Spatial Equity Data Tool Frequently Asked Questions

Last updated November 18, 2021.

General Questions

Who should use this tool?

Anyone interested in understanding the representativeness of data and/or programs in the US could use this tool. We think it has two main uses.

First, the Spatial Equity Data Tool can identify whether a given dataset is representative of the target population. It's important to check data are representative before using them for analysis or decisionmaking. For example, a government official interested in using 311 request data to target public works spending could use this tool to learn whether any neighborhoods or groups are underrepresented in the 311 data.

Second, the tool can identify whether place-based interventions are equitably distributed. Such interventions include any program or service that can be tied to a physical location: parks, bike share stations, Wi-Fi hotspots, electric vehicle charging stations, food distribution sites—and many more! Government officials can use this tool to examine whether a planned intervention equitably reaches affected neighborhoods and groups. Community organizations and residents can use this tool to advocate for more equitable distribution of resources. Nonprofits can use this tool to target their programs to areas underserved by other programs.

Where do the baseline and demographic data come from?

The data come from the American Community Survey five-year tract-, county-, state-, and national-level estimates for 2015–19. We chose these demographic variables and baseline populations because we believe they are common variables and target populations that municipalities and service providers are interested in, based on conversations with key stakeholders and early beta testers. Please see the technical appendix for more information.

Where can I get more help using this tool?

Information about the data and methods used in our tool is available in the technical appendix, and the code can be found on GitHub. For other questions, please email anarayanan@urban.org.

How has the tool changed over time?

The Spatial Equity Data Tool version 1 was released on September 24, 2020. The version 1 tool conducted analysis only at the city level. In response to user feedback, the version 2 tool was launched on November 18, 2021, introducing functionality to conduct analyses at the county, state, and national levels.

The version 2 tool also introduced the ability to select a baseline population for the demographic disparity chart. In version 1, the demographic disparity analysis always used total population as the baseline. In version 2, users can select the total population, the population with extremely low incomes, or the child (< 18) population. For example, if a user were to select the child population as the baseline, the value for the "Asian residents" point in the chart should be interpreted as the difference between the proportion of children in the neighborhoods that the data come from who are Asian and the proportion of all children in the geography who are Asian. See the "How does changing the baseline dataset affect my results?" FAQ for more information on changing baselines.

We also added two new baseline datasets to the geographic disparity map: residents with extremely low incomes and child (< 18) residents. For more information, see the "Why are there different baselines for the geographic disparity map and demographic disparity chart?" FAQ.

The version 2 tool also uses updated data from the 2015–19 five-year ACS, instead of the 2014–18 five-year ACS data the version 1 tool used.

Finally, the version 2 tool changed the order in which the geographic disparity map and the demographic disparity chart appear on the page. In version 1, the geographic disparity map appears first. In version 2, the demographic disparity chart appears first. We made this change in response to user feedback.

Choosing a Geography

Which geographies can I choose from? And why did you select these options?

You can use the Spatial Equity Data Tool to analyze city-level, county-level, state-level, and national-level data. The original version 1 tool only included city-level analysis. County-, state-, and national-level analysis was introduced in version 2 of the tool in November 2021. We chose to add more geographies because they were the most frequently requested by users. We recognize that the tool doesn't cover all use cases, such as regional and metropolitan-area analyses, and hope to introduce additional geographic functionality in future updates.

How do I know which geography is right for my data?

We recommend determining which geography you should use on the basis of the geography your data cover. For example, if your dataset covers the entire US, we recommend using national-level analysis. And if your dataset covers a single county, we recommend using the county-level analysis. To figure out the geography your data cover, we recommend reading the documentation associated with your data and considering information such as the data-collecting agency and target population.

If your dataset doesn't fit cleanly into one of our four geographies, we recommend using a geography covered entirely by your data. For example, if your data cover a region of several states, we would recommend using the state-, county- or city-level analysis. We would not recommend using national-level analysis, as your data will be compared against the entire US, so results may not be accurate.

By default, the tool will select the most frequently occurring geography in your data. So, if you uploaded regional (multistate) data and select state-level analysis, the tool will by default run the analysis for the most frequently occurring state in your data. You can manually select a different geography by using the advanced filter feature (if your dataset has a column corresponding to state) or filter your data to a specific state before uploading.

Can I use the tool for data from US territories?

We do not include the US territories in the tool, as the American Community Survey does not collect data in American Samoa, Guam, the Northern Mariana Islands, and the US Virgin Islands and does not report all of the data used in the tool for Puerto Rico. We hope in future iterations of the tool to include US territories.

What geography should I choose for data from Washington, DC?

We recommend analyzing data for Washington, DC, using city- or county-level analysis. While it is possible to analyze Washington, DC, data using state-level analysis, because the state and county boundaries for Washington, DC are equivalent, the geographic disparity map will only show a single value for Washington, DC.

Using Sample Data

How do I use the sample data?

We recommend that new users start by using one of our sample datasets, so you understand the functionality of the tool before uploading your own data. We selected sample datasets for each geographic level that we feel represent the kinds of data users are likely to upload and that demonstrate the tool's functionality.

Taking the city-level sample datasets as examples, the New York City Wi-Fi hotspot sample dataset demonstrates how changing the baseline dataset (from the default of total population to the population without internet access) enables users to evaluate the distribution of a resource relative to a specific target user.

The New Orleans 311 dataset shows how using the filter functionality can help users focus on a specific subset of data—in this case, requests logged between January 1, 2014, and January 1, 2019. Finally, the Minneapolis bike share station data illustrate how using weights affects the results. By weighting the data by the number of bikes available, our analysis will capture that each station serves a different number of people.

For the county, state, and national-level analyses, we have selected the following sample datasets and preset advanced options:

- National
  - Electric vehicle charging stations
  - Public libraries (weighted by the public service hours per year, or the `*HOURS*` variable)
- State
  - Substance use and mental health facilities in Washington (filter `State` variable to "WA")
  - Low-Income Housing Tax Credit projects in Alabama (weighted by number of low-income units, or the `*LI_UNITS*` variable)

- County
  - Playgrounds in Miami-Dade County, FL (filter `*TOTLOT*` variable to "Yes" to subset to playgrounds)
  - Polling places in Bucks County, PA

For more information on these particular sample datasets and how we compiled them, please see our Urban Institute Data Catalog entry.

Uploading Your Own Data

What data can I use with the Spatial Equity Tool?

You can use any CSV file with geographic point location data as long as it satisfies the following requirements:

- The file must have column headers in the first row.
- Two columns must correspond to longitude and latitude (in the EPSG:4326 or WGS84 coordinate reference system).
- The data file must be smaller than 2 GB.
- The geographic point locations must be from the US (50 states plus the District of Columbia).
- The file should use UTF-8, UTF-16, or ISO-8859-1 (i.e., Latin-1) encoding. For help saving your CSV with UTF-8 encoding, please see this web page.

If you have point data in a shapefile (.shp), you can convert that file to a CSV using QGIS or this simple R script.

How does the tool treat null values?

Null values in the longitude/latitude columns, the weight column, or any of the selected filter columns will cause the tool to discard that row. Our tool uses the Pandas default CSV reader, which treats the following values as NA:

- ` ` (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`

I have a dataset of polygons (e.g., census blocks). How can I use it with this tool?

You need to assign a geographic (longitude/latitude) point to each polygon to use that dataset with this tool. We recommend doing this only with polygons that map cleanly to census tracts—namely tracts, block groups, and blocks.

My dataset only has addresses, not longitude and latitude. What do I do?

You need to geocode the addresses by assigning each one a longitude-latitude point to use this tool. You can find more information about available geocoders and factors to consider when selecting a geocoder here.

My file is larger than 2 GB. What do I do?

First, try getting rid of unnecessary columns. The only columns the tool needs are your longitude/latitude columns and any columns you are using for filters and weights. If your file size is still over 2 GB, we recommend taking a random sample of your data and uploading that to the tool.

Where can I find data to use with the tool?

For city- and county-level datasets, a great place to start is municipal open data portals. All of our sample city-level and county-level datasets come from such portals. The US City Open Data Census, created by the Open Knowledge Foundation and the Sunlight Foundation, gives an overview of the numerous city- and county-level datasets available. Data.gov, a central repository for the US government's open data, also maintains a list of state, city, and county open data sites. These state open data portals are a great place to look for state-level data. If you are interested in a specific city, county, or state not listed in these centralized locations, plugging"[geography name] open data portal" into a search engine is a good next step.

For national-level data, a great place to start is Data.gov. If you are interested in data from a specific federal agency, we recommend searching their website. Several sample datasets come from such websites, including electric vehicle charging stations (US Department of Energy), public libraries (Institute of Museum and Library Services), Low-Income Housing Tax Credit projects (US Department of Housing and Urban Development), and substance use and mental health facilities (the Substance Abuse and Mental Health Services Administration within the US Department of Health and Human Services). Note that we feature the Low-Income Housing Tax Credit projects and substance use and mental health facilities data at the state level, illustrating how nationwide data can be filtered to examine results for a specific state.

Can I upload confidential or private data to this tool?

Per our terms of use, you should not upload confidential, private, and/or sensitive data to this tool. All user-uploaded data and results are stored in publicly accessible cloud storage. While it is unlikely, another user or a bad actor could access and download your uploaded data.

The Urban team is working on a set of tools to make it easier for confidential data users to run spatial equity analysis in a secure environment. If you have confidential data you would like to run through our tool now, please reach out to anarayanan@urban.org.

I received a warning that my data are only located in a few states (for national-level analysis) or counties (for state-level analysis). What does that mean?

This warning indicates that your data points fall within less than 50 percent of the states in the US (for national-level analysis) or less than 50 percent of the counties in the identified state (for state-level analysis). For both the geographic and the demographic disparity analyses, we compare the distribution of your data against the baseline population distribution in the full geography. For example, if your data only pertain to the northeast region of the United States but you selected national-level analysis, the geographic disparity map would likely show all of the northeast states as significantly overrepresented, because the share of data points in each state would be compared against that state's share of the entire US population (instead of its share of the northeast region population, which would be more appropriate).

If you get this warning, ask yourself whether your dataset is intended to represent the entire geography in question. If the answer is no, we would recommend selecting a smaller geographic level of analysis, which will by default run the analysis for the most frequently occurring geography in your data.

Using the Advanced Options

How do I use the filters on my data?

You can use the filters to analyze a subset of the points in the dataset that you upload to the tool. The tool allows the following filter types:

- **Text:** Filter by text values in a selected column that are equal to or not equal to one or more values. Multiple values can be entered separated by commas. In such a case, the values will be evaluated as "or" conditions for which rows will be kept if the selected column equals or doesn't equal any selected values.
- **Numeric:** Filter numeric columns by equal, not equal, less than, less than or equal to, greater than, or greater than or equal to a number. You can add multiple numeric filters to the same column and filter multiple columns.
- **Date:** Filter to keep rows with a particular date or date range.

If you set multiple filter conditions, the tool will only use rows that meet all conditions. For example, if you select two numeric filters and two date filters, only rows that match all four filtering conditions will be returned. This means that certain filtering operations, such as filtering to data in either of two date ranges, are not possible.

If you want to filter your data in a way that is not enabled by the tool (such as regular expressions or geographic filters), then you need to filter your data before uploading.

The tool detects whether your column is text, numeric, or date on the basis of the first 10 rows in your dataset. The tool will only recognize a column as a date column if it follows the ECMAScript date time string format (for example, YYYY-MM-DD). Columns with just years (for example, 2014) may be recognized as text columns. For help understanding how our tool recognizes column types, please see this page.

How does changing the weights affect my results?

The weights determine how each point is counted when measuring representativeness. If your dataset is bike share stations and you select to weight by number of bikes, then the tool will weight a bike share station with 10 bikes 10 times as much as a bike share station with 1 bike when constructing geographic and demographic disparity measures. If you do not select a weight, then each row (i.e., geographic point)

in the data is weighted equally. Rows with a weight of 0 are treated as null/NA values and discarded from the analysis. Bear in mind that weighting your data will affect both the geographic disparity score shown in the map and the demographic disparity scores shown on the chart.

Interpreting Your Results

I've run my analysis; what do the results mean?

While the data tool can tell you what geographic areas and groups are under- or overrepresented in your data, it cannot tell you why. Using the tool to help you identify disparities is a first step to understanding why they exist and how to address them. Here are some potential drivers of unrepresentativeness you could explore:

- **Data collection issues:** The design and implementation of data collection systems can yield unequal representation. For example, resident-generated datasets, such as 311 requests, may reflect higher usage of the system by some groups (PDF) of residents. Therefore, the data may not accurately represent the true need for city services. We encourage you to use the results of the tool to discuss how to improve data collection efforts among underrepresented groups and geographic areas.
- **Program implementation:** The program captured in the data may not have been designed for the equity objective our tool is assessing. Some cities put public Wi-Fi hotspots in government buildings or downtown commercial centers to cater to the tourist and business populations. As a result, Wi-Fi hotspots would be overrepresented in commercial neighborhoods but underrepresented in less-central residential neighborhoods. In this case, you may decide that a subset of the data is more relevant to equity and use the tool's filter function to examine the hotspots not located in government buildings. We encourage you to use the results of this tool to discuss how the design or implementation of a place-based program could yield more equitable results.
- **Historical inequities:** Data reflect the biases of the systems that generate them. For example, police arrest data are often concentrated in low-income communities of color because of previous policy decisions to overpolice these communities. These biased data are often fed into predictive policing algorithms, which, in turn, send even more police officers into these neighborhoods, generating even more disproportionate arrest records. We encourage you to use the results of this tool to discuss how historical inequities inform current policies and data.
- **Mismatched baseline datasets:** While our tool offers several baseline datasets, it may not offer the baseline that best represents the most equitable distribution of your data. For example, when analyzing disparities in pothole-repair requests, the correct baseline dataset to compare against might be a dataset on traffic flow or some other measurement of likelihood of potholes. Unfortunately, traffic flow is not one of the datasets available in our tool. We encourage you to select the baseline dataset that most closely represents the ideal distribution of your data, but we recognize that in some cases our available baseline may still be ineffective.
- **Mismatched geographies:** For both the geographic and demographic disparity analysis, we compare the distribution of your data against the baseline population distribution in the full geography. If your dataset is not intended to represent that entire geography (e.g., if your data only cover a single region of the US but you use national-level analysis), the results may not accurately reflect disparities in your data. See the "I received a warning that my data are only located in a few states (for national-level analysis) or counties (for state-level analysis). What does that mean?" FAQ for more detail.

Why did my analysis return very few or no rows?

Check the "SUMMARY" section at the top of the tool. Your results may have fewer rows than expected for a few reasons:

- Your data contained many null values for the longitude/latitude columns, weight columns, or filter columns.
- Your selected filter conditions are overly restrictive.
- A large portion of your data falls outside the main geography identified in the data.

My dataset has multiple geographies; why do the results show only one? (Or, my data are for a region; why am I only seeing the central city/county/state?)

The tool can analyze data for one geography at a time at the selected geographic level (e.g., one city, one county, one state, or for the entire US). By default, the tool shows the data for the most frequently occurring geography in the data. If you want to see the results for another geography, use the filter functionality. Also, bear in mind that our tool does not cover US territories, and the city-level analysis only works on cities with more than 50,000 residents.

The boundaries shown in the city-level tool differ from the official city boundaries. What could be causing this?

Our 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 tracts and census places don't overlap perfectly, so parts of some tracts fall outside the place boundary. Because of our overinclusive definition, the tool thinks that many cities—particularly small and medium-sized ones—are bigger than they actually are, in both geographic size and population. Please see the technical appendix for more information.

Why are there different baselines for the geographic disparity map and the demographic disparity chart?

In short, we have different baseline populations because of different data availability from the American Community Survey. *To include a baseline population for the demographic disparity scores, we require detailed cross-tabulations of ACS variables, which are not available for some baseline populations.* For example, to include children (< 18) as a baseline dataset, we used ACS estimates of Asian, Black, Latinx, white, and other racial/ethnic child populations at the tract level. The ACS does not include such demographic cross-tabulations for all of the different baseline populations that we include for the geographic disparity chart. Accordingly, only a subset of those baselines—total population, child (< 18) population, and population with extremely low incomes—are included for the demographic disparity chart.

How does changing the baseline dataset affect my results?

The baseline dataset represents the "ideal" distribution of your data; If the baseline dataset is total population, then your data are perfectly representative when the proportion of points in your dataset matches the proportion of the total population in each tract. When you change the baseline dataset to, say, the population with low incomes, the proportion of your data points in a tract is compared with the proportion of the geography's population with low-incomes living in that tract. Select the baseline dataset that you think best represents the ideal distribution of your data.

Users can change the baseline dataset for both the geographic disparity map and the demographic disparity chart; however, changing the baseline dataset for one *does not* automatically change the

baseline dataset for the other. Moreover, each analysis has different baseline datasets available. For information on why that is the case see the "Why are there different baselines for the geographic disparity map and the demographic disparity chart?" FAQ above.

How is the geographic disparity score calculated?

The geographic disparity score is the percentage-point difference between the share of your dataset falling within a particular subgeography and the share of the baseline dataset (e.g., the total population) in that subgeography. The geographic subareas used for calculating this score differ based on the geographic level of analysis:

Overall geography	Subgeography for geographic disparity
US	State
State	County
County	Tract
City	Tract

We chose to use different subgeographies for each level of analysis based on feedback from users on the level of analysis of geographic bias that would be most meaningful and usable for each overall geography.

For example, if you are using the state-level tool, and a county accounts for 10 percent of your dataset and 20 percent of the overall state's population, that county's disparity score is 10% − 20% = − 10%. This measure is calculated for every subgeography in the geography (e.g., for every county in the state) and for every baseline dataset to give users a sense of which parts of the city are under- or overrepresented.

How is the demographic disparity score calculated?

The short version: We first estimate the demographic groups your data depict by taking a weighted average of the demographics across all the tracts your data come from. We compare that average with the overall demographics of the selected baseline population and report the difference as the demographic disparity score.

The long version: The demographic disparity score is the percentage point difference between the representation of a demographic group in the data (*the data-implied average percentage*) and the representation of a demographic group in the city, county, state, or nation (*the geography-wide average percentage*). Importantly, the demographic disparity score is always calculated with tract-level data, regardless of the geography.

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 Latinx and Tract 2 is 40 percent Latinx, then the geography-wide average percentage of Latinx residents is (0.5)(0.2) + (0.5)(0.4) = 0.3. The geography-wide average percentage answers the question, "What is the share of Latinx residents in an average tract of the city?" Note that the geography-wide percentage for the county-, state-, and national-level analysis is either directly reported by the ACS or can be calculated from directly reported counts for the numerator (Latinx population) and denominator (total population).

Now imagine 80 percent of the points in your dataset are associated with Tract 1 and 20 percent are associated with Tract 2 while the Latinx proportion of the population remains the same as above. Then the data-implied average percentage of Latinx residents would be (0.8)(0.2) + (0.2)(0.4) = 0.24. The data-implied average percentage of Latinx residents answers the question, "What is the share of Latinx residents in the average tract from which the data originate?"

Finally, the *demographic* disparity score is the difference between the two percentages, or 0.24 − 0.3 = −0.06. In this example, Latinx residents seem to be underrepresented by 6 percentage points. We essentially repeat this calculation for all our demographic variables of interest using all census tracts in the city, county, state, or nation. For more information on the data used or the limitations of this methodology, please see the technical appendix.

Why do I see multiple dots for each variable on the demographic disparity chart for national-level and state-level analysis?

In version 2 of the tool, the demographic disparity chart shows in grey the demographic disparity scores for each state in the US for national-level analysis and each county in the state for state-level analysis. These scores should be interpreted as the difference between the percentage of the demographic group in the neighborhoods that the data in *the given state or county* come from and the percentage of the demographic group in the *given state or county* overall.

This information can help users better understand the distribution of the disparity within their geography of interest (e.g., whether all counties in the focus state have similar rates of disparity for Latinx residents), identify an area that may be driving the overall disparity score (e.g., a state in the US with a particularly high disparity for a given group), and identify areas that are more equitably allocating resources (where the disparity is closer to 0). Users can then use the state-level or county-level analysis to further examine results in states or counties of interest.

How is statistical significance calculated?

Census-reported figures for tract-level population and demographic statistics are estimates and subject to sampling error. We use the census-reported margins of error for these estimates to calculate 95 percent confidence intervals for the geographic and demographic disparity scores. If 0 does not fall within this confidence interval, then we report this bias as statistically significant. In other words, after taking into account the variability in the census-reported estimates, the data still significantly over- or underrepresent certain groups. For more detail on our statistical significance calculations, please see the technical appendix.

How do I export my results?

You can export either the data displayed, or an image of the demographic disparity charts and the geographic disparity maps by clicking the export data button located at the bottom of the chart or the map.

Terms of Use

By using our tool, you agree to the following terms of use:

- **General terms:** Each time you use or cause access to this website (a) you acknowledge that you have read, understand, and agree to our Terms of Use; (b) you acknowledge that you allow us to use your data and/or content for tuning, research, and diagnostic purposes of our services; (c) you acknowledge that your data and/or content may be deleted from our servers at any time, at our discretion; (d) you acknowledge that submitting information to this site is your option and you do so at your own risk; (e) you explicitly agree to not provide any confidential data, including personally identifiable health data as defined by the HIPAA Privacy Rule.
- **Limitation of liability:** The Urban Institute will not be liable for any damages of any kind arising out of or relating to the use of your data. The Urban Institute shall not have any liability or responsibility for your acts, omissions, or conduct or the conduct of any user or other third party.
- **Indemnity:** You agree to indemnify and hold harmless the Urban Institute and its Board members, directors, officers, employees, agents, and contractors from and against any and all claims, damages, losses, costs (including without limitation reasonable attorneys' fees), or other expenses that arise directly or indirectly out of or from (a) your breach of any provision of our Terms of Service; (b) your activities in connection with the website; or (c) unsolicited information you provide to the Urban Institute through the website.

This feature 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-, state-, and county-level analysis. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.