

Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity

November 2021

Why ethics and empathy in imputation?

Why impute disaggregated data?

- Many areas of concern – including wealth, taxes, health records, justice involvement – have substantial gaps where detailed racial and ethnic information is either largely missing or unavailable in the otherwise strongest data sources
- Race-disaggregated data is key to identify long-standing racial disparities and develop policies and programs that address those disparities

Changemakers need race-disaggregated data

- Government agencies have made big commitments to creating and using race-disaggregated data
 - Biden Administration Executive Order for Racial Equity (e.g. equity reviews of programming by federal agencies)
 - Local governments creating Equity Offices and equity data and assessment standards
- Advocates and movement leaders need better data to prioritize and organize efforts
- Funders have committed at least \$50 billion to racial equity efforts in the last year.

Policymakers use different approaches to fill data gaps

- Directly collecting new race and ethnicity data
 - Expensive, time consuming, and difficult/impossible in some cases
- Imputing missing values on existing race and ethnicity variables
 - Long-standing practice, used by many agencies on wide variety of data
- Generating new race and ethnicity variables on existing data
 - Growing body of methods can increase accuracy, recently used by several agencies for equity analysis (e.g. CFPB, EEOC, HHS, etc.)

Standards and Recommendations

What do we mean by ethics and empathy?

■ Ethics

- Balancing uses of disaggregated data with potential harms
- Minimizing risks associated with disaggregation
- More just, equitable distribution of resources

■ Empathy

- Thoughtfulness in engaging with and responding to communities represented in the data
- Acknowledgement of the personhood of individuals and their experiences
- Agency - efforts to make data available and useful to communities represented in the data

Asking whether imputation is the right approach

- Do the potential benefits of imputation outweigh the risks?
 - **Opportunity cost:** Would resources for imputation be better used to improve data collection? What is the next-best available data?
 - **Fitness for purpose:** How will the imputed data be used? Does the available data support the use case?
 - **Outcomes:** How do the applications advance racial equity? (e.g. identifying disparities, case-making and corrective action, targeting resources)

Standards and Recommendations Guide

- *Whether* imputation is the right approach for disaggregating data for a given use case
- *Who* should be involved in the process for review and accountability – with a particular emphasis on community partners
- *How* to develop community-led standards for data sharing that protect privacy and harm from use by bad actors

Standards for Imputing Data Disaggregated by Race and Ethnicity: Relevance, Interpretability, Coherence, Accuracy, Privacy, and Institutional Environment

Standard	Description of standard for evidence	Recommended actions for analysts
Relevance	<p>The degree to which the disaggregated data meet the needs of the user in regard to data level, timeliness, etc.</p> <p>How suitable the data are for the research purposes.</p>	<p>Be clear about use cases. Who needs these imputed data and why?</p> <p>Using accessible language, be honest and transparent about the benefits and weaknesses of your imputation or data integration effort. It's possible that the imputation solved a critical missing data problem, or that it's a short-term patch in a situation where additional on-the-ground data collection is what is needed most.</p>
Interpretability	<p>The clarity of information to ensure that the data are used appropriately.</p>	<p>Be explicit about how the data are represented, what model was used, and why. Consider how the data might be limited or biased.</p> <p>Ensure that the data do not reinforce or enhance any biases. Be intentional about not just reporting what the results are but also describing why they are what they are, which can help researchers better understand the underlying or contributing racialized problems or challenges.</p> <p>Accurately calculate and clearly communicate uncertainty of any estimates derived from imputed data.</p>
Coherence	<p>Making sure the datasets used in the imputation represent the same population as the dataset being imputed.</p>	<p>Describe in detail the reasons for selecting the data source being used to append race and ethnicity data. Be clear about how the data source is similar to or different from the dataset of interest and why it was selected to draw on to impute missing race and ethnicity values.</p> <p>Also provide detail about how the data were imputed, clarifying how the imputation process using these data fit the intended purpose of the imputation.</p>

Standards for Imputing Data Disaggregated by Race and Ethnicity: Relevance, Interpretability, Coherence, Accuracy, Privacy, and Institutional Environment

Accuracy

The closeness of the imputed data to their unknown true values.

Use data analysis best practices, including making every effort to be rigorous and careful and documenting each step and decision point used in the imputation process. Make the methodological approach available for other analysts and researchers to review.

At each step of the imputation process, examine the potential for methodological choices to produce inaccurate results—including differential accuracy by racial and ethnic group.

Mitigate risks of inaccuracy where possible, communicate any unmitigated risk, and engage community in determining when unmitigated risk warrants terminating the imputation process.

Benchmark estimates calculated from the imputed data against trusted statistics disaggregated by race and ethnicity. Estimates can also be benchmarked against the same dataset used to create the imputation, but at different levels of aggregation (e.g., national versus state).

Standards for Imputing Data Disaggregated by Race and Ethnicity: Relevance, Interpretability, Coherence, Accuracy, Privacy, and Institutional Environment

Standard	Description of standard for evidence	Recommended actions for analysts
Privacy	The ability of the imputed data and any output or analyses that could come from the data to preserve the privacy and not risk identifying people.	<p>Establish clear data governance so ownership and management of the data (and any sensitive information that may be part of the imputation, particularly surnames) so that data are preserved and not seen or handled by anyone without proper protocols and protections in place.</p> <p>Establish guardrails so outputs and results won't risk exposing anyone's identity or otherwise violating privacy.</p> <p>Work with community members from smaller and less data-visible groups to agree on approaches to balance releasing information that is helpful for providing information on disparities or policy impact, while minimizing individual-level and community-level privacy risks and other potential harms.</p>
Institutional environment	The credibility of the researchers for producing high-quality and reliable data, and quality standards in the organization for assuring adherence to ethical practice.	<p>Build in accountability throughout the process—both on the technical side and the potential (policy/ethical) impact side. As part of this, develop a structure and process for consensus building and decisionmaking, which can include deciding not to do the imputation.</p> <p>With the aid of key stakeholders, develop a set of policies around release and restriction. Who gets access to the data when it's done and under what conditions?</p> <p>Data-privacy laws limit the processing of sensitive data categories including race, ethnicity, and sexual orientation. Data analysts should be aware of these restrictions and to the extent possible engage with their institutional review board or other ethical-protections entities to ensure the necessary steps are taken to ensure data privacy.</p>

Case Study: Ethically Imputing Race/Ethnicity on Credit Bureau Data

Ethical Imputation Case Study

- **Motivation:** Risk that using imputation to disaggregate by race/ethnicity can yield inaccurate, racially biased results
- **Goal:** Identify lessons learned for how researchers can think about incorporating equity into imputation
- **Method:** Case study imputing a combined race/ethnicity variable on 2013 credit bureau data using the zip code and age variables in data and American Community Survey (ACS) data

Checkpoint 1: before imputation,
audit input data for bias

Checkpoint 1:

- Does the dataset accurately represent the target population? How could structural racism drive unrepresentativeness?
- Do all datasets used in imputation represent the same population?
- Does missing data disproportionately affect certain groups?

Checkpoint 1: before imputation, audit input data for bias


Use datasets that most accurately represent target population, adjusting datasets as needed to align populations

Impute missing values with most accurate available data

Communicate impact of data limitations and recognize where data are not sufficient to support imputation

Checkpoint 2: during imputation,
examine where bias could be
introduced at each step

Imputation Methodology Overview

 **START:** zip code 12345, age range 1, race/ethnicity unknown 

 **END:** zip code 12345, age range 1, race/ethnicity implicates   

Gather data: collect ACS race and age range totals for an individual's zip code.

Randomly sample: take a sample to account for uncertainty in the ACS totals.

Rake counts: estimate the zip code's population by race and age range using a process called raking.

Adjust counts: scale the population counts downward to exclude the credit invisible population (a definition of this population is available on page 12).

Calculate probabilities: convert the counts into probabilities for each age range.

Assign race: randomly select a race value using probabilities for age range.

Checkpoint 2 Example: Rake Counts

Rake counts: estimate the zip code's population by race and age range using a process called raking.

X	X	X	~
X	X	X	~
X	X	X	~
~	~	~	

This step can be skipped for white and Hispanic groups as the ACS directly reports those estimates.

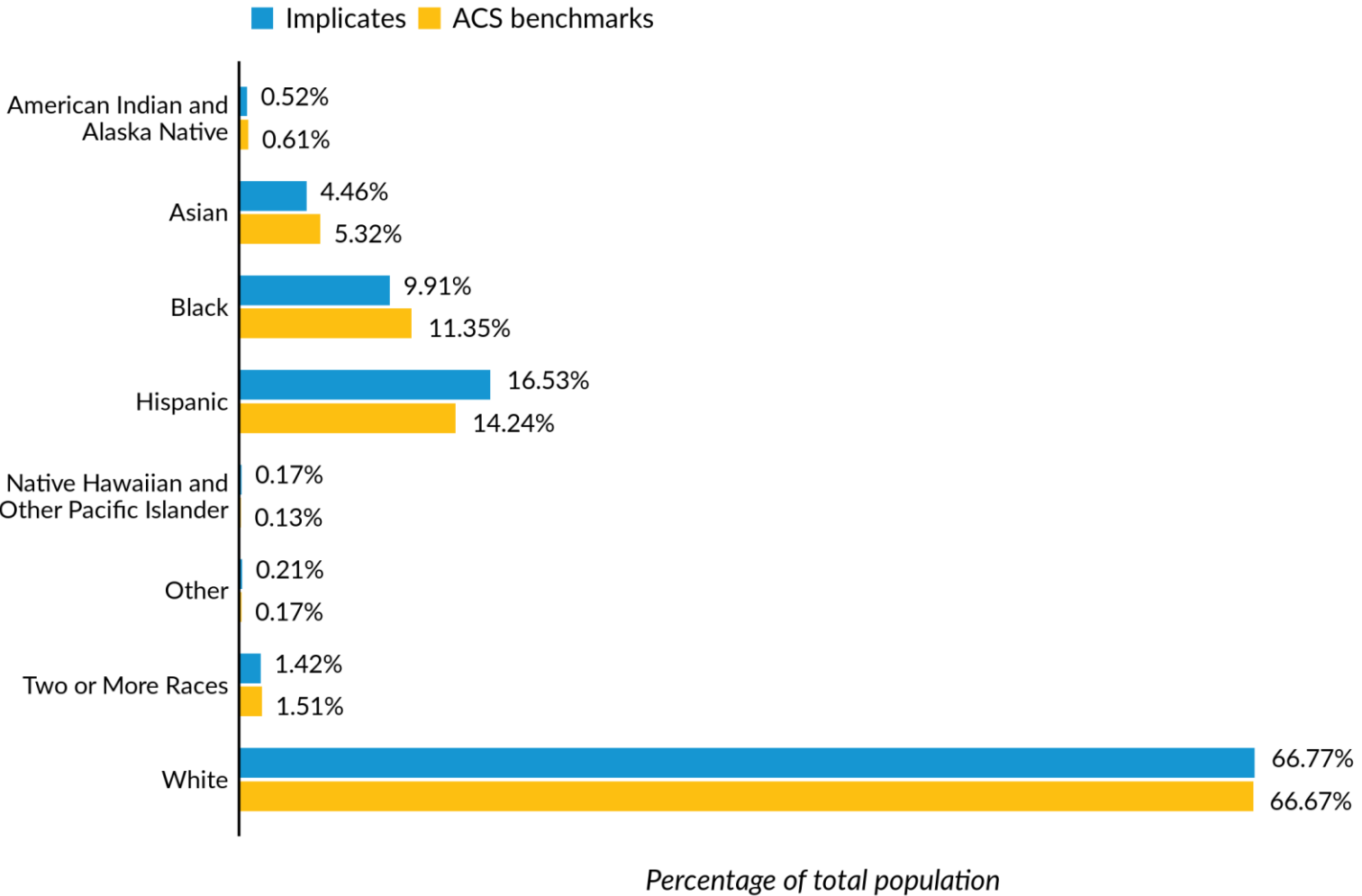
Summary of Checkpoint 2:

During imputation, examine where bias could be introduced at each step

- Assess impacts differentially
- Communicate potential limitations
- Capture uncertainty the best you can

Checkpoint 3: after imputation,
assess whether imputed data are
accurate enough to be used ethically
for your analytic purpose

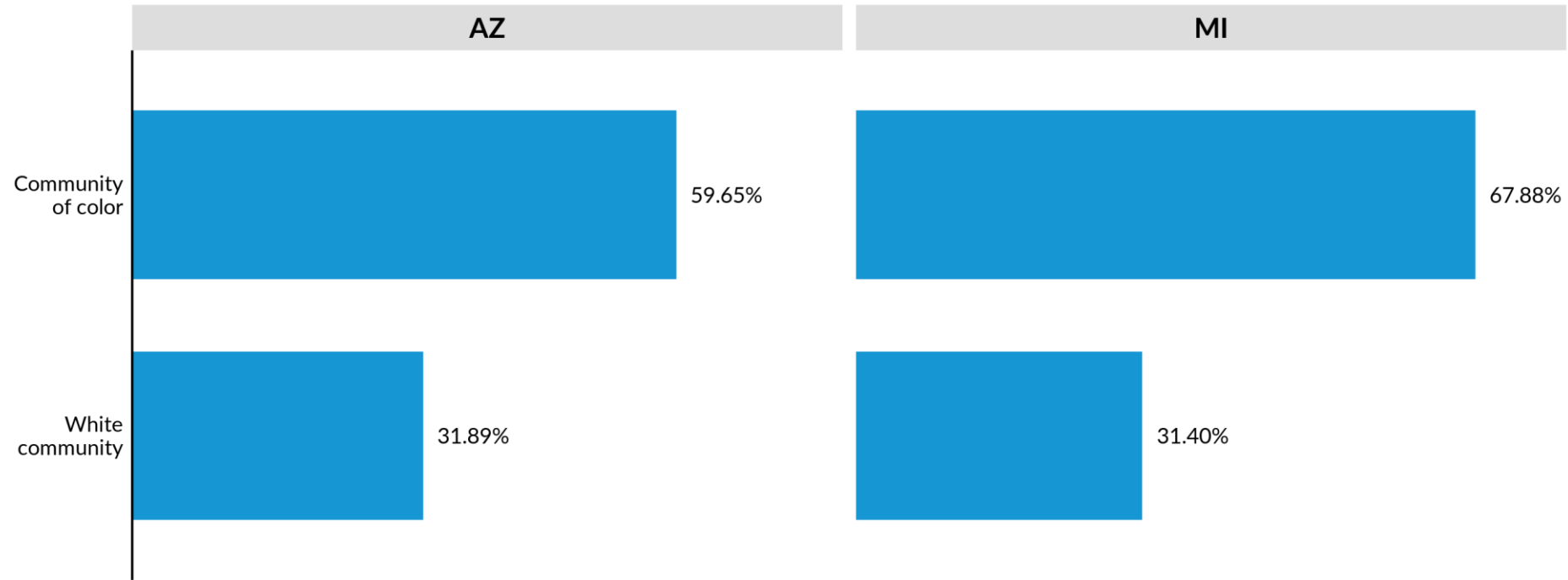
Checkpoint 3: benchmark against trusted aggregate statistics



Challenges + Potential Improvements

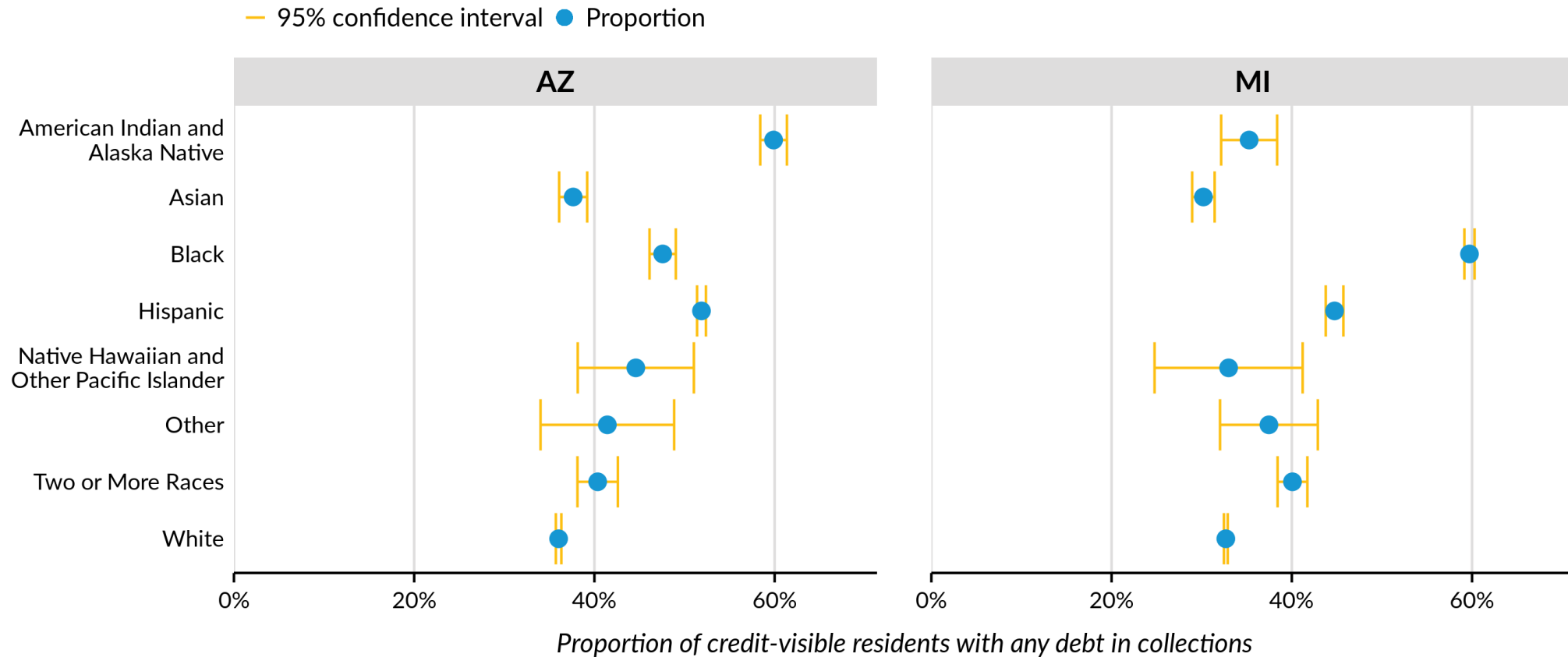
- Consistently underestimate Black and Asian populations and overestimate Hispanic and white populations
- Method is most accurate in racially homogeneous ZCTAs and less accurate in more racially diverse ZCTAs
- Raking is potential source of inaccuracy for non-Hispanic non-white groups, exploring alternate approaches
- Incorporate other data sources with information on financial outcomes by race (SCF, SIPP) into model and/or benchmarks

Checkpoint 3: Examine Fitness for Purpose



Proportion of credit-visible residents with any debt in collections

Checkpoint 3: Examine Fitness for Purpose



Key Takeaways

- Equity must be considered in every decision
- Examine differential outcomes by race and ethnicity
- Clearly communicate limitations (including accurately estimating margins of error of statistics)
- Examine fitness for purpose in context of specific analytic case

Spatial Equity Data Tool

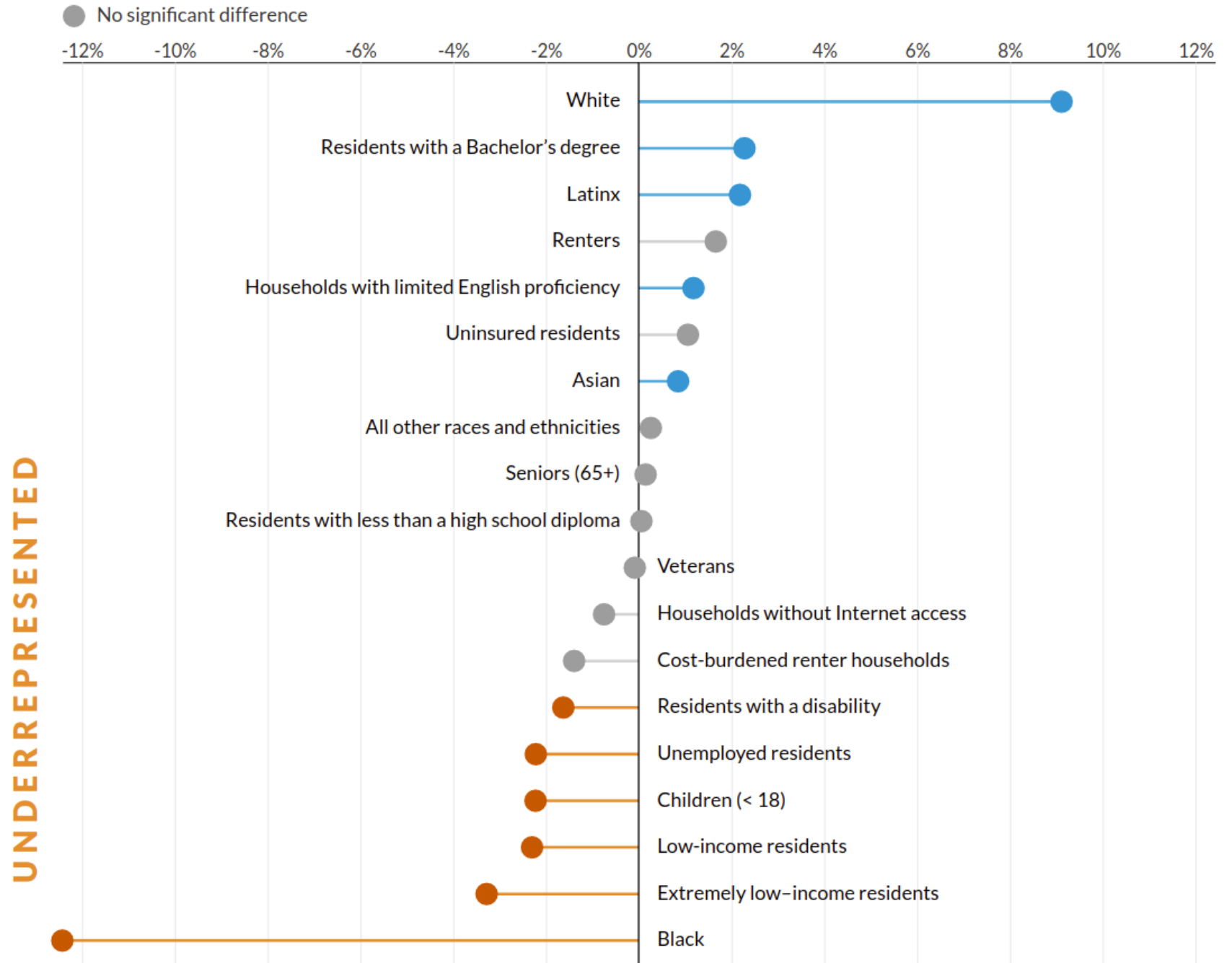
Spatial Equity Data Tool: What does it do?

- Automatically assesses racial, economic, and geographic representativeness of user uploaded point data
- Can be used with any geographic point data (ie datasets with lat/lon points)

What can it be used for?

- Equity in allocation of place-based programs or resources (e.g. grocery stores, bike share stations, tax credits, funding allocation)
- Examine representativeness of program participants
- Identifying areas for future investment (e.g. indicators of need, previous investment)

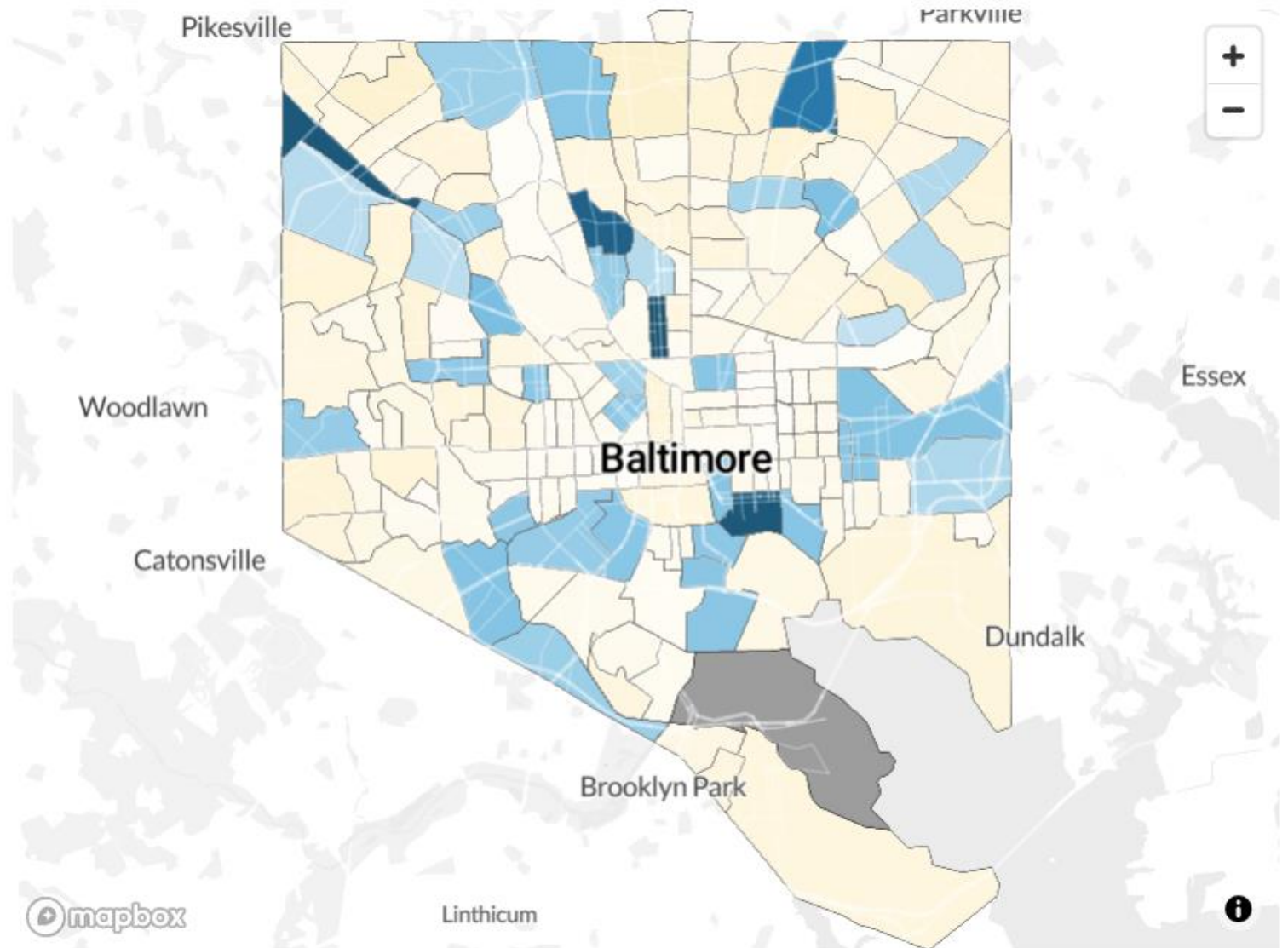
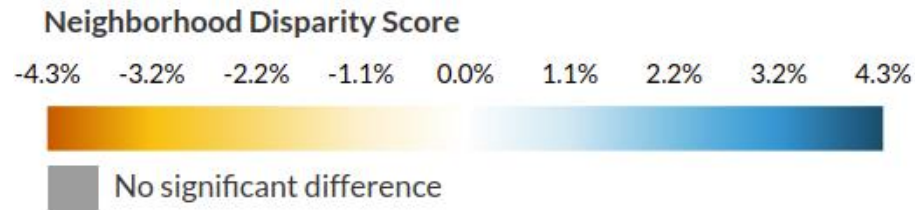
An example: Baltimore Grocery Stores



OVERREPRESENTED

UNDERREPRESENTED

An example: Baltimore Grocery Stores



Launching Tomorrow... Spatial Equity Data Tool Version 2!

- Expand to counties, states and national level data
- Add additional baseline datasets to compare data against

<https://apps.urban.org/features/equity-data-tool/>

