Toward an Open Data Bias Assessment Tool

Measuring Bias in Open Spatial Data

Ajjit Narayanan
Urban Institute

Graham MacDonald
Urban Institute

February 2019

The authors welcome feedback on this working paper. Please send all inquiries to gmacdonald@urban.org.

Thank you to Urban's Statistical Methods Group for their review of our statistical tests.

Abstract

Data is a critical resource for government decisionmaking, and in recent years, local governments, in a bid for transparency, community engagement, and innovation, have released many municipal datasets on publicly accessible open data portals. In recent years, advocates, reporters, and others have voiced concerns about the bias of algorithms used to guide public decisions and the data that power them.

Although significant progress is being made in developing tools for algorithmic bias and transparency, we could not find any standardized tools available for assessing bias in open data itself. In other words, how can policymakers, analysts, and advocates systematically measure the level of bias in the data that power city decisionmaking, whether an algorithm is used or not?

To fill this gap, we present a prototype of an automated bias assessment tool for geographic data. This new tool will allow city officials, concerned residents, and other stakeholders to quickly assess the bias and representativeness of their data. The tool allows users to upload a file with latitude and longitude coordinates and receive simple metrics of spatial and demographic bias across their city.

The tool is built on geographic and demographic data from the Census and assumes that the population distribution in a city represents the “ground truth” of the underlying distribution in the data uploaded. To provide an illustrative example of the tool’s use and output, we test our bias assessment on three datasets—bikeshare station locations, 311 service request locations, and Low Income Housing Tax Credit (LIHTC) building locations—across a few, hand-selected example cities.

Across the small sample of cities we studied, we consistently find that bikeshare stations are concentrated in downtown areas, overserve neighborhoods with high numbers of non-Hispanic white, non-Hispanic Asian, and college-educated residents, and underserve neighborhoods with large numbers of non-Hispanic Black, Hispanic, unemployed, and poor residents. The results from our analysis of bias in 311 service requests and LIHTC building location data are much more mixed across cities. Of particular note: 311 service requests from Boston and DC overrepresent white and college-educated residents while 311 service requests from Philadelphia overrepresent non-Hispanic Black and poorer neighborhoods. LIHTC location

data from Raleigh demonstrate that buildings tend to be located in neighborhoods with higher shares of black and poor residents and lower shares of white and college-educated residents relative to the city average, in contrast to the other cities we studied, which tended to have much smaller differences.

Introduction

Open data is defined as data that can be freely used, reused, and redistributed by anyone (Ubaldi 2013). Over the past few years, there has been a significant increase in the number of cities releasing open datasets to the public (Kitchin 2014). Researchers have observed a number of benefits to the growing availability of open data, namely increased government transparency and accountability, higher community engagement, and more opportunities for economic and civic innovation (Jansen et al. 2012). Open data is being used by governments, the private sector, and individual citizens to create predictive models, develop visual dashboards of city health, and inform public understanding of the people that use government services. However, as the open data movement expands, researchers and government are gaining a growing understanding of the risks to data-driven decisionmaking.

For example, a first-of-its-kind report written by the Future of Privacy Forum recently attempted to assess the risks of open data in the city of Seattle (City of Seattle 2018). In the report, the team identified two primary risks: re-identification of individual data and data equity. While the report identified tools for combatting re-identification risk, such as synthetic data-generating techniques, it also points out that the field has not yet developed any standardized tools or best practices for assessing equity or bias in open data.

The report identifies data equity and bias as an important issue because the data that city governments collect and publish are sometimes not representative of the ground truth. For example, policymakers would like arrest data to tell us about the true geographic distribution of crimes committed in a city. However, low-income and minority neighborhoods tend to be over-surveilled compared with majority-white neighborhoods; as a result, low-income people and people of color end up being disproportionately represented in arrest data (Piquero 2008). Similarly, citizen-generated data, like pothole service and snow removal requests, tend to be biased toward wealthier and more computer-literate residents (Feigenbaum 2015). A dataset about pothole service requests may not align with the actual location of potholes in a city if certain populations are more likely to report them. If left undiscovered, bias in open datasets could translate directly into biased policy decisions.

As municipalities turn to algorithmic decisionmaking and machine-learning models to support administrative and policy decisions, biased data may intensify existing disparities. Criminal sentencing algorithms, for example, have been shown to be racially biased because they depend on already biased input data, namely historic arrest and recidivism data (Angwin et al. 2016). This algorithmic bias can further reinforce the original bias in the data in a pernicious feedback loop, as people with arrest records are given longer sentences, find it harder to integrate back into society, are more likely to be targeted for patrols, and are again incarcerated (O'Neill 2016).

Recently, researchers have made inroads to understanding and measuring algorithmic bias in standardized ways; for example, the University of Chicago recently developed the Aequitas tool to help policymakers and analysts assess algorithmic bias. However, though we were able to find several one-off studies that quantified bias in open data in the academic literature, we were unable to locate tools that systematically measure and analyze bias within the underlying datasets. While we believe studying algorithmic decisionmaking is important, perhaps even more important is the possible impact of biased data on non-algorithmic decisionmaking—data-driven choices made by government staff. Even without the assistance of machine learning or complex models, individual city officials and policymakers may unknowingly use biased data to make and justify policy decisions.

The Future of Privacy Forum report came to a similar conclusion: that Seattle should "develop or obtain tools for evaluating the representativeness of the city's open data (including whether underserved or vulnerable populations are over or underrepresented in certain ways)." It is critical that city governments have the tools to quantify bias in any dataset, used algorithmically or not, so they can

- evaluate the accuracy and representativeness of the dataset;
- decide whether to release data on open data portals;
- improve data-collection processes, especially from under- or overrepresented populations; and
- increase awareness of the limitations of the data when using it for analysis.

With this paper, we hope to take the first step in answering this call by providing a prototype of an automated dataset bias assessment tool for geographic data.

Literature Review

Before beginning our literature review, we scanned the open data portals of four of the largest or fastest growing cities in the United States—New York City, Chicago, Austin, and Nashville—and found that many datasets tended to describe government services provided, such as fire or police activity, water supply, transportation, sanitation, libraries, and public parks. When we reviewed the literature, we found that there have been quite a few studies measuring the equity and accessibility of government services. For example, Nicholls (2001) used Census socioeconomic data and GIS network analysis to study the spatial and demographic distribution of public parks in Bryan, Texas. She found that while a large percentage of the city did not have access to any parks, on average, non-white, Black, and poor residents had better access to public parks than their white and higher-income neighbors. A similar study was conducted by Talen (1997) in Pueblo, CO, and Macon, GA. It found that nonwhite residents and residents with high housing values were less likely to be located near parks in Macon, while the reverse was true in Pueblo.

Many studies outside the US have also analyzed bias in government service delivery. Delmelle, Cahille, and Casas (2012), for example, analyzed the spatial distribution of the Bus Rapid Transit (BRT) system in Cali, Colombia, and found that while middle-income groups have adequate access to the BRT stations, low- and high-income groups tended to be excluded. A study of public playgrounds in Edmonton, Canada, by Smoyer-Tomic, Hewko, and Hodgson (2004) found that lower-income residents had the highest accessibility to playgrounds. However, this effect diminished when isolated to high-quality playgrounds. Overall, the studies we examined found that government service delivery data demonstrate patterns of inaccessibility for certain groups, but these effects differ among cities, affected groups, and the government service in question.

In our scan of city data portals, we also found that resident-generated data, such as pothole service and snow removal requests, were fairly common. In recent years, we found a few studies conducted on the “digital divide” and its effect on resident-generated data. The digital divide refers to disparities in Internet and technology access along demographic lines, and in the literature, we found a growing understanding that this applies to citizen contacts with e-government services, such as applications for licenses or 311 service requests. For example, Hall and Owens (2011) analyze Pew polling data on citizen interaction with e-government services and found that lower levels of income and education have significant and large negative effects on propensity to use e-government services. They also found moderate negative impacts for older citizens, Blacks, and Hispanics. Thomas and Streib (2003) find similar economic and demographic disparities in citizen contact with government services in Georgia. A host of other studies indicate similar results; in general, the digital divide falls along the lines of socioeconomic status, gender, race, language, and disability (Bélanger and Carter 2009; Zillien and Hargittai 2009; Riggins and Dewan 2005).

One such resident-generated dataset available in our scan of data portals that has received particular attention in the research literature is calls for 311 service requests. 311 is a municipal service that allows residents to call in nonemergency requests for city services like pothole and streetlight repair, or park and graffiti cleanup. Several recent studies have analyzed spatial patterns and socioeconomic disparities in 311 call data. One particularly comprehensive study by Cavallo, Lynch and Scull (2014) analyzed 311 data from New York, San Francisco, and Washington, DC, and used spatial regression techniques to predict the number of 311 calls in a census tract based on the socioeconomic characteristics of the tract. While effects varied across cities, the percentage of Black and Hispanic residents, percentage of households with children, and percentage of foreign-born residents in a tract were associated with fewer 311 requests. The percentage of Asian residents and higher mean income in a tract were associated with a higher number of 311 requests. Kontokosta, Hong, and Korsberg (2017) measured bias in 311 reporting propensities by comparing "ground truth" data in heating and water building violations in New York City to complaint volume in 311. They found that neighborhoods that tend to underreport to 311 have higher unemployment rates, larger nonwhite populations, a higher proportion of unmarried residents, and a larger number of limited English speakers. Neighborhoods that tend to overreport have higher rents and incomes and a higher proportion of female, elderly, non-Hispanic white and non-Hispanic Asian residents. Similar to the findings for government service delivery, the literature generally agrees that many citizen-generated datasets on open data portals could be biased relative to "ground truth," though effects may differ among cities and affected groups.

Methods

We present a prototype of a bias assessment tool that can be used by policymakers, analysts, and advocates to easily measure the level of bias in open geographic data. To determine the features our bias assessment tool should include, we preliminarily reviewed the 10 most popular datasets on the municipal open data portals of New York City, Chicago, Austin, and Nashville. We chose these four places as they were geographically diverse and contained both small and large cities. We looked at 40 of the most popular geographic datasets and found that

- most datasets had a geographic variable such as address, latitude/longitude, or zip code (approximately 75% of the 40 datasets);
- very few datasets had demographic data, such as race, age or gender (less than 5% of all datasets); and
- Most datasets were available as CSVs or other spatial file formats like GeoJSON or Shapefiles.

As a result, we decided to build a bias assessment tool that would take as input datasets with geographic data and present simple, interpretable measures of bias as its output. To simplify the process, we would only require users to upload a file with latitude and longitude coordinates—in either CSV or GeoJSON format—and design the tool to compute and output the relevant spatial and demographic bias metrics. Our prototype tool is designed to use the uploaded point-level data and combine it with geographic neighborhood-level data from the Census American Community Survey to determine demographic representation. Our methodology allows users to get a sense of the demographic profile and bias of their data even when demographic data isn't directly available within the dataset itself.

How the Tool Works

Our tool takes six steps to generate bias assessment statistics.

1. *Determine the dataset's source city.*

   Our audience is users of city open data portals. However, to minimize the burden on users, we do not require users to select their city and determine the city automatically from the data provided. To do this, the tool randomly samples 10 percent of the dataset and geocodes those points to a city. (Our definition of a city is all census tracts contained within the Census Place boundaries. This may be larger than the official city boundary. See appendix B for more details.) If the data has points from more than one city, the tool chooses the city that appears most often in the sample.

2. *Read in all geographic and demographic data for that city.*
   The tool reads in geographic data—the dataset itself, along with the spatial boundaries of all census tracts in the city—and demographic data—from the 2011–15 five-year American Community Survey's (ACS) Data profile. The specific demographic variables we pull from the ACS for each Census tract in the city are as follows:
   » percent of non-Hispanic White residents
   » percent of non-Hispanic Black residents
   » percent of non-Hispanic Asian residents
   » percent of other race non-Hispanic residents
   » percent of Hispanic residents
   » percent of residents with a bachelor's degree or higher
   » unemployment rate
   » percent of families and individuals whose income in the past 12 months fell below the poverty level

3. *Compute the Census tract to which each data point belongs.*
   The tool spatially joins each datapoint to the set of all census tracts in the city. All points that don't fall within a census tract—points outside of the city bound—are discarded.

4. *Calculate spatial bias metrics, which we call tract reporting bias.*
   The tract reporting bias is the percentage-point difference between the share of the dataset falling within a particular tract and the share of a city's population living in that tract. So if a particular census tract accounts for 10 percent of a dataset and 20 percent of a city's population, that tract's reporting bias would be 10 - 20 = -10%. This measure is calculated for every tract in the city and gives users a sense of which parts of the city are under- or overrepresented.

5. *Calculate demographic bias metrics, which we call demographic reporting bias.*
   The demographic reporting bias 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?"

   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 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 originates?"

   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. This measure is also calculated for our other demographic statistics of interest: the share of the population with a bachelor's degree or higher, the unemployment rate, the poverty rate, and our racial and ethnic population shares.

6. *Assess the statistical significance of our bias metrics.*
   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 99.7% confidence intervals for the tract reporting bias and the demographic reporting bias. If 0 does not fall within this confidence interval, then we report this bias as statistically significant. In other words, after you take into account the variability in the Census reported estimates, there is still statistically significant bias in the data. For more detail on our sampling procedure used to compute statistical significance see Appendix D.

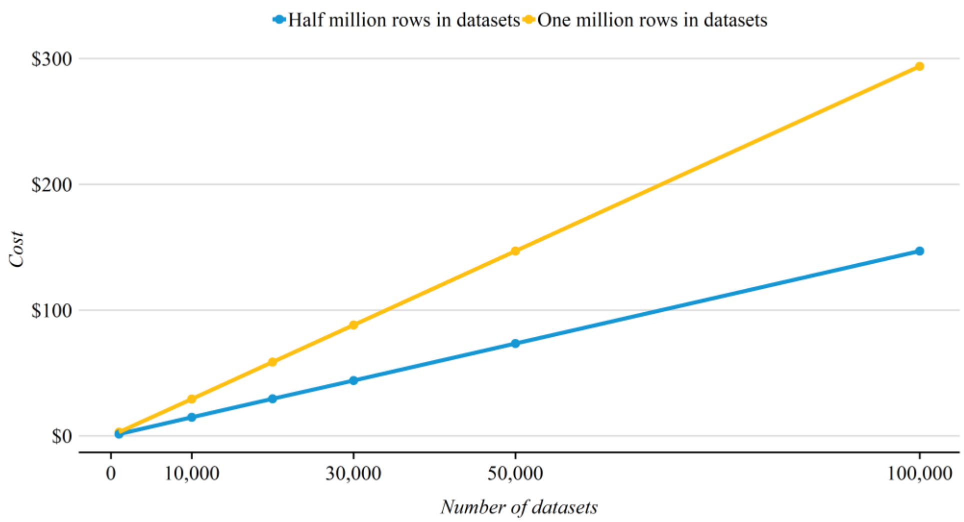
How we built the tool

The six steps from the previous section are encapsulated in a Python script and run using a serverless cloud architecture in Amazon Web Services. Building a web app like this typically involves setting up a computing service that is "always on" waiting for jobs to be submitted. With a serverless architecture, however, computing resources are used only for the length of time the code runs. The three main advantages of our serverless framework are the following:

- You don't need to be worried about setting up servers, allocating computing resources, updating security vulnerabilities, and maintaining infrastructure.
- Scaling is simple. It doesn't matter if you're using the tool on 10 datasets or 1,000,000. The cloud service provider will ensure that the increased load is handled with minimal interruption for the user and little effort from the IT team.
- It's very low cost. You're paying only for active computing time, not idle time, and the architecture we use in AWS has an extremely low cost structure. For example, if our tool received 1,000 datasets per month, each with 500,000 rows, we estimate the total cost of the tool as $1.67 a month, or $20 a year. Costs scale

linearly with the number of datasets, so 50,000 datasets per month costs $83.26 a year while 100,000 datasets costs $166.51.

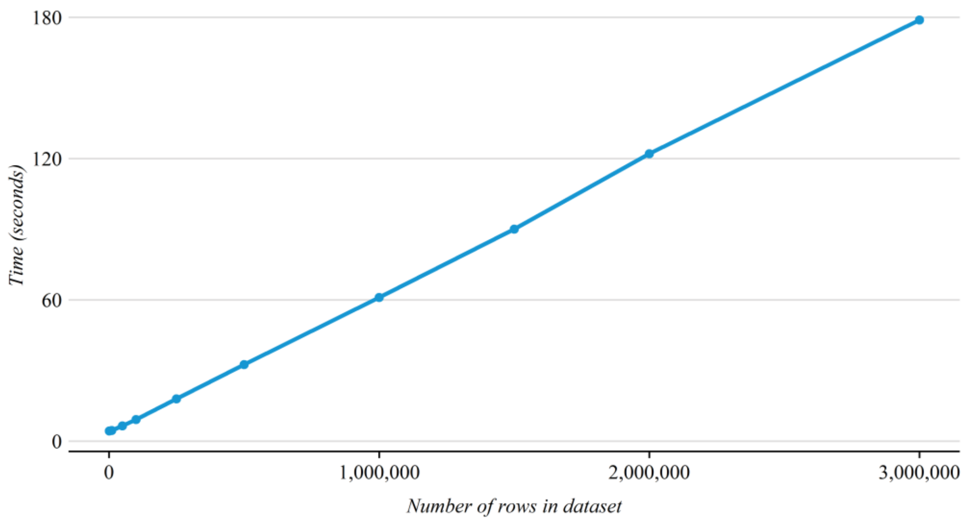
Figure 1. Total System Cost



Source: https://dashbird.io/lambda-cost-calculator.
Notes: Testing done on most powerful AWS configuration with 3.8 GB of RAM. Lower-cost but slower configurations are available on AWS.

- It's fast. We can run our code on the fastest infrastructure that the cloud service provider offers and have it execute much faster than it could on a laptop or desktop. For reference, a 10,000-row dataset processes in around 4 seconds, a 100,000-row dataset processes in about 17 seconds and a 1,000,000-row dataset processes in approximately 60 seconds. Speed generally scales linearly with size. Even the largest geographic datasets on open data portals that we examined contained less than 10 million records, which would typically process within minutes using our prototype tool.

Figure 2. Total System Speed



Source: Urban Institute testing.
Notes: Testing done on most powerful AWS configuration with 3.8 GB of RAM. Lower cost but slower configurations are available on AWS.

Limitations

Quantifying bias in data requires some measure of the ground truth for comparison. For the purposes of our tool, we use the population distribution within a city as the ground truth. In other words, our application assumes that a "bias-free" dataset would have data generated from each part of the city in proportion to the population that lives there. This approach has many limitations but is a sensible starting point for many open datasets, especially those we expect to come equally from all parts of a city, such as the location of public parks. However, it may not work as well with data for which population is a poor indicator of the underlying data generation process, such as 311 calls. While we might expect to receive more 311 calls in neighborhoods with more people, we might also expect to receive more 311 calls in areas with more issues such as uncollected garbage, and these issues may not be highly related to how many people live there.

We impute demographic data from geographic data, which assumes that all data points that come from a particular 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.

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 bias estimates generated by our tool are less reliable, as the vast majority of tracts will only contain one or two 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.

This tool currently supports assessing bias in a single city. If a dataset spans multiple cities, the tool will only operate on the most frequent city represented 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.

Finally, as noted above, the tool's operational definition for the boundary of a city might differ slightly from the boundary that the Census uses. A census place is the Census analog for a city while census tracts are Census analogs for neighborhoods. Often the boundaries of census places and census tracts don't overlap perfectly, meaning some tracts are only partially covered by the Place boundary. Our tool defines a "city" as all census tracts whose area is at least 1 percent covered by the by the relevant census place. This overinclusive definition will cause our tool to think that many cities—particularly small and medium-sized one—are bigger than they actually are, in both geographic size and population. See appendix B for a much more detailed discussion of this limitation.

Data

To test the bias assessment tool, we apply it to three datasets—bikeshare station location data, 311 data, and LIHTC data—across a few example cities. We chose these datasets because intercity comparisons are simple—these datasets are common and fairly standard across cities—and because equitable access to transportation, city services, and affordable housing are important priorities for many cities.

Bikeshare Station Data

Over the past few years, bikeshare systems have grown in popularity. At the end of 2017, there were 100,00 bikes in bikeshare systems, more than double the 42,500 bikes estimated to be available at the end of 2016 ("Bike Share" 2017). While bikeshare is a growing transportation option for city residents, many have questioned the equity of bikeshare systems. For example, in many major cities, bikeshare riders tend to be have higher incomes and are more likely to be white than the underlying population (Smith, Oh, and Lei 2015).

Many cities whose open data portals we examined have published data on the location of bikeshare stations. Our bias assessment tool provides a quick and easy way to answer questions around who bikeshare stations are serving, where they are, and how these patterns differ across cities.

As a test of our tool, we downloaded data on bikeshare station ocations from four cities that published the data on their open data portals: Boston, Chicago, Washington, DC, and Philadelphia.

Table 1. Bikeshare Locations Data: Overview

City	Number of stations	URL	Date accessed
Boston	1,964	http://bostonopendata-boston.opendata.arcgis.com/datasets/d02c9d2003af455fbc37f550cc53d3a4_0.geojson	08/05/2018
Chicago	559	https://data.cityofchicago.org/resource/aavc-b2wj.geojson	08/05/2018
Washington, DC	269	https://opendata.arcgis.com/datasets/a1f7acf65795451d9f0a38565a975b3_5.geojson	08/05/2018
Philadelphia	126	https://api.phila.gov/bike-share-stations/v1	08/05/2018

Source: Open Data portals for Boston, Chicago, Washington DC, and Philadelphia

311 Service Request Data

The 311 service allows residents to report requests for non-emergency city services such as snow removal or street light repair. Previous studies have found that there is significant bias in 311 data: residents who have the time and resources to create service requests are generally more highly educated, have higher incomes, and are more likely to be white

(Feigenbaum 2015). The bias assessment tool provides a quick way to assess the representativeness of 311 service requests and allows us to answer questions surrounding where service requests come from and which demographic groups are more or less likely to submit requests. As we note in our limitations section, however, 311 data may not represent the underlying need for service, and so readers should be cautious when interpreting results.

We downloaded data on 311 service requests from Boston, San Francisco, Washington, DC, and Philadelphia. To standardize our intercity comparisons, we only looked at data for calendar year 2017.

Table 2. 311 Data: Overview

City	Number of requests	URL	Date accessed
Boston	1,099,707	https://data.boston.gov/dataset/311-service-requests	08/05/2018
San Francisco	487,142	https://data.sfgov.org/resource/ktji-gk7t.csv?$where=requested_datetime%20>%20”2017-01 01”%20AND%20requested_datetime%20<%20”2017-12 31”%20&$limit=30000000%20&$select=lat, long,requested_datetime,service_name	08/05/2018
Washington DC	309,542	https://opendata.arcgis.com/datasets/19905e2b0e1140ec9ce8437776feb595_8.csv	08/05/2018
Philadelphia	194,703	https://phl.carto.com/api/v2/sql?q=SELECT%20requested_datetime,lat,lon,service_name%20FROM%20public_cases_fc%20WHERE%20requested_datetime%20%3E=%20%272017-01-01%27%20AND%20requested_datetime%20%3C%20%272017-12-31%27	08/05/2018

Source: Open Data portals for Boston, San Francisco, Washington DC, and Philadelphia.
Note: 311 requests were limited to the calendar year of 2017 to standardize intercity comparisons.

LIHTC Data

LIHTC is a tax credit program funded by the federal government to incentivize the building of low income housing. The LIHTC program provides billions of dollars in tax credits for the acquisition, rehabilitation, and construction of new affordable rental housing units (Keightley 2013). The bias assessment tool can help us quickly answer questions surrounding the types of neighborhoods in which LIHTC housing is being built. We downloaded LIHTC data on all cities in the US from https://lihtc.huduser.gov. We then filtered the data to the 6 example cities with the largest number of LIHTC units—New York, Philadelphia, Chicago, Raleigh, Detroit, and Seattle.

Table 3. LIHTC Data: Overview

City	Number of buildings	Date accessed
New York	1,923	08/06/2018
Philadelphia	445	08/06/2018
Chicago	416	08/06/2018
Raleigh	379	08/06/2018
Detroit	305	08/06/2018
Seattle	281	08/06/2018

Source: https://lihtc.huduser.gov**.**
Note: These cities were chosen because they had the highest number of LIHTC buildings

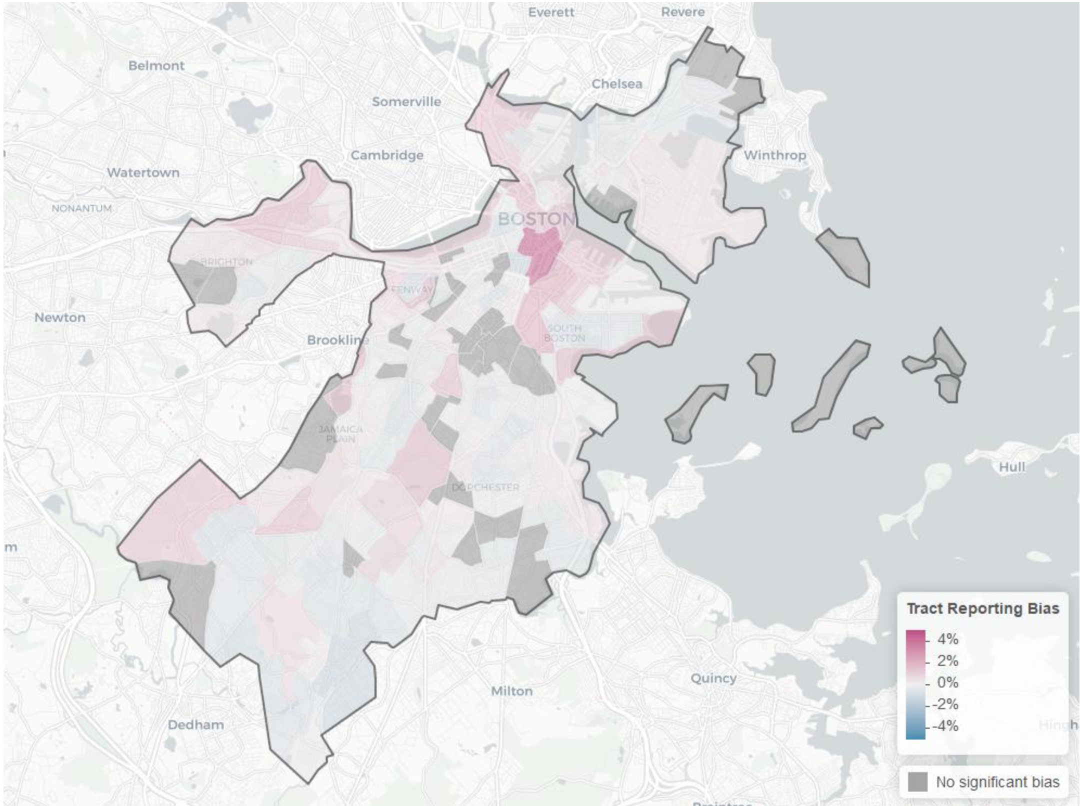
Results

We walk through the results of each set of our example datasets in the following sections.

Bikeshare Station Analysis

We start out by analyzing the results of the Boston bikeshare locations dataset as an illustrative example, and then display the full set of results across the four cities we examined. The first chart is a visualization of the geographic bias in the Boston bikeshare data as measured by our Tract Reporting Bias metric.

Figure 3. Boston Bikeshare Locations: Tract Reporting Bias

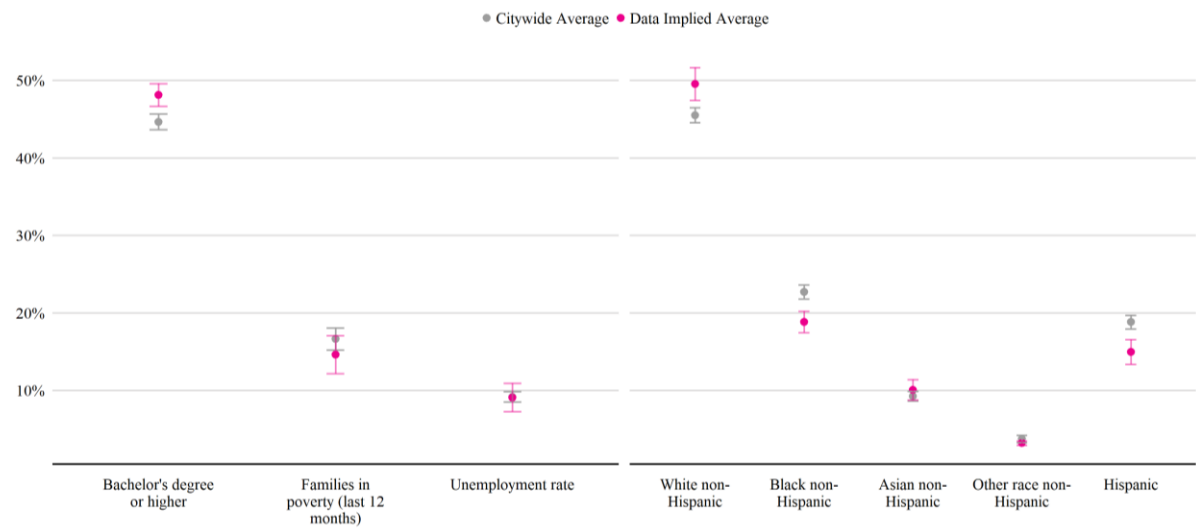


Source: Urban Institute testing.
Note: Leaflet | © OpenStreetMap contributors, © CartoDB

The gray tracts in the map represent areas where the tool found no significant bias. In other words, the bikeshare data in the tract is represented in the same proportion as the population in that tract, or at most within the margin of error for the population estimates. The pink areas represent tracts that are overreported in the data, while the blue areas represent tracts that were underreported. In the case of Boston, the tract with the highest positive reporting bias of +4% is the dark pink tract located in downtown Boston near City Hall. We might assume that overrepresentation is reasonable in this case; downtown Boston should be a popular destination for bikeshare riders, so it makes sense to have more stations there. The tracts with negative reporting bias are primarily concentrated in South and Southwest Boston, though the levels of underrepresentation are relatively small in comparison.

The next chart we produced is a visualization of the demographic bias in the Boston bikeshare data.

Figure 4. Boston Bikes Citywide versus Data-Implied Demographic Averages

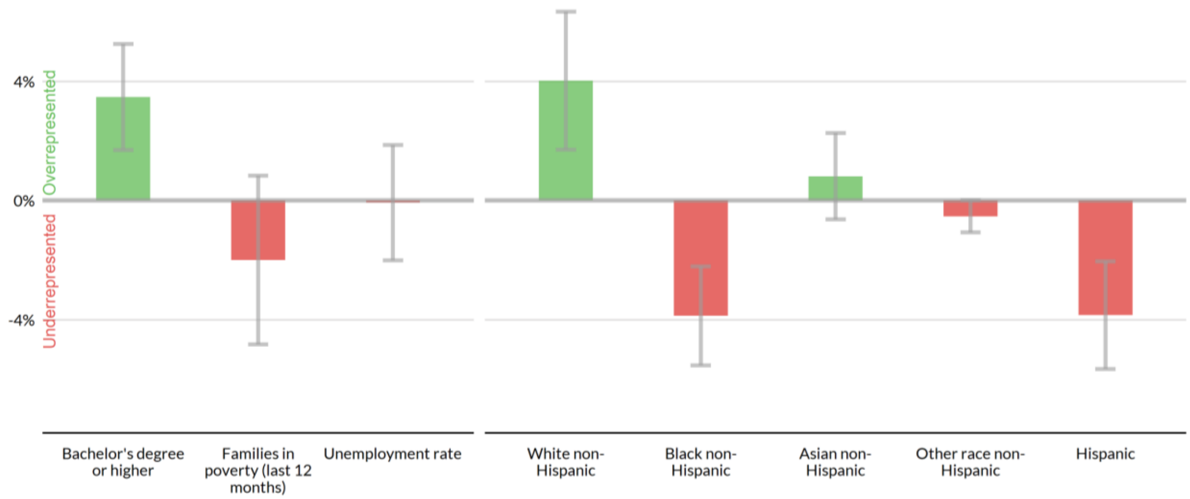


Source: Hubway Stations—Boston Open Data Portal.

The gray dots represent citywide averages and the pink dots represent the data implied averages for bikeshare station locations. We also calculated 99.7% confidence intervals around both measures and show them as bands (See appendix D for more information about our random sampling procedure for creating the confidence intervals). If the pink dots are above the gray dots, then that demographic group is overrepresented in the data. For example, the city's population is approximately 45 percent white non-Hispanic, while bikeshare stations are located in areas whose population is 50 percent white non-Hispanic on average. This means that we estimate white non-Hispanic residents are overrepresented by about 5 percent in the data. And if the gray and pink bands do not overlap, then this means that this demographic bias (or the difference between the two) is statistically significant at the 99.7% level once you take into account the variation in Census-provided estimates. In the case of the white non-Hispanic column, because the confidence interval bands don't overlap, we deem the overrepresentation to be statistically significant.

The final graph we produced is the demographic reporting bias, or the percentage-point difference between the data implied average and the citywide average centered at 0. We also plot the confidence interval around the difference. This is essentially a normalized version of the previous graph that gives readers a clearer sense of under and overrepresentation relative to the confidence band.

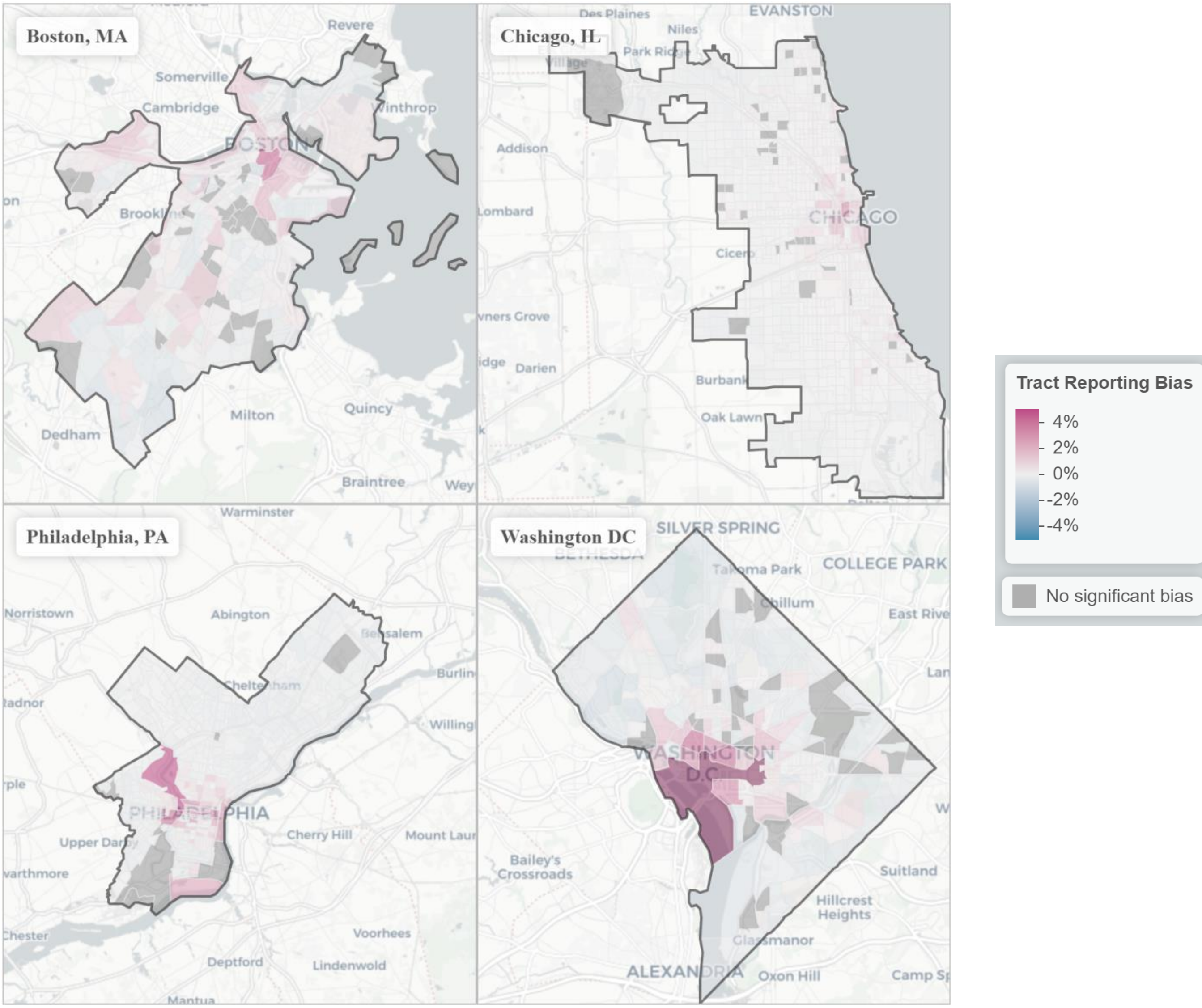
Figure 5. Boston Bikes Demographic Reporting Bias (percentage difference)



Source: Hubway Stations—Boston Open Data Portal.

White non-Hispanic and higher-educated residents are overrepresented by around 4 percent while black and Hispanic residents are underrepresented by around 3.5 percent. And this demographic bias is statistically significant, as the confidence bands don't touch zero. This dataset seems to suggest that in Boston, bikeshare stations are more accessible to higher-educated white neighborhoods than communities of color. These differences represent relatively small margins, however, compared with the other cities and datasets we investigated. We now repeat the above analysis for all the cities we have data on. From now on, we only display this normalized graph when discussing demographic bias. If readers want to see the underlying citywide versus data-implied demographic average graphs with dual confidence intervals, they can consult appendix F.

Figure 6. Bikeshare Stations Tract Reporting Bias: Intercity Comparison



Source: Urban Institute Testing
Note: Leaflet | © OpenStreetMap contributors, © CartoDB

Across all cities, downtown areas tend to be overrepresented and areas farther from the city center tend to be modestly underrepresented. A notable exception is Boston, which has several overrepresented tracts far from the city center. This suggests that bikeshare systems are mainly located in downtown areas and don't provide coverage to large swaths of the city's outlying areas. So people who live outside the city center may have little incentive or ability to use these bikeshare systems.

Figure 7. Bikeshare Stations: Demographic Reporting Bias (percentage difference)

Source: Hubway Stations, Divvy Bicycle Stations, Capitol Bikeshare Locations, Indego Bikeshare Stations—Respective Municipal Open Data Portals
Note: Other race non-Hispanic includes Alaskan Natives, American Indians, Native Hawaiians, Pacific Islanders, multiracial, and all other racial categories

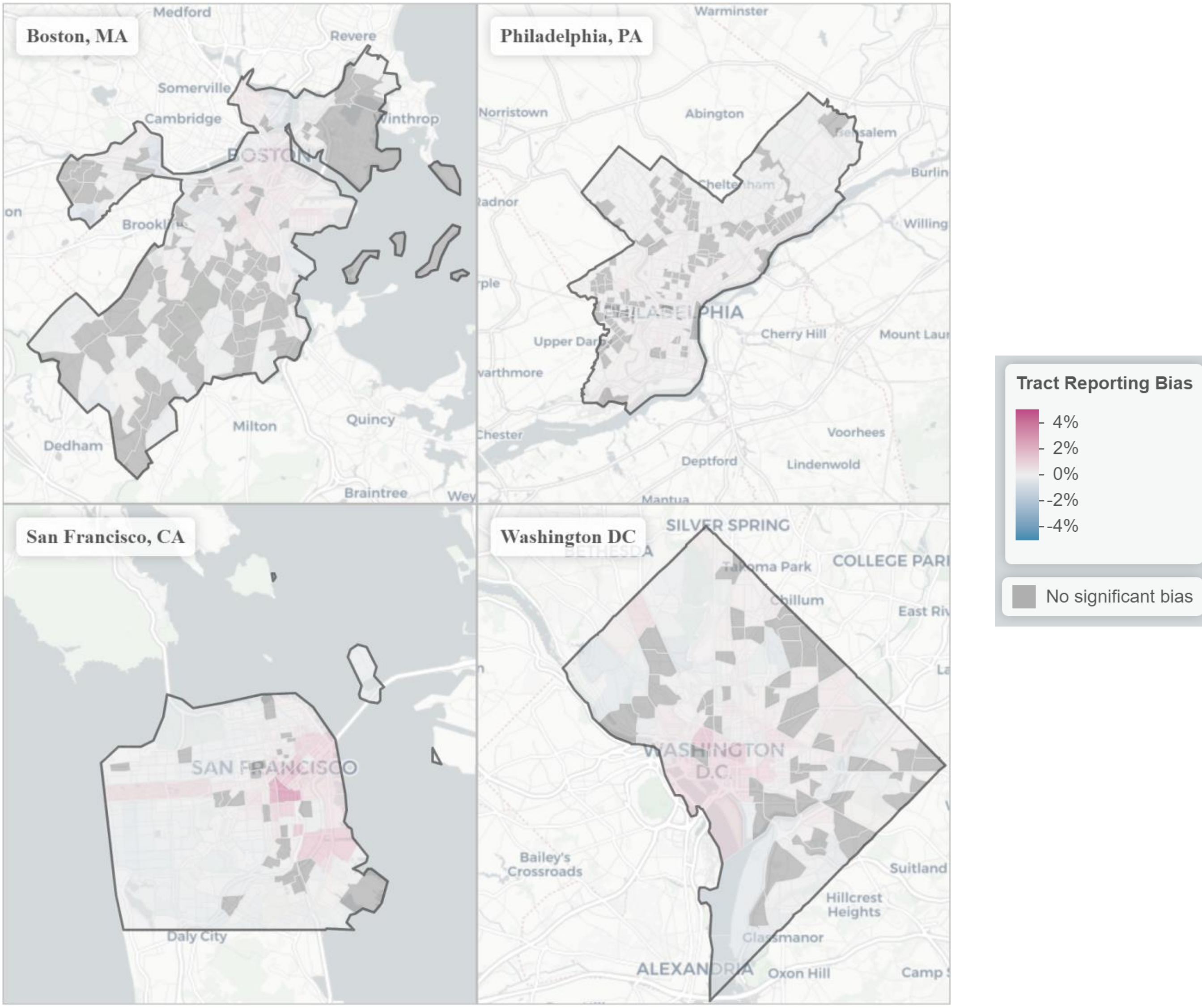
We see that the demographic biases are remarkably directionally consistent across all cities. All four cities exhibit statistically significant overrepresentation of white residents and residents with a bachelor's degree or higher, and moderate overrepresentation of Asian residents. All four cities also exhibit underrepresentation of Black, Hispanic, unemployed, and impoverished residents, though in Boston and DC, a few of these results are not statistically significant. Comparatively, Boston displays the lowest amount of demographic bias—in other words, the lowest amount of collective over and underrepresentation of certain demographic groups—among the cities we chose. Boston has the lowest levels of overrepresentation when it comes to white and higher-educated residents, and it has some of the lowest levels of underrepresentation when it comes to Black, Hispanic, unemployed, and impoverished residents. This may be because Boston has one of the largest bikeshare systems—559 total stations—especially when compared to Philadelphia and Chicago, who each have less than 200. But it also may be because the bikeshare system in Boston is less concentrated in the city center, with additional stations throughout the city.

As a final note, it's important to remember that the bikeshare stations dataset had the smallest sample sizes; Philadelphia had only 126 stations in the data. As we mentioned in the methods section, we recommend using datasets of a few hundred points or larger to derive valid, more rigorous interpretations of bias. However, because the bias patterns are remarkably consistent across cities with both small and large numbers of stations, we can be reasonably confident in the direction of the results.

311 Service Request Analysis

In this section, we analyze 311 data across Washington, DC, Boston, Philadelphia, and San Francisco. Across all cities, a large number of tracts showed no significant level of bias, indicating that 311 requests were being generated in accordance with the population in that tract. We see some modestly overrepresented census tracts around the downtown areas, especially for Boston, Washington, DC, and San Francisco. A majority of the census tracts demonstrate very small levels of underrepresentation in the cities we analyzed.

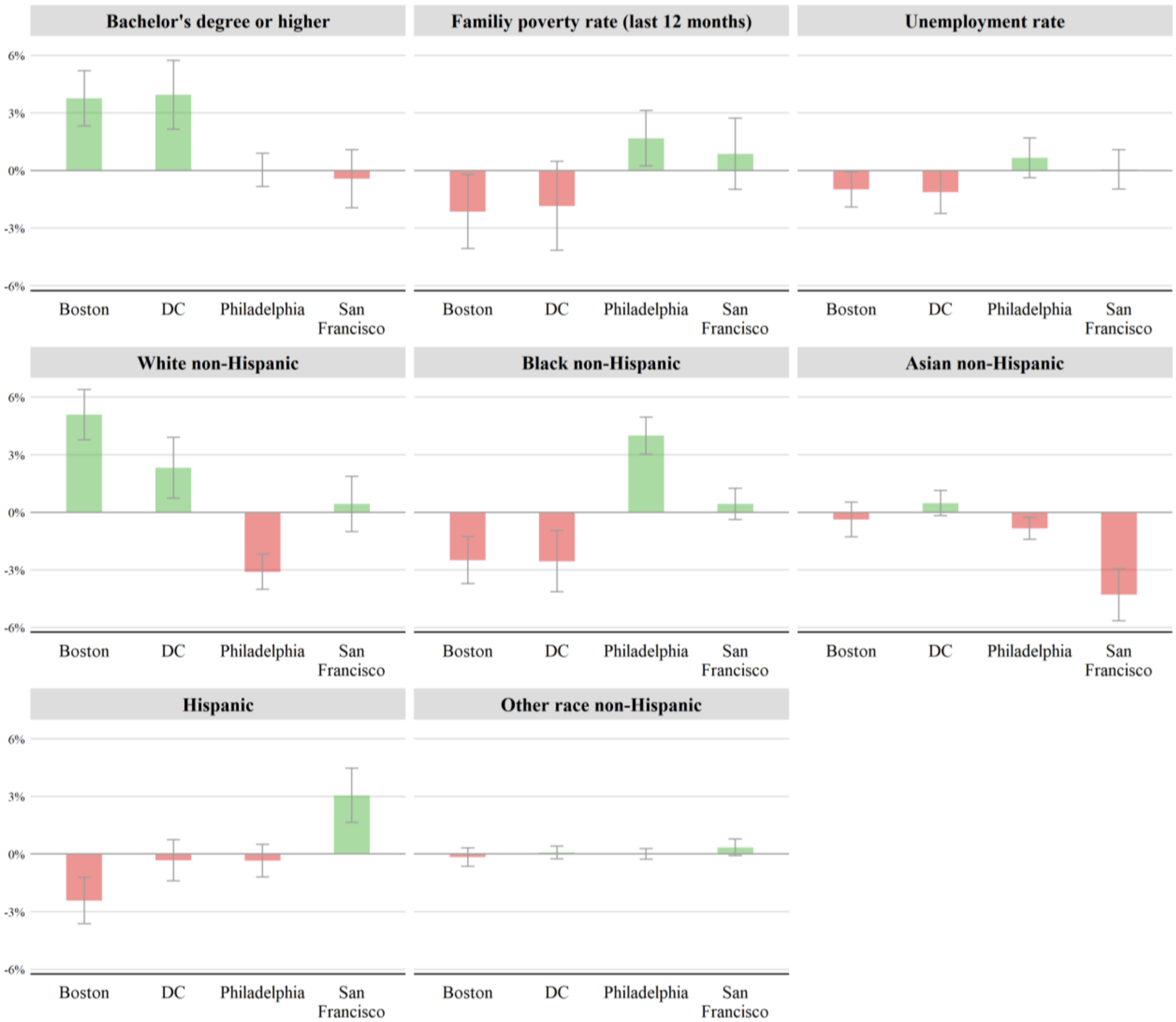
Figure 8. 311 Requests: Tract Reporting Bias



Source: Urban Institute testing.
Note: Leaflet | © OpenStreetMap contributors, © CartoDB.

In terms of demographic bias, there does not seem to be a consistent pattern across the cities we analyzed. Boston and DC both exhibit modest and statistically significant overrepresentation of residents with a bachelor's degree or higher and white non-Hispanic residents. These cities also exhibit underrepresentation of Black and impoverished residents. Philadelphia and San Francisco both exhibit more unique patterns. Philadelphia shows significant overrepresentation of Black non-Hispanic residents and underrepresentation of white non-Hispanic residents. San Francisco, on the other hand, demonstrates statistically significant overrepresentation of Hispanic residents and underrepresentation of Asian non-Hispanic residents.

Figure 9. 311 Requests: Demographic Reporting Bias (percentage difference)



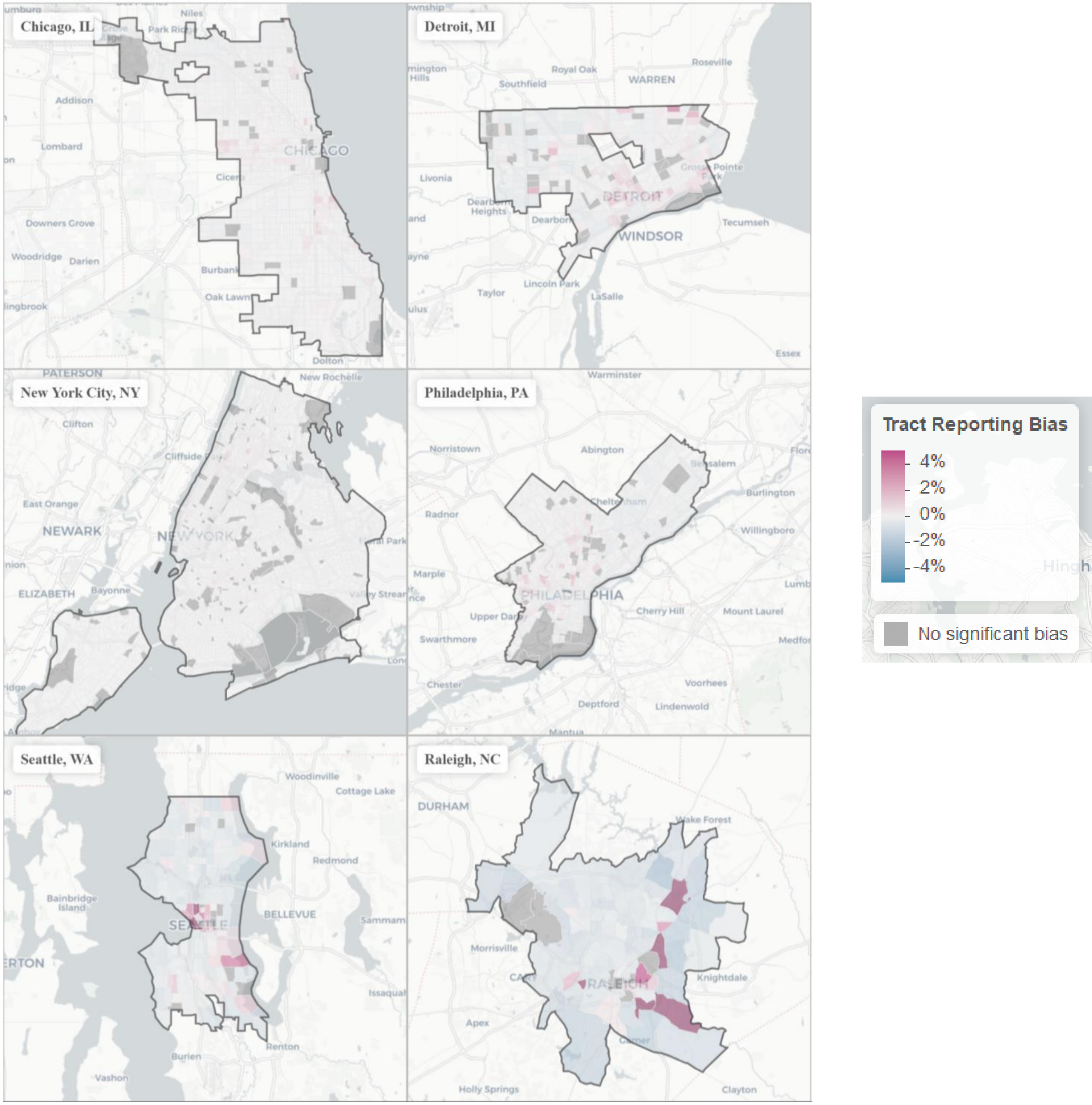
Source: 311 Service Requests, 311 Service and Information Requests, City Service Requests in 2017, 311 Cases—Municipal Open Data Portals.
Note: Other race non-Hispanic includes Alaskan Natives, American Indians, Native Hawaiians, Pacific Islanders, and all other racial categories

However, readers should be careful when interpreting the results of the bias assessment tool on 311 data. Demographic bias in 311 data can stem from the underuse of 311 services by certain communities or reflect the fact that government services are less urgently needed in certain communities. In other words, it's hard to tell if underrepresentation of white non-Hispanic residents in Philadelphia is because white non-Hispanic residents tend to use 311 less or because white non-Hispanic neighborhoods tend to need fewer government services in the first place. To control for this discrepancy, previous studies of 311 data limit the data to reports that are generally universal and equally distributed across neighborhoods, such as snow removal or pothole service requests. Future iterations of this tool would need to account for this discrepancy (perhaps by introducing other variables besides the population distribution to use as the ground truth), and users should generally be wary when they expect the data-generating process may differ significantly from the population distribution, as is the case with 311.

LIHTC Analysis

In this section, we analyze the location of low-income tax credit funded affordable housing project locations across Detroit, New York City, Philadelphia, Chicago, Raleigh, and Seattle. Philadelphia, New York, and Chicago have relatively few tracts outside of the city center with overrepresentation of LIHTC housing units. In Philadelphia, these overrepresented tracts are located primarily in North and West Philadelphia, in Chicago primarily in the Southside, and in New York, primarily in the Harlem and upper Manhattan area. In Raleigh, areas of underrepresentation tend to be adjacent to areas of overrepresentation and are all on the East side of the city. And Seattle and Detroit both contain overrepresented tracts around the city center.

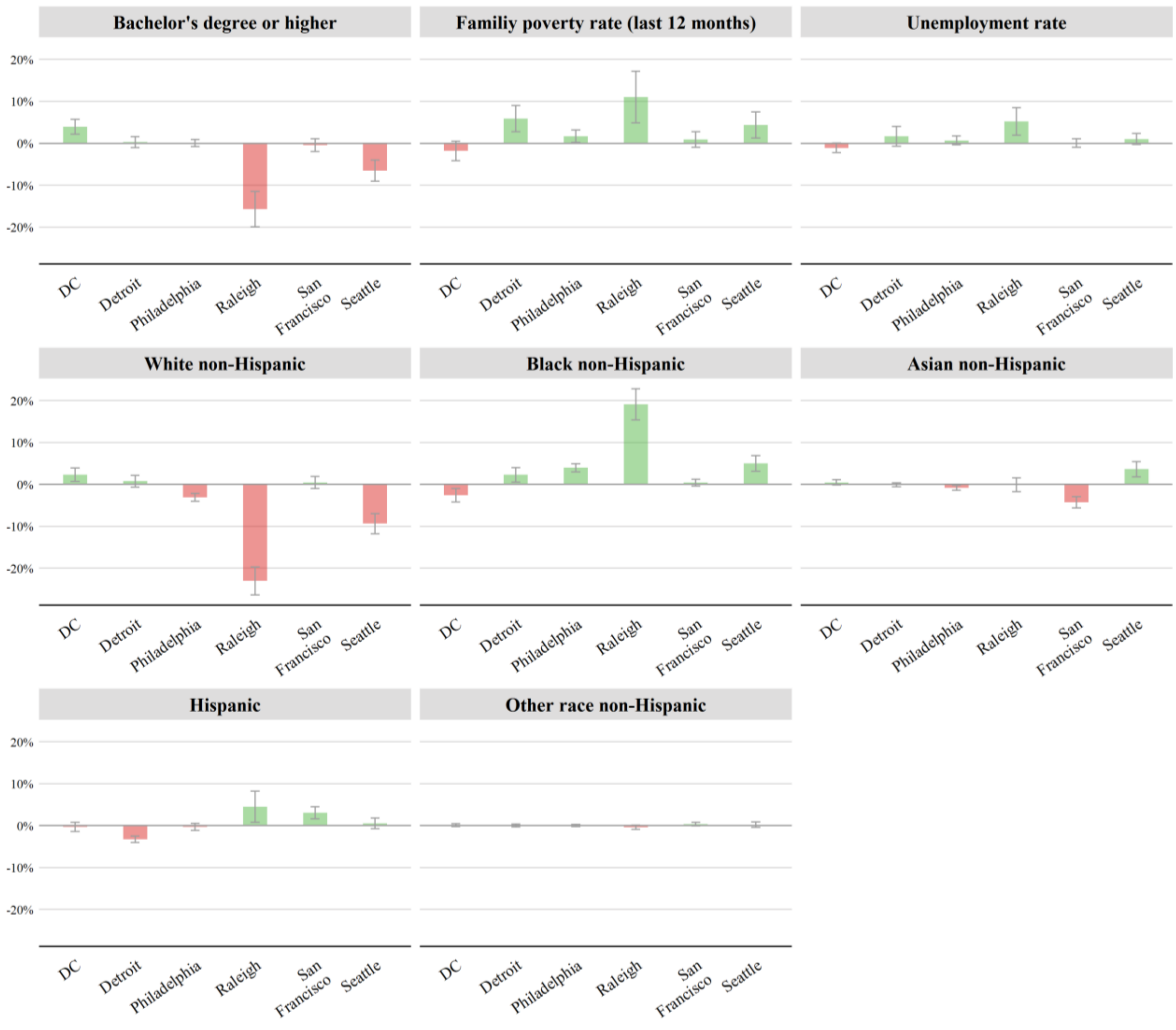
Figure 10. LIHTC Buildings: Tract Reporting Bias



Source: Urban Institute testing.
Note: Leaflet | © OpenStreetMap contributors, © CartoDB.

LIHTC data show higher levels of over and underrepresentation relative to the 311 data. Raleigh and Seattle, in particular, exhibit large and significant underrepresentation of white non-Hispanic residents and overrepresentation of Black non-Hispanic residents. Raleigh and Seattle also show large and significant underrepresentation of residents with a bachelor's degree or higher. Across all cities, LIHTC buildings seem to be slightly overrepresented in impoverished and high-unemployment neighborhoods. Other significant biases include overrepresentation of college-educated and white non-Hispanic residents in DC and underrepresentation of Asian non-Hispanic residents in San Francisco. Overall, cities tend to have LIHTC housing in poorer parts of the city, except Washington, DC. The city with the most amount of demographic bias is Raleigh, with high overrepresentation of Black and impoverished residents and high underrepresentation of white and higher-educated neighborhoods especially compared to the other cities studied.

Figure 11. LIHTC Data: Demographic Reporting Bias (percentage difference)



Source: Urban Institute analysis.
Note: Other race non-Hispanic includes Alaskan Natives, American Indians, Native Hawaiians, Pacific Islanders, multiracial, and all other racial categories.

Conclusion

As more city governments embrace data driven decision making, it is important that cities have readily available methods for measuring and understanding bias in the underlying datasets. Our automated bias assessment tool prototype provides a simple but important first step in this effort. By combining user uploaded point-level data and geographic neighborhood-level data from the Census American Community Survey, the tool provides simple measures of spatial and demographic bias.

We found many interesting patterns when we applied our tool to three types of test data: bikeshare station locations, 311 service requests, and LIHTC location data. When we looked at bikeshare stations, we found that non-Hispanic white, non-Hispanic Asian, and higher-income neighborhoods are routinely overrepresented while Black, Hispanic, and low-income and poor neighborhoods are routinely underrepresented. When we looked at 311 requests for service, we found that white and college-educated residents in DC and Boston are overrepresented while non-Hispanic Black and poor neighborhoods are overrepresented in Philadelphia. And finally, we found that although LIHTC data demonstrated relatively lower levels of bias overall, Raleigh exhibited particularly high overrepresentation of buildings in Black and lower-income neighborhoods.

We plan to build out a public interface and dashboard to the tool so anyone can easily explore bias in example data or upload and analyze bias in their own data. An accessible public interface would allow the public to understand what the tool can do and to answer all sorts of questions surrounding bias in open datasets. This interface would be completely free and benefit not just city leaders, but nonprofits, local programmers, equity advocates, local Code for America brigades, data intermediaries in cities, and a host of other actors. Given the low cost of the serverless architecture, it would be possible for Urban to host the web interface with only modest support. In addition, we plan to add features to the tool:

- The ability to analyze bias by census block group for a more granular view
- The ability to specify variables other than population distribution as the "ground truth" dataset. Examples include the "24-hour population," which takes into

account the people that live or work in a given tract, or custom user-uploaded data.

Appendix A. Open Data Portal Analysis

After scanning a few municipal open data portals, we decided to focus on the cities of New York, Austin, Chicago, Nashville, and Los Angeles. We chose these cities because they were of varying sizes, located across the United States and each provided the ability to sort by the most popular datasets on their open data portals. For each city, we identified the 10 most popular datasets on the city's open data portal, and for each dataset recorded the following variables:

1. Dataset name: The title of the dataset on the open data portal
2. Geo var: A dummy variable indicating if the dataset had any geographic variables like zip code, address, latitudes/longitudes or WKT geometry columns.
3. Demo var: A dummy variable indicating if the dataset had any demographic variables like age, race, ethnicity or gender.
4. Geographic variables list: A list of all the geographic variables listed. Possible options are address, lat/lon, state plane, and wkt
5. Demographic variables list: A list of all the geographic variables listed. Possible options are age, gender, race, and veteran status. Note that only one dataset reported any demographic variables.

The results of our open data portal scan are below. Overall the titles suggest that datasets about transit, 311, crime, building permits are some of the most popular. We also see that most datasets have geographic variables and only few datasets contain demographic information.

Table A.1. Open Data Portals Analysis

City Portal	Dataset Name	Geo var	Demo var	Geographic variables list	Demographic variables list
NYC	DOB Job Application Filings	1	0	address, lat/lon	
NYC	TLC New Driver Application Status	0	0		
NYC	For Hire Vehicles (FHV) - Active	1	0	address	
NYC	Civil Service List (Active)	0	1		veteran status
NYC	311 service requests	1	0	state plane, lat/lon	
NYC	Subway entrances	1	0	lat/lon, wkt	
NYC	Medallion Drivers	0	0		
NYC	NYPD Motor Vehicle Collisions	1	0	lat/lon	
NYC	Street Hail Livery (SHL) Permits	0	0		
NYC	City Record Online	1	0	address	
Austin	Austin Animal Center Found Pets Map	1	0	address, lat/lon	
Austin	Off Leash Areas	1	0	address, lat/lon	
Austin	Issued Construction Permits	1	0	lat/lon	
Austin	Map of Declared Dangerous Dogs	1	0	address, lat/lon	
Austin	Food Establishment Inspection Scores	1	0	address	
Austin	APD CRIME INCIDENTS	1	0	address, lat/lon	
Austin	Neighborhood Groups Community Registry	1	0	address, zip	
Austin	Real-Time Traffic Incident Reports	1	0	lat/lon	
Chicago	Crimes - 2001 to present	1	0	state plane, lat/lon	
Chicago	Current Employee Names, Salaries, Position Titles	0	0		
Chicago	Building Permits	1	0	lat/lon	
Chicago	Lobbyist Data - Historical - Lobbyist Registry - 2010	1	0	address	
Chicago	Affordable Rental Housing Developments	1	0	state plane, lat/lon	
Chicago	Business Licenses - Current Active	1	0	lat/lon	
Chicago	Problem Landlord List - Map	1	0	state plane, lat/lon	
Nashville	General Government Employees Titles, Base Annual Salaries	0	0		
Nashville	Building Permits Issued	1	0	address, lat/lon	
Nashville	Metro Water Services Outages	1	0	address, lat/lon	
Nashville	Property Standards Violations	1	0	address, lat/lon	
Nashville	Historic Nashville City Cemetery Interments	0	0		
Nashville	Residential Short Term Rental Permits	1	0	address, lat/lon	
Nashville	General Government Employees Demographics	1	0	race, gender, age	
Nashville	hubNashville (311) Service Request Data	1	0	address, lat/lon	
Los Angeles	Building, Safety Permit Information	1	0	address, lat/lon	
Los Angeles	Listing of Active Businesses	1	0	address, lat/lon	
Los Angeles	2014 Registered Foreclosure Properties	1	0	address, lat/lon	
Los Angeles	MAP OF HCIDLA MANAGED PIPELINE PROJECTS BEGINNING IN 2003 TO PRESENT	1	0	address, lat/lon	
Los Angeles	New building permit	1	0	address, lat/lon	
Los Angeles	Electrical permits	1	0	address, lat/lon	
Los Angeles	MyLA311 Service Request Data 2016	1	0	lat/lon	

Source: Urban analysis of most accessed datasets on open data portals. NYC (https://opendata.cityofnewyork.us/), Austin (https://data.austintexas.gov/), Chicago (https://data.cityofchicago.org/), Nashville (https://data.nashville.gov/), LA (https://data.lacity.org/)
Notes: When geo var = 1, this means a geographic variable was present in the dataset. If demo var =1, this means a demographic variables (race, ethnicitiy, age, or gender) was present in the data

Appendix B. 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 defined as a concentration of population which 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 we want our tool to use is available only at the census tract level. So we needed an operational definition of a city that spans whole census tracts.

Our tool uses the following inclusive definition of a city: 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, 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 terms of area and population.

We decided on a 1% tract area cutoff for two reasons:

1. After visual inspection of a few cities, the 1% cutoff gave reasonable results for city boundaries.
2. The 1% cutoff allowed us to exclude tracts that were right on the border of Census places. If we imposed no cutoff (ie we defined a city as all tracts that were contained even partially within the Place boundary), then the city boundaries included border tracts and the cities were much larger than expected.

Applying a more accurate city definition would require using block- or block group–level data, which would significantly increase computation time. Improvements in computational capacity would enable us to use more granular block group level data in the future, though smaller granularity has the downside of larger margins of error.

To understand why we adopted this tract based overinclusive definition of a city, we present a few maps and figures. Below are the names and estimated populations of four small to medium sized cities: Cupertino, CA; Flagstaff, AZ; Aurora, CO; and Madison, WI.

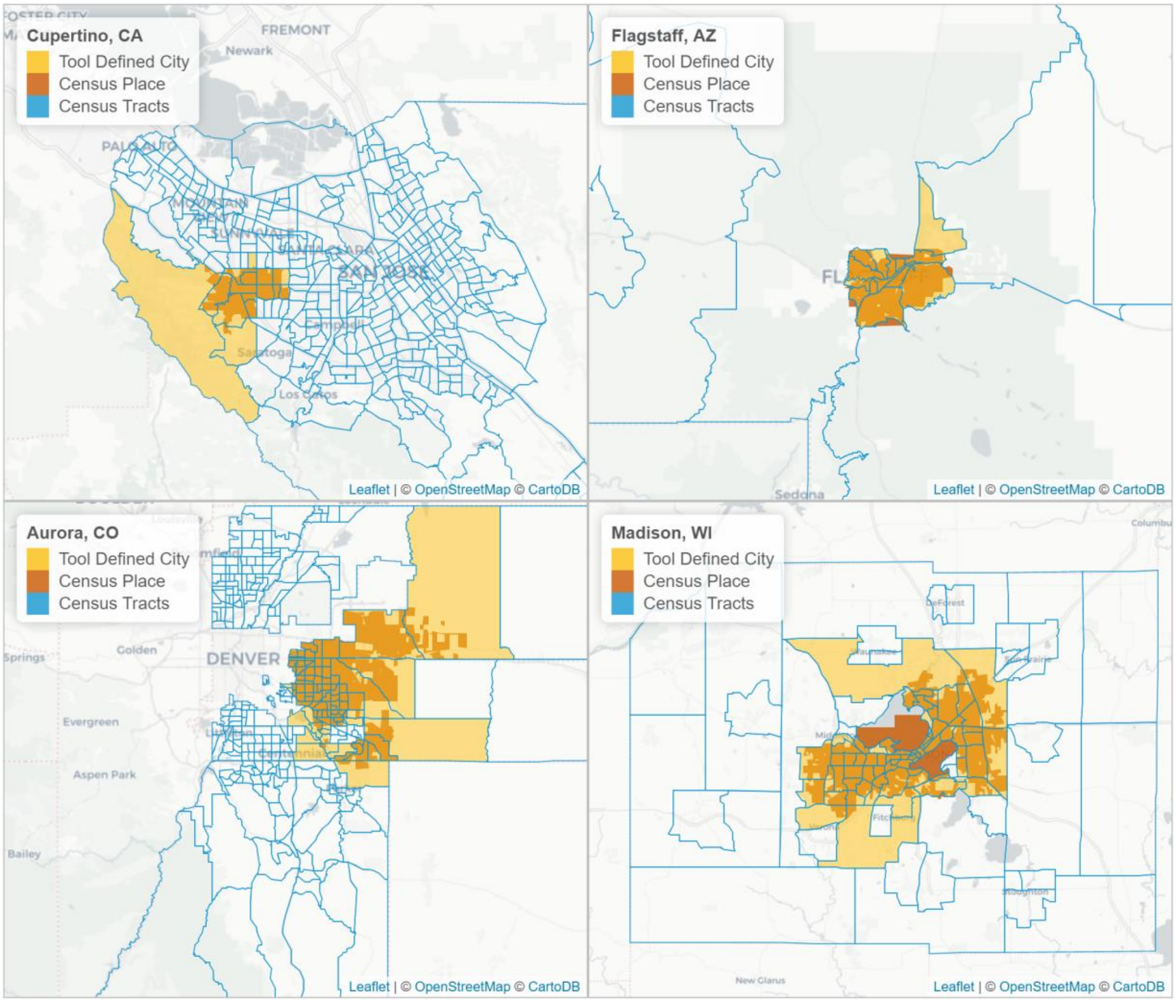
Table B.1. Small to Medium-Sized Census Places

Place GEOID	Place name	Population	Standard Error (population)
0617610	Cupertino city, California	60,297	63
0423620	Flagstaff city, Arizona	69,270	53
5548000	Madison city, Wisconsin	246,034	98
0804000	Aurora city, Colorado	351,131	274

Sources: Place names and GEOID pulled from https://www2.census.gov/geo/docs/reference/codes/files/national_places.txt. Population data pulled from 2012–16 five-year American Community Survey

And below are maps of the boundaries of these cities. The boundaries of Census tracts in the general vicinity are drawn in blue, the Census Place boundaries are shaded in orange, and what the tool defines as the city bounds (all tracts with at least 1 percent of their area covered by the place boundary) are shaded in yellow.

Figure B.1. Boundaries for Cupertino, Flagstaff, Aurora, and Madison



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

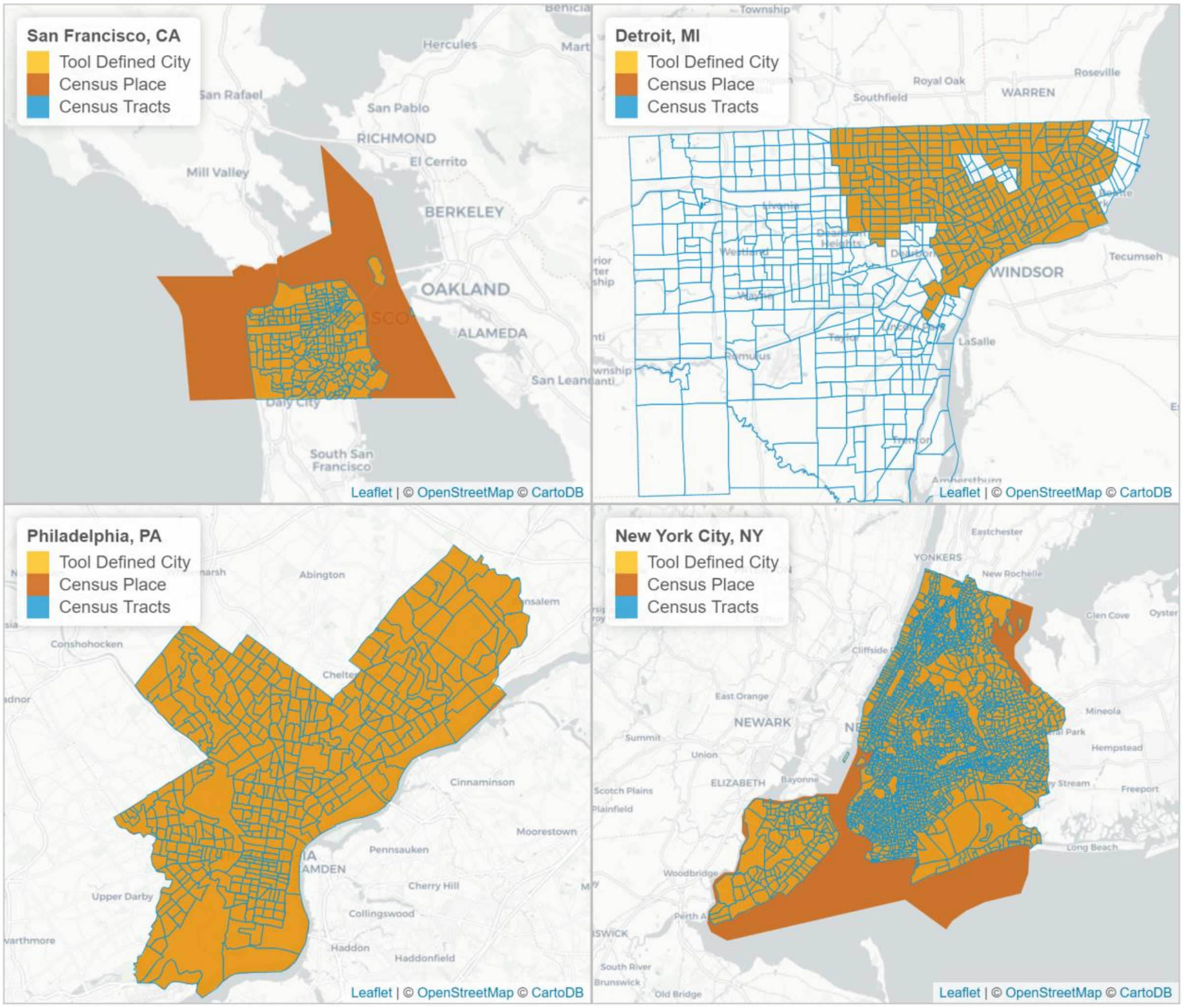
For these four small cities, the census place boundaries are irregular and only partially cover many census tracts. This leads our tool to think these cities are larger than the Place boundary implies, both in terms of areal size and population. In other words, the yellow area is far bigger than the orange area. For example, the city of Aurora has many small specks of orange scattered throughout tracts, meaning the census place has very irregular boundaries. As a result the area of Aurora as defined by our tool in yellow is greater than the actual area of Aurora in orange.

The city of Flagstaff also illustrates how our 1% cutoff works. While there are some small areas in orange that extend out to neighboring tracts, these tracts are so large that the orange part constitute less than 1% of the tract's total area. As a result, the large neighboring tracts are not considered a part of Flagstaff by our tool. However there are some tracts that are partially covered by the place boundary and fall above the 1% cutoff, so the tool counts the whole tract as a part of Flagstaff (those tracts are shaded in yellow). Generally, most small and medium sized cities have Place boundaries that do not exactly correspond to Census tracts and are subject to the over inclusive definition city definition problem, so users should be very careful when using our tool on small and medium size cities.

While the maps above illustrate the impact of the boundary discrepancy on the size of cities, it is also important to consider the effect of the boundary discrepancy on population estimates. Take the example of Cupertino, CA above. Our tool defines Cupertino as a collection of 16 census tracts and estimates that the total population is 77,166 with an accompanying standard error of 1376. However, as seen in Table XX, the actual population of Cupertino is 60,297 with an associated standard error of 63. So our tool thinks that the population is around 17,000 larger than it actually is. The standard error of the estimate is also larger because the Census utilizes address level data in cities to make more precise estimates)

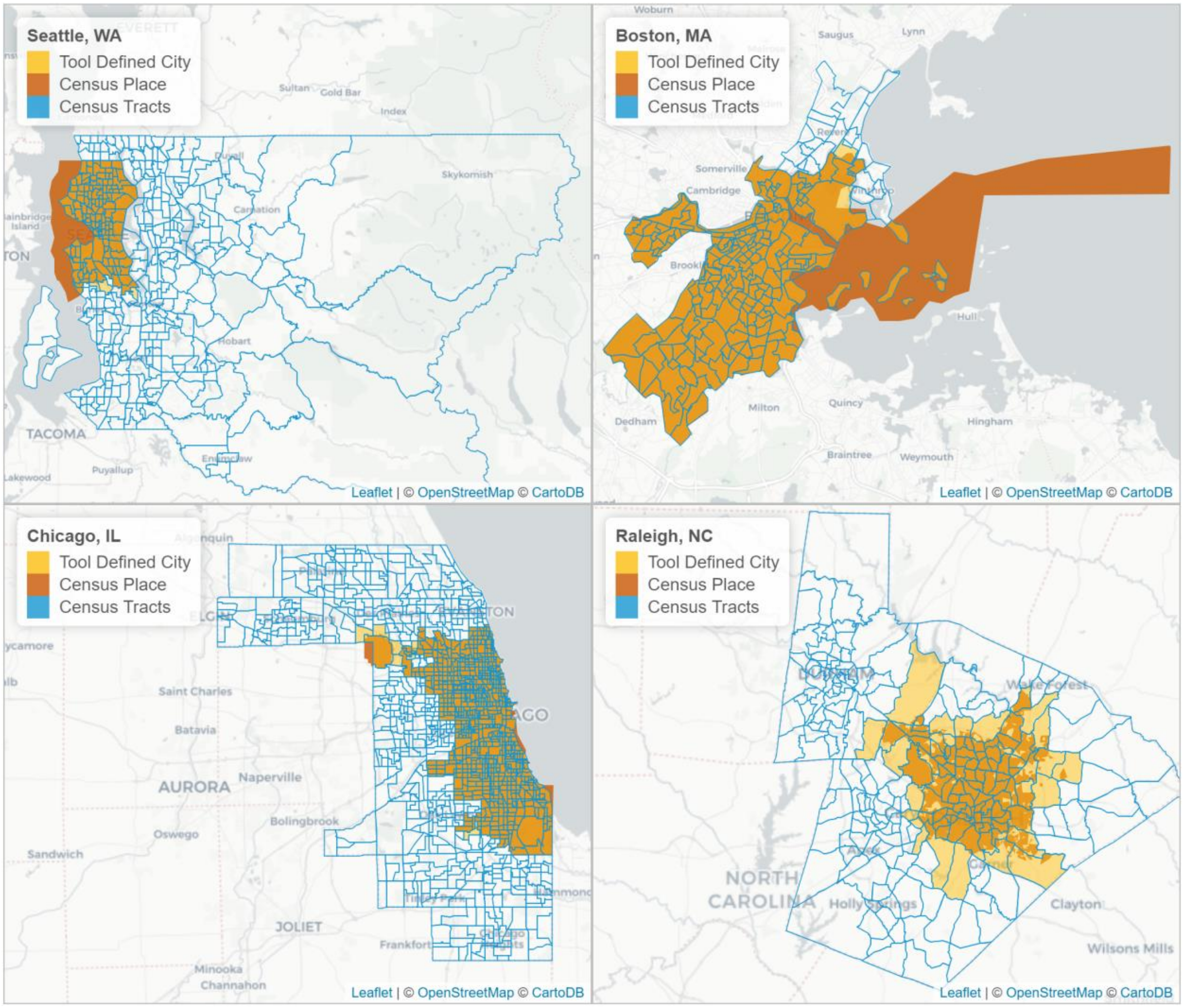
Luckily, this problem mostly disappears in large cities where the Census Place boundaries usually correspond with Census tract boundaries. Below are maps for all the large cities studied in this report, namely San Francisco, Detroit, Philadelphia, New York, Seattle, Chicago, Boston, and Raleigh. San Francisco, Detroit, Philadelphia, and New York show perfect correspondence between Census place and tract boundaries. Seattle, Chicago and Boston are mostly well behaved, with a few partially covered Census tracts within the place boundaries. Raleigh is the one exception and has much more irregular place boundaries which partially cover many tracts. As a result, the tool thinks Raleigh is larger than it actually is.

Figure B.2. Boundaries for San Francisco, Detroit, Philadelphia, and New York City



Source: Urban Institute analysis
Notes: Made with Leaflet | © OpenStreetMap contributors, © CartoDB. The darkest orange in the graph are bodies of water that are a part of the census place boundaries but are not assigned a census tract.

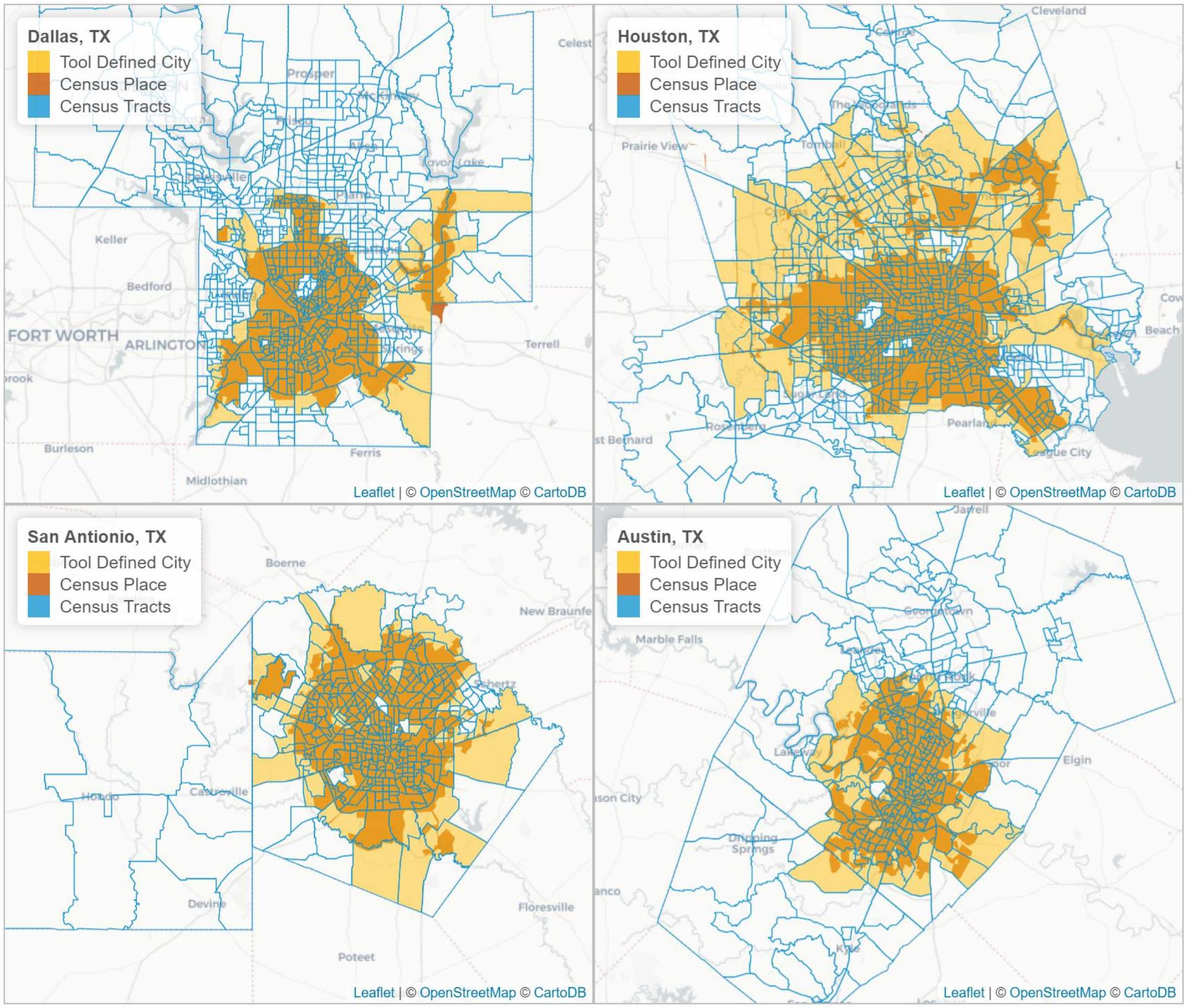
Figure B.3. Boundaries for Seattle, Boston, Chicago, and Raleigh



Source: Urban Institute analysis.
Notes: Made with Leaflet | © OpenStreetMap contributors, © CartoDB. The darkest orange in the graph are bodies of water that are a part of the census place boundaries but are not assigned a census tract.

One important exception to the trend of well-defined large cities are Texan cities. Below are the boundary maps for Austin, San Antonio, Houston and Dallas.

Figure B.4. Boundaries for Austin, San Antonio, Houston and Dallas



Source: Urban Institute analysis.
Notes: Made with Leaflet | © OpenStreetMap contributors, © CartoDB. The darkest orange in the graph are bodies of water that are a part of the census place boundaries but are not assigned a census tract.

Texas cities have very irregular place boundaries, which causes the tool to think these cities are a lot bigger than they actually are. Extra caution should be applied when using our tool on data from Texas cities.

Appendix C. Cities Covered by Our Tool

To minimize the storage and computing costs for our application, we only allowed the tool to store data on places with populations greater than 50,000 as measured by the 2011–16 five-year American Community Survey. Below are the four largest and four smallest places by population that the tool will work for so readers can get a sense of what cities are covered by our tool. If the uploaded data come from a city with less than 50,000 population, the tool will not work.

Table C.1. Largest and Smallest Census Places That Work with Our Tool

Place GEOID	Place name	Population (2015)
3651000	New York city, New York	8,426,743
0644000	Los Angeles city, California	3,900,794
1714000	Chicago city, Illinois	2,717,534
4835000	Houston city, Texas	2217,706
2701900	Apple Valley city, Minnesota	50,309
2670520	Saginaw city, Michigan	50,288
2642820	Kentwood city, Michigan	50,286
4173650	Tigard city, Oregon	50,276

Sources: Place names and GEOID pulled from https://www2.census.gov/geo/docs/reference/codes/files/national_places.txt. Population data pulled from 2011–16 American Community Survey

Appendix D. American Community Survey Data

We use the 2012–16 five-year ACS Data Profile Estimates as the source for socioeconomic information on every census tract in our list of cities. The Data Profile provides broad social, economic, housing, and demographic data at the level of the Census tract. For more information on the Data Profile, visit https://www.census.gov/ acs/www/data/data-tables-and-tools/data-profiles/2016/. The variable ID's and labels for each of the variables we analyzed are included in the table below. It is important to note that in addition to the estimates, we also pull the margins of error (MOE) for each variable. And with the exception of the total population variable, all variable estimates and margins of error are reported in percentages. It would have also been possible to use the ACS detail tables to calculate each of these percentages, but it would have required pulling far more variables as the denominators for many of the percentage based variables listed below are different. For example, the percent of people with a bachelor's degree or higher is computed in relation to all adults 25 or older while the unemployment rate is computed in relation to those in the labor force. So for ease of use and ease of sampling (discussed in detail in Appendix X) we elected to use the percentage based ACS Data Profile estimates. For the full list of 2500 Data Profile variables that are available, one can visit.

https://api.census.gov/data/2016/acs/acs5/profile/variables.html

Table D.1. Five-Year ACS (2012–16) Variables Used

Variable Name	Variable Label	Name in report
DP02_0086E	Estimate!!PLACE OF BIRTH!!Total population	Population
DP02_0067PE	Percent!!EDUCATIONAL ATTAINMENT!!Percent bachelor's degree or higher	Bachelors Degree or Higher
DP03_0119PE	Percent!!PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL!!All f.amilies	Family Poverty rate (last 12 months)
DP03_0009PE	Percent!!EMPLOYMENT STATUS!!Civilian labor force!!Unemployment Rate	Unemployment Rate
DP05_0072PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!White alone	White non-Hispanic
DP05_0073PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!Black or African American alone	Black non-Hispanic

DP05_0074PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!American Indian and Alaska Native alone	(part of other race non-Hispanic)
DP05_0075PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!Asian alone	Asian non-Hispanic
DP05_0076PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!Native Hawaiian and Other Pacific Islander alone	(part of other race non-Hispanic)
DP05_0077PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!Some other race alone	(part of other race non-Hispanic)
DP05_0078PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Not Hispanic or Latino!!Two or more races	(part of other race non-Hispanic)
DP05_0066PE	Percent!!HISPANIC OR LATINO AND RACE!!Total population!!Hispanic or Latino (of any race)	Hispanic

Source: Five-year ACS (2012–16) Variable Documentation, https://api.census.gov/data/2016/acs/acs5/profile/variables.html
Notes: When geo var = 1, this means a geographic variable was present in the dataset

Appendix E. Sampling Procedure for Assessing Significance

Significance of Spatial Bias Metrics

The spatial bias metric is defined as the proportion of the data originating in a tract (the data proportion) minus the proportion of the city's population living in a tract (the population proportion). The data proportion is a fixed number since the tool only evaluates one dataset at a time. However the population in each tract (and therefore the population proportion in each tract) is an estimate from the ACS 5-year survey, with associated margins of error. We developed a random sampling procedure that takes into account the variability in the population estimates and tell us whether the spatial bias metric is statistically significant—in other words if the data proportion in a tract is significantly different from the population proportion in that tract. The sampling procedure was as follows:

1. Generate new population samples for every tract in a city using the Census reported estimates and margins of error. We assume that the population estimates are normally distributed with mean equal to the reported estimate and standard deviation equal to reported margin of error (MOE) divided by 1.645 as Census default margins of error are calculated at the 90% confidence level. If the sampled population in a tract is negative (which happens in very few cases because we analyzed relatively large cities with large populations), we manually truncate the value to 0.

$$Population\ Sample_{tract\ i} \sim N\left(\mu = \ Pop\ estimate_{tract\ i}, \sigma = \frac{MOE_{tract\ i}}{1.645}\right)$$

2. Obtain population proportions for each tract by dividing the 'new' population in each tract by the total 'new' population across the city.

$$Population\ proportion_{tract\ i} = \ Population\frac{Population\ sample_{tract\ i}}{\sum_{\{j=1\}}^{\{j=\text{n}\}} Population\ sample_{tract\ j}}$$

3. Repeat 10,000 times to get 10,000 samples of population proportions in each tract
4. Create a 99.7% confidence interval for the sample population proportions for each tract. We chose the more stringent 99.7% (3 standard deviations from the mean) to be

sure that the vast majority of bias the tool reports as "statistically significant" does not represent noise. In other words, for practical reasons, we wish to minimize false positives.

For each tract, if the data proportion falls inside the confidence interval for the population proportion, we report the spatial bias metric for that tract as not statistically significant. In other words, the data proportion is not statistically different from the population proportion once the variability in the Census population estimates is taken into account.

Significance of Demographic Bias Metrics

Our demographic bias metric reports the percentage difference between a citywide average demographic statistic and the data implied average demographic statistic. For example, if we analyze the demographic bias of the share of black residents in a dataset, we can represent the calculation mathematically as:

$$Bias = \sum_{\{j=1\}}^{\{j=n\}} Data_{prop_{tract\ j}} * pct_{black_{tract\ j}} - \sum_{\{j=1\}}^{\{j=n\}} Pop_{prop_{tract\ j}} * pct_{black_{tract\ j}}$$

Data-implied average percent black Citywide average percent black

In this case the population proportion in each tract and the share of black residents in each tract are only estimates and have associated margins of error. In order to take into account the variability around Census estimates in assessing the significance of our demographic bias statistic, we use the following sampling procedure:

1. Generate new random population samples for each tract in the city assuming that the population estimates are normally distributed with mean equal to the reported estimate and standard deviation equal to reported margin of error (MOE) divided by 1.645.

2. Use these samples to generate new population proportions for each tract by dividing the 'new' population in each tract by the total 'new' population across the city.
3. Generate new samples of the demographic percentage of interest—in this example the share of black residents in each tract. We again assume that all estimates are normally distributed with mean equal to the reported estimate and standard deviation equal to reported margin of error divided by 1.645. If any of the population estimates or other demographic statistics generated are negative, we manually truncate them to 0.
4. Generate new samples of the data implied averages. To do this, we take the (constant) data proportion in each tract and multiply it by the newly sampled demographic percentage in each tract, then sum across all tracts.
5. Generate new samples of the citywide averages. To do this, we take the newly sampled population proportions in each tract and multiply it by the newly samples demographic percentage in each tract, then sum across all tracts.
6. Repeat 10,000 times to times to get 10,000 samples of the citywide average and data implied averages.
7. Take the difference between the data implied averages and the citywide average to get 10,000 samples of the demographic bias statistic
8. Create a 99.7% confidence interval for the demographic bias statistic samples that contains the middle 99.7% of the data.

For each demographic bias statistic of interest, we conduct a significance test—we check if the confidence interval contains 0. If 0 falls within the confidence interval, we say that the demographic bias is not statistically significant. In other words, the data implied average demographic statistic is not statistically different from the citywide average statistic and there is no evidence of demographic bias in the user provided dataset after taking into account the variability in the Census provided statistics. If 0 doesn't fall within the confidence interval, we call the resulting bias statistically significant.

It is also important to note that this random sampling approach is *not* the Census approved method for calculating standard errors of derived estimates like the population proportion or the share of black residents in a tract. The Census provided formula[1] for calculating the standard error of a proportion P - where P is A/B - is as follows:

$$SE(P) = \frac{1}{B}\sqrt{SE(A)^2 - P^2 * SE(B)^2}$$

In the case of the population proportion, the numerator A is the population estimate of a tract and the denominator B is the sum of population estimates across all Census tracts in a city. The Census provided formula for calculating the standard error of a sum of estimates is as follows:

$$SE(S_1 + S_2 + \cdots) = \sqrt{SE(S_1)^2 + SE(S_2)^2 + \cdots}$$

Combining these formulas, we could obtain standard errors for the population proportions for each tract and calculate the respective 99.7% confidence intervals without having to go through the random sampling procedure. But the Census approach has two primary drawbacks:

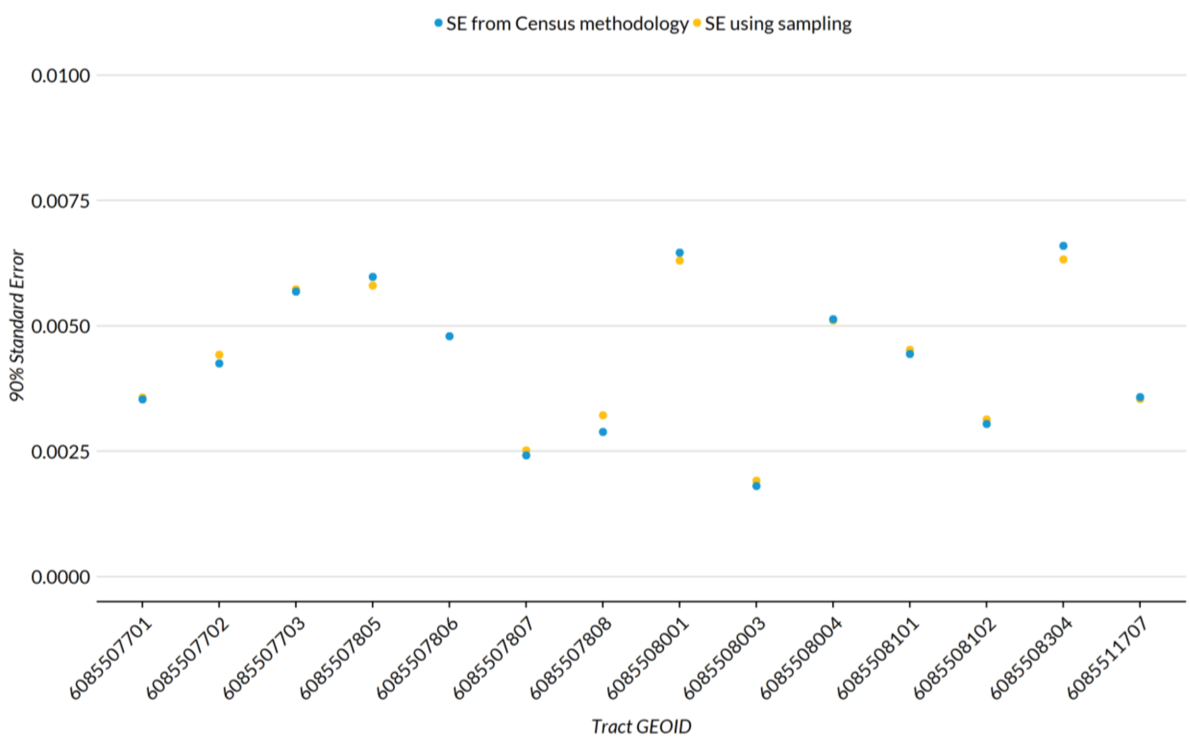
1. It's not clear how to take into account variation in multiple Census estimates. For example, when assessing the significance of our demographic bias metric, we need to take into account both the variation in the population proportion and the variation in the demographic statistic in each tract. Our random sampling approach makes this easy by leveraging the fact that we can sample the total population independently of the demographic percentage of interest.
2. It's not computationally efficient. As mentioned in Appendix C, most of our demographics statistics of interest have different denominators—for example the percent of people with a bachelor's degree or higher is computed in relation to all adults 25 or older while the unemployment rate is computed in relation to those in the labor force. And some of these denominators are really sums over other

[1] https://www2.census.gov/programs-surveys/acs/tech_docs/statistical_testing/2016StatisticalTesting5year.pdf?

Census variables. Using the Census approach, we would have to store and compute standard errors for all the variables that make up each of the numerators and denominators in each tract. With our normal sampling approach, we only need to take two normal random samples of the population and of the demographic statistic of interest in each tract.

To confirm our random sampling approach aligned with the Census approved methodology, we computed standard errors for the population proportions of all Census tracts in the small town of Cupertino, CA using both our random sampling approach and the Census formula approach. The standard errors in both cases were almost identical. Below is a graph of the standard errors of the population proportion in each Census tract computed both ways.

Figure E.1. Comparing Standard Errors of Population Proportions



Appendix F. Citywide vs Data-Implied Demographic Averages

Figure F.1. Bikeshare Stations: Citywide vs Data Implied Demographic Averages

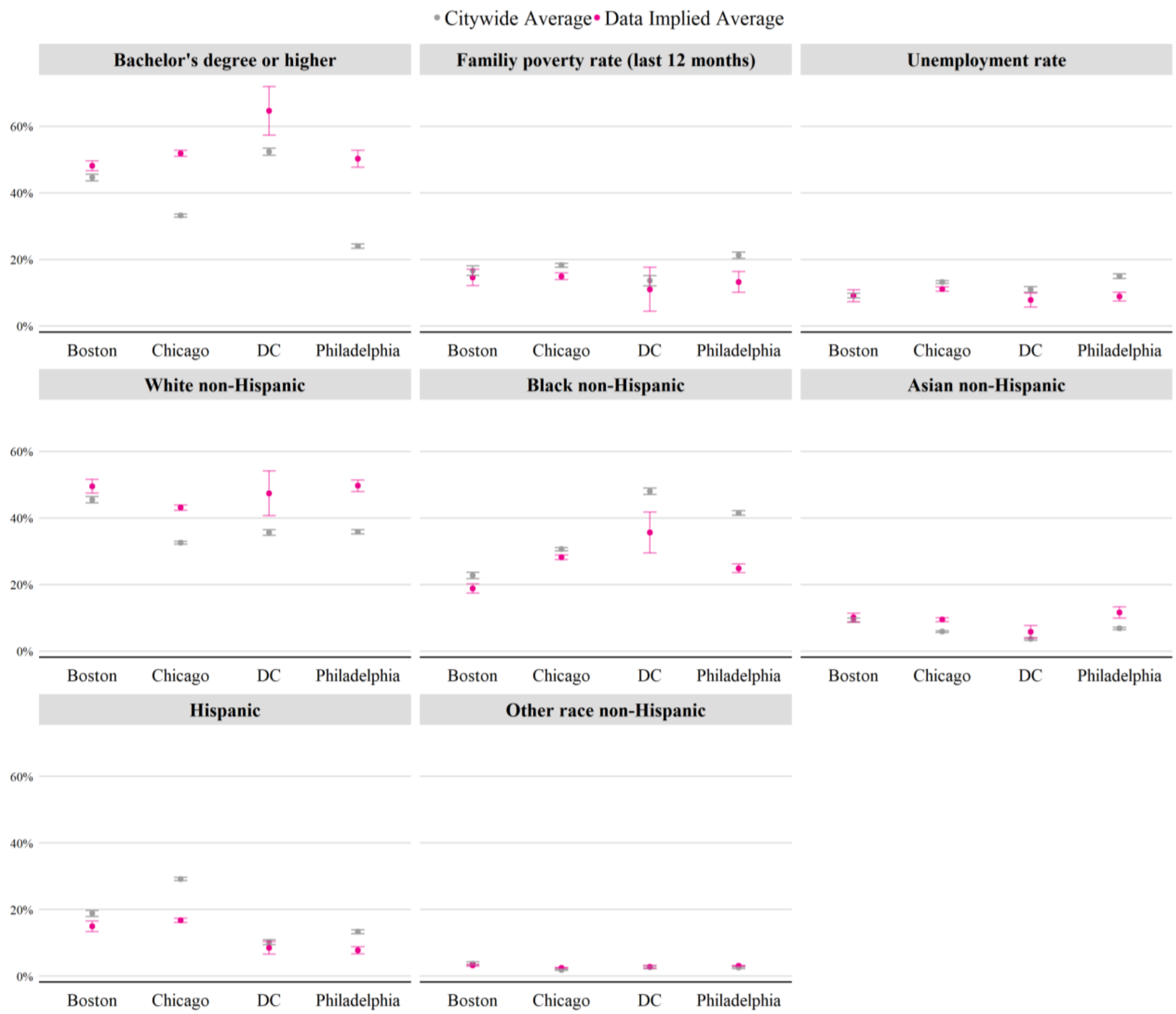


Figure F.2. 311 Requests: Citywide vs Data Implied Demographic Averages

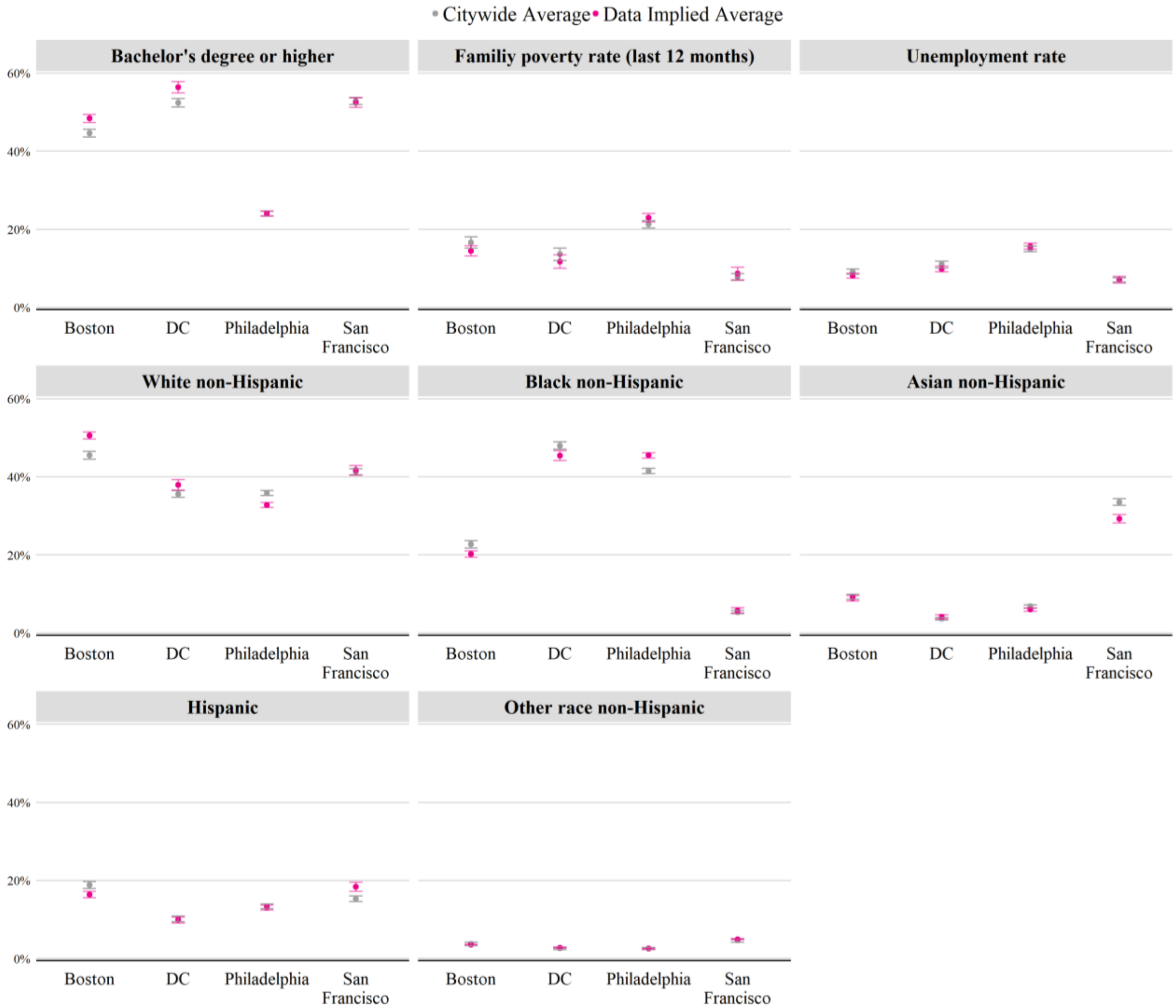
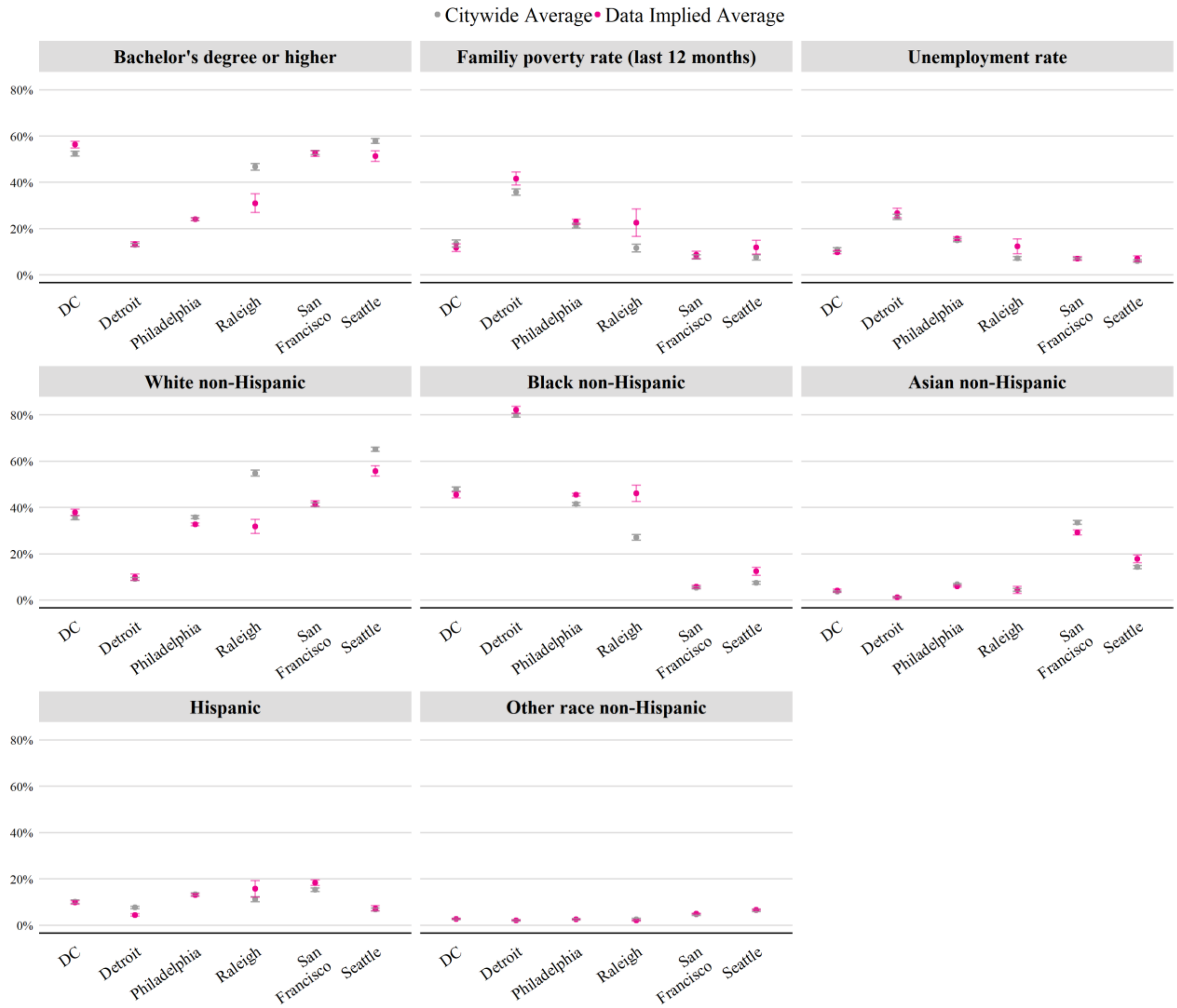


Figure F.3. LIHTC Buildings: Citywide vs Data Implied Demographic Averages



References

Angwin, Julia, et al. “Machine Bias.” *ProPublica*, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

“Bike Share in the U.S.: 2017.” *NACTO Bike Share Initiative*, 2017, National Association of City Transportation Officials, nacto.org/bike-share-statistics-2017

Cavallo, Sara, Joann Lynch, and Peter Scull. “The digital divide in citizen-initiated government contacts: A GIS approach.” *Journal of Urban Technology* 21.4 (2014): 77-93.Keightley, Mark P. “An Introduction to the Low-Income Housing Tax Credit.” *Congressional Research Service, February* 12 (2013).

Kontokosta, Constantine, Boyeong Hong, and Kristi Korsberg. “Equity in 311 reporting: Understanding socio-spatial differentials in the propensity to complain.” *arXiv preprint arXiv:1710.02452* (2017).

Polonetsky, Jules, Omer Tene, and Kelsey Finch. “Shades of Gray: Seeing the Full Spectrum of Practical Data De-Intentification.” *Santa Clara L. Rev.* 56 (2016): 593.

Piquero, Alex R. “Disproportionate minority contact.” *The future of children* (2008): 59-79.

Feigenbaum, James J., and Andrew Hall. “How high-income areas receive more service from municipal government: Evidence from city administrative data.” (2015).

O’Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

Nicholls, Sarah. “Measuring the accessibility and equity of public parks: A case study using GIS.” *Managing leisure* 6.4 (2001): 201-219.

Delmelle, Elizabeth Cahill, and Irene Casas. “Evaluating the spatial equity of bus rapid transit-based accessibility patterns in a developing country: The case of Cali, Colombia.” *Transport Policy* 20 (2012): 36-46.

Hall, Thad E., and Jennifer Owens. “The digital divide and e-government services.” *Proceedings of the 5th International Conference on Theory and Practice of Electronic Governance*. ACM, 2011.

Bélanger, France, and Lemuria Carter. “Digitizing government interactions with constituents: an historical review of e-government research in information systems.” *Journal of the Association for information Systems* 13.5 (2012): 363.

Kitchin, Rob. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014.

Cavallo, Sara, Joann Lynch, and Peter Scull. “The digital divide in citizen-initiated government contacts: A GIS approach.” *Journal of Urban Technology* 21.4 (2014): 77-93.

“City of Seattle Open Data Risk Assessment.” Future of Privacy Forum, Jan. 2018, https://fpf.org/wp-content/uploads/2018/01/FPF-Open-Data-Risk-Assessment-for-City-of-Seattle.pdf

Janssen, Marijn, Yannis Charalabidis, and Anneke Zuiderwijk. “Benefits, adoption barriers and myths of open data and open government.” *Information systems management* 29.4 (2012): 258-268.

Riggins, Frederick J., and Sanjeev Dewan. “The digital divide: Current and future research directions.” *Journal of the Association for information systems* 6.12 (2005): 4.

Smith, C. Scott, Jun-Seok Oh, and Cheyenne Lei. “Exploring the Equity Dimensions of US Bicycle Sharing Systems.” *Transportation Research Center for Livable Communities*. Western Michigan University Kalamazoo, MI, 2015.

Smoyer-Tomic, Karen E., Jared N. Hewko, and M. John Hodgson. “Spatial accessibility and equity of playgrounds in Edmonton, Canada.” Canadian Geographer/Le Géographe canadien 48.3 (2004): 287-302.

Talen, Emily. “The social equity of urban service distribution: An exploration of park access in Pueblo, Colorado, and Macon, Georgia.” *Urban geography* 18, no. 6 (1997): 521-541.

Ubaldi, B. (2013), “Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives”, OECD Working Papers on Public Governance, No. 22, OECD Publishing, Paris, https://doi.org/10.1787/5k46bj4f03s7-en

Zillien, Nicole, and Eszter Hargittai. “Digital distinction: Status-specific types of internet usage.” Social Science *Quarterly* 90.2 (2009): 274-291.