Spatial Equity Data Tool: Frequently Asked Questions

Last updated September 23, 2020

General Questions

Who should use this tool?

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

First, it 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 decision-making. For example, a city official interested in using 311 request data to target public works spending should use this tool to learn whether any neighborhoods or groups are underrepresented in the 311 data.

Second, it can identify whether place-based interventions are equitably distributed. That includes any program or service that can be tied to a physical location: parks, bike share stations, wi-fi hotspots, digital literacy trainings, food distribution sites—and many more! City 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 tract-level American Community Survey five-year estimates for 2014–18. 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, 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 anaraynan@urban.org.

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 three sample datasets that we feel represent what users are likely to upload and that demonstrate different portions of the tool's functionality. 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 illustrates how using weights affects the results. By weighting the data by the number of bikes available, we can capture that each station serves a different number of people in our analysis. For more information on these particular sample datasets and how we compiled them, please see our Urban 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 WGS 84 coordinate reference system)
- The data file must be smaller than 2 GB.
- The geographic point locations must be from a city in the US with population above 50,000.
- The file should use UTF-8, UTF-16, or ISO- 8859-1 (i.e., Latin1) 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.

How does the tool treat null values?

Null values in either the latitude/longitude columns, the weight column, or any of the selected filter columns will cause that row to be discarded by the tool. 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 (latitude/longitude) 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 latitude and longitude. What do I do?

You need to geocode the addresses by assigning each one a latitude-longitude point to use that dataset with 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 latitude/longitude 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?

A great place to start is municipal open data portals. All three of our sample datasets come from such portals. The U.S. City Open Data Census, created by the Open Knowledge Foundation and Sunlight Foundation, is a great place to get an overview of the numerous datasets available.

Can I upload confidential/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. If you have confidential data you would like to run through our tool, please reach out to anarayanan@urban.org.

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:

- **Numeric:** Filter numeric columns by less than, less than or equal to, equal to, greater than or equal to, or greater than a number. You can add multiple numeric filters on the same column, and filter multiple columns.
- **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 which case they will be

evaluated as "or" conditions where rows will be kept if the selected column equals or doesn't equal any selected values.
- **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, like filtering to data that is 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 numeric, text, or, date based on 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 was bike share stations and you select to weight by number of bikes, then a bike share station with 10 bikes would be weighted 10 times as much a bike share station with one bike when we construct the geographic 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 only affects the geographic disparity score shown in the map; it does not affect 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 neighborhoods 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 unrepresented groups and neighborhoods.

- **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 population. As a result, wi-fi hotspots would be very 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-request data, the correct baseline dataset to compare against might be a dataset on traffic flow or some other measurement of likelihood of potholes. Unfortunately, that 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.

Why did my analysis return very few or no rows?

Check the "Summary: Your data on..." 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 latitude/longitude columns, weight columns, or filter columns.
- Your selected filter conditions are overly restrictive.
- A large portion of your data falls outside the main city identified in the data.

My dataset has multiple cities; why do the results show only one? (Or, my data are for a metropolitan area; why am I only seeing the central city?)

The tool can analyze data at the city level for one city at a time. By default, the tool shows the data for the most frequently occurring cityin the data. If you want to see the results for another city, use the filter functionality. Also, bear in mind that our tool only works on cities with more than 50,000 residents.

The city boundaries shown in the 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 places and census tracts 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.

How does changing the baseline dataset affect my results?

The baseline dataset represents the "ideal" distribution that your data are compared to. If the baseline dataset is total population, then your data are perfectly representative if 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, low-income population), the proportion of your data points in a certain tract is compared with the proportion of the city's low-income population in that tract. Select the baseline dataset that you think best represents the ideal distribution of your data.

Changing the baseline dataset *only* affects the geographic disparity scores shown in the map. It does *not* affect the demographic disparities displayed in the chart because that section always compares your data with the demographics of the city's total population.

How is the disparity score calculated?

The disparity score is the percentage-point difference between the share of your dataset falling within a particular census tract and the share of the baseline dataset (e.g., the city's total population) in that tract. For example, if a census tract accounts for 10 percent of a dataset and 20 percent of a city's population, that tract's disparity score is 10 - 20 = - 10%. This measure is calculated for every tract in the city and for every baseline dataset to give users a sense of which parts of the city are under- or overrepresented.

How is demographic representativeness calculated?

The demographic representativeness 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 (*the citywide average 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 Hispanic and tract 2 is 40 percent Hispanic, then the citywide average percentage of Hispanic residents is (0.5)(0.2) + (0.5)(0.4) = 0.3. The citywide average percentage answers the question: "What is the share of Hispanic residents in an average tract of the city?"

Now imagine 80 percent of the points in your dataset are associated with tract 1 and 20 percent are associated with tract 2. Then the data-implied average percentage of Hispanic residents would be (0.8)(0.2)+ (0.2)(0.4) = 0.24. The data-implied average percentage of Hispanic residents answers the question: "What is the share of Hispanic residents in the average tract from which the data originate?" Finally, the *demographic reporting bias* is the difference between the two percentages, or 0.24 - 0.3 = -0.06. In this example, Hispanic 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 a city. For more information on the data used or the limitations of this methodology, please see the technical appendix.

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 representativeness and demographic representativeness. 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, see the technical appendix.

How do I export my results?

You can export the data displayed in the geographic disparity map and the demographic disparity charts by clicking the export data button at the bottom of the map.

You can export an image file of the demographic disparity charts by clicking the export charts button at the bottom of the chart.

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 Service; (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.