URBAN
INSTITUTE · ELEVATE · THE · DEBATE

Spatial Equity Data Tool

Technical Appendix

Last updated November 18, 2021.

Ajjit Narayanan, Alena Stern, and Graham MacDonald

This document describes the methods and data the Spatial Equity Data Tool uses to evaluate the demographic and geographic representativeness of datasets. While the version 1 tool (released September 24, 2020) performed analysis only at the city level, the version 2 tool (released November 18, 2021) can perform analysis at the city, county, state, and national levels. In short, the user selects a geographic level of analysis (city, county, state, or national) and the tool geocodes the uploaded data to a geography of that level (a single census place [city], county, state, or the entire US) and compares the distribution of the data within the geography with the distribution of a set of baseline measures from the five-year American Community Survey (ACS). It then visualizes these differences for smaller geographies (tract, tract, county, or state, respectively) on a map and at the geography-wide (citywide, countywide, statewide, or nationwide, respectively) level on a chart.

Methods

Once users upload their data, the Spatial Equity Data Tool goes through nine steps to produce the demographic and geographic representation metrics.

1. Filter Out Null Values

The tool filters out rows in which the longitude, latitude, weight, or filter columns have null values. All the following are treated as null values:

- ` ` (i.e., blank values)
- `#N/A`
- `#N/A N/A`

- `#NA`
- `-1.#IND`
- `-1.#QNAN`
- `-NaN`
- `-nan`
- `1.#IND`
- `1.#QNAN`
- `<NA>`
- `N/A`
- `NA`
- `NULL`
- `NaN`
- `n/a`
- `nan`
- `null`

2. Apply Filters

If the user selects filters for their data, the tool applies those filters. The tool can accept three types of filters (text, numeric, and date) that are applied in the following order: text filters, numeric filters, and date filters. If multiple filter conditions are given, the filter conditions are chained and evaluated together. For example, if a user chooses the filters

1. zip_code = 20024 and
2. date_opened = 01/01/2020,

then the tool would filter to rows in which the zip_code column is 20024 *and* the date_opened column is 01/01/2020.

3. Apply Weights

If the user selects a weight column, then the tool notes it. If not, the tool creates a dummy weight column with values of 1 for every row, thus weighting every row in the dataset equally.

4. Determine the Dataset's Source Geography

To minimize the burden on users, the tool determines the geography automatically from the data provided. Before uploading the data, the user selects a geographic level of analysis: city, county, state, or national. The tool uses this selection to identify the relevant set of potential geographic boundaries (e.g., all cities, all counties, all states, or the nation). For county, state, and national analyses, we use the boundaries provided by the US Census Bureau (noting that we do not include US territories as part of the US boundary, given the lack of available demographic data). We define city boundaries as all census tracts contained within the census place boundary. In some cases, our definition may be larger than the official city boundary; see the "City Definition" section for more details. For city-level analysis, the tool uses a precomputed list of all 835 US cities with populations above 50,000 in 2019.

The tool then spatially joins the points in the data to a precompiled dataset of the boundaries corresponding to the selected geographic level of analysis. If the dataset has more than 50 points, then the tool takes a 5 percent sample for the spatial join to limit computation time. If the points in the data fall within the boundaries of multiple geographies, the tool only keeps the points within the most frequently occurring geography in the data. For example, if a user has selected county-level analysis and the user's data fall within multiple counties, only data from the most frequently occurring county in the dataset will be kept.

5. Read in All Geographic and Demographic Data for That Geography

Once the tool identifies a geography, it reads in all the precompiled geographic and demographic data for that geography. These data include

1. the boundaries for all census tracts in the geography
2. the ACS demographic variables for every tract in the geography
3. precomputed ACS demographic statistics for the entire geography

National:

4. the boundaries for all states in the US (for geographic disparity score)
5. precomputed ACS demographic statistics for all states in the US (for demographic disparity score)

State:

4. the boundaries for all counties in the state (for geographic disparity score)
5. precomputed ACS demographic statistics for all counties in the state (for demographic disparity score)

For more information on the specific ACS demographic variables used in the tool, see the "Data" section.

6. Compute Which Census Tract Each Data Point Falls Into

The tool spatially joins each point (i.e., row) in the user-uploaded data to the set of all census tracts in the geography. Any points that do not fall within any census tract—that is, points located outside the main geography boundary—are discarded and not included in future calculations.

7. Compute Geographic Disparity Scores

The tool calculates geographic disparity scores to assess which areas are over- and underrepresented relative to a baseline population. The geographic disparity scores are calculated and visualized at different subgeographies, depending on the selected geography:

- National: the tool calculates the geographic disparity score for each ***state*** in the US.
- State: the tool calculates the geographic disparity score for each ***county*** in the state.
- County: the tool calculates the geographic disparity score for each ***tract*** in the county.
- City: the tool calculates the geographic disparity score for each ***tract*** in the city.

We use different subgeographies for each level of analysis because user feedback indicated that the different geographic levels of analysis required different "subgeographies" to make the geographic disparity score map meaningful (e.g., users suggested that a nationwide tract level map was not meaningful).

The geographic disparity score is the difference between the proportion of the user-uploaded data within that subgeography and the proportion of the main geography's baseline population within that subgeography. For example, if a state-level analysis is selected, the geographic disparity score is the difference between the proportion of the user-uploaded data within that county and the proportion of the state's baseline population within that county. We compute this score for the following baseline populations:

1. total population
2. population with low incomes
3. population with extremely low incomes
4. population without internet access
5. population of cost-burdened renter households
6. senior (≥ 65) population
7. child (< 18) population

These geographic disparity scores tell us, at a high level, how the data are over- and underrepresented for each subgeography relative to each baseline population. For the state-level analysis example, if a county contains 3.3 percent of the user-uploaded dataset but 1.4 percent of the state's senior population, then the spatial disparity score for the senior population in that county would be 3.3 − 1.4, or 1.9 percent. The tool repeats this calculation for every subgeography and for every baseline population. These geographic disparity scores are displayed by the interactive maps on the tool.

8. Compute Demographic Disparity Scores

The demographic disparity score is the percentage point difference between the representation of a demographic group in the data (the data-implied percentage) and the representation of that group in the geography (the geography-wide percentage). The demographic disparity score is always calculated using tract-level data, regardless of the selected geography. This is because we consider the tract level the appropriate unit of analysis to assess the demographics of the individuals with access to a given data point. We illustrate how we calculate this metric through the example of calculating the demographic disparity score for Black residents in a simple geography with two census tracts.

The first step is to calculate the geography-wide percentage, which answers the question, "What is the share of Black residents in the entire geography?" For city-level analysis, we calculate the geography-wide percentages from the tract-level census data for the tracts in the city. To illustrate, assume that our example geography is a city and that each of the two tracts is home to 50 percent of the city's population. If Tract 1 is 20 percent Black and Tract 2 is 40 percent Black, then the geography-wide (in this case, citywide) percentage of Black residents is (0.5)(0.2) + (0.5)(0.4) = 0.3. For county, state, and national analysis, the geography-wide percentages are directly reported by the census, so we can imagine the census reported that 30 percent of the population of our geography is Black.

Now imagine 80 percent of the data uploaded by the user are located in Tract 1 and 20 percent are located in Tract 2. Then the data-implied percentage of Black residents would be (0.8)(0.2) + (0.2)(0.4) = 0.24. This is calculated identically across national, state, county, and city-level analysis. The data-implied percentage of Black residents answers the question: "What is the share of Black residents in the tracts that the data come from?"

Finally, the demographic disparity score is the difference between the data-implied percentage and the geography-wide percentage, or 0.24 − 0.3 = −0.06. In this example, Black residents are underrepresented by 6 percentage points in the data relative to the geography population. These demographic disparity scores are calculated for all our demographic variables of interest (see the "Data" section for a full list) and visualized by the interactive chart on the tool.

For some demographic variables, we use the appropriate universe (or denominator) instead of the total population for the geography-wide calculations. For example, to generate the demographic disparity score for veterans, we use the proportion of the civilian population 18 years and older as the denominator. For more information on the universes used for each demographic variable, see the "Data" section.

For the state- and national-level analyses, we choose to show the demographic disparity score for the relevant geography (state or US) and the demographic disparity score for the smaller subgeographies that make up the main geography (counties and states, respectively). We describe how we calculate these demographic disparity scores for each level of analysis:

- National: We calculate the national demographic disparity scores as well as the demographic disparity scores for each state. To calculate the scores for each state, the geography-wide percentage is the census-published share of the given demographic group in the state (e.g., the share of Black residents in Illinois). The data-implied percentage is calculated as described above for the subset of points that fall in each state (e.g., the share of Black residents in the tracts that the data *in Illinois* come from). The state disparity score is the difference between the geography-wide percentage and the data-implied percentage for each state.
- State: We calculate the state demographic disparity scores as well as the demographic disparity scores for each county in the state. To calculate the scores for each county, the geography-wide percentage is the census-published share of the given demographic group in the county (e.g., the share of Black residents in Cook County). The data-implied percentage is calculated as described above for the subset of points that fall in each county (e.g., the share of Black residents in the tracts that the data *in Cook County* come from). The county disparity score is the difference between the geography-wide percentage and the data-implied percentage for each county.

Note that in both cases, we do not display disparity scores for subgeographies that contain no points in the user-uploaded dataset. The subgeography disparity scores are displayed in grey in the tool chart.

9. Compute Statistical Significance

Both the demographic and the geographic disparity scores above rely on ACS demographic variables, which have associated margins of error. We use the margins of error for the relevant ACS variables to construct 95 percent confidence intervals for both the geographic and the demographic disparity scores and determine if the scores differ significantly from 0. To calculate the margins of error, we use the formulas for user-derived proportions and percentages from a recent ACS handbook (US Census Bureau 2018). We perform this significance calculation for all our demographic and geographic disparity scores. Insignificant scores are reported as dark grey in our tool to differentiate them from significant scores. For more detailed information, see the "Statistical Significance Calculations" section.

Limitations

The methodology behind this tool has at least five limitations that users should take into account:

1. Quantifying geographic representativeness of data requires some measure of ground truth (i.e., baseline measures) for comparison. The tool allows different baseline measures as laid out in step 8. Though these baseline measures should be a sensible starting point for many datasets (particularly datasets that concern the specific baseline populations), it may occur that none of our baseline measures could be suitable for some user uploaded datasets.
2. The tool imputes demographic data from user-uploaded geographic data, which assumes that all data points coming from a census tract inherit the same attributes of that census tract. This is problematic because data points from a majority-white census tract could have been generated by nonwhite residents, and vice versa. Likewise, a program or resource located in a tract may not be used equally by all residents of the tract.
3. This tool works best on medium to large datasets because the geographic unit of analysis for the demographic disparity score is the census tract. While there's no fixed rule, a good rule of thumb is to aim for a number of points in the dataset at least as large as the number of tracts in the geography (e.g., approximately 73,000 for the US or hundreds of census tracts in a large city). With a small number of points, the tool generates less reliable disparity scores, as the vast majority of tracts will contain no points or just a few points. While we believe this tool may be useful on smaller datasets in certain cases, we recommend users rely on it for datasets that follow the rule of thumb above.
4. This tool currently supports assessing disparities in a single geography (city, state, county, or the US). If a dataset spans multiple geographies, the tool will only operate on the most frequently occurring geography in the data and remove the remainder of the observations from the dataset. This is particularly problematic for regional analyses that span multiple geographies but do not cover the entire US.
5. Finally, as noted under "City Definition," the tool's operational definition for the boundary of a city might differ slightly from the official city boundary that the Census Bureau uses.

The tool defines a city as all census tracts whose area is at least 1 percent covered by the relevant census place. Often the boundaries of census places and census tracts don't overlap perfectly, meaning some tracts are only partially covered by the place boundary. This overinclusive definition will cause our tool to think that many cities—particularly small and medium-sized ones—are bigger than they are, in both geographic size and population.

Data

In addition to the user-uploaded data, the tool pulls in and uses census tract, county, state, and nationwide data from the 2015–19 ACS. Below are the ACS variables used to calculate the demographic and geographic disparity scores. We also note in parentheses the universe (i.e., denominators) used for each variable of interest.

Geographic Disparity Scores

The tool uses seven ACS baseline variables to calculate geographic disparity scores. We use different subgeographies across the different levels of analysis noted in parentheses: national (state), state (county), county (tract), and city (tract).

1. total population: percentage of the geography's total population living in each subgeography (universe: total geography population)
2. population with low incomes: percentage of the geography's residents with incomes below 200 percent of the federal poverty level living in each subgeography (universe: total number of residents with low incomes in the geography)
3. population with extremely low incomes: percentage of the geography's residents with incomes below the federal poverty level living in each subgeography (universe: total number of residents with extremely low incomes in the geography)
4. senior population: percentage of the geography's population ages 65 and older living in each subgeography (universe: total number of seniors in the geography)
5. child population: percentage of the geography's population under 18 living in each subgeography (universe: total number of children in the geography)
6. population of cost-burdened renter households: percentage of the geography's renter households who pay more than 35 percent of their income on rent living in each subgeography (universe: total number of cost-burdened renter households in the geography

7. population without internet access: percentage of the geography's households who do not have internet access in each subgeography (universe: total number of households without internet access in the geography)

Demographic Disparity Scores

The tool computes a demographic disparity score for each demographic group as defined below. Note that the tool uses "Latinx" instead of Latino or Hispanic to remain inclusive of gender-nonconforming and nonbinary individuals. We use "Hispanic" in the technical appendix to be consistent with the terminology the ACS uses. We have split the demographic groups by the baseline population to which they apply. Specifically, you can choose from three baseline populations on the demographic disparity chart.

Total population:

- non-Hispanic Asian residents (as a share of total population)
- non-Hispanic Black residents (as a share of total population)
- Hispanic residents (as a share of total population)
- non-Hispanic white residents (as a share of total population)
- non-Hispanic residents of all other races or ethnicities (as a share of total population)
- senior residents ages 65 and older (as a share of total population)
- children, or residents under the age of 18 (as a share of total population)
- veteran residents (as a share of civilian population 18 years old or older)
- uninsured residents (as a share of civilian noninstitutionalized population)
- residents with disabilities (as a share of civilian noninstitutionalized population)
- renters (as a share of occupied housing units)
- cost-burdened renter households, in which renters spend more than 35 percent of income on rent (as a share of renter-occupied households paying rent, when gross rent can be determined)
- households without internet access (as a share of occupied housing units)
- households with limited English proficiency (as a share of occupied housing units)
- residents with low incomes (as a share of population for whom poverty status is determined)
- residents with extremely low incomes (as a share of population for whom poverty status is determined)
- residents with a bachelor's degree or higher (as a share of total population over 25 years old)
- residents with less than a high school diploma (as a share of total population over 25 years old)

- residents who are unemployed (as a share of total population over 16 years old in the civilian labor force)

Child population:

- Hispanic and non-Hispanic Asian children (as a share of total child population)
- Hispanic and non-Hispanic Black children (as a share of total child population)
- Hispanic children (as a share of total child population).
- Hispanic and non-Hispanic white children (as a share of total child population)
- Hispanic and non-Hispanic children of all other races or ethnicities (as a share of total child population)
- children in households with extremely low incomes (as a share of child population for whom poverty status is determined)
- children with a disability (as a share of child civilian noninstitutionalized population)
- uninsured children under the age of 19 (as a share of civilian noninstitutionalized population under the age of 19)
- households with limited English proficiency (as a share of occupied housing units with children)

Population with extremely low incomes
Note: The extremely low-income population is the population with incomes below the federal poverty level.

- Hispanic and non-Hispanic extremely low–income Asian residents (as a share of total extremely low–income population)
- Hispanic and non-Hispanic extremely low–income Black residents (as a share of total extremely low–income population)
- Hispanic extremely low-income residents
- Hispanic and non-Hispanic extremely low–income white residents (as a share of total extremely low–income population)
- Hispanic and non-Hispanic extremely low–income residents of all other races or ethnicities (as a share of total extremely low–income population)
- extremely low–income residents who are unemployed (as a share of total population over 16 years old in the civilian labor force and who are extremely low income)
- extremely low–income residents with a bachelor's degree or higher (as a share of total population over 25 years old who are extremely low income)

- extremely low–income residents with less than a high school diploma (as a share of total population over 25 years old who are extremely low income)
- extremely low–income senior residents ages 65 and older (as a share of total extremely low–income population)
- extremely low–income children under the age of 18 (as a share of total extremely low income population)
- extremely low–income veteran residents (as a share of total civilian population over 18 years old who are extremely low income)
- extremely low–income residents with a disability (as a share of civilian non–institutionalized population who are extremely low income)
- extremely low–income uninsured residents (as a share of civilian non-institutionalized population who are extremely low income)

City Definition

The census analog of a city is a *census place*, and the census analog of a neighborhood is a *census tract*. A census place is a concentration of population that has a name, is locally recognized, and is not part of any other place. Census tract boundaries often, but not always, coincide with the boundaries of places. In some cities—particularly small and medium-sized ones—the city boundaries only partially cover some census tracts. This is problematic because the neighborhood-level demographic data our tool uses are available only at the census tract level. Creating a more accurate city definition would require using more granular block- or block group–level data, which would increase computation time and have far larger margins of error for our demographic variables. So, we needed an operational definition of a city that spans whole census tracts.

We decided on the following census tract–based definition for city boundaries: All tracts that had at least 1 percent of their area contained within the place boundary were considered part of that respective city. Because of this overinclusive definition, the tool will think that many cities, particularly small or medium-sized cities and a handful of irregularly shaped large cities, are bigger than they actually are, in both area and population. We decided on this 1 percent overinclusive cutoff for two reasons:

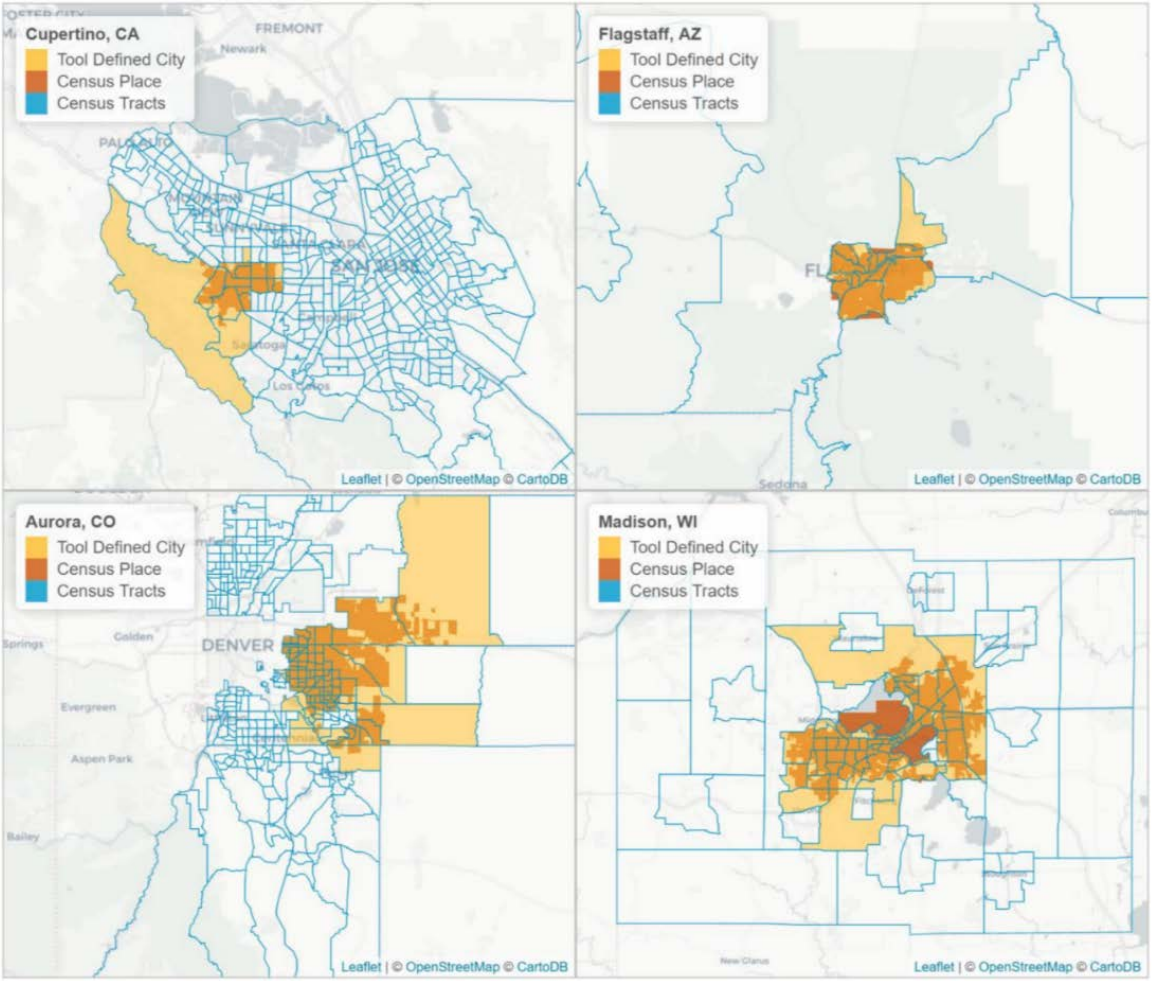
1. After visual inspection of a few cities, the 1 percent cutoff gave reasonable results for city boundaries.
2. The 1 percent cutoff allowed us to exclude tracts on the border of census places. If we imposed no cutoff (i.e., we defined a city as all tracts contained even partially within the

place boundary), then the city definitions would include tracts that shared a small portion of the border and the cities would be much larger than expected.

To visualize how this definition affects small and medium-sized cities, consider figure 1. We can see that small cities such as Cupertino, California, and Aurora, Colorado, have irregular place boundaries that partially overlap a few census tracts. As a result, the area of the cities as defined by our tool (in yellow) is greater than the actual area of the cities (in orange).

FIGURE 1

Boundaries for Cupertino, California; Flagstaff, Arizona; Aurora, Colorado; and Madison, Wisconsin



Source: Urban Institute analysis of US Census Bureau TIGRIS/Line 2018 shapefiles.
Notes: Made with Leaflet | © OpenStreetMap contributors, © CartoDB. The darkest orange in the Madison map is bodies of water that are a part of the census place boundaries but are not assigned a census tract.

County, State, and United States Boundaries

We define the county, state, and US boundaries using the TIGRIS/Line boundary shapefiles from 2019, to align with the data we pull from the 2015–19 five-year ACS. Within the census geography hierarchy, tracts always fit cleanly within county, state, and US boundaries. As a result, we can use directly reported county, state, or US estimates from the census for the national-, state-, and county-level tools. For the city-level tool, we have to derive the comparable city-level estimate from the constituent census tracts, as described in the "City Definition" section above. The states in our tool include the 50 US states and the District of Columbia. We do not include the US territories in our analysis as the ACS does not collect data in American Samoa, Guam, Northern Mariana Islands, and the US Virgin Islands and does not report all of the data used in the tool for Puerto Rico.

Statistical Significance Calculations

To calculate the statistical significance of our demographic and geographic disparity scores, we make use of several Census-provided standard error and margin of error formulas (US Census Bureau 2018, chapter 8). We use formula 1 for aggregating data across geographic areas:

$$\begin{gathered}\mathrm{MOE}(\hat{X}_1 + \hat{X}_2 + \cdots + \hat{X}_n) = \\ \pm\sqrt{[\mathrm{MOE}(\hat{X}_1)]^2 + [\mathrm{MOE}(\hat{X}_2)]^2 + \cdots + [\mathrm{MOE}(\hat{X}_n)]^2}\end{gathered} \tag{1}$$

and formula 6 for computing standard errors and margins of error of user-derived percentages:

$$\mathrm{MOE}(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[\mathrm{MOE}(\hat{X})]^2 - \left(\hat{P}^2 * [\mathrm{MOE}(\hat{Y})]^2\right)} \tag{6}$$

For the rest of this section, we will refer to these formulas as formulas 1 and 6. Since we are interested in computing standard errors, we will also make use of the fact that

$$SE(X) = \frac{MOE(X)}{1.645}$$

Geographic Disparity Scores

The geographic disparity score is the percentage point difference between the proportion of the data originating in a subgeography (state for national-level analysis, county for state-level analysis, and tract for county- and city-level analysis) and the proportion of a geography's total baseline population (total population, child population, etc.) residing within a subgeography. The proportion of the data originating from a subgeography (the data proportion) is a fixed number, but the proportion of a geography's total baseline population residing in a subgeography (baseline proportion) is an estimate from the ACS five-year survey, with associated margins of error for both the numerator (the total baseline population in the subgeography) and the denominator (the total baseline population in the geography). To take into account the variability in the ACS measures, we computed standard errors for the baseline proportions using formulas 1 and 6 to construct 95 percent confidence intervals for the geographic disparity score. If the confidence interval contains 0, then we report that subgeography's disparity score as insignificant.

As a simple example, imagine a geography with two census subgeographies, A and B. Subgeography A contains 20 people and 10 data points. Subgeography B contains 25 people and 10 data points. The geographic disparity score for the total population in subgeography A would be (10/20) − (20/45) = +0.0556. But is this overrepresentation significant?

Now assume that the census-provided margins of error for the total population is 10 people in Subgeography A and 5 people in Subgeography B. And the corresponding standard errors are $\frac{10}{1.645}$ and $\frac{5}{1.645}$, or 6.08 and 3.04, respectively. We also need the standard error for the total population in the geography.

- For the national-, state-, and county-level analyses, the census directly reports estimates and margins of error for the total population in the geography. Assume the census-reported MOE for the total population in the geography is 11.18.
- For city-level analysis, we must derive the MOE for the total population in the geography ourselves. We would use formula 1 to calculate the margin of error for the total population in the city as $\sqrt{(5^2 + 10^2)}$, or 11.18.

Now that all required MOEs are calculated, we use formula 6 to calculate the standard error of the population proportion in subgeography A as

$$SE(population\ proportion) = \frac{MOE(population\ proportion)}{1.645} = \frac{\sqrt{10^2 - \left(\frac{20}{45}^2 * 11.18^2\right)}}{45(1.645)} = 0.117$$

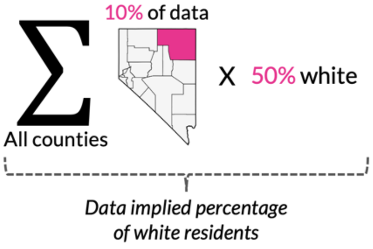
To construct the 95 percent confidence interval, we simply set the upper and lower bounds at 1.96 times the standard error. In this case, the population proportion confidence interval for subgeography A would be

$$\mathit{Confidence\ interval\ for\ population\ proportion} = \frac{20}{45} \pm 1.96(0.117) = [0.215, 0.674]$$

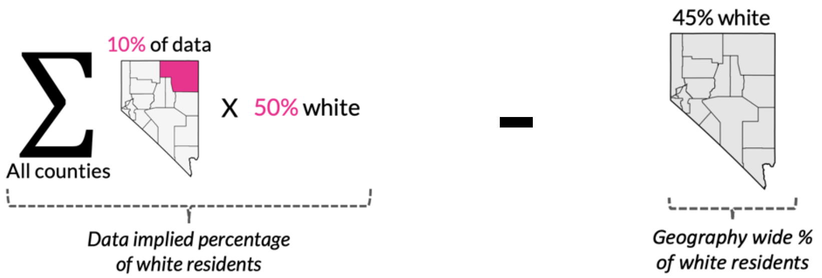
To calculate the corresponding confidence interval for the geographic disparity score, we subtract the confidence interval bounds from the data proportion of 0.5, leaving us with [−0.285, 0.174]. Note that this interval contains 0, meaning we do not have enough evidence to conclude the disparity score differs from 0. Therefore, we mark subgeography A's geographic disparity score as not statistically significant and would grey it out on the map.

Demographic Disparity Scores

The demographic disparity score is the percentage-point difference between the representation of a demographic group in the data (the data-implied percentage) and the representation of that group in the geography (the geography-wide percentage). The data-implied percentage is calculated identically across all geographic levels of analysis: we take a weighted average across all census tracts in the geography, where the weights are the proportion of the data points falling into each tract. Below is a pictorial representation of the data-implied percentage as calculated for the percentage of white residents in the state of Nevada.

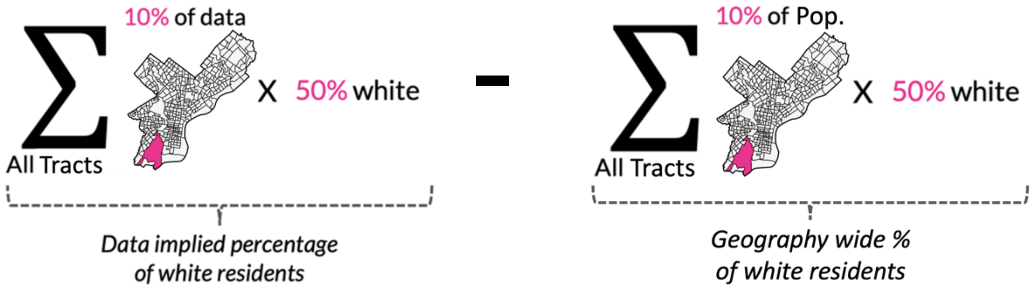


And the geography-wide percentage is the actual reported value for the percentage of white residents across all of Nevada. For the county-, state-, and national-level tools, these percentages are directly reported by the census (or calculated from census-reported counts for the appropriate numerator and denominator). This value is subtracted from the data-implied percentage to calculate the demographic disparity score. Below is a pictorial representation for how the full demographic disparity score calculation is done for white residents in Nevada:


10% of data
All counties
X 50% white
45% white
Data implied percentage of white residents
Geography wide % of white residents

For just the city-level analysis, the geography-wide percentage is calculated from the tracts that compose the city, as our city boundary definitions may differ from the official Census boundaries.

Below is a mathematical and pictorial representation of the demographic disparity score as calculated for white residents in the city of Philadelphia. Note that the data-implied percentage is calculated the same way as for the county, state, and national tool, but the way the geography-wide percentage is calculated has changed.


10% of data
All Tracts
X 50% white
10% of Pop.
All Tracts
X 50% white
Data implied percentage of white residents
Geography wide % of white residents

For the data-implied percentage, the share of white residents in each tract is an ACS variable with associated margin of error. Similarly, the geography-wide percentage uses the total geography population and the total white population in the geography, both of which are ACS variables with associated margins of error. To visualize this, you can rewrite the geography-wide percentage in the diagrams above as follows:

$$\textit{geography-wide percentage (city)} = \frac{\sum \textit{number of white residents}_i}{\sum \textit{total population}_i}$$

$$\textit{geography-wide percentage (county, state, US)} = \frac{\textit{number of white residents}}{\textit{total population}}$$

To take into account the variability in all these ACS figures, we calculate standard errors for the data-implied percentage and the geography-wide percentage using the relevant formulas from the ACS handbook. We then perform a test for statistical significance for the difference between these two ACS measures using a 95 percent confidence level. We use the formula 3 from the ACS handbook (US Census Bureau 2018, chapter 7) to compute statistical significance, where X_1 is the data-implied percentage, X_2 is the geography-wide percentage, and Z_{CL} is 1.96 (because we set a 95 percent confidence level).

$$\text{If} \quad \left| \frac{\widehat{X}_1 - \widehat{X}_2}{\sqrt{\left[SE\left(\widehat{X}_1\right)\right]^2 + \left[SE\left(\widehat{X}_2\right)\right]^2}} \right| > Z_{CL} \qquad (3)$$

Next, we work through an example of the significance calculations for one of our demographic variables—the percentage of the total population that is white:

Imagine a geography with two census tracts, A and B. Census tract A contains 20 people, 15 of whom are white. Census tract B contains 25 people, 5 of whom are white. Also assume that tract A contains 30 percent of the data and tract B contains 70 percent of the data. The data-implied percentage for white residents would be (0.3) * (15/20) + (0.7) * (5/25), or 0.365. If the selected geography were a city, the geography-wide percentage would be (15 + 5) / (20 + 25), or 0.444. Note that for county, state, and national analysis, the numerator and denominator for the geography (or in some cases, the percentage itself of 0.444) would be directly reported. The demographic disparity score for white residents is 0.365 − 0.444 = −0.079.

Now assume that the margin of error for the percentage of white residents is 2 percent (0.02) in tract A is and 5 percent (0.05) in tract B. Using formula 1 from the ACS handbook, and the fact that the data proportion is a constant number for each tract, the standard error for the data-implied percentage would be calculated as follows:

$$SE(data\ implied\ percentage) = \frac{MOE(data\ implied\ percentage)}{1.645} = \frac{\sqrt{(0.3 * 0.02)^2 + (0.7 * 0.05)^2}}{1.645} = 0.021$$

We then need to calculate the standard error of the geography-wide percentage. For the county, state, and national analyses, the standard error may be directly reported (if the census directly reports the percentage estimate and associated margin of error) or can be calculated for the reported margins of error for the numerator (total white population) and denominator (total population) using formula 6 as illustrated below. For the city-level analysis, we need to use the individual tract margins of error for the numerator and denominator to calculate the standard error of the geography-wide percentage as follows. Let's assume that the margins of error for the population counts and white resident counts are as shown below:

	Population MOE	White resident MOE
Tract A	5.5	3.5
Tract B	5	1.5

Using formula 1 from the ACS handbook for aggregating data across geographic areas, the MOE for the total city population would be$\sqrt{5.5^2 + \ 5^2}$, or 7.43, and the MOE for the total white resident population would be $\sqrt{3.5^2 + 1.5^2}$, or 3.81. Then using formula 6, the standard error calculation for the geography-wide percentage of white residents would be

$$SE\ (geography\text{-}wide\ percentage) = \frac{MOE\ (geography\text{-}wide\ percentage)}{1.645} = \frac{\sqrt{3.81^2 - \left(\frac{20^2}{45} * \ 7.43^2\right)}}{45\ (1.645)} = 0.0255$$

Now that we have the value and the standard errors for both the data-implied percentage and the geography-wide percentage, we perform a significance test with a 95 percent confidence level.

$$Z_{CL} = \frac{0.444 - 0.365}{\sqrt{0.021^2 + 0.0255^2}} = \ 2.391$$

This is larger than the critical value of 1.96, so we would report the demographic disparity of white residents at −0.079 as statistically significant.

Reference

US Census Bureau. 2018. *Understanding and Using American Community Survey Data: What All Data Users Need to Know.* Washington, DC: US Government Printing Office.

Acknowledgments

This research was originally funded in collaboration with the Mastercard Center for Inclusive Growth. The Bill & Melinda Gates Foundation funded the tool's expansion from city-level analysis to national-, county-, and state-level analysis. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

We thank our partners—Zachary McDade from the Denver Department of Public Safety; Eva Pereira and Preston Mills from the Los Angeles Mayor's Office of Budget and Innovation; Cass Wilkinson Saldaña from Children's Hospital of Philadelphia; Gbenga Ajilore, Robert Dinterman, Kasey Martin, Pat Patterson, and Tim Wu from the US Department of Agriculture Rural Development; Logan Thomas, Regina Van Horne, and Keith Witt from the Appalachian Regional Commission; and Erica Meade from the US Department of Health and Human Services—for their contributions and feedback on this project.

For more information on this project, see apps.urban.org/features/equity-data-tool/.

URBAN INSTITUTE

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org