URBAN INSTITUTE · ELEVATE · THE · DEBATE

Spatial Equity Data Tool

Technical Appendix

Ajjit Narayanan, Alena Stern, and Graham MacDonald
September 2020

This document describes the methods and data used by the Spatial Equity Data Tool to evaluate the demographic and geographic representativeness of datasets. At a high level, the tool geocodes the uploaded data to a census place and compares the distribution of the data within the city to the distribution of a set of baseline measures from the five-year American Community Survey. It then visualizes these differences at the census tract level on a map and at the city 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 where the latitude, longitude, 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
- NULLs
- 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 (numeric, text, and date) that are applied in the following order: numeric filters, text 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 where 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 City

To minimize the burden on users, the tool determines the city automatically from the data provided. To do this, the tool spatially joins the points in the data to a precompiled dataset of all 831 US cities with populations above 50,000 in 2018. If the points in the data fall within the boundaries of multiple cities, the tool only keeps the points within the most frequently occurring city in the data. Our definition of a city is all census tracts contained within the census place boundary. In some cases, our definition may be larger than the official city boundary; see the "Defining a City" section (page 7) for more details.

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

Once a city has been identified, the tool reads in all the precompiled geographic and demographic data for that city. These data include

1. the boundaries for all census tracts in the city, and
2. the ACS demographic variables for every tract in the city.

For more information on the specific ACS demographic variables used in the tool, see the "Data" section (page 6).

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 city. Any points that do not fall within any census tract—that is, points located outside the main city boundary—are discarded and not included in future calculations.

7. Compute Spatial Disparity Scores

For every tract in the city, the tool calculates a spatial disparity score. The disparity score is the difference between the proportion of the user-uploaded data within that tract and the proportion of the city's baseline population within that tract. We compute this score for each of five baseline populations:

1. total population
2. low-income population
3. population without Internet access
4. cost-burdened renter population
5. senior population

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

For the baseline variables, the proportions are calculated in terms of the total city. For example, the proportion of low-income people in a tract is the number of low-income residents in that tract divided by the total number of low-income residents in all tracts within that city.

8. Compute Demographic Disparities

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 city (the citywide percentage). Take a simple example city with two census tracts, each 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 citywide percentage of Black residents is (0.5)(0.2) + (0.5)(0.4) = 0.3. The citywide percentage answers the question, "What is the share of Black residents in the entire city?"

Now imagine 80 percent of the data uploaded by the user is associated with tract 1 and 20 percent is associated with tract 2. Then the data-implied percentage of Black residents would be (0.8)(0.2)+ (0.2)(0.4) = 0.24. The data-implied percentage of Black residents answers the question: "What is the share of Black residents in the average tract from which the data originate?"

Finally, the demographic disparity score is the difference between the data-implied percentage and the citywide 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 city 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 instead of the total population for the citywide 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 (page 6).

9. Compute Statistical Significance

Both the geographic and demographic disparity scores above rely on ACS tract-level 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 demographic disparity scores and determine if the scores differ significantly from 0. To calculate the margins of error, we use the Census Bureau–provided 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 gray in our tool to differentiate them from significant scores. For more detailed information on our statistical significance calculations, see page 9.

Limitations

The methodology behind this tool has 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 a couple of different baseline measures as laid out on page 6. Though these baseline measures should be a sensible starting point for many municipal datasets (particularly datasets that concern the specific baseline populations), there might be some datasets for which none of our baseline measure are suitable.
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.
3. This tool works best on medium to large datasets with at least a few hundred data points. This is because the geographic unit of analysis we use is the census tract, and there are typically hundreds of census tracts in a city. With a small number of points, the disparity scores generated by our tool are less reliable, 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 containing at least a few hundred rows.
4. This tool currently supports assessing disparities in a single city. If a dataset spans multiple cities, the tool will only operate on the most frequent city in the data and remove the remainder of the observations from the dataset. This is particularly problematic for regional or county-level analyses that span multiple cities.
5. Finally, as noted under "Defining a City," 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–level data from the 2014–18 ACS. Below are the ACS variables used to calculate the geographic and demographic disparity scores. We also note in parentheses the universe (i.e., denominators) used for each variable of interest.

Geographic Disparity Scores

The tool uses five ACS baseline variables to calculate geographic disparity scores:

1. total population: percentage of the city's total population living in each tract (universe: total city population)
2. low-income population: percentage of the city's residents with incomes below the poverty level living in each tract (universe: total number of low-income residents in the city)
3. senior population: percentage of the city's population ages 65 and older living in each tract (universe: total number of seniors in the city)
4. population of cost-burdened renters: percentage of the city's renter households who pay more than 35 percent of their income on rent living in each tract (universe: total number of cost-burdened renter households in the city)
5. population without Internet access: percentage of the city's households that do not have Internet access (universe: total number of households without Internet access in the city)

Demographic Disparity Scores

The tool computes a demographic disparity score for each demographic variable below:

- non-Hispanic white residents (as a share of total population)
- non-Hispanic Black residents (as a share of total population)
- non-Hispanic Asian residents (as a share of total population)
- Hispanic residents (as a share of total population).
- residents of all other races or ethnicities (as a share of total population)
- senior residents (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 over 18 years old)
- uninsured residents (as a share of civilian non-institutionalized population)
- residents with a disability (as a share of civilian non-institutionalized population)
- renters (as a share of occupied housing units)

- cost-burdened renter households (as a share of renter-occupied households paying rent where 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)
- low-income residents (as a share of population for whom poverty status is determined)
- extremely low–income residents (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)

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 used by the American Community Survey.

Defining a City

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 is 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/medium sized cities and a handful of irregularly shaped large cities, are bigger than they actually are, both in area and population. We decided on this 1 percent overinclusive cutoff for two reasons:

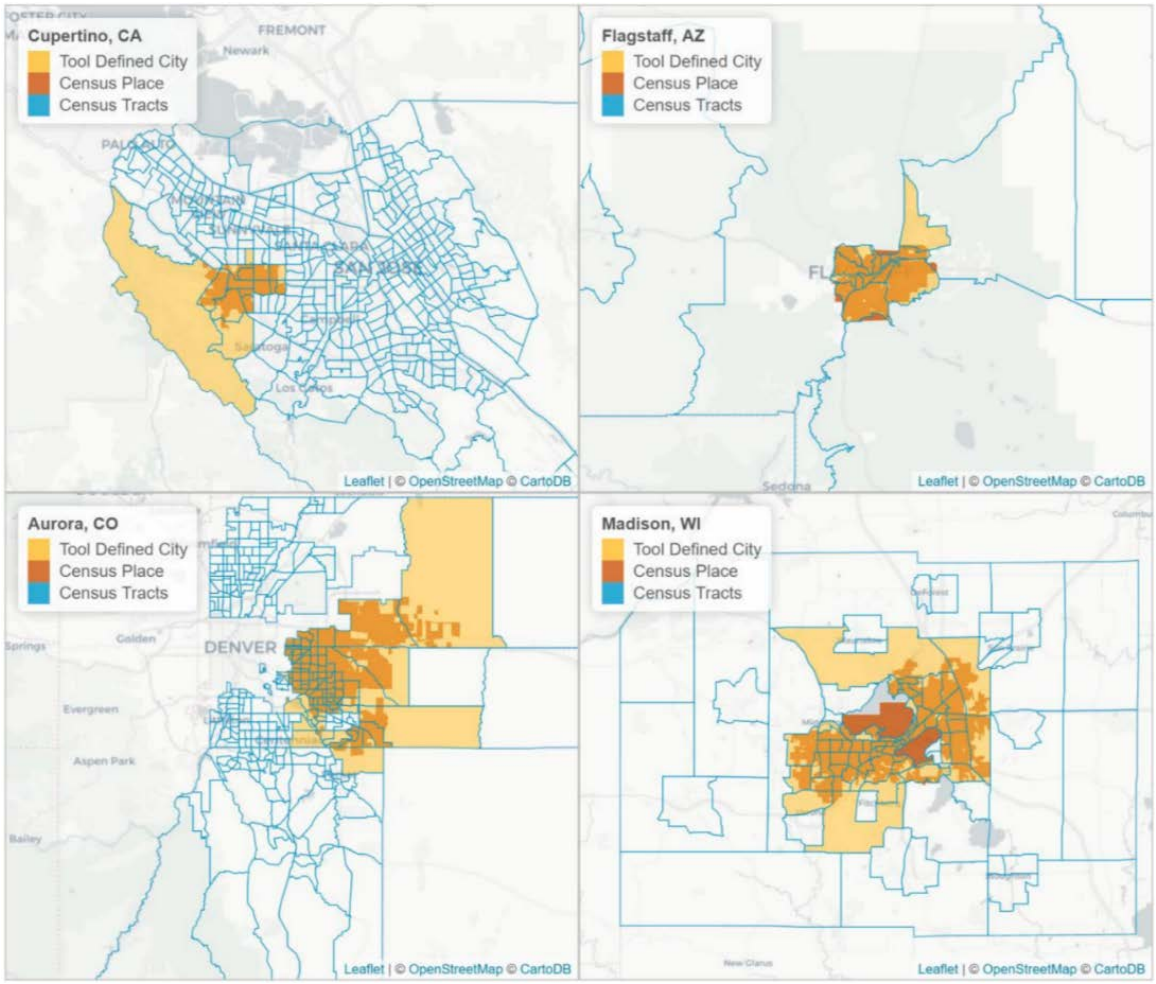
1. After visual inspection of a few cities, the 1 percent cutoff gave reasonable results for city boundaries.

2. The 1 percent cutoff allowed us to exclude tracts that were on the border of census places. If we imposed no cutoff (i.e., we defined a city as all tracts that were contained even partially within the place boundary), then the city definitions included tracts that shared a small portion of the border, and the cities were much larger than expected.

To visualize how this definition affects small and medium sized cities, consider figure 1. We can see that small cities like 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, CA; Flagstaff, AZ; Aurora, CO; and Madison, WI



Source: Urban Institute analysis of Census 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.

Statistical Significance Calculations

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

$$\mathrm{MOE}(\widehat{X}_1 + \widehat{X}_2 + \cdots + \widehat{X}_n) = \pm\sqrt{[\mathrm{MOE}(\widehat{X}_1)]^2 + [\mathrm{MOE}(\widehat{X}_2)]^2 + \cdots + [\mathrm{MOE}(\widehat{X}_n)]^2} \quad (1)$$

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

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

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 (SE), 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 census tract and the proportion of a city's total baseline population (total population, low-income population, etc.) residing within a census tract. The proportion of the data originating from a census tract (the data proportion) is a fixed number, but the proportion of a city's total baseline population residing in a tract (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 tract) and the denominator (the total baseline population in the city). 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 tract's disparity score as insignificant.

As a simple example, imagine a city with two census tracts, A and B. Census tract A contains 20 people and 10 data points. Census tract B contains 25 people and 10 data points. The spatial disparity score for tract A would be (10/20) - (20/45) = +0.0556. But is this overrepresentation significant?

Now assume that the census-provided margins of errors for the population is 10 people in tract A and 5 people in tract B. Using formula 1, the standard error of the total population in the city is $\frac{\sqrt{(5^2+10^2)}}{1.645}$, or 6.8, and the standard error of the population in tract A is $\frac{10}{1.645}$, or 6.08. Using formula 6, the standard error of the population proportion in tract A is

$$SE(population\ proportion) = \frac{MOE(population\ proportion)}{1.645} = \frac{\sqrt{6.08^2 - \left(\frac{20^2}{45} * 6.8^2\right)}}{45(1.645)} = 0.071$$

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 tract A would be

$$Confidence\ interval\ for\ population\ proportion = \frac{20}{45} \pm 1.96(0.071) = [0.31, 0.58]$$

To calculate the corresponding confidence interval for the spatial disparity score, we subtract the bounds from the data proportion, leaving us with [-0.08 0.19]. Note that this interval contains 0, meaning we do not have enough evidence to conclude the disparity score differs from 0. Therefore, we mark this tract's geographic disparity score as statistically insignificant.

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 city (the citywide percentage). Below is a mathematical and pictorial representation of the demographic disparity score for the percentage of white residents:

Data-implied percentage Citywide percentage

For the data-implied percentage, the share of white residents in each tract is an ACS variable with associated margin of error. Similarly, the citywide percentage uses the total city population and the

total white population in the city, both of which are ACS variables with associated margins of error. To visualize this, you can rewrite the citywide percentage in the diagram above as follows:

$$Citywide\ Percentage = \frac{\sum number\ of\ white\ residents_i}{\sum total\ population_i}$$

To take into account the variability in all these ACS figures, we calculate standard errors for the data-implied percentage and the citywide 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 following formula 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 citywide percentage, and Z_{CL} is 1.96 (since we set a 95 percent confidence level).

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

In the next section, we work through an example of the significance calculations for one of our demographic variables: the percentage of white residents:

Imagine a city 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. The citywide percentage would be (15+5)/(20+25), or 0.444. The demographic disparity score for white residents is 0.365 - 0.444 = -0.079.

Now assume that the margin of error (MOE) 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:

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

For the standard error of the citywide percentage, assume that the margin of errors 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 standard error for the total city population would be $\frac{\sqrt{5.5^2+5^2}}{1.645}$, or 4.51, and the standard error for the total white resident population would be $\frac{\sqrt{3.5^2+1.5^2}}{1.645}$, or 2.32. Then using formula 6, the standard error calculation for the citywide percentage of white residents would be

$$SE(citywide\ percentage) = \frac{MOE(citywide\ percentage)}{1.645} = \frac{\sqrt{2.32^2 - \left(\frac{20^2}{45} * 4.51^2\right)}}{45(1.645)} = 0.016$$

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

$$Z_{CL} = \frac{0.444 - 0.365}{\sqrt{0.021^2 + 0.016^2}} = 2.99$$

This is larger than the critical value of 1.96, so we would report the demographic disparity of white residents as -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 funded in collaboration with the Mastercard Center for Inclusive Growth. 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, and Cass Wilkinson Saldaña from Children's Hospital of Philadelphia—for their contributions and feedback on this project.

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.

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

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org

ABOUT THE URBAN INSTITUTE

The nonprofit Urban Institute is a leading research organization dedicated to developing evidence-based insights that improve people's lives and strengthen communities. For 50 years, Urban has been the trusted source for rigorous analysis of complex social and economic issues; strategic advice to policymakers, philanthropists, and practitioners; and new, promising ideas that expand opportunities for all. Our work inspires effective decisions that advance fairness and enhance the well-being of people and places.