fair and sustainable practices in transit-oriented development, including affordable housing and a focus on environmental justice groups.

Because VDOT transitioned its SMART SCALE accessibility analysis in house in 2018, data on disadvantaged populations are sourced from the American Community Survey (ACS). VDOT's Transportation and Mobility Planning Division (TMPD) observed potential unreliability in the current disadvantaged population data, stemming from three primary concerns: data granularity, collinearity, and accuracy. First, one of these data elements, LEP, is only publicly available at the tract level. Virginia has 1,907 tracts compared with 5,332 block groups. Only one U.S. Census Bureau Table (C16002) exhibits the number of LEP households at the block group level, but the table does not indicate the actual LEP population (U.S. Census Bureau, n.d.a.). Second, for all three data elements, tabulations can generally only be obtained independent of the other two data elements. In other words, census data at the tract level allow users to determine the number of persons meeting each criterion independently of the other two criteria (e.g., the number of low-income individuals in census tract A and, separately, the number of racial minorities in census tract B). If any correlation among these data elements exists (e.g., an individual is both below the poverty level and has LEP), double-counting can occur when summing these values for a particular location. Third, at small units of geography, the margin of error can be relatively large, as shown in Table 1, which provides these attributes for one particular location in Accomack County. For example, at a 90% confidence level, the true number of LEP households in Tract 902 is between 0 and 118..

Table 1. Excerpt of 2015–2019 Data for a Location in Accomack County

Table No.	Data Element	Estimate (margin of error)[b]	Applies to
B16002	Limited English-speaking household	52 (±66)	All block groups within Tract 902 (unavailable for a single block group)
B02009	Black or African American alone or in combination with one or more other races [a]	405 (±279)	Block Group 2 within Tract 902
B17017	Income in the past 12 months below poverty level	38 (±45)	

[a] Comparable table for other races can be tabulated to find total racial minority population.
[b] Margin of error uses a 90% confidence level.

A previous VTRC technical assistance effort (Li et al., 2023) developed an approach for estimating the disadvantaged population across Virginia using 56 Public Use Microdata Areas (PUMAs). PUMAs, which are designed to include populations of about 100,000 people, provide a scale where overlapping criteria can be assessed with custom queries. The model developed by Li et al. (2023) estimated the percentage of the disadvantaged population within a given block group based on the weighted sum of the three criteria percentages: LEP, poverty, and racial minority status. A limitation of this approach is that the reliance on only 56 PUMAs in Virginia for the model development may not capture all the variance that exists at a more granular level

within different tract or block-level groups. The small sample size may limit the generalizability of the findings and the accuracy of predictions. Li et al. (2023) found that "adding the percentages of LEP, poverty, and racial minority populations will overestimate the percentage of the 'disadvantaged population," which necessitates searching for alternate data sources to minimize potential overlap when aggregating these three populations.

Given the challenges associated with correcting errors in ACS data, TMPD sought alternative methods to ensure the accurate extraction of population data, which is an immediate need for VDOT's SMART SCALE project prioritization process and which was pursued as a Transportation Planning Research Advisory Committee research need (Ohlms, 2024). In addition, recent Federal initiatives have raised the possibility that states may be interested in alternative ways of defining these populations. For example, exploring alternatives that may directly influence grant funding opportunities stemming from the IIJA (GFOA, 2023) and helping define the equity areas for VDOT's Office of Intermodal Planning and Investment, which are applicable to Virginia's Transportation Plan, are both needed.

PURPOSE AND SCOPE

In alignment with the need to estimate the disadvantaged population defined in VDOT's SMART SCALE, the Justice40 Initiative under IIJA (GFOA, 2023), and the refinements to the Office of Intermodal Planning and Investment's equity areas for Virginia's Transportation Plan, this project sought to devise a methodology for extracting data on disadvantaged populations from publicly accessible sources. The scope of this endeavor was bound by the currently available data, encompassing open sources and a special tabulation provided by the Census Bureau at the census tract level. The researchers expanded the analysis to the census block group level as data became available through requests for additional Census Bureau tabulations.

METHODS

To meet the objectives of this project, the following tasks were undertaken:

1. Review relevant literature.
2. Obtain disadvantaged population data from a Census Bureau custom tabulation.
3. Cleanse the custom tabulation.
4. Adjust the population universe.
5. Develop the estimation model at the census tract level.
6. Develop the estimation model at the census block group level.

Literature Review

This literature review explored the existing methodologies for identifying disadvantaged populations and implementing best practices in mapping disadvantaged communities. The review included exploring the field of disadvantaged population study and national surveys that organize

key factors and their relationships with disadvantaged populations. The review also included examining various methods, including socioeconomic indicators, environmental justice, and geospatial analyses, used to identify, estimate, and map these populations.

Methods for Identifying Disadvantaged Populations

The identification of disadvantaged populations has significantly improved through the use of geospatial analysis and statistical methods, using a combination of health assessments, demographic surveys, and transportation data. Studies employ techniques such as hotspot analysis, Geographic Information System (GIS)-based models, and the Enhanced Two-Step Floating Catchment Area method to reveal spatial disparities in access to healthcare and transportation. Integrating data from sources like the National Household Travel Survey (NHTS) (FHWA, n.d.) and ACS reveals detailed patterns of mobility challenges and neighborhood disadvantages, which can be used to address the needs of vulnerable groups effectively.

Geospatial Analysis for Analyzing Disadvantaged Populations

Geospatial analysis has emerged as a critical tool for identifying and addressing the needs of disadvantaged populations, as demonstrated in diverse studies. Bejleri et al. (2018) addressed the critical issue of limited transportation options available to vulnerable populations in Florida, including older adults, individuals with disabilities, and low-income groups. They developed a GIS-based geospatial model using data from Alachua County, incorporating the locations and attributes of transportation service providers, travel origins and destinations, the street network, and demographic profiles of transportation-disadvantaged populations. The model computes accessibility scores by considering both the number of destinations and the travel impedance to these destinations within each service area, differentiated by route type. The model demonstrates how geospatial technology can identify and address the transportation needs of vulnerable populations, offering a crucial tool for planning, policy development, and targeted interventions to enhance transportation accessibility.

Gilliland et al. (2019) examined disparities in access to primary healthcare services within the city of London, Ontario, Canada, particularly among vulnerable populations. The study also used a geospatial analysis method, specifically the Enhanced Two-Step Floating Catchment Area method. This approach involved calculating accessibility scores by considering both the number of primary care providers and the population within defined catchment areas. The method assessed the geographic accessibility of primary care providers for specific linguistic minorities, including French, Arabic, and Spanish speakers.

Use of National Household Travel Survey Data

NHTS is a valuable source of national data that allows analysis of personal and household travel trends. It covers daily, noncommercial travel by all modes, including characteristics of travelers, households, and vehicles. NHTS has been widely used to determine transportation patterns and disparities, especially for disadvantaged populations.

Mattson (2012) used 2001–2009 NHTS data to analyze travel behavior, including driving frequency, mode choice, trip purpose, and distance. The study defined transportation-disadvantaged groups as older adults, people with disabilities, individuals in low-income households, and those living in rural areas who faced mobility challenges. The study found that (1) older adults (aged 65 and older) reported fewer trips and a lower likelihood of driving, which indicated a significant reduction in mobility with age, especially for those older than 85; (2) disadvantaged groups exhibited significantly fewer trips and a lower likelihood of driving compared with those without disabilities; and (3) individuals from low-income households (below the poverty line) made fewer trips and had lower automobile access. These results revealed the critical mobility challenges transportation-disadvantaged groups face.

Mattson and Molina (2022) then presented a comprehensive analysis of travel behavior among transportation-disadvantaged groups—including older adults, individuals with disabilities, low-income households, and rural residents—using data from the 2017 NHTS. The study reviewed the mobility challenges faced by these groups and revealed differences in travel behavior based on age, disability, income, and geography. For instance, the study explored trip rates, miles driven, mode shares, and other behaviors and compared findings with data from 2009 and 2001 to identify trends. The analysis found a decline in driving and trip frequency among disadvantaged groups, a heavy reliance on automobiles in rural areas, and the effect of socioeconomic factors on travel decisions.

Brumbaugh (2018) performed a similar analysis using data from the 2017 NHTS to analyze the daily travel patterns of American adults with travel-limiting disabilities, highlighting trends over time by comparisons with 2001 and 2009 NHTS data. This study discussed the challenges faced by individuals with disabilities, including lower employment rates, lower household incomes, fewer trips made per day, reliance on others for transportation, and reduced access to technology and ride-hailing services. The report also explored compensatory strategies adopted by people with disabilities, such as limiting travel to daytime, using special transportation services, and reducing day-to-day travel. In addition, it evaluated the potential of technology, including autonomous vehicles, to mitigate transportation limitations for people with disabilities.

Utilization of American Community Survey Data

Spielman and Singleton (2015) explored using ACS data to analyze neighborhood characteristics. Recognizing the challenges of large margins of error in ACS data for small geographic areas, the authors proposed shifting from a variable-based mode of inquiry to a composite, multivariate analysis of census tracts. This method proposed a collection of variables to average out these errors, providing a more accurate and detailed picture of neighborhood contexts. This study developed a geodemographic typology (i.e., a method of classifying and organizing geographic areas or populations based on a combination of demographic and geographic characteristics) of the neighborhood characteristics for all U.S. census tracts validated with public domain data from the City of Chicago and the U.S. Federal Election Commission. This approach addressed ACS's limitations by using its rich dataset to construct a nuanced understanding of neighborhood characteristics.

Kind and Buckingham (2018) focused on addressing health disparities in the United States using the Area Deprivation Index and ACS data. The methodology involved updating the Area Deprivation Index with recent ACS data and making this information accessible through the Neighborhood Atlas, a platform that visualizes neighborhood disadvantage. The U.S. Health Resources and Services Administration originally created the Area Deprivation Index using long-form census data. Kind and Buckingham (2018) updated the index to incorporate more recent ACS data, comprising 17 measures related to education, employment, housing quality, and poverty. This effort aimed to promote an increased understanding of how neighborhood disadvantage affects health by providing detailed cross-references of more than 69 million nine-digit ZIP Codes. This effort allows the index to be merged with a wide variety of other data resources.

Statistical Method for Analyzing Disadvantaged Population

Mattson (2012) analyzed travel behavior and mobility among transportation-disadvantaged groups based on NHTS data incorporating two statistical analyses: regression analysis and cluster analysis. Regression analysis uses binary logit models to identify characteristics of individuals who are interested in going out more frequently, particularly focusing on those who have not taken a trip in more than a week. Cluster analysis categorized NHTS respondents into 12 groups based on factors such as household income, age, gender, household size, and medical conditions affecting travel ability. Mattson analyzed travel behavior within each cluster to identify distinct patterns among transportation-disadvantaged groups so as to reveal the mobility challenges faced by those groups.

Case (2009) explored methods for estimating the residential locations of nondrivers at the census block level, focusing particularly on those in zero-vehicle households because of their vulnerability during evacuation events and their need for mobility enhancements like nearby bus stops and activity locations. This study used regression techniques on earlier nondriver location data and applicable census data to achieve these estimates for more than 20,000 blocks in Hampton Roads. This process involved subdividing transportation analysis zone-level nondriver data, applying linear regression to establish relationships between census-provided data at the block level and nondriver numbers, and refining models to accurately represent the distribution of nondrivers in zero-vehicle households and those with vehicles.

Practices of Mapping Disadvantaged Communities

The researchers explored disadvantaged community mapping through GIS, which enables the visualization of geographic concentrations of inequality and social needs. GIS has been widely used to map disadvantaged areas using spatial analysis techniques to overlay the transportation infrastructure with demographic data.

Equity and Justice40 analysis tools developed by USDOT (2024) are designed to help identify and assist disadvantaged communities through grant programs and initiatives, ensuring that benefits from transportation investments reach those most in need. Key tools include the following:

1. Climate and Economic Justice Screening Tool (Council on Environmental Quality, 2022) was designed by the White House Council on Environmental Quality and includes an interactive map and datasets that indicate burdens in eight categories at the census tract level: climate change, energy, health, housing, legacy pollution, transportation, water and wastewater, and workforce development. These indicators help identify overburdened communities that are underserved and thereby disadvantaged.
2. Equitable Transportation Community Explorer (USDOT, 2023a) helps users understand how communities or project areas experience disadvantages related to transportation investments or opportunities. This tool aims to identify how projects can reverse or mitigate these disadvantages.
3. Areas of Persistent Poverty and Historically Disadvantaged Communities mapping tool (USDOT, 2023b) published GIS maps, including the disadvantaged population data in the attribute table, to help users determine if a grant project is within an Area of Persistent Poverty or a Historically Disadvantaged Community.
4. Environmental Justice Screen was developed by the U.S. Environmental Protection Agency (2024) and provides nationally consistent demographic and environmental information for project areas. This tool combines environmental and demographic socioeconomic indicators into environmental justice indices, offering capabilities like color-coded mapping, standard reports for selected areas, and comparisons with state or national levels.
5. PLACES, developed by the CDC (2023), provides population-level analysis and community estimates of health measures across the United States.
6. Screening Tool for Equity Analysis of Projects (USDOT, 2023c) is a GIS tool for assessing data layers, including race, color, and national origin. Developed by the Federal Highway Administration, the tool enables rapid screening of potential project locations for Title VI and environmental justice considerations using ACS data.
7. Justice40 Rail Explorer (USDOT, 2023d) facilitates understanding how rail infrastructure intersects with communities and the potential benefits of rail investments, allowing users to explore existing and future rail infrastructure in the context of community experiences and burdens.

The Office of Economic Impact and Diversity (2023) created the Energy Justice Mapping Tool (shown in Figure 1), which categorizes disadvantaged communities based on 36 burden indicators that reflect fossil fuel dependence, energy burden, environmental and climate hazards, and socioeconomic vulnerabilities at the census tract level. This tool adopts a cumulative burden approach to define disadvantaged communities and calculates the indicators at the census tract level. The methodology involves calculating percentile values for each indicator, aggregating these values into a composite score, and selecting tracts based on these scores and income criteria. This process ensures that the communities identified not only face significant environmental and socioeconomic challenges but also include a diverse representation across states. The tool provides a visual and analytical platform to explore disadvantaged communities that offers users various search functionalities—including geography, tract number, tribal name, or territory name—and displays detailed burden indicator data for selected tracts.

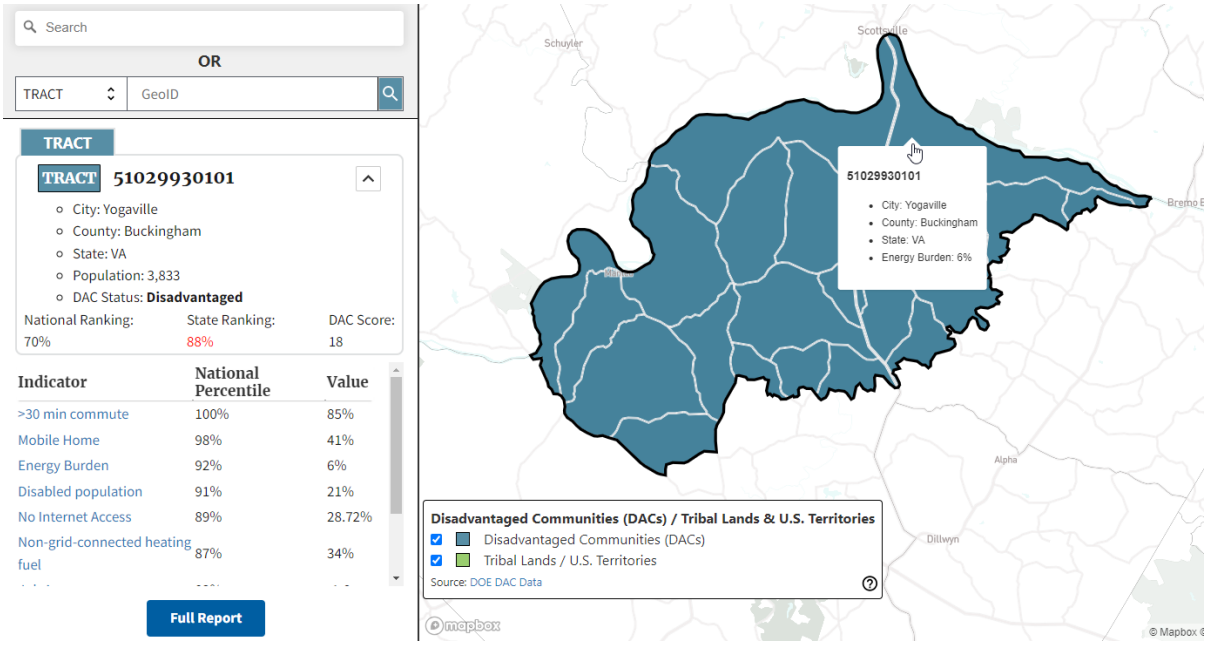


Figure 1. Screenshot of the Energy Justice Mapping Tool (Office of Economic Impact and Diversity, 2023; reprinted with permission)

Other states and regions are implementing GIS methods to map and understand the composition and needs of their communities, aiding in equitable transportation planning and policymaking. For example, the Memphis Metropolitan Planning Organization (2020) produced maps (Figure 2) that show the concentration of vulnerable communities within the region, comparing these demographics with regional averages. This comparison enabled the identification of areas with higher percentages of minorities, LEP individuals, seniors, low-income households, persons with disabilities, people without vehicle access, and people without internet access.

Figure 2. Memphis Metropolitan Planning Organization (MPO) Census Block Groups and Tracts that Exceed the Regional Average of the Following Groups: (a) Racial and Ethnic Minorities; (b) limited English proficiency; (c) Aged 65 and Older; (d) Poverty and Low-Income Households; (e) Persons with Disabilities; (f) Persons without Vehicle Access; and (g) Persons without Internet Access. Reprinted with permission from Memphis MPO.

The California Office of Environmental Health Hazard Assessment developed a mapping tool (Figure 3) to visually represent the areas identified as disadvantaged communities within California (California Environmental Protection Agency, 2021). These communities are designated based on criteria that include geographic, socioeconomic, public health, and environmental hazard factors. The map is used to guide the allocation of funds from California's Cap-and-Trade Program, ensuring that investments are directed toward improving the quality of

life, public health, and economic opportunities in these communities while also contributing to the reduction of pollution and the mitigation of climate change. The tool is designed to be user-friendly. By clicking on a census tract, users can view additional information about that area, helping stakeholders understand why certain areas are designated as disadvantaged and the specific challenges they face. Users can export a map view that includes the legend and any open pop-up windows, facilitating the sharing and analysis of the information.

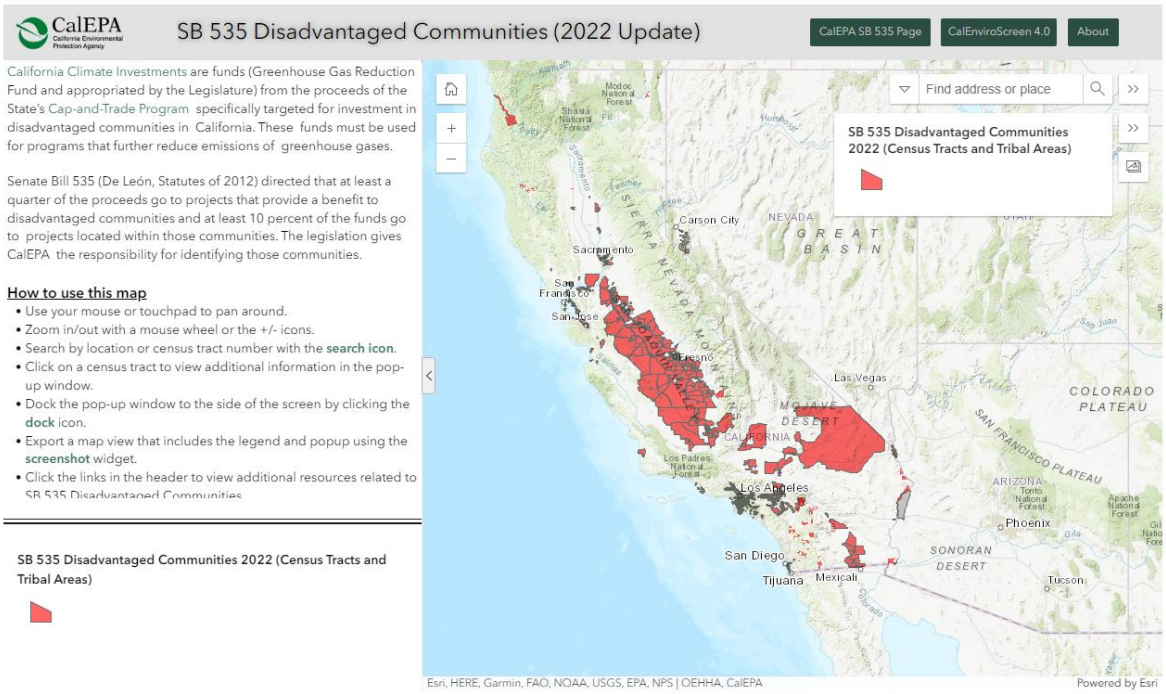


Figure 3. Screenshot of the SB 535 Disadvantaged Communities Map. Reprinted with permission from California Environmental Protection Agency.

New York State recognizes that climate change does not affect all communities equally and aims to ensure that frontline and underserved communities benefit from the transition to cleaner and greener energy sources, reduced pollution, cleaner air, and economic opportunities. The Climate Justice Working Group in New York finalized criteria on March 27, 2023, to identify these disadvantaged communities in an interactive map (New York State Energy Research and Development Authority, 2023). This interactive map (Figure 4) allows users to identify areas throughout the state that meet the criteria. Users can determine whether an address is in a disadvantaged community by entering it in a search box or by zooming in and out to view different parts of counties, cities, towns, and neighborhoods. This map is designed to help visualize the locations of these communities and understand the geographic scope of the Climate Act's initiatives. For example, New York State's Climate Act (New York State Climate Action Council, 2022) requires New York to reduce economywide greenhouse gas emissions.

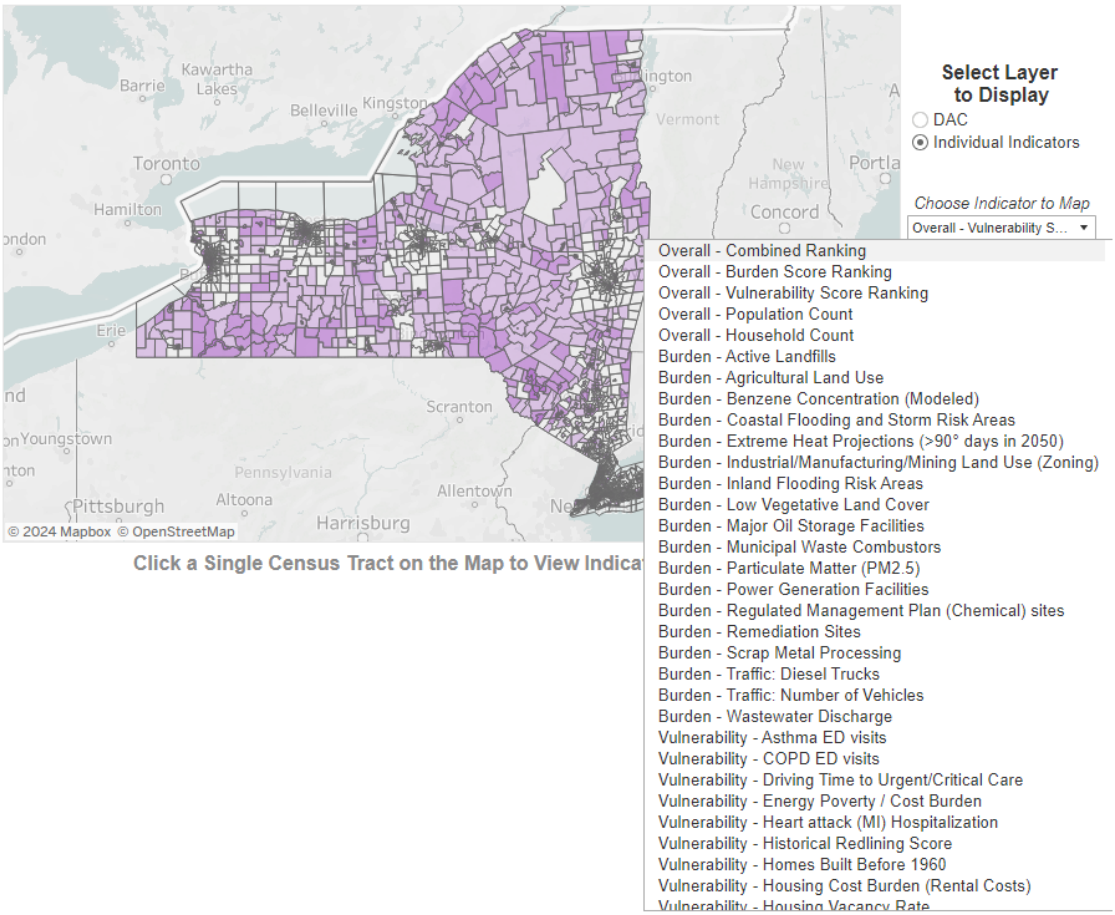


Figure 4. Screenshot of the Disadvantaged Communities Criteria Maps Published by the New York State Energy Research and Development Authority (NYSERDA, 2023). Reprinted with permission from NYSERDA.

Obtain Disadvantaged Population Data from a Census Bureau Custom Tabulation

During an 18-month period, the research team worked with the Census Bureau to purchase a custom tabulation of disadvantaged persons that met the input criteria used in SMART SCALE while also following disclosure rules established by the Census Disclosure Review Board. That tabulation yielded the number of disadvantaged persons by census tract and census block group in Virginia, where the designation of disadvantaged was based on three criteria:

- LEP: Refers to persons for whom English is not their primary language and who have a limited ability to read, write, speak, or understand English. LEP includes people who reported to the U. S. Census that they speak English at a level less than very well, not well, or not at all (Rogoff, 2012; U.S. Department of Health and Human Services, 2024; U.S. Department of Justice, 2020).
- Racial minority: Defined as Black, American Indian, Asian, Pacific Islander, Other, and/or Hispanic. Individuals who indicated two or more races were automatically included.

- Low-income: Refers to individuals whose household income in the past 12 months was below the poverty level.

Cleanse the Custom Tabulation

The research team identified three issues in the custom tabulation that may influence the estimation of the disadvantaged population. These issues—limited population universe, deliberate introduction of error by the Census Bureau, and large marginal errors—were identified through a comparison between the custom tabulation and publicly available tables.

Limited Universe of Population for the Custom Tabulation

The custom tabulation limited the table universe to populations 5 years and older for whom poverty status is determined (Spanos, 2023a) for all census tracts or census block groups in Virginia and excluded institutionalized persons (e.g., persons in prisons), people in military group quarters, people in college dormitories, and people whose poverty status is not determined (Spanos, 2023b, 2024b). In this context, the total population of the custom tabulation is smaller than the total population of the publicly available census tables.

For the census tract tables, Table 2 shows a population difference between Census Table S0601 and the custom tabulation in six census tracts where total population and racial minority population are different populations than those listed in Census Table S0601. For example, four census tracts have both total population and racial minority population as zero, and two census tracts have low total population and zero racial minority population in the custom tabulation (U.S. Census Bureau, n.d.b.). The Census Bureau "suppressed" these six census tracts when generating the custom tabulation. That is, these populations are left out of the universe of the custom tabulation (Spanos, 2024a). Therefore, they have been excluded in the modeling at the census tract level.

Table 2. Example of Errors Caused by Suppression in the Custom Tabulation

GEO_ID	Name	Table S0601		Custom Tabulation	
Geography	Geographic Area Name	Total Population	Racial Minority Population	Total Population	Racial Minority Population
51740980100	Census Tract 9801, Portsmouth City	661	356	0	0
51081880102	Census Tract 8801.02, Greensville County	3,166	2,232	5	0
51175200300	Census Tract 2003, Southampton County	1,301	703	0	0
51059440504	Census Tract 4405.04, Fairfax County	1,012	469	0	0
51121020100	Census Tract 201, Montgomery County	8,832	1,528	4	0
51183870202	Census Tract 8702.02, Sussex County	855	674	0	0

GEO ID = geographic identifier.

For the census block group tables, 20 census block groups from the custom tabulation have a disadvantaged population of zero, and both the total population and disadvantaged population deviate from the total populations in the publicly available data because of the limited

universe of the custom tabulation used. Therefore, these 20 census block groups (shown in Appendix A) have been excluded from the modeling at the census block group level.

Data Error Introduced by the Discrete Gaussian Noise Mechanism

The custom tabulation involves the application of a discrete Gaussian noise mechanism to the census data, which introduces noise into the data to ensure confidentiality. As noted by Spanos (2023c), this noise mechanism for privacy protection is added to data to maintain the privacy of individual records while still allowing for accurate statistical analysis. For example, small values are adjusted to add noise, with a specific rounding method applied where values between 1 and 7 are rounded to 4, and 0 remains as 0. This technique ensures that while the data retain their utility for analysis, individual responses are protected, albeit at the cost of some accuracy due to the added noise and rounding adjustments. In this context, this rounding was done for both tabulations for census tracts and census block groups. Therefore, errors were introduced to both tabulations the researchers received.

In 15 out of 2,198 census tracts in this dataset, the disadvantaged population exceeds the total population within the same dataset (i.e., the custom tabulation; Table 3 shows five examples). For analysis, the disadvantaged population of those 15 census tracts was adjusted to match the total population. For example, for census tract 51087201405, the disadvantaged population was changed from 2,258 to 2,253 to match its total population in the analysis.

Table 3. Example of Errors Caused by the Noise Mechanism in the Custom Tabulation

GEO_ID	Name	Title	Census Estimate	Disadvantaged Population/ Total Population
51700030800	Census Tract 308, Newport News City, Virginia	Total population	1,576	1.002
		Disadvantaged population	1,579	
51760020200	Census Tract 202, Richmond City, Virginia	Total population	3,184	1.000
		Disadvantaged population	3,185	
51650011400	Census Tract 114, Hampton City, Virginia	Total population	303	1.069
		Disadvantaged population	324	
51087201405	Census Tract 2014.05, Henrico County, Virginia	Total population	2,253	1.002
		Disadvantaged population	2,258	
51590980100	Census Tract 9801, Danville City, Virginia	Total population	0	—
		Disadvantaged population	24	

— = no data. GEO ID = geographic identifier.

For the census block group tables, 30 of the 5,963 census block groups have a disadvantaged population larger than the total population (see Appendix B). Therefore, for analysis, the disadvantaged populations for these census block groups have been adjusted to match their total populations.

Large Marginal Error for the Custom Tabulation

In some cases, the disadvantaged population in the census tract custom tabulation was not equal to zero, but the total disadvantaged population (comprising total racial minority plus total

LEP plus total poverty) from the publicly available data was zero. Table 4 shows this error, which can be explained by the census margin of error (CME). For example, the marginal error indicated by the DP_CME in Table 1 is close to the total disadvantaged population in the custom tabulation, whereby the total disadvantaged population is close to zero when factoring in the error plus or minus the DP_CME. Therefore, these four census tracts in Table 4 were excluded from the analysis. The researchers found no such error in the census block group tables.

Table 4. Example of Large Marginal Errors in the Custom Tabulation at Census Tract Level

GEO ID	Custom Tabulation			Census Table S0601		Census Table S1701: Total White Poverty	Census Table B16009: Total LEP
	Total Population	Total Disadvantaged Population	DP_CME	Total Population	Total Racial Minority		
51670980100	7	6	±13	0	0	0	0
51103990100	11	1	±13	0	0	0	0
51119990100	12	4	±13	0	0	0	0
51053980100	10	1	±13	0	0	0	0

DP_CME = disadvantaged population census marginal error in the custom tabulation; GEO ID = geographic identifier; LEP = limited English proficiency.

Adjust the Population Universe

The limited population universe used for the custom tabulation means that the population used in this study is 7.2% less than the ground truth population on average, the 2016–2020 ACS. Accordingly, the disadvantaged population variable used in model development was calculated using Equation 1.

$$\mathrm{DP}_{\mathrm{dependent}} = \mathrm{TPPT} \times \frac{\mathrm{DPCT}}{\mathrm{TPCT}} \quad \text{(Eq. 1)}$$

Where:
$\mathrm{DP}_{\mathrm{dependent}}$ = disadvantaged population used as the dependent variable in modeling
TPPT = total population from the publicly available table.
DPCT = disadvantaged population of the custom tabulation.
TPCT = total population of the custom tabulation.

This approach is employed to account for potential disadvantaged populations that are not included in the custom tabulation universe.

Develop the Estimation Model at the Census Tract Level

The researchers developed a three-step methodology to estimate disadvantaged populations at the census tract level, as illustrated in Figure 5. The first step involves identifying the required table and filtering the necessary columns. The second step entails developing the estimation models. Least squares regression is employed for developing these models and processed in the statistical software package SPSS (IBM, n.d.). The third step is evaluating the models based on four metrics: coefficient of determination (adjusted R^2), standard error of

estimate (SEE), mean testing error, and plots of residuals.

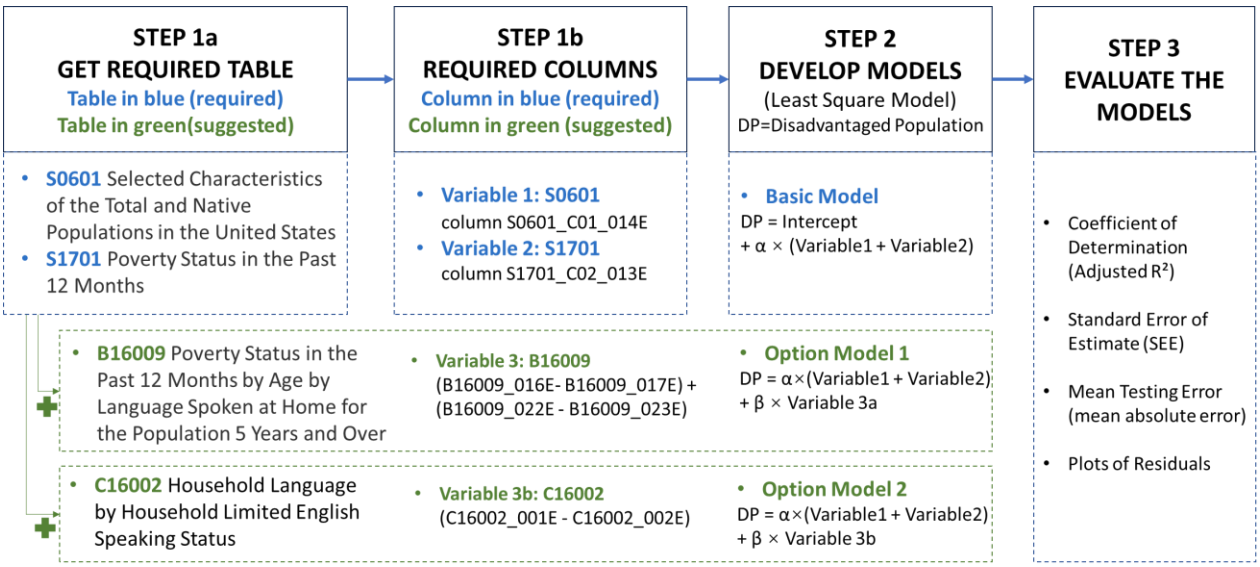


Figure 5. Steps for Modeling

Step 1: Obtain the Required Table and Columns at the Census Tract Level

Given the limited data availability because some census tracts lack population data from 1-year census tables, researchers used the 2020 ACS 5-year (2016–2020) estimates to determine the total population of racial minorities and the White population below the poverty level (U.S. Census Bureau, n.d.b., n.d.c.).

The independent variables consist of three discrete components (Figure 6): racial minority (in the orange rectangle), White population below the poverty level (in the green rectangle), and White LEP not below the poverty level (in the blue rectangle). Note that for these data elements, 5-year (2016–2020) ACS estimates are used in this analysis to match what the custom tabulation used. Although the Census Bureau provides extensive publicly available data on various demographic groups, no specific census table directly presents information on the White LEP population not below the poverty level at both the census tract and census block group level.

Variable 1—Total Racial Minority at the Census Tract Level

For the racial minority population, two data columns are needed. The first is column S0601_C01_014E in Table S0601, which is the total percentage of White alone population within the census tract. The second is column S0601_C01_001E in table S0601, which indicates the total population within the census tract. The total racial minority population within the census tract = the total population within the census tract S0601_C01_001E × (1 − the total percentage of White alone population within the census tract S0601_C01_014E) (U.S. Census Bureau, n.d.b.).

Variable 2—White Population Below the Poverty Level at the Census Tract Level

One data column is needed to represent the population of White individuals below the poverty level. The column labeled S1701_C02_013E in Census Table S1701 contains the total population of White individuals below the poverty level (U.S. Census Bureau, n.d.c.).

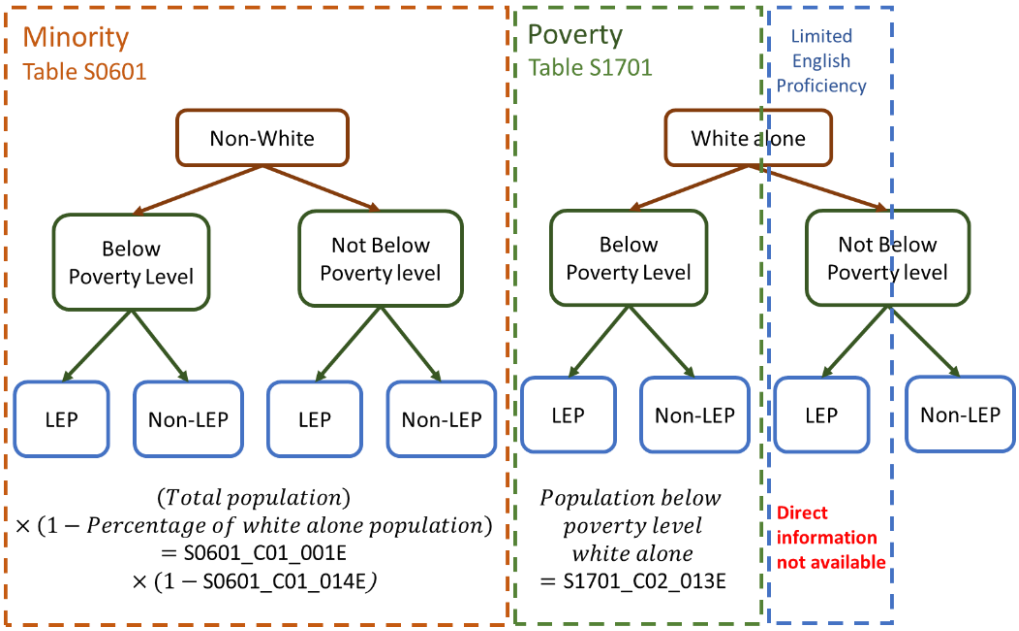


Figure 6. Disadvantaged Population Components. LEP = limited English proficiency

Variable 3—Total LEP Population at the Census Tract Level

Variables 1 and 2 are non-overlapping population segments that can be directly used to determine disadvantaged populations. Ideally, Variable 3 should provide exact data about the White LEP population that is not below the poverty level. However, as these data are unavailable at the tract level, the researchers applied modeling to estimate the LEP population as closely as possible to the target demographic.

To address the absence of exact data for the White LEP population not below the poverty level, relying on alternative data sources and modeling techniques was necessary. Neither the 1- nor 5-year census tables provide precise statistics for this specific population segment. As a result, the researchers used alternative data from the 2020 ACS 5-year estimates, such as Census Tables B16009 or C16002, to approximate this group within the LEP population (U.S. Census Bureau, n.d.a., n.d.d.).

Census Tables B16009 and C16002 categorize individuals as either speaking only English or speaking a language other than English based on income level, but they do not fully align with the definition of LEP (U.S. Census Bureau, n.d.a., n.d.d.). The specific concern is that individuals who speak only English but at a proficiency level below “very well” may be excluded from the data represented by these tables. Individuals who speak a language other than English but also speak English “very well” may be included inappropriately.

One key assumption in using Census Tables B16009 and C16002 is that all nonpoverty individuals who do not exclusively speak English are considered the closest available approximation of the nonpoverty LEP population (U.S. Census Bureau, n.d.a., n.d.d.). This assumption relies on the only two available tables at the census tract level, where individuals are categorized based on both their primary language and income. Although this category may not perfectly align with the true definition of LEP, it is the closest approximation available for modeling purposes.

The four columns required from Census Table B16009 (U.S. Census Bureau, n.d.d.) needed for the LEP estimation include the following:

- B16009_016E: Total population with income above the poverty level, ages 5 to 17.
- B16009_017E: Total population speaking only English with income above the poverty level, ages 5 to 17.
- B16009_022E: Total population with income above the poverty level, ages 18 and older.
- B16009_023E: Total population speaking only English with income above the poverty level, ages 18 and older.

The total population above the poverty level that do not exclusively speak English = (total population with income above the poverty level, ages 5 to 17 B16009_016E − total population speaking only English with income above the poverty level, ages 5 to 17 B16009_017E) + (total population with income above the poverty level, ages 18 and older B16009_022E − total population speaking only English with income above the poverty level, ages 18 and older B16009_023E).

The two columns required from Census Table C16002 (U.S. Census Bureau, n.d.a.) needed for the LEP estimation include the following:

- C16002_001E: Total number of households.
- C16002_002E: Total number of households speaking only English.

The total number of households that do not exclusively speak English = (total number of households C16002_001E) − (total number of households speaking only English C16002_002E) .

Step 2: Develop Least Squares Models at the Census Tract Level

The researchers developed three models. The first is the Basic model, which includes only the variables of racial minority (Variable 1) and White population below the poverty level (Variable 2). The researchers also developed two optional models based on the Basic model, each incorporating an additional variable that represents the LEP population.

Basic Model

The dependent variable is the total number of disadvantaged populations at the tract level, which was provided by the Census Bureau in the custom tabulation (Spanos, 2023a), whereas the

independent variables include Variables 1 and 2 in Step 1:

$$\text{Disadvantaged population} = \text{intercept} + \alpha \times (\text{Variable 1} + \text{Variable 2}) \quad \text{(Eq. 2)}$$

Where:
α = coefficient.
Variable 1 = total racial minority population from Census Table S0601 (U.S. Census Bureau, n.d.b.).
Variable 2 = total White population below the poverty level from Census Table S1701 (U.S. Census Bureau, n.d.c.).

Theoretically, the total disadvantaged population should be no less than the sum of the population for racial minorities (Variable 1) and the White population below the poverty level (Variable 2), as the total disadvantaged population also includes the LEP population above the poverty level. In this scenario, an intercept needs to be introduced to represent the missing LEP population. For example, this model is logically transformed from the model that predicts Variable 3, which treats Variables 1 and 2 as a combined factor: Variable 3 = constant + coefficient * (Disadvantaged population – Variable 1 − Variable 2). Therefore, the coefficient in the final model can potentially explain the percentage that Variable 3 represents of the total population on average.

Considering the existence of this intercept, this model is not applicable to census tracts with zero population counts for the two variables and the dependent variable. Therefore, the researchers have excluded 28 census tracts from the development of the Basic model, leaving 2,160 census tracts for modeling.

Option Models

An option model was built twice using all 2,188 census tracts. One model involved adding a variable to the basic model based on the LEP household data from Census Table C16002, whereas the other replaced that variable with the LEP population data from Census Table B16009 (U.S. Census Bureau, n.d.a., n.d.d.). These two models have the same form:

$$\text{Disadvantaged population} = \alpha \times (\text{Variable 1} + \text{Variable 2}) + \beta \times \text{Variable 3} \quad \text{(Eq. 3)}$$

Where:
α and β = coefficients.
Variable 1 = total racial minority population from Census Table S0601 (U.S. Census Bureau, n.d.b.).
Variable 2 = total White population below the poverty level from the Census Table S1701 (U.S. Census Bureau, n.d.c.).
Variable 3 = total population above the poverty level who do not exclusively speak English from either Census Table B16009 or C16002 (U.S. Census Bureau, n.d.a., n.d.d.).

Logically, the total disadvantaged population is equal to the sum of Variable 1, Variable 2, and Variable 3, without the coefficient and the intercept. If all the variables are zero, no disadvantaged population would be present. That is, the dependent variable should also be zero. Therefore, regression without intercept is used for the option models.

The researchers did not conduct a test for nonconstant variance because the model's form is predetermined by the definition of a disadvantaged population. As a result, the researchers sought more accurate data for the representation of Variable 3, rather than making modifications to the model's structure. For this reason, the researchers tested two different data tables in this analysis.

Step 3: Evaluate the Census Tract Models

The researchers randomly spit the data into a training set (70% of the data) and a testing set (the remaining 30%). The first two metrics in the following list were used to evaluate the goodness of fit of the model to the training data, and the latter two metrics were used to evaluate how the model performed when applied to a new dataset.

- Coefficient of determination (adjusted R^2): This metric measures the proportion of the variance in the dependent variable that can be explained by the independent variables (features) in the model. A higher adjusted R^2 indicates a better fitting model.
- SEE: This metric quantifies the average distance between the observed values and the predicted values by the least squares regression model. A lower SEE indicates that the model's predictions are closer to the actual data points, suggesting better accuracy.
- Mean testing error (mean absolute error): This metric refers to the average error of the model when tested on the remaining 30% of the dataset. A lower mean testing error suggests better predictive performance.
- Plots of residuals: Residuals represent the differences between the observed values and the predicted values from the model. Plotting the residuals helps identify patterns in the errors. A model with residuals having a mean value close to zero is considered unbiased. A consistent spread of the residuals across the plot indicates that the model has constant variance and is homoscedastic. Generally, an unbiased and homoscedastic model is preferred.

Develop the Estimation Model at the Census Block Group Level

Following a similar modeling process as Figure 5 shows, three tables containing census block group data are required. Only one category of data (i.e., White LEP population not below the poverty level) is not available at the census tract level. However, at the census block group level, two categories of data are missing (i.e., White poverty population and White LEP population not below the poverty level). Therefore, different tables are used in the model building for the block group level versus the tract level. In total, 5,866 census block groups, excluding block groups with data errors, are involved in the modeling.

Step 1: Obtain the Required Table and Columns at the Census Block Group Level

Census block groups have less available data than the census tracts. For example, many census block groups lack population data from 1-year census tables, and no table separates the poverty population from the racial minority population at the census block group level. The researchers used the 2020 ACS 5-year (2016–2020) estimate to determine the total population of racial minorities (U.S. Census Bureau, n.d.e.), the total population below the poverty level (U.S. Census Bureau, n.d.f.), and the total LEP population (U.S. Census Bureau, n.d.g.) for all the census block groups.

Variable 4—Total Racial Minority Population at the Census Block Group Level

The researchers used two columns in Census Table B02001 (the total population in column B02001_001E and the total White alone population in column B02001_002E) for calculating the total racial minority population for each census block group (U.S. Census Bureau n.d.e.):

> Total Racial Minority Population for Each Census Block Group = B02001_001E (Total Population) − B02001_002E (Total Population White Alone)

Variable 5—Total Poverty Population at the Census Block Group Level

The researchers used the column of total poverty population, B17021_002E, in Census Table B17021 (Poverty Status of Individuals in the Past 12 Months by Living Arrangement), to represent the total poverty population for each census block group:

> Total Population Below the Poverty Level for Each Census Block Group = B17021_002E (Total Poverty Population)

Variable 6—Total LEP Population at the Census Block Group Level

The researchers used the following 18 columns in Census Table B16004 (Age by Language Spoken at Home by Ability to Speak English for the Population 5 Years and Over) to calculate the total LEP population for each census block group (U.S. Census Bureau n.d.g.). Total LEP Population for each census block group =

+	B16004_002E	Total population 5 to 17 years
-	B16004_003E	Population 5 to 17 years speaking only English
-	B16004_005E	Population 5 to 17 years speaking Spanish; speaking English very well
-	B16004_010E	Population 5 to 17 years speaking other Indo-European languages; speaking English very well
-	B16004_015E	Population 5 to 17 years speaking Asian and Pacific Island languages; speaking English very well
-	B16004_020E	Population 5 to 17 years speaking other languages; speaking English very well
+	B16004_024E	Total population 18 to 64 years
-	B16004_025E	Population 18 to 64 years speaking only English
-	B16004_027E	Population 18 to 64 years speaking Spanish; speaking English very well

- B16004_032E Population 18 to 64 years speaking other Indo-European languages; speaking English very well
- B16004_037E Population 18 to 64 years speaking Asian and Pacific Island languages; speaking English very well
- B16004_042E Population 18 to 64 years speaking other languages; speaking English very well
+ B16004_046E Total population 65 years and over
- B16004_047E Population 65 years and over speaking only English
- B16004_049E Population 65 years and over speaking Spanish; speaking English very well
- B16004_054E Population 65 years and over speaking other Indo-European languages; speaking English very well
- B16004_059E Population 65 years and over speaking Asian and Pacific Island languages; speaking English very well
- B16004_064E Population 65 years and over speaking other languages; speaking English very well

Step 2: Develop Least Squares Models at the Census Block Group Level

Given that the three variables in the publicly available data have overlapping populations (e.g., the total racial minority population includes some individuals from the poverty population, and the total poverty population includes some individuals from the LEP population), the researchers adjusted the model structure based on the census tract model. Specifically, the census tract model combines Variables 1 and 2 into a single variable in Equations 2 and 3 because no populations overlap between these two variables. Conversely, at the census block group level, where all variables share overlapping populations, Equation 4 treats each variable separately, assigning a distinct coefficient to each:

$$\text{Disadvantaged population} = \alpha \times \text{Variable 4} + \beta \times \text{Variable 5} + \gamma \times \text{Variable 6} \qquad \text{(Eq. 4)}$$

Where:
α, β, γ = coefficients.
Variable 4 = total racial minority population from the Census Table B02001 (U.S. Census Bureau, n.d.e.).
Variable 5 = total poverty population from the Census Table B17021 (U.S. Census Bureau, n.d.f.).
Variable 6 = total LEP population from the Census Table B16004 (U.S. Census Bureau, n.d.g.).

If all the variables are zero, no disadvantaged population would be present. That is, the dependent variable should also be zero. Therefore, regression without intercept is used for this model.

Step 3. Evaluate the Models at the Census Block Group Level

Like the evaluation of the census tract model, the dataset was split into a 70% training set for model training and a 30% testing set for performance evaluation. Evaluation metrics that aim to address model fit, accuracy, predictive performance, and bias include adjusted R^2, SEE, mean absolute error, and residual plots.

RESULTS

The results provide findings from the disadvantaged population distribution in the custom tabulation, along with the developed models and their performance with respect to training and testing data at both the census tract and census block group levels.

Disadvantaged Population Distribution in the Custom Tabulation

Census Tract Level

In total, Virginia has 2,198 census tracts, and disadvantaged population data were received for all these tracts through custom tabulation. The number of persons classified as disadvantaged populations ranged from a low of 0 to a high of 6,398, with a mean value of 1,527 and a median value of 1,264. Figure 7 shows a distribution map where the darker shades represent higher disadvantaged populations.

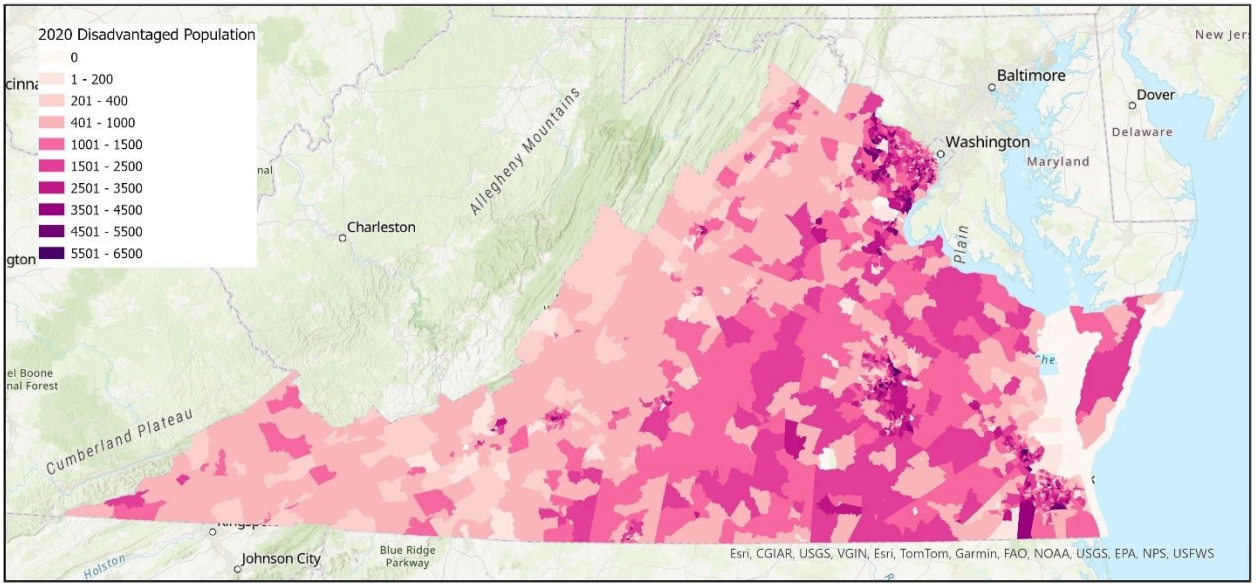


Figure 7. Disadvantaged Population for all Census Tracts in Virginia (2020)

Figure 8 indicates the distribution of the disadvantaged population index across census tracts in Virginia as of 2020. The index measures the disadvantaged population divided by the total population. Darker colors represent higher concentrations of disadvantaged populations, revealing that the disadvantaged population index is higher near urbanized regions, notably in areas of the Fredericksburg, Hampton Roads, Richmond, and Northern Virginia Districts. In contrast, the western part of the state, which includes more rural regions, such as the Cumberland Plateau and areas in the Bristol and Salem Districts, shows a generally lower index of disadvantaged populations.

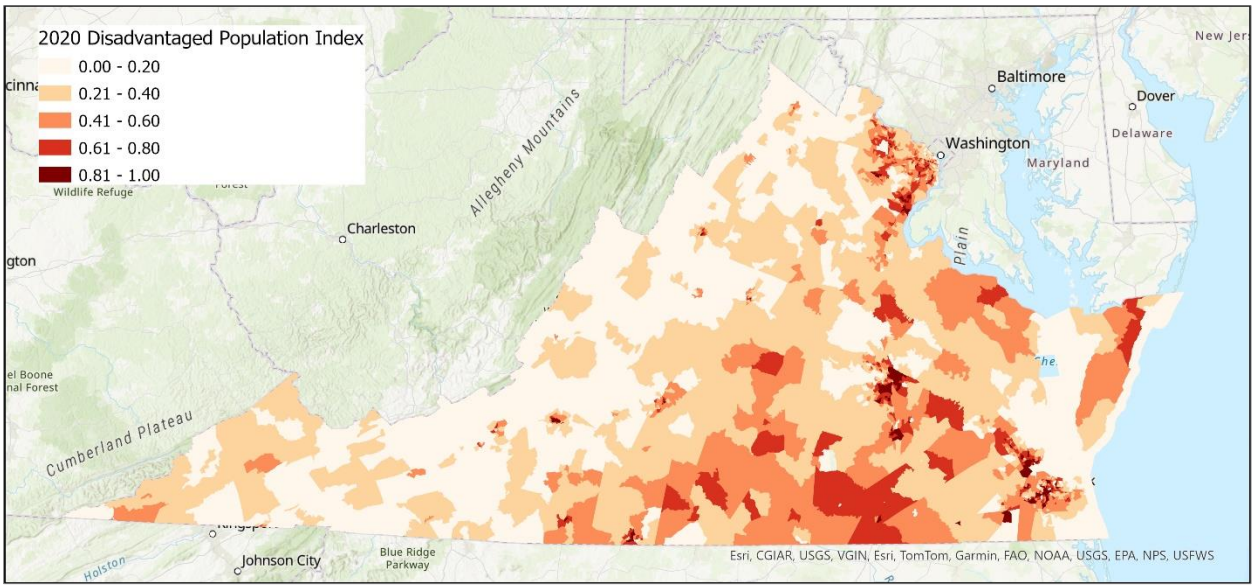


Figure 8. Disadvantaged Population Index for all Census Tracts in Virginia (2020)

Census Block Group Level

Virginia has 5,963 census block groups, and disadvantaged population data were received for all these census block groups through custom tabulation. The number of persons defined categorically as a disadvantaged population ranged from a low of 0 to a high of 3,896 for all block groups, with a mean value of 563 and a median value of 433.

Figure 9 shows the 2020 disadvantaged population for all census block groups in Virginia. This map provides a granular perspective of demographic distribution and presents a more precise visualization of where disadvantaged populations are concentrated. In this map, the color scale progresses from light pink to deep purple, corresponding to the increasing number of disadvantaged individuals within each block group. For example, urban areas, such as Northern Virginia, Richmond, and Hampton Roads, show a mix of color intensities, indicating a varied distribution of disadvantaged populations.

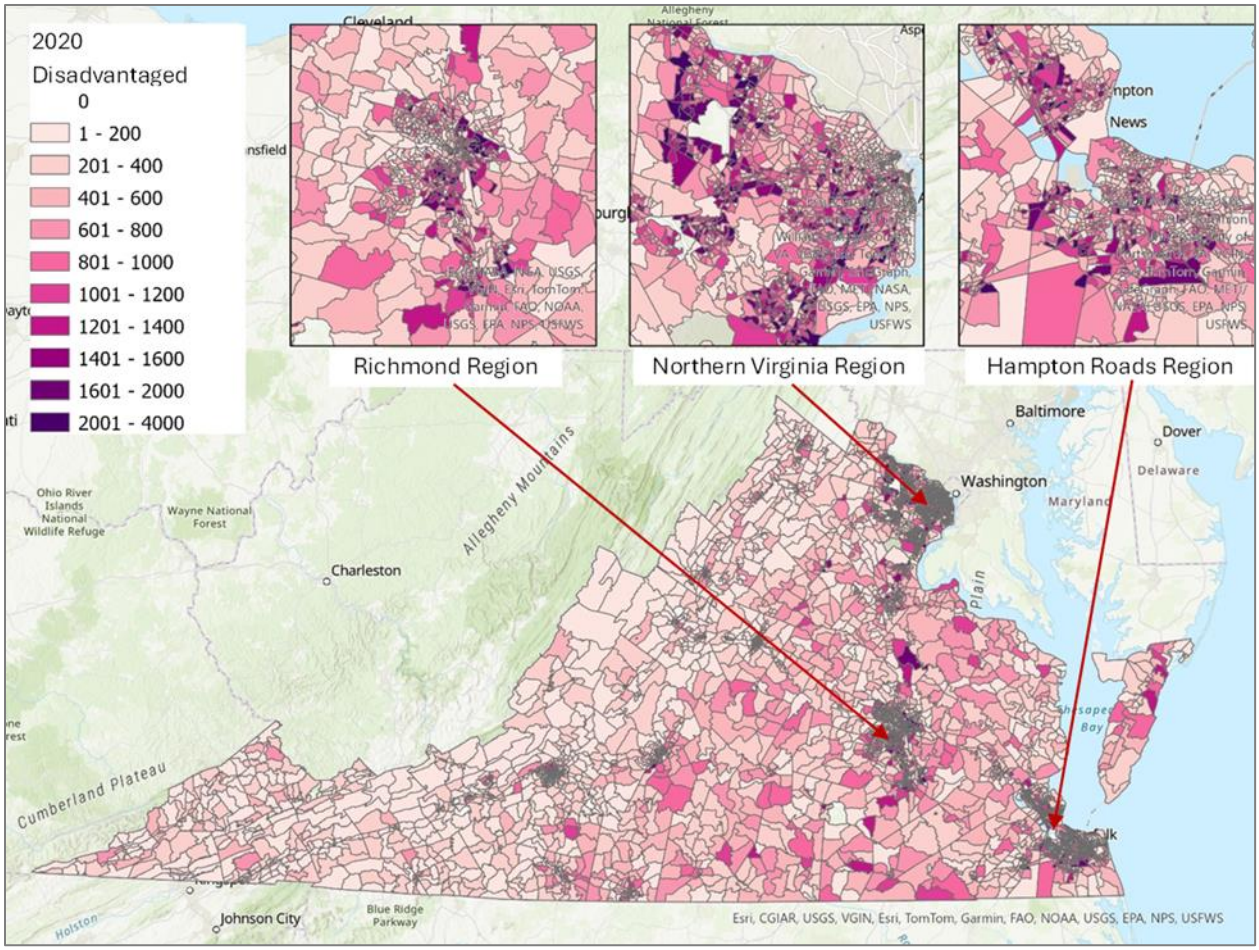


Figure 9. Disadvantaged Population for all Census Block Groups in Virginia (2020)

Figure 10 shows the 2020 disadvantaged population index at the census block group level. Because block groups are smaller, they reveal more variations within the larger census tracts. As shown in Figures 8 and 10, the Lynchburg, Richmond, Hampton Roads, Fredericksburg, and Northern Virginia districts display higher levels of the disadvantaged population index, indicating a greater concentration of disadvantaged populations. Within the Fredericksburg district, the block group map provides more detailed patterns of where disadvantaged populations are concentrated due to its higher resolution. For example, as Figure 11 shows, block group 510330301003 (outlined in yellow) has an increased disadvantaged population index of 0.62 compared with 0.32 for the census tract in which it is located. Block group 511330201003 (outlined in blue) has a decreased disadvantaged population index of 0.18 compared with 0.44 for the census tract in which it is located.

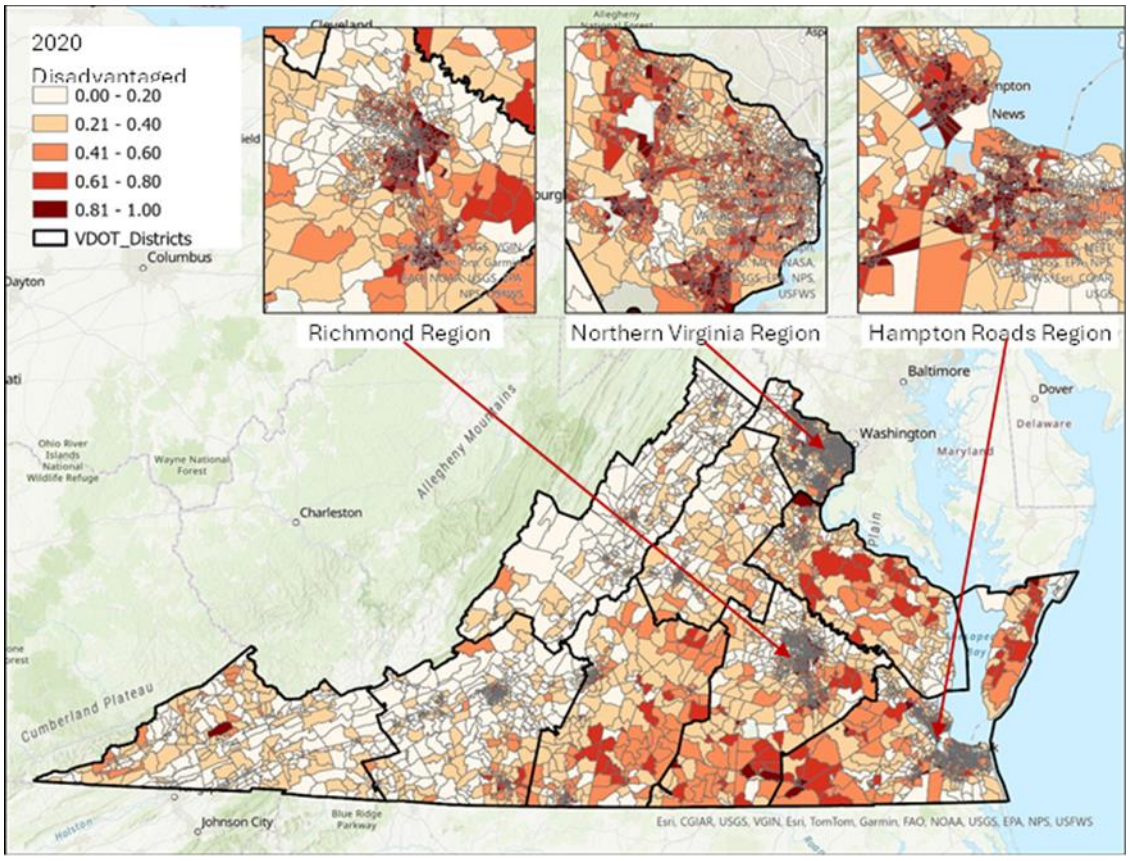


Figure 10. Disadvantaged Population Index for all Census Block Groups in Virginia (2020)

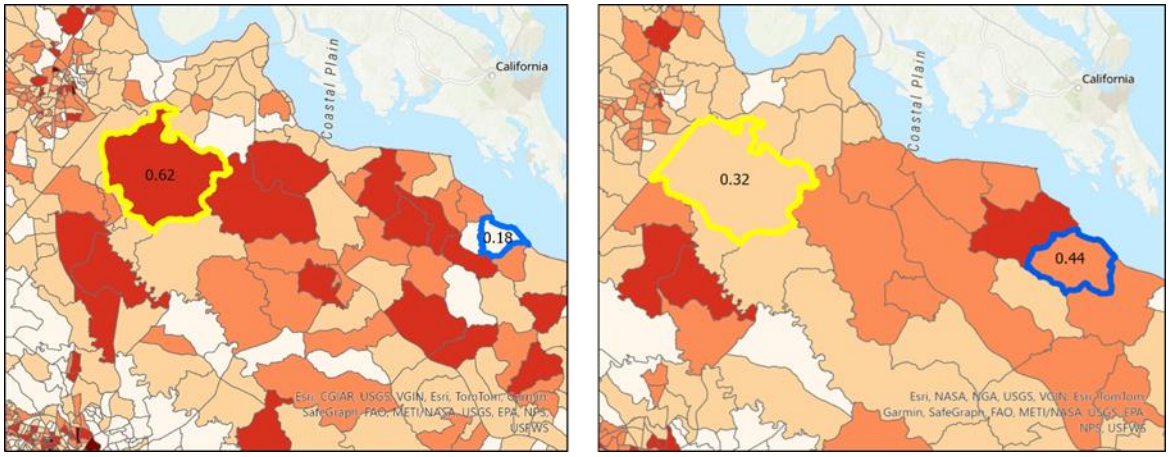


Figure 11. Comparing the Disadvantaged Population Index for Census Block Group (Left) and for Census Tract (Right) in the Fredericksburg District

Estimation Models at the Census Tract Level

Table 5 displays the three models developed (Basic, Option 1, and Option 2). Each model has a corresponding formula for calculating the disadvantaged population. The coefficient for the combined independent variable (S0601_Variable 1 + S1701_Variable 2) perfectly matches the logic of the definition of the dependent variable. For example, the coefficient of 0.999 suggests

that the total disadvantaged population includes nearly all racial minority and White populations below the poverty level, which aligns exactly with the definition of the disadvantaged population.

Table 5. Developed Models

Models	Formula
Basic	DP[a] = 19.932 + 1.105 * (S0601_Variable 1[b] + S1701_Variable 2[c])
Option 1[f]	DP =0.997 * (S0601_Variable 1 + S1701_Variable 2) + 0.745 * C16002_Variable 3[d]
Option 2[f]	DP =0.999 * (S0601_Variable 1 + S1701_Variable 2) + 0.330 * B16009_Variable 3[e]

[a] Disadvantaged Population.
[b] Total racial minority population from Census Table S0601 (U.S. Census Bureau, n.d.b.).
[c] Total White population below the poverty level from Census Table S1701 (U.S. Census Bureau, n.d.c.).
[d] Number of limited English proficiency (LEP) households from Census Table C16002 (U.S. Census Bureau, n.d.a.).
[e] Total LEP population from Census Table B16009 (U.S. Census Bureau, n.d.d.).
[f] Regression without intercept.

Generally, all three models have high adjusted R^2 values (greater than 0.95), indicating that the variables selected in the modeling explain more than 95% of the variance of the total disadvantaged population. R^2 values are not useful in this case because they cannot differentiate among the models.

Any of the three models can be used for estimating the disadvantaged population because the mean absolute error for all three is relatively close. Table 6 summarizes the performance and characteristics of each model. Taking the Basic model as an example, the mean size of the disadvantaged population was 1,534 people per tract for all the 2,188 census tracts involved in the analysis, such that the mean absolute error (169 people) represents approximately an 11% error (169/1,534). In other words, the relative benefit of using custom tabulation census tract-level data from the Census Bureau is approximately an 11% reduction in error.

Table 6. Model Performance

Models	Training Data	Testing Data		
	Standard Error of Estimate	Mean Absolute Error	Plots of Residuals	
Basic	281	169	Biased	Heteroscedastic
Option 1	226	125	Unbiased	Homoscedastic
Option 2	193	118	Unbiased	Homoscedastic

In scenarios where access to detailed LEP population data from the Census Bureau becomes limited or unavailable, the Basic model, or Option 1 model, serves as a reliable alternative for estimation purposes, maintaining a relatively high level of accuracy.

The blue polygons in Figure 12 display that, out of 2,188 census tracts in Virginia, 171 have an estimation error for the Basic model that exceeds the specified CME (e.g., the potential range of error associated with estimates derived from the census data, which is typically expressed as a plus or minus value around an estimate) for the total disadvantaged population data considered as an incorrect estimation. For the Option models, fewer census tracts exist where the estimation error surpassed this margin: 115 census tracts for the Option 1 model and 97 census tracts for the Option 2 model. In other words, more than 95% of the census tracts

(2,091 out of 2,188) will have correct estimations if Option 2 model is applied. For those 97 incorrect tracts, the average estimation error is 599 people, and the average estimation error minus CME is 258 people.

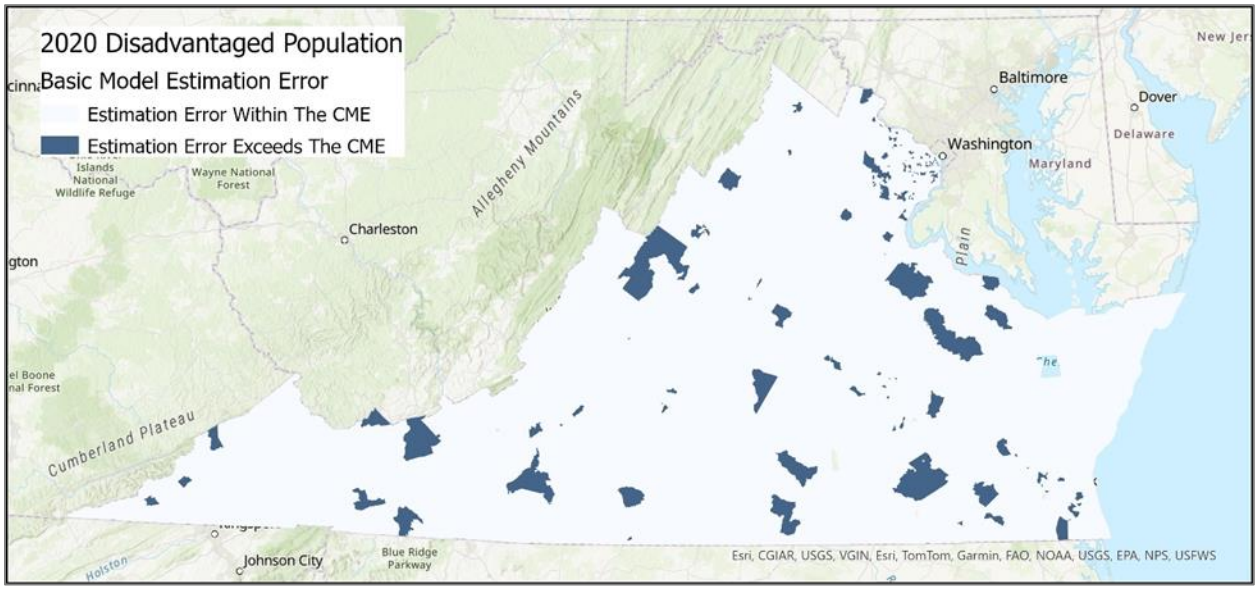
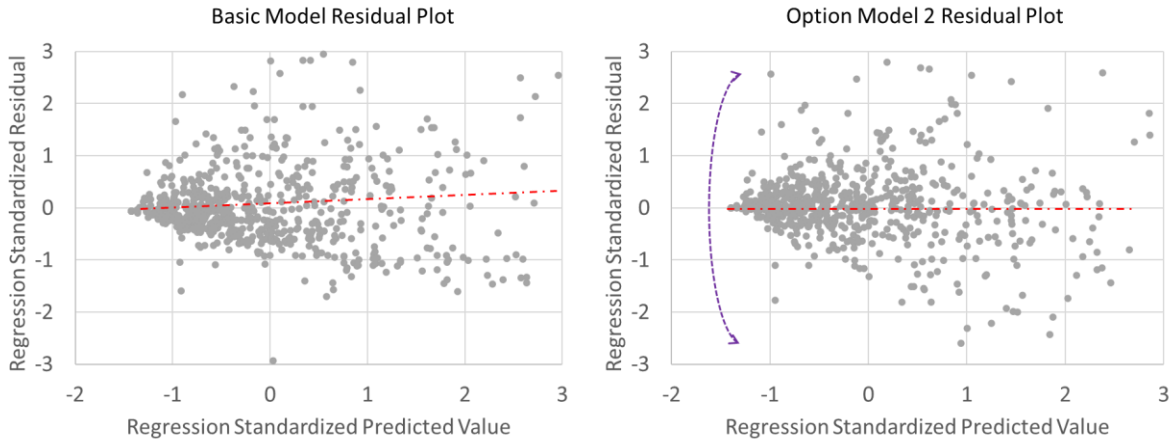


Figure 12. Areas in Virginia Where the Basic Model Estimation Error is Larger than the Census Margin of Error (CME)

When comparing the Option models with the Basic model, adding one variable to represent the LEP population improves only slightly the model's performance. For example, the difference in the mean absolute error is less than 44 persons per census tract. The reason for this outcome is that most of the disadvantaged population in the ground truth data is represented by the total population of the racial minority and the White population below the poverty level. In this case, Variable 3, either from Census Table B16009 or Table C16002, supplements the missing population of the White LEPs not below the poverty level, which slightly increases the model's performance (U.S. Census Bureau, n.d.a., n.d.d.). Based on the 0.330 coefficient of Variable 3 from Census Table B16009, a rough inference suggests that approximately 70% of the LEP population overlaps with those experiencing poverty and belonging to racial minority groups.

Among all three models, Option 2 (with the variable of LEP population from Census Table B16009) demonstrates the best performance, and the researchers recommended this option for use as it has the lowest SEE (193 persons) and mean absolute error (118 persons per census tract) (U.S. Census Bureau, n.d.d.). Most importantly, when comparing Option 2 model with the Basic model, the introduction of Variable 3 from Census Table B16009 corrected the residual plot from biased and heteroscedastic to unbiased and homoscedastic. As shown in Figure 13, the variance of the residuals became more evenly scattered and constant across different values of the independent variables, thus enhancing the model's reliability and predictability for different census tracts.



(a) **(b)**

Figure 13. (a) Residual Plots of the Basic Model and (b) Option Model 2. The red line indicates the average of the residuals. The residuals of the Option 2 model have an average of zero, indicating the model is unbiased. The purple arrows indicate that those residuals are more evenly scattered, and the model is homoscedastic.

Estimation Model at the Census Block Group Level

The researchers developed the following model for estimating the disadvantaged population at the census block group level:

$$DP = 0.886 * \text{Variable 4} + 0.422 * \text{Variable 5} + 0.125 * \text{Variable 6} \quad \text{(Eq. 5)}$$

Where:
DP = disadvantaged population.
Variable 4 = total racial minority population from Census Table B02001 (U.S. Census Bureau n.d.e.).
Variable 5 = total poverty population from Census Table B17021 (U.S. Census Bureau n.d.f.).
Variable 6 = total LEP population from Census Table B16004 (U.S. Census Bureau n.d.g.).

Table 7 presents the model performance according to Equation 5. The adjusted R^2 is 0.97, indicating that 97% of the variance in the dependent variable can be explained by the independent variables. This analysis makes sense because the total disadvantaged population consists of all three disadvantaged populations as the independent variables. Of note, because all three populations are used at the block group level, the researchers developed only one model.

Table 7. Performance of the Census Block Group Model

Training Data		Testing Data		
Adjusted R^2	**Standard Error of Estimate**	**Mean Absolute Error**	**Plots of Residuals**	
0.97*	138	88	Unbiased	Heteroscedastic

*Regression without intercept and cannot be compared with regression having intercept.

The SEE and mean absolute error are 138 persons and 88 persons, respectively, corresponding to 24.5% and 15.6% of the mean value (564 persons) of the disadvantaged population for all 5,943 census block groups involved in the analysis. The relative benefit of custom tabulation census block level data from the Census Bureau is an approximately 15.6% (88/564) reduction in error.

Out of the 5,943 census block groups, 593 have an estimation error that exceeds the specified CME, considered as an incorrect estimation, shown as blue polygons in Figure 14. This finding indicates that more than 90% of the census block groups will have correct estimations if the block group model is applied. For those 593 block groups with incorrect estimation, the average estimation error is 227 people, and the average value of the estimation error minus CME (i.e., the exceeded portion of the estimation error) is 107 people.

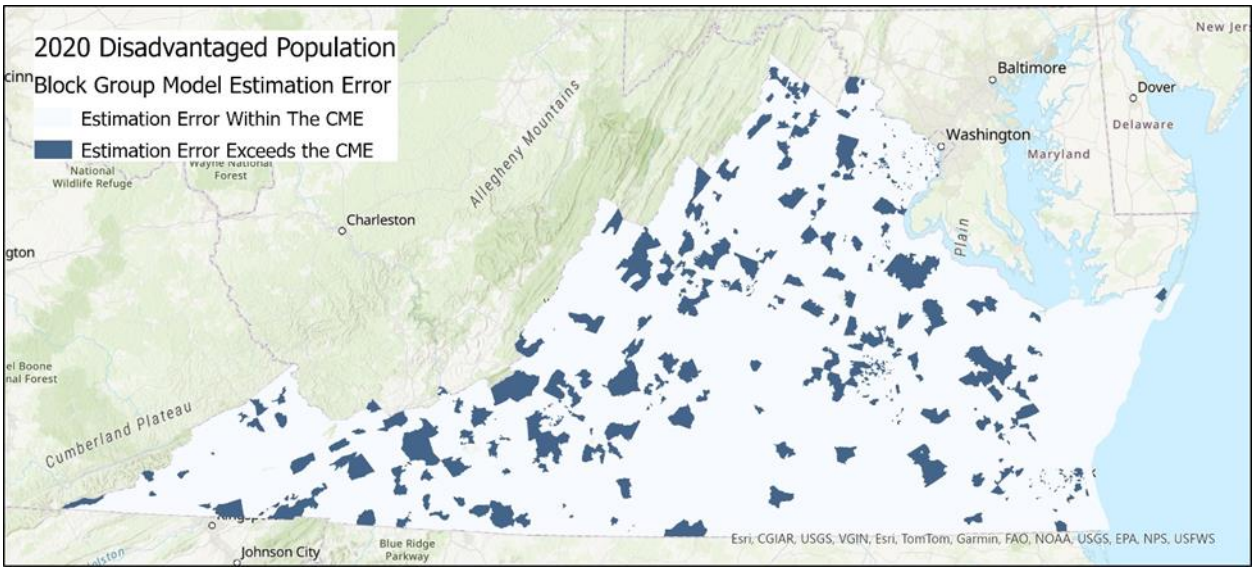


Figure 14. Areas in Virginia Where the Block Group Model Estimation Error is Larger than the Census Margin of Error (CME)

Figure 15 depicts the residual plot of the estimation model, which is unbiased and slightly heteroscedastic for the portion between −1.5 and −0.5 standardized predicted values. Typically, a possible solution for correcting heteroscedasticity is to redefine the variables, such as changing the dependent variable from the actual disadvantaged population to the disadvantaged population index (e.g., total disadvantaged population/total population). However, this method is not applicable to this specific analysis for the following three reasons:

1. Using the disadvantaged population index as the dependent variable will decrease the adjusted R^2 from 0.97 to 0.82.
2. The definition of a disadvantaged population predetermines the model's form, and the unit of the dependent variable needs to match the unit of the independent variable because the model is estimating the population (e.g., a unit of persons).
3. The disadvantaged population index is randomly distributed across urban areas (e.g., regardless of the total population size, some census block groups have 100% of their population categorized as disadvantaged, whereas others have only a small portion), and this distribution does not clearly indicate a positive correlation between the

disadvantaged population index and the actual number of disadvantaged individuals in those areas.

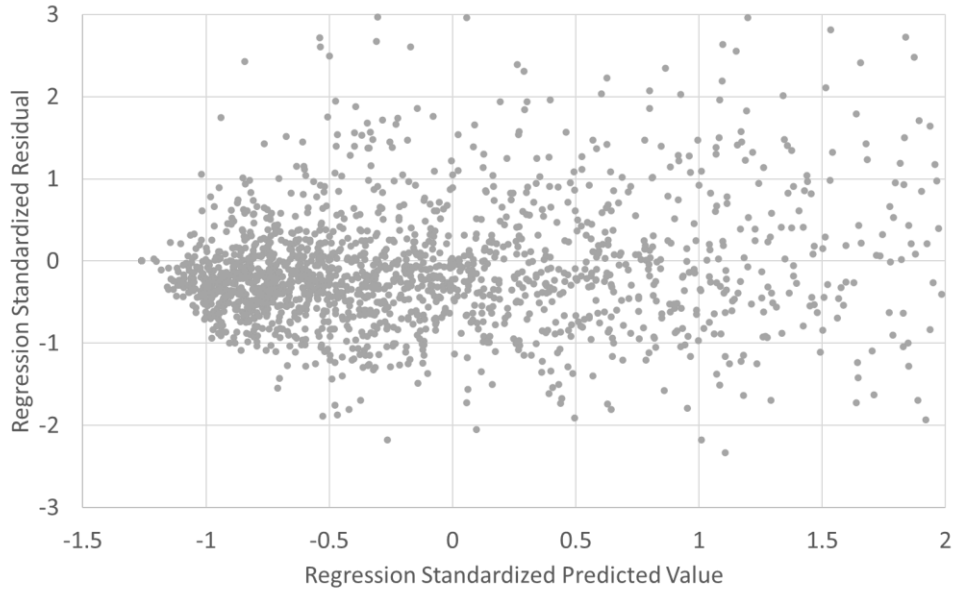


Figure 15. Residual Plot of the Estimation Model at the Census Block Group Level

DISCUSSION

Cross-Verification of Estimation Results

The estimation results from the model at the census tract level are statistically valid (e.g., the model has 0.99 adjusted R^2 with an unbiased and homoscedastic residual plot) and can serve as a reference point for validating the estimation performance of the census block group model. Because direct data from custom tabulations might not be available in the future, using reliable estimates at the tract level to verify the estimates for block groups provides an indirect but robust method of cross-verification. If the estimates at the block group level align well with the estimates at the tract level, then confidence in the accuracy and reliability of the block group model's outputs is enhanced, even in the absence of direct tabulated data for verification.

Census block groups are smaller geographic units compared with census tracts. By summing up the estimation errors from block groups to tracts, the errors are aggregated to a higher level of geography. This finding can help in understanding the overall accuracy at the tract level. Comparing the aggregated error with CME of the census tract data assesses the reliability of the estimates at the tract level.

Figure 16 shows the cross-verification process for a census tract. The research team applied the census block group model (Equation 5) to all the census block groups and summed the estimated disadvantaged populations from block groups to tracts. When comparing the total disadvantaged population from the model with that directly from the custom tabulation, the estimation is considered accurate if the difference between these two populations does not

exceed CME from the custom tabulation.

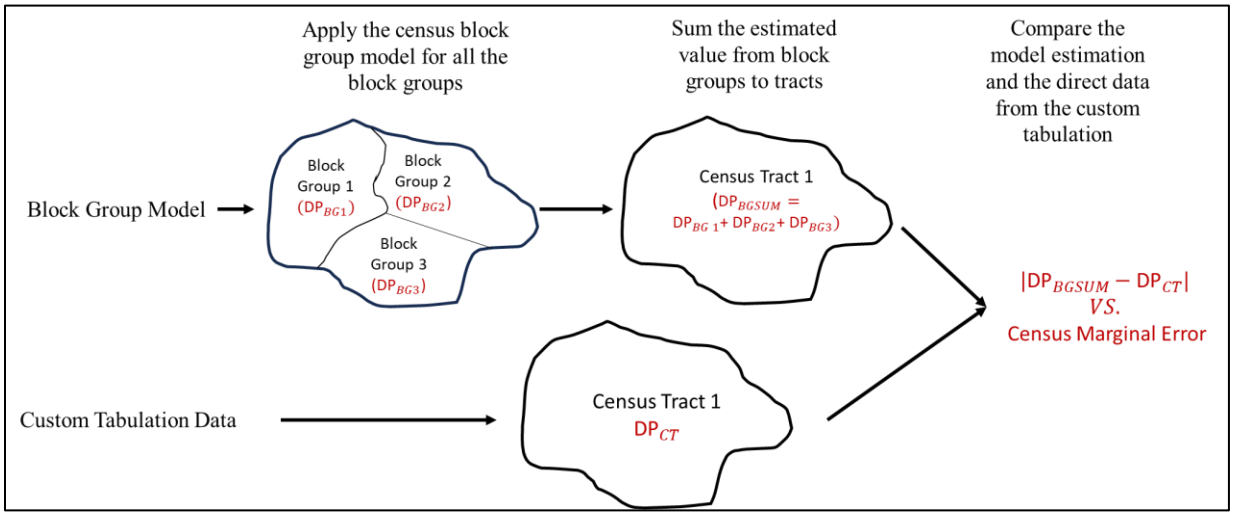


Figure 16. The Steps of Cross-Verification. BG = block group; CT = census tract; DP = disadvantaged population.

The accuracy rate for the census block group model is 84.6% based on the custom tabulation. That is, 15.4% (336 out of 2,188) of the census tracts have an estimation error summation exceeding CME, with an average estimation error of 474 people. The average value for the portion of the estimation error exceeding CME is 181 people, corresponding to 11.8% of the average disadvantaged population (1,534 people) at the census tract level.

Given that there are 2,198 census tracts, the percentage of tracts whose error exceeds the marginal error shrinks from 15% (i.e., 336/2,198) to 4% (i.e., 97/2,198) if one replaces the cross-verification for the census block group model with the Option 2 model (shown in Table 8). However, the average estimation errors for those census tracts with incorrect estimations based on the cross-verification of the block group model are smaller than when the Option 2 model is directly applied at the census tract level. That is, the average estimation error is 474 versus 599 people, and the average estimation error exceeding CME is 181 versus 258 people. An *F*-test was conducted to compare whether the errors for the 336 census tracts are statistically different from the errors for the 97 census tracts based on the Option 2 model. The result yielded a *p*-value of 0.75 from the *F*-test, which fails to reject the null hypothesis that the two groups have the same variance. This result suggests that the variances between the two groups of errors are not significantly different from each other. Overall, this analysis confirms that the estimates from the block group model at the tract level are reliable and statistically acceptable.

Table 8. Comparing the Block Group Model and Option 2 Model in Cross-Verification

Models	Census Tracts Having Estimation Error Exceeding the Marginal Error		
	Total Number	Average Error	Average Value for (Estimation Error−CME)
Block Group Model Cross-Verification	336	474	181
Option 2 Model	97	599	258

CME = census margin of error.

The Necessity of Having Detailed Data

Improving modeling precision requires selecting data that balance accessibility and estimation accuracy. The previous section (The Estimation Model at the Census Block Group Level) discussed complications caused by the overlapping populations among the variables, specifically heteroscedasticity in the block group model. This complication leads to a critical question: Is investing more effort in obtaining highly detailed yet less accessible data worthwhile, or can users rely on the accuracy of data that are more easily obtained but perhaps lack fine detail?

To address this question, the research team compared two models developed from two distinct datasets, with a particular emphasis on the census tract level, where both sets of data are applicable: (1) a dataset that minimizes overlapping populations but is difficult to find at a finer geographical resolution and (2) a dataset that is easily accessible at various levels, such as PUMA, census tracts, or census block groups, but includes overlapping populations. The goal was to determine whether the choice of dataset significantly affects estimation error.

The first model is the Option 2 model (Equation 6), which is based on three variables. Variable 1 is the total racial minority population from the table. Variable 2 is the total White population below the poverty level, and variable 3 is the total population above the poverty level who do not exclusively speak English. Although some populations in Variable 3 overlap with those in Variable 1, the portion of the overlapping population in Variable 3 can be determined by the coefficient of Variable 3. For example, the coefficient of 0.323 means approximately 70% of the population in Variable 3 is overlapped with the combined Variables 1 and 2.

$$\text{Disadvantaged population} = 0.999 \times (\text{Variable 1} + \text{Variable 2}) + 0.323 \times \text{Variable 3} \quad \text{(Eq. 6)}$$

Where:
Variable 1 = total racial minority population from Census Table B0601.
Variable 2 = total White population below the poverty level from Census Table S1701 (U.S. Census Bureau, n.d.c.).
Variable 3 = Total population above the poverty level who do not exclusively speak English from Census Table B16009 (U.S. Census Bureau, n.d.d.).

The researchers developed the second model (Equation 7) based on the same census tracts in the training data as the Option 2 model but with different variables: Variable 7 is the total racial minority population, Variable 8 is the total population below the poverty level, and Variable 9 is the total LEP population. These three variables have overlapping populations, and the portion of the overlapping population cannot be detected from the coefficient.

$$\text{Disadvantaged population} = 0.924 \times \text{Variable 7} + 0.385 \times \text{Variable 8} + 0.114 \times \text{Variable 9} \quad \text{(Eq. 7)}$$

Where:
Variable 7 = total racial minority population from Census Table B02001 (U.S. Census Bureau n.d.e.).

Variable 8 = total population below the poverty level from Census Table B17021 (U.S. Census Bureau n.d.f.).
Variable 9 = total LEP population from Census Table B16004 (U.S. Census Bureau n.d.g.).

Table 9 presents the performance of these two models. Although the adjusted R^2 values for both are close (i.e., the difference is less than 0.01), Equation 7 exhibits a substantially higher (i.e., more than 1.5 times) SEE and mean absolute error compared with Equation 6. Similarly, as observed in the census block group model, Variables 7–9 result in a heteroscedastic residual plot for Equation 7 due to their overlapping populations.

Table 9. Performance of the Two Census Tract Models

Models	Training Data		Testing Data		
	Adjusted R^2	Standard Error of Estimate	Mean Absolute Error	Plots of Residuals	
Equation 6	0.99*	193	118	Unbiased	Homoscedastic
Equation 7	0.98*	284	187	Unbiased	Heteroscedastic

*Regression without intercept and cannot be compared with regression having intercept.

Because the absolute error for all the testing data exhibits a right-tailed distribution and the variances of these errors differ between the two models, a one-tailed *t*-test with unequal variance was employed to examine whether a significant difference exists between the mean values of the testing errors (mean absolute error) for these models. The resulting *p*-value for the *t*-test is <0.05, indicating that the mean absolute error for Equation 6 is statistically significantly smaller than that for Equation 7. This significant difference in the estimation errors between the two models validates the hypothesis that more detailed data lead to more reliable and preferable outcomes in estimating disadvantaged populations.

Testing the Reliability of the Census Block Group Model

The estimation models have been developed based solely on 2020 block group data, which is publicly available on the Census Bureau website (U.S. Census Bureau, n.d.e., n.d.f., n.d.g.). If the variables used in those developed models become unavailable in the future, the researchers are investigating alternative methods to ensure continuity and effectiveness. In this effort, the researchers requested the custom tabulation for the 2021 and 2022 block group data from the Census Bureau. By incorporating the 2020, 2021, and 2022 block group data, the goal is to verify if the developed model is comprehensive and reliable, better capturing the nuances and changes over time in disadvantaged populations.

In reviewing the 2020–2022 custom tabulation data provided by the Census Bureau, the researchers found that the number of census block groups containing the disadvantaged population varied yearly. For example, as Table 10 shows, the 2020 data included 5,963 block groups with total population data and an identical number of block groups with disadvantaged population data (e.g., LEP, minority, or below poverty level population data). However, discrepancies were noted in the 2021 and 2022 data. Specifically, the 2021 custom tabulation showed 5,892 block groups with total population data but only 5,867 block groups with

disadvantaged population data. Similarly, the 2022 custom tabulation had 5,887 block groups with total population data and 5,862 block groups with disadvantaged population data.

Table 10. Yearly Data Discrepancies of the Custom Tabulation at the Block Group Level

Year	Number of Census Block Groups Data Received from U.S. Census Bureau	
	Having Total Population Data	Having Limited English Proficiency, Minority, or Below Poverty Level Population Data
2020	5,963	5,963
2021	5,892	5,867
2022	5,887	5,862

These discrepancies can be attributed to the ACS sampling methodology, where smaller block groups may lack data, and the necessity to suppress some estimates for disclosure purposes (Spanos, 2024c). Notably, stricter disclosure rules were applied to the 2021 and 2022 tables, and the Census Bureau has indicated that these rules will become increasingly stringent in the future. This trend will make obtaining detailed block group data increasingly difficult, thus emphasizing the importance of this research. The model's ability to provide accurate estimations without relying on custom table requests from the Census Bureau is crucial for future analyses. Considering the potential influence of the pandemic on the disadvantaged population in 2020, the researchers requested a custom tabulation for the 2021 and 2022 block group data from the Census Bureau. The researchers did not request census tract data because block group data can be transferred to the census tract level, and block group data provide a finer level of detail.

Tukey's multiple comparisons of means were applied to compare the differences among the 2020, 2021, and 2022 disadvantaged population data. A 95% familywise confidence level was used in this comparison, whereby the confidence intervals for the comparisons are adjusted to maintain a 95% simultaneous confidence level. The results show that the mean value in 2021 is 13.45 persons higher than in 2020 with an adjusted *p*-value of 0.28, which is not statistically significant, and the mean value in 2022 is 12.81 persons, which is significantly higher than in 2021 with an adjusted *p*-value of 0.01. This result suggests that a significant increase occurred in the disadvantaged population from 2020 to 2022, but the change from 2021 to 2022 was not large enough to be statistically significant.

The block group model was applied to the 2021 and 2022 data to evaluate its reliability. After removing census block groups without disadvantaged population data from the custom tabulation, the researchers found that 553 out of 5,865 block groups in 2021 (9.4% inaccurate estimation rate) and 419 out of 5,848 block groups in 2022 (7.2% inaccurate estimation rate) had estimation errors that exceeded CME threshold when applying the block group model developed based on 2020 data. These rates are less than the 10% inaccurate estimation rate observed in 2020, indicating consistent estimation across different years and suggesting that the model may not be influenced by special events in those specific years.

Yearly Change of Disadvantaged Population

The researchers applied the census block group model for 2013 to 2022 data. Table 11 presents the key statistics for the disadvantaged population from 2013 to 2022, including the average value of the disadvantaged population in each year for all census block groups with population data and their 95% confidence interval limits, representing the range where the true mean likely falls with 95% certainty.

Table 11. Yearly Change of the Disadvantaged Population From 2013 to 2022

Year	Mean	95% Confidence Interval	
		Lower Bound	Upper Bound
2013	482	469	496
2014	491	478	505
2015	497	483	511
2016	503	489	517
2017	508	494	523
2018	513	499	528
2019	518	503	532
2020	494	480	508
2021	515	501	529
2022	534	520	549

The researchers obtained a *p*-value of 4.23×10^{-6} by performing a one-way analysis of variance (ANOVA) test on the means of the disadvantaged population for each year from 2013 to 2022. The ANOVA test compares the means of multiple groups to determine if any statistically significant differences between the group means are present. For example, ANOVA analyzes the within-group variance (i.e., the variability of the data points within each group) and the between-group variance (i.e., the variability of the group means relative to the overall mean), then calculates the ratio of between-group variance to within-group variance. In this case, a large ratio was found, suggesting that the variation between groups is significantly greater than within groups. In other words, the resulting *p*-value of <0.05 suggests that a statistically significant difference is present in the means of the disadvantaged population across the years. This result implies that the population trends observed from 2013 to 2022 are not random but show meaningful shifts during this period (i.e., a continuously increasing trend for the disadvantaged population from 2013 to 2022).

Noting that COVID-19 may have influenced the population data in 2020 and 2021, researchers generated the trend line displayed in Figure 17 without data from these 2 years. This trend line can be used as the yearly estimate starting in 2013 based on Equation 8. For example, the average disadvantaged population in 2022 equals 484.79 + 5.5882 × (2022 − 2013) = 535 persons.

$$\text{Disadvantaged population in year } X = 484.79 + 5.5882 \times (\text{year } X - 2013) \qquad \text{(Eq. 8)}$$

Because the trend line appears linear, the technical review panel asked if the data had been calculated using a regression-based approach derived from the total population rather than being collected through localized surveys. Examination of the average disadvantaged population

for each block group showed that a regression-based approach was not used. Although some block groups saw their populations grow, others saw their population drop (Table 12).

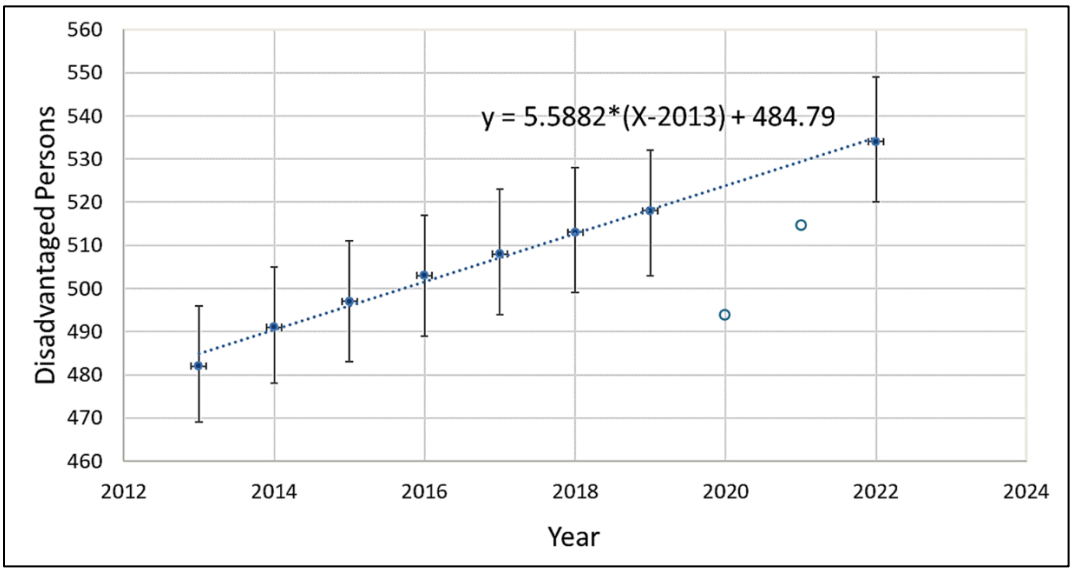


Figure 17. Yearly Change of the Average Disadvantaged Population from 2013 to 2022

Table 12. Block Groups with Decreased Disadvantaged Population from 2013 to 2019

GEO ID	Disadvantaged Population in Each Year									
	2013	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**
510872003052	443	383	306	301	209	189	184	258	279	502
510230402002	141	128	81	51	26	25	24	37	17	50
516500101034	484	469	450	439	357	277	264	185	199	139

GEO ID = geographic identifier.

This dynamic behavior—where populations in certain areas decline rather than increase uniformly—strongly indicates that the disadvantaged population figures were not derived from the Census Bureau simply applying a regression model to the total population data. If such a regression-based approach had been used, the researchers would expect more uniform trends across all block groups, primarily increasing or remaining constant over time. The observed fluctuations and decreases in certain block groups suggest that these figures were calculated using actual, localized data rather than a generalized regression on overall population numbers. For each block group, a rough estimation of the disadvantaged population for a future year can be calculated as follows:

$$\frac{Disadvantaged\ population\ in\ year\ X}{Disadvantaged\ population\ in\ 2013} = \frac{average\ disadvantaged\ population\ in\ year\ X}{average\ disadvantaged\ population\ in\ 2013} = \frac{484.79 + 5.5882 \times (\text{year X} - 2013)}{482} \quad \text{(Eq. 9)}$$

For example, assuming the disadvantaged population for block group A in 2013 was 1,050 persons, the predicted disadvantaged population for block group A in 2025 is:

$$1{,}050 \times \frac{484.79 + 5.5882 \times (2025 - 2013)}{482} = 1{,}202 \text{ persons}$$

Equation 9 can be used as a rough estimation tool for predicting future disadvantaged population trends during 5-, 10-, 20-year, or even longer periods. However, the equation does not provide the level of accuracy offered by more detailed models, such as Equation 5, which accounts for variations across specific block groups. The disadvantaged population is dynamic, often shifting from one location to another, with some block groups experiencing a decline (e.g., block group 511455003001 saw a decrease in the disadvantaged population from 1,940.20 to 1,033.79), whereas others experience growth. As a result, Equation 9 captures the general trend but may not reflect local population changes in individual block groups.

Post-Model Adjustment

The proposed methodology uses least squares regression without an intercept in some models because the dependent variable (disadvantaged population) theoretically should equal the sum of the individual categories that define a disadvantaged population (e.g., racial minorities, low-income, and LEP individuals). In addition, the total population for each census tract or census block group is not explicitly included as an independent variable in the model development. The models are based on specific variables, such as racial minority population, poverty levels, and LEP, rather than the total population. Therefore, adjusting the coefficients based on the total population would be inappropriate because it would introduce a factor that is not part of the model's original structure and might distort the statistical significance of the relationships between the variables, which could introduce new inaccuracies.

One possible method for controlling the disadvantaged population so that it does not exceed the total population could be using postmodel adjustment, which is a practical solution to ensure logical consistency in the results. For example, the researchers recommend using the Option 2 model for estimating the disadvantaged population at the census tract. The post-model adjustment for Option 2 would include a constraint whereby the disadvantaged population is less than or equal to the total population:

Disadvantaged population = minimum {Total population, Option 2} (Eq. 10)

Similarly, for the census block group model, the post-model adjustment is:

Disadvantaged population = minimum {Total population, census block group model} (Eq. 11)

The adjustment of using the minimum function to constrain the disadvantaged population estimates ensures mathematical integrity by preventing the calculated disadvantaged population from exceeding the total population when the model overestimates. This approach maintains statistical validity by allowing the original regression coefficients to remain unchanged, preserving the relationships between the variables without distorting the model's structure. In addition, the method is flexible, as it can be applied to both census tract and block group models, providing a consistent solution to address any overestimation.

Given that the population data from the Census Bureau contain large margins of error, adjusting the disadvantaged population estimates based on potentially inaccurate total population figures is only needed for a small number of block groups. For example, after accounting for marginal error, none of the census block groups in 2020, 2021, or 2022 had an estimated disadvantaged population larger than the total population. Even without considering the marginal error, only a small percentage (0.5%) of census block groups showed such discrepancies: 30 out of 5,963 in 2020, 32 out of 5,867 in 2021, and 28 out of 5,862 in 2022.

If the total population exceeds the estimated population from either the Option 2 model or the census block group model, an additional post-model adjustment is suggested so that the estimated disadvantaged population is not less than the total population in any individual category:

> Disadvantaged Population with Additional Adjustment = maximum {Option 2, Total racial minority population, Total low-income population, Total LEP population} (Eq. 12)

Similarly, for the census block group model, the additional post-model adjustment is:

> Disadvantaged Population with Additional Adjustment = maximum {Census block group model, Total racial minority population, Total low-income population, Total LEP population} (Eq. 13)

Future Research Needs

Currently, only the three population datasets used in the modeling are available at the census block group level. This gap points to the need for future research to explore alternative population datasets until new data become available.

CONCLUSIONS

- *The preferred estimation models demonstrate high reliability*. At the census tract and block group levels, the best models explained 99% and 97%, respectively, of the variance in the dataset. Furthermore, when these models were tested on data not used to build the model, the models showed mean absolute error rates of 8% at the tract level and 16% at the block group level. For instance, the best tract level model had an error of 118 disadvantaged persons per tract, which is 8% of the mean value of 1,534 disadvantaged individuals per tract.

- *Detailed, publicly available data are necessary to accurately estimate disadvantaged populations at a statistically significant level ($p < 0.05$).* At the census tract level, a model that did not account for the missing LEP population, described previously as the Basic model, would have yielded a mean absolute error of 11%. In practical terms, accounting for overlapping variables is more important than having the results at a finer level of geographic detail.

- *More than 95% of the census tracts and more than 90% of the census block groups will have accurate estimations, considering the marginal error of the census data, if the best model is applied*. That is, the testing error resulting from these models is less than the specified CME for nearly all tracts and block groups.

- *Cross-verification demonstrates consistent estimates across time and geographic areas, indicating that the block group model's reliability is not significantly affected by special events (e.g., the COVID-19 pandemic) and that it produces reliable estimates when aggregating the error at the tract level.* Applying the block group model (developed from 2020 data) to 2021 and 2022 data showed that the testing error exceeded the specified CME for 9.4 and 7.2% of block groups, respectively, which is an improvement over the 10% figure for 2020. By applying the block group model and aggregating the estimation error at the tract level, the block group model has an accuracy rate of 84.6%, which is approximately 10% less than the direct application of the Option 2 model at the tract level.

- *The average disadvantaged population shows a statistically significant upward trend, as confirmed by the p-value less than 0.05*. These observed changes are meaningful and not random, as captured by Equation 7, which offers a useful tool for rough estimations of future population, although the equation lacks the precision needed to account for local variations in specific block groups. Although the overall trend points to growth, the disadvantaged population is dynamic, with some areas experiencing declines and others seeing increases.

RECOMMENDATIONS

1. *VDOT's TMPD should estimate the statewide disadvantaged populations using publicly available data at the block group level to support SMART SCALE project prioritization.* The block group model has been validated for multiyear use with ACS data from 2013 to 2022, demonstrating more than 90% accuracy and reliability for long-term applications. VDOT's TMPD should apply the block group model to the most recent 5-year ACS data, including racial minority populations, poverty levels, and individuals with LEP, to estimate disadvantaged populations at the block group level for the entire state of Virginia. This estimated disadvantaged population should then be used in VDOT's SMART SCALE project prioritization.

2. *VDOT's TMPD should continue using the block group model for future years while monitoring population changes over time.* Equations 8 and 9 are used for long-term forecasting of the disadvantaged population in a specific block group. To ensure their accuracy and relevance, TMPD should compare yearly changes in disadvantaged population estimates by applying these equations to the trends derived from yearly changes using the block group model. Based on these observations, appropriate adjustments to the estimation formulas in Equations 8 and 9 should be made to enhance future forecasting.

3. *VDOT's TMPD should consider using the proposed disadvantaged population model to provide more accurate data in the Pathways for Planning (P4P).* The proposed block group model provides more accurate estimates of disadvantaged populations while reducing

double-counting compared with the existing data in P4P. VDOT's TMPD should apply the block group model to the most recent ACS 5-year data to provide users with the most up-to-date disadvantaged population information in P4P.

IMPLEMENTATION AND BENEFITS

Researchers and the technical review panel (listed in the Acknowledgments) for the project collaborate to craft a plan to implement the study recommendations and to determine the benefits of doing so. This process is to ensure that the implementation plan is developed and approved with the participation and support of those involved with VDOT operations. The implementation plan and the accompanying benefits are provided here.

Implementation

To implement Recommendations 1, 2, and 3, TMPD will use the step-by-step process given in this subsection for estimating and updating disadvantaged population data statewide. These recommendations will be implemented by December 31, 2027. To implement Recommendations 1 and 3, TMPD should apply the block group model (Equation 5) and estimate disadvantaged populations using the most recent 5-year ACS data. The most recent ACS data are for years 2019–2023 and can be obtained from U.S. Census Bureau (n.d.e., n.d.f., n.d.g.). To implement Recommendation 2, this process will be piloted during at least a 3-year period because the 2023–2025 data are needed, and these data are expected to be released in Spring 2027.

Implementation for Recommendation 1

Recommendation 1 involves TMPD's Modeling and Accessibility Program applying a block group-level model on the most recent 5-year ACS data to estimate disadvantaged populations across Virginia, with a focus on three publicly available census tables: Census Table B02001 for racial minority populations, Census Table B17021 for poverty levels, and Census Table B16004 for individuals with LEP (U.S. Census Bureau n.d.e., n.d.f., n.d.g.). VDOT's TMPD will establish a workflow to update disadvantaged population data for each block group in Virginia by applying the proposed block group model (Equation 5) and post-model adjustment (Equations 11 and 13) every year and then use that estimated disadvantaged population data in VDOT's SMART SCALE project prioritization process.

Implementation for Recommendation 2

Recommendation 2 focuses on refining the forecasting models through annual reviews. This process will be piloted during at least 3 years of data (2023–2025), encompassing data sourcing and model validation. TMPD's Modeling and Accessibility Program will analyze yearly changes in disadvantaged population data for 3 more years of data (2023–2025) by applying the block group model and verifying the results using Equations 8 and 9 to calibrate these equations and identify significant trends. The 2025 data are the 5-year (2020–2025) ACS data, which are

expected to be released by the Census Bureau at the end of 2026.

Following is a step-by-step process for refining Equation 9 based on the updated coefficients derived from the refined linear model (Equation 8):

1. Collect 13 years of publicly available census tables (2013–2025): 5-year Census Table B02001 for racial minority populations, 5-year Census Table B17021 for poverty levels, and 5-year Census Table B16004 for individuals with LEP (U.S. Census Bureau n.d.e., n.d.f., n.d.g.).
2. Apply the block group model to the data from the collected census table and generate the disadvantaged population data for each block group for each year (2013–2025).
3. Filter block groups with constant area and continuous data (2013–2025): Identify block groups with continuous disadvantaged population data across all years from 2013 to 2025, exclude block groups with missing or inconsistent data, and ensure the spatial area of block groups remains constant throughout the period (i.e., no boundary changes).
4. Calculate the mean disadvantaged population for the filtered block groups for each year from 2013 to 2025 and then use the yearly averages to analyze the trend.
5. Fit a simple linear regression model using the yearly average disadvantaged population data in which the dependent variable is the yearly average disadvantaged population and the independent variable is the year (2013–2025).
6. Derive the equation of the form like Equation 8:
   Averaged disadvantaged population in year $X = a + b *$ (year X−2013)
   Where:
   a = the intercept (base population in 2013).
   b = the slope (average annual change in disadvantaged population).
7. Extend the linear model for individual block groups like Equation 9. Generate future estimates for disadvantaged populations at the block group level using the following relationship for each block group:

$$\frac{\text{Disadvantaged population in year } X}{\text{Disadvantaged population in 2013}} = \frac{\text{Average disadvantaged population in year } X}{\text{Average disadvantaged population in 2013}}$$

The new coefficients improve accuracy when forecasting disadvantaged populations at the block group level for future years if needed.

Implementation for Recommendation 3

Recommendation 3 involves updating the disadvantaged population data used for P4P annually. By applying the block group model to the most recent ACS 5-year data, TMPD's Planning Data Solutions Team will enhance the existing demographic map with more accurate disadvantaged population information. This update will ensure that P4P users can access the latest data for their planning efforts.

Benefits

This section addresses two benefits of using the developed models to estimate disadvantaged populations at the block group level for project prioritization and funding allocation under the SMART SCALE system: (1) accelerating project prioritization, thereby reducing cost escalation, and (2) maintaining the integrity of the project selection process.

Accelerating Project Prioritization

Obtaining more accurate and granular block group data regarding disadvantaged populations from Census Bureau custom tabulations can take up to 18 months from the time of the request.

If VDOT wanted these more accurate data, then the primary benefit of using these models would be the ability to prioritize and complete high-impact projects faster than this 18-month period. The top 30 SMART SCALE projects in 2024 had a total cost of approximately $106 million (CTB, 2024b). By accelerating project prioritization and implementation by 18 months, VDOT can reduce the opportunity cost associated with delaying spending for 18 months. The benefit of faster prioritization can be captured by the difference between the rate of return on idle funds waiting to be spent and the rate of return on investments made in SMART SCALE projects, as calculated in Equation 14:

> Benefit of faster prioritization = (rate of return on SMART SCALE investments – rate of return on idle funds) × time saved (1.5 years) × total project spending ($106 million) (Eq. 14)

Regarding the rate of return on idle funds, SMART SCALE projects are required to remain fully funded, even in the event of delays, with potential adjustments to the timing of funding between fiscal years. According to the *Code of Virginia* §33.2-214 E, funding for projects must be sufficient to complete them within a 6-year horizon unless certain exceptions apply. Importantly, any delay, even up to 18 months, will not generate interest benefiting the project, meaning the rate of return on idle funds is effectively 0% (Farmer, 2024).

The rate of return on SMART SCALE investments is based on a benefit-cost analysis of three separate El Paso transportation projects as part of a grant program application at the U.S. Department of Transportation. The internal rate of return for long-term benefits during 20 years, without discounting the benefits to present value, is approximately 38.903% for a bridge project, 2.048% for a transit terminal and parking garage project, and 8.522% for a highway corridor connectivity and improvement project (Glover, 2021). Excluding one transit and parking-related project and one corridor improvement project among the top 30 SMART SCALE projects in fiscal year 2024, an 8.522% rate of return is used for calculating the benefit of faster prioritization for the remaining 28 projects, as shown in Equation 15.

> Benefit of faster prioritization for the remaining 28 projects = (8.522% − 0%) × time saved (1.5 years) × total project spending ($102 million) = $13 million (Eq. 15)

In summary, using the models developed in this project to obtain more accurate

disadvantaged populations at the block group level, VDOT can avoid delays in the SMART SCALE process (that would otherwise occur if VDOT required the accurate data associated with Census Bureau custom tabulations). Accelerating the prioritization and implementation of high-impact projects saves time and reduces the opportunity cost associated with idle funds. With a total project cost of $106 million for the top 30 SMART SCALE projects in 2024, excluding two specific projects, as noted previously, a faster prioritization could yield a benefit of $13 million based on the difference between the rate of return on idle funds (0%) and the rate of return on SMART SCALE investments (8.522%).

Maintaining Integrity of the Project Selection Process

The disadvantaged population is a key factor in measuring accessibility in SMART SCALE project prioritization. For example, 20% of the accessibility factor's weight is allocated to the "Access to Jobs for Disadvantaged Populations" measure. This measure is calculated by dividing the average change in "decayed" jobs by the disadvantaged population in each analytic zone at the census block group level (CTB, 2024a).

According to Equation 5 (disadvantaged population = 0.886 * total racial minority population + 0.422 * total poverty population + 0.125 * total LEP population), the actual disadvantaged population for each block group is between 12.5% and 67.1% of the total population for the three categories currently used in SMART SCALE.

Because the current SMART SCALE method overestimates the disadvantaged population for each block group, applying the more accurate methodology presented here will result in a smaller total disadvantaged population figure used to calculate the score for disadvantaged access to jobs. This reduction in overlapping population counts will improve the overall accessibility score and, consequently, influence the SMART SCALE score, even though the 20% weighting remains the same. This enhancement in accuracy increases the precision of the overall SMART SCALE score. As a result, projects that improve job access for disadvantaged communities will be favored to the degree identified by the CTB (2024b), leading to more equitable outcomes without altering the established weight.

ACKNOWLEDGMENTS

This study benefited from the guidance of a technical review panel that provided critical feedback throughout the project: Peng Xiao, Modeling and Accessibility Program Manager, VDOT TMPD (Project Champion); Jitender Ramchandani, Program Manager, Office of Intermodal Planning and Investment; Jungwook Jun, Planning Data Solutions Manager, VDOT TMPD; Angela Effah-Amponsah, Program Administrative Specialist, VDOT Hampton Roads District; Vahid Moshtagh, Senior Project Manager, VDOT Northern Virginia District; and Peter Ohlms, Senior Research Scientist, VTRC. A special tabulation of disadvantaged populations was provided by Nick Spanos, Branch Chief, and Kristin Wimbrow, IT Specialist, both of the Custom Tabulation Branch, American Community Survey Office, U.S. Census Bureau.

REFERENCES

Bejleri, I., Noh, S., Gu, Z., Steiner, R.L., and Winter, S.M. Analytical Method to Determine Transportation Service Gaps for Transportation Disadvantaged Populations. *Transportation Research Record*, Vol. 2672(8), 2018, pp: 649–661.

Brumbaugh, S. *Travel Patterns of American Adults with Disabilities*. Bureau Of Transportation Statistics, U.S. Department of Transportation, Washington, DC, 2018. https://www7.bts.dot.gov/sites/bts.dot.gov/files/docs/explore-topics-and-geography/topics/passenger-travel/222466/travel-patterns-american-adults-disabilities-11-26-19.pdf. Accessed February 28, 2024.

California Environmental Protection Agency. SB 535 Disadvantaged Communities. State of California, California Office of Environmental Health Hazard Assessment, 2021. https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40. Accessed March 11, 2023.

Case, R.B. *Non-Driver Residential Locations at the Census Block Level by Vehicle Availability*. Hampton Roads Transportation Planning Organization, Chesapeake, VA, 2009. https://www.hrtpo.org/DocumentCenter/View/1990/T09-05-Non-Driver-Final-Report-PDF. Accessed March 1, 2024.

Centers for Disease Control and Prevention. PLACES: Local Data for Better Health. 2023. https://www.cdc.gov/places/. Accessed March 4, 2024.

Code of Federal Regulations. 13 C.F.R. § 124.103, 2011. https://www.ecfr.gov/current/title-13/chapter-I/part-124/subpart-A/subject-group-ECFR4ef1291a4a984ab/section-124.103. Accessed October 20, 2023.

Commonwealth Transportation Board. *SMART SCALE Technical Guide*. Commonwealth Transportation Board, Richmond, VA, 2021.

Commonwealth Transportation Board. *SMART SCALE Technical Guide.* Commonwealth Transportation Board, Richmond, VA, 2024a. https://smartscale.virginia.gov/media/smartscale/documents/508_R6_Technical-Guide_FINAL_FINAL_acc043024_PM.pdf. Accessed September 3, 2024.

Commonwealth Transportation Board. *Project Scores*. Commonwealth Transportation Board, Richmond, VA, 2024b. https://smartscale.virginia.gov/media/smartscale/documents/current-projects-round-5/project_scores_fy24_20230117.xlsx. Accessed September 10, 2024.

Council on Environmental Quality. Climate and Economic Justice Screening Tool (CEJST). 2022. https://screeningtool.geoplatform.gov/en/#3/33.47/-97.5. Accessed March 4, 2024.

Farmer, L. Email to Yiqing Xu. October 6, 2024.

Federal Highway Administration. National Household Travel Survey, n.d. https://nhts.ornl.gov/. Accessed April 30, 2025.

Federal Transit Administration. 2021 *Planning Emphasis Areas*. FTA, Washington, DC, 2021. https://www.transit.dot.gov/regulations-and-programs/transportation-planning/2021-planning-emphasis-areas. Accessed February 17, 2023.

Gilliland, J.A., Shah, T.I., Clark, A., Sibbald, S., and Seabrook, J.A. A Geospatial Approach to Understanding Inequalities in Accessibility to Primary Care Among Vulnerable Populations. *PloS One*, Vol. 14(1), 2019, p: e0210113.

Glover, B. Return on Investment: Transportation Projects Can More than Pay for Themselves in Benefits. Texas A&M Transportation Institute. El Paso, Texas, 2021. https://ciitr.tti.tamu.edu/2021/02/23/return-on-investment-transportation-projects-can-more-than-pay-for-themselves-in-benefits/#:~:text=Intelligent%20Transportation%20Research-,Return%20on%20Investment%3A%20Transportation%20Projects%20Can%20More,Pay%20for%20Themselves%20in%20Benefits&text=Roads%2C%20bridges%2C%20bike%20lanes%20and,dollars%20in%20the%20long%20run. Accessed October 7, 2024.

Government Finance Officers Association. "Infrastructure Investment and Jobs Act (IIJA) Implementation Resources." 2023. https://www.gfoa.org/the-infrastructure-investment-and-jobs-act-iija-was. Accessed February 17, 2023.

IBM. IBM SPSS Software, n.d. https://www.ibm.com/spss. Accessed April 30, 2025.

Kind, A.J., and Buckingham, W.R. Making Neighborhood-Disadvantage Metrics Accessible—The Neighborhood Atlas. *The New England Journal of Medicine*, Vol. 378(26), 2018, pp: 2456–2458. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6051533/pdf/nihms979553.pdf. Accessed March 1, 2024.

Li, K., Miller, J.S., and Ohlms, P.B. An Approach for Estimating the Size of the Disadvantaged Population Based on Virginia's Public Use Microdata Areas. March 19, 2023. http://vdotlibrary/taprojects/122889/122889_FinalDoc2.docx. Accessed November 14, 2023.

Mattson, J.W. *Travel Behavior and Mobility of Transportation-Disadvantaged Populations: Evidence from the National Household Travel Survey*. DP-258. Upper Great Plains Transportation Institute, Fargo, ND, 2012. https://www.dvrpc.org/getinvolved/publicparticipation/pdf/disadvantaged_travelers_patterns.pdf. Accessed February 28, 2024.

Mattson, J., and Molina, A. *Travel Behavior of Transportation-Disadvantaged Populations: Trends and Geographic Disparities*. SURTCOM22-10. U.S. Department of Transportation, Washington, DC, 2022. https://www.chinautc.com/upload/fckeditor/surtcom22-10.pdf. Accessed February 28, 2024.

Memphis Metropolitan Planning Organization. *2020 Public Participation Plan*. Memphis Metropolitan Planning Organization, Memphis, TN, 2020.

New York State Climate Action Council, Scoping Plan Full Report, 2022. https://climate.ny.gov/resources/scoping-plan. Accessed March 29, 2024.

New York State Energy Research and Development Authority. Disadvantaged Communities Criteria. 2023. https://climate.ny.gov/Resources/Disadvantaged-Communities-Criteria. Accessed March 29, 2024.

Office of Economic Impact and Diversity. *Energy Justice Mapping Tool: Disadvantaged Communities Reporter*. U.S. Department of Energy, Washington, DC, 2023. https://energyjustice.egs.anl.gov/. Accessed October 20, 2023.

Ohlms, P. Status of Projects Working 10-16-24. 2024. https://virginia.app.box.com/s/xgdqvgvtvlaupbj6eaed0j9a1955f6lk. Accessed December 10, 2024.

Rogoff, P. *Title VI Requirements and Guidelines for Federal Transit Administration Recipients, FTA Circular C 4702.1B*. Federal Transit Administration, Washington, DC, 2012. https://www.transit.dot.gov/sites/fta.dot.gov/files/docs/FTA_Title_VI_FINAL.pdf. Accessed December 6, 2024.

Spanos, N. Email to John Miller. October 3, 2023a.

Spanos, N. Email to John Miller. April 1, 2023b.

Spanos, N. Email to John Miller. April 3, 2023c.

Spanos, N. Email to John Miller. March 13, 2024a.

Spanos, N. Email to Yiqing Xu. February 5, 2024b.

Spanos, N. Email to Lance Dougald. June 6, 2024c.

Spielman, S.E., and Singleton, A. Studying Neighborhoods Using Uncertain Data From the American Community Survey: A Contextual Approach. *Annals of the Association of American Geographers*, Vol. 105(5), 2015, pp: 1003–1025.

U.S. Census Bureau. American Community Survey Data. Washington, DC, July 31, 2024. https://www.census.gov/programs-surveys/acs/data.html. Accessed April 30, 2025.

U.S. Census Bureau. Table C16002: Household Language by Household Limited English Speaking Status. 2021 ACS 5-year Estimates Detailed Tables. Washington, DC, n.d.a. https://data.census.gov/table/ACSDT1Y2023.C16002?q=C16002. Accessed November 14, 2023.

U.S. Census Bureau. Table S0601: Selected Characteristics of the Total and Native Populations in the United States. 2020 ACS 5-year Estimates Subject Tables. Washington, DC, n.d.b. https://data.census.gov/table?q=S0601. Accessed September 21, 2023.

U.S. Census Bureau. Table S1701: Poverty Status in the Past 12 Months. 2020 ACS 5-year Estimates Subject Tables. Washington, DC, n.d.c. https://data.census.gov/table?q=S1701. Accessed September 21, 2023.

U.S. Census Bureau. Table B16009: Poverty Status in the Past 12 Months by Age by Language Spoken at Home for the Population 5 Years and Over. 2020 ACS 5-year Estimates Detailed Tables. Washington, DC, n.d.d. https://data.census.gov/table?q=B16009. Accessed September 21, 2023.

U.S. Census Bureau. Table B02001: Race. 2020 ACS 5-year Estimates Detailed Tables. Washington, DC, n.d.e. https://data.census.gov/table?q=B02001. Accessed September 21, 2023.

U.S. Census Bureau. Table B17021: Poverty Status of Individuals in the Past 12 Months by Living Arrangement. 2020 ACS 5-year Estimates Detailed Tables. Washington, DC, n.d.f. https://data.census.gov/table?q=B17021. Accessed September 21, 2023.

U.S. Census Bureau. Table B16004: Age by Language Spoken at Home by Ability to Speak English for the Population 5 Years and Over. 2020 ACS 5-year Estimates Detailed Tables. Washington, DC, n.d.g. https://data.census.gov/table?q=B16004. Accessed September 21, 2023.

U.S. Department of Transportation. Equitable Access to Transportation Systems: What is Transportation Equity? U.S. Department of Transportation, Bureau of Transportation Statistics, 2022. https://transportation.libguides.com/Transportation-Equity/What-Is-Equity. Accessed November 9, 2023.

U.S. Department of Transportation. USDOT Equitable Transportation Community (ETC) Explorer. 2023a. https://experience.arcgis.com/experience/0920984aa80a4362b8778d779b090723/page/Transportation-Insecurity-Analysis-Tool/. Accessed March 4, 2024.

U.S. Department of Transportation. Grant Project Location Verification. 2023b. https://www.transportation.gov/RAISEgrants/raise-app-hdc. Accessed March 4, 2024.

U.S. Department of Transportation. Screening Tool for Equity Analysis of Projects (STEAP). 2023c. https://maps.dot.gov/fhwa/steap/. Accessed March 4, 2024.

U.S. Department of Transportation. Justice40 Rail Explorer. 2023d. https://usdot.maps.arcgis.com/apps/webappviewer/index.html?id=fd9810f673b64d228ae072bead46f703. Accessed March 4, 2024.

U.S. Department of Transportation. Equity and Justice40 Analysis Tools. 2024. https://www.transportation.gov/grants/dot-navigator/equity-and-justice40-analysis-tools. Accessed March 4, 2024.

U.S. Environmental Protection Agency. EJScreen – EPA's Environmental Justice Screening and Mapping Tool. 2024. https://19january2021snapshot.epa.gov/ejscreen_.html. Accessed on March 4, 2024.

U.S. Department of Justice. Source and Methodology. Washington, DC, 2020. https://www.lep.gov/source-and-methodology. Accessed December 5, 2024.

U.S. Environmental Protection Agency. EJScreen–EPA's Environmental Justice Screening and Mapping Tool. 2024. https://ejscreen.epa.gov/mapper/. Accessed March 4, 2024.

Virginia Department of Transportation. Pathways for Planning, n.d. https://vdotp4p.com/. Accessed April 30, 2025.

Virginia Office of Intermodal Planning and Investment. VTrans-Transportation Plan. Megatrend 4: Socio-demographic Changes. January 21, 2022. https://vtrans.virginia.gov/long-term-planning/megatrend-sociodemographic. Accessed April 30, 2025.

APPENDIX A: DATA ERROR INTRODUCED BY THE LIMITED UNIVERSE OF THE CUSTOM TABULATION

Table A1. Data Error Caused by the Limited Universe for the 20 Census Block Groups

GEO ID	Geoname	Custom Tabulation		Table B02001		Table B17021: Total Poverty	Table B16004: Total LEP
		Total Population	DP	Total Population	Total Racial Minority		
511790102181	BG 1, Census Tract 102.18, Stafford County, Virginia	152	0	1,574	533	0	1,582
511330203012	BG 2, Census Tract 203.01, Northumberland County, Virginia	739	0	738	0	0	702
511670304034	BG 4, Census Tract 304.03, Russell County, Virginia	945	0	1,032	0	0	809
511670304032	BG 2, Census Tract 304.03, Russell County, Virginia	370	0	421	10	0	350
510150707011	BG 1, Census Tract 707.01, Augusta County, Virginia	1,142	0	1,157	0	0	1,039
511850203023	BG 3, Census Tract 203.02, Tazewell County, Virginia	958	0	979	0	0	906
518100440082	BG 2, Census Tract 440.08, Virginia Beach city, Virginia	59	0	58	0	0	58
510030109041	BG 1, Census Tract 109.04, Albemarle County, Virginia	5	0	1,487	568	0	1,541
510670201042	BG 2, Census Tract 201.04, Franklin County, Virginia	616	0	676	0	0	586
511710402011	BG 1, Census Tract 402.01, Shenandoah County, Virginia	673	0	737	0	0	475
517402130021	BG 1, Census Tract 2130.02, Portsmouth city, Virginia	645	0	675	0	0	585
511099501012	BG 2, Census Tract 9501.01, Louisa County, Virginia	716	0	859	0	0	480
511770203142	BG 2, Census Tract 203.14, Spotsylvania County, Virginia	37	0	39	0	0	39
511770203151	BG 1, Census Tract 203.15, Spotsylvania County, Virginia	34	0	31	0	0	31
510019801001	BG 1, Census Tract 9801, Accomack County, Virginia	26	0	27	0	0	27
511552105001	BG 1, Census Tract 2105, Pulaski County, Virginia	1,080	0	1,098	0	0	901

GEO ID	Geoname	Custom Tabulation		Table B02001		Table B17021: Total Poverty	Table B16004: Total LEP
		Total Population	DP	Total Population	Total Racial Minority		
510150703003	BG 3, Census Tract 703, Augusta County, Virginia	206	0	210	0	0	199
511910102003	BG 3, Census Tract 102, Washington County, Virginia	736	0	750	0	0	620
510599801001	BG 1, Census Tract 9801, Fairfax County, Virginia	5	0	5	0	0	5
511390303004	BG 4, Census Tract 303, Page County, Virginia	754	0	754	0	0	655

BG = block group; DP = disadvantaged population; GEO ID = geographic identifier; LEP = limited English proficiency.

APPENDIX B: DATA ERROR INTRODUCED BY THE DISCRETE GAUSSIAN NOISE MECHANISM

Table B1. Data Error Caused by the Noise Mechanism for the 30 Census Block Groups

GEO_ID	Geoname	Total Population	CME	Total DP	DP_CME
510594516011	BG 1, Census Tract 4516.01, Fairfax County, Virginia	2,131	579	2,132	579
510872008053	BG 3, Census Tract 2008.05, Henrico County, Virginia	988	345	989	345
510872010014	BG 4, Census Tract 2010.01, Henrico County, Virginia	1,899	722	1,900	722
510872014051	BG 1, Census Tract 2014.05, Henrico County, Virginia	2,265	1,148	2,272	1,148
511539004036	BG 6, Census Tract 9004.03, Prince William County, Virginia	22	23	25	23
511539014098	BG 8, Census Tract 9014.09, Prince William County, Virginia	83	135	84	135
515500202001	BG 1, Census Tract 202, Chesapeake city, Virginia	1,060	418	1,063	418
515500202002	BG 2, Census Tract 202, Chesapeake city, Virginia	823	449	825	449
515900006001	BG 1, Census Tract 6, Danville city, Virginia	735	201	741	201
515958901001	BG 1, Census Tract 8901, Emporia city, Virginia	955	363	958	363
515958901002	BG 2, Census Tract 8901, Emporia city, Virginia	302	142	303	142
517000308001	BG 1, Census Tract 308, Newport News city, Virginia	712	302	715	302
517000308003	BG 3, Census Tract 308, Newport News city, Virginia	320	118	325	118
517100043002	BG 2, Census Tract 43, Norfolk city, Virginia	1,307	264	1,308	264
517100043003	BG 3, Census Tract 43, Norfolk city, Virginia	328	124	333	124
517100050003	BG 3, Census Tract 50, Norfolk city, Virginia	971	327	972	327
517100051003	BG 3, Census Tract 51, Norfolk city, Virginia	758	281	763	281
517308105004	BG 4, Census Tract 8105, Petersburg city, Virginia	708	232	711	232
517308106001	BG 1, Census Tract 8106, Petersburg city, Virginia	1,128	251	1,130	251
517402118004	BG 4, Census Tract 2118, Portsmouth city, Virginia	424	161	427	161
517402124003	BG 3, Census Tract 2124, Portsmouth city, Virginia	1,304	502	1,306	502
517600202001	BG 1, Census Tract 202, Richmond city, Virginia	1,403	409	1,404	409
517600204005	BG 5, Census Tract 204, Richmond city, Virginia	927	345	929	345
517600301001	BG 1, Census Tract 301, Richmond city, Virginia	550	205	552	205
517600301002	BG 2, Census Tract 301, Richmond city, Virginia	1,536	460	1,537	460
517600302002	BG 2, Census Tract 302, Richmond city, Virginia	16	32	20	32
517600608004	BG 4, Census Tract 608, Richmond city, Virginia	17	34	20	34
517600706011	BG 1, Census Tract 706.01, Richmond city, Virginia	1,313	416	1,318	416
517600706013	BG 3, Census Tract 706.01, Richmond city, Virginia	1,467	414	1,470	414
517700025011	BG 1, Census Tract 25.01, Roanoke city, Virginia	863	292	864	292

BG = block group; CME = census margin of error; DP = disadvantaged population; GEO ID = geographic identifier.