THE OHIO STATE UNIVERSITY

Developing Synthetic Individual-Level Population Datasets: The Case of Contextualizing Maps of Privacy-Preserving Census Data

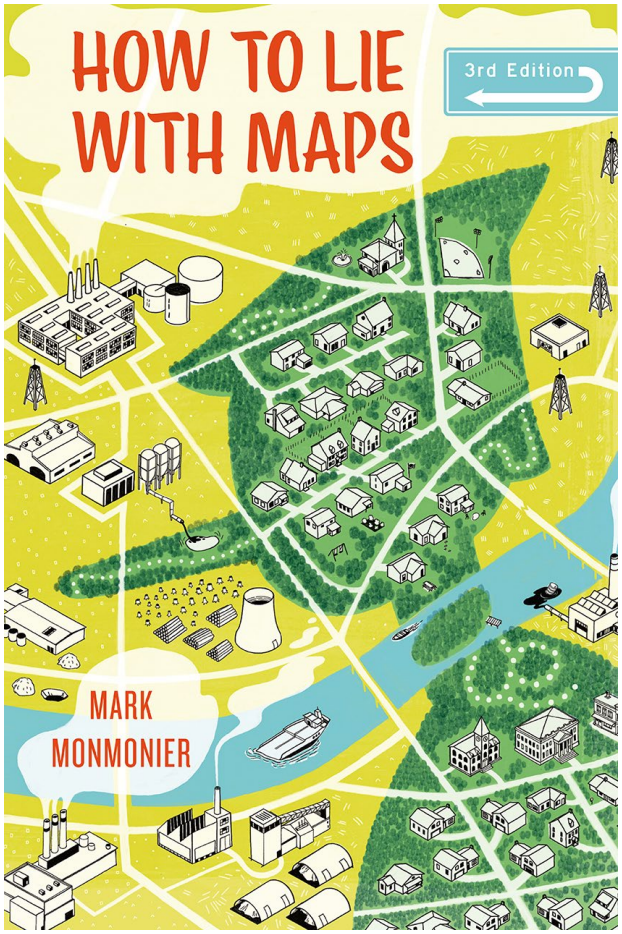
Yue Lin*, Ningchuan Xiao

Department of Geography

The Ohio State University

lin.3326@osu.edu

Maps as Social Constructions



Maps are **artifacts (instead of facts)** that must be interpreted in their social, cultural, and political contexts (Harley, 1989, 1990; Crampton, 2001).

Contextualizing mapmaking means understanding the contexts in which maps are created in order to interpret them appropriately.

Census Data and Differential Privacy

The New York Times

TheUpshot

To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data

Guaranteeing people's confidentiality has become more of a challenge, but some scholars worry that the new system will impede research.

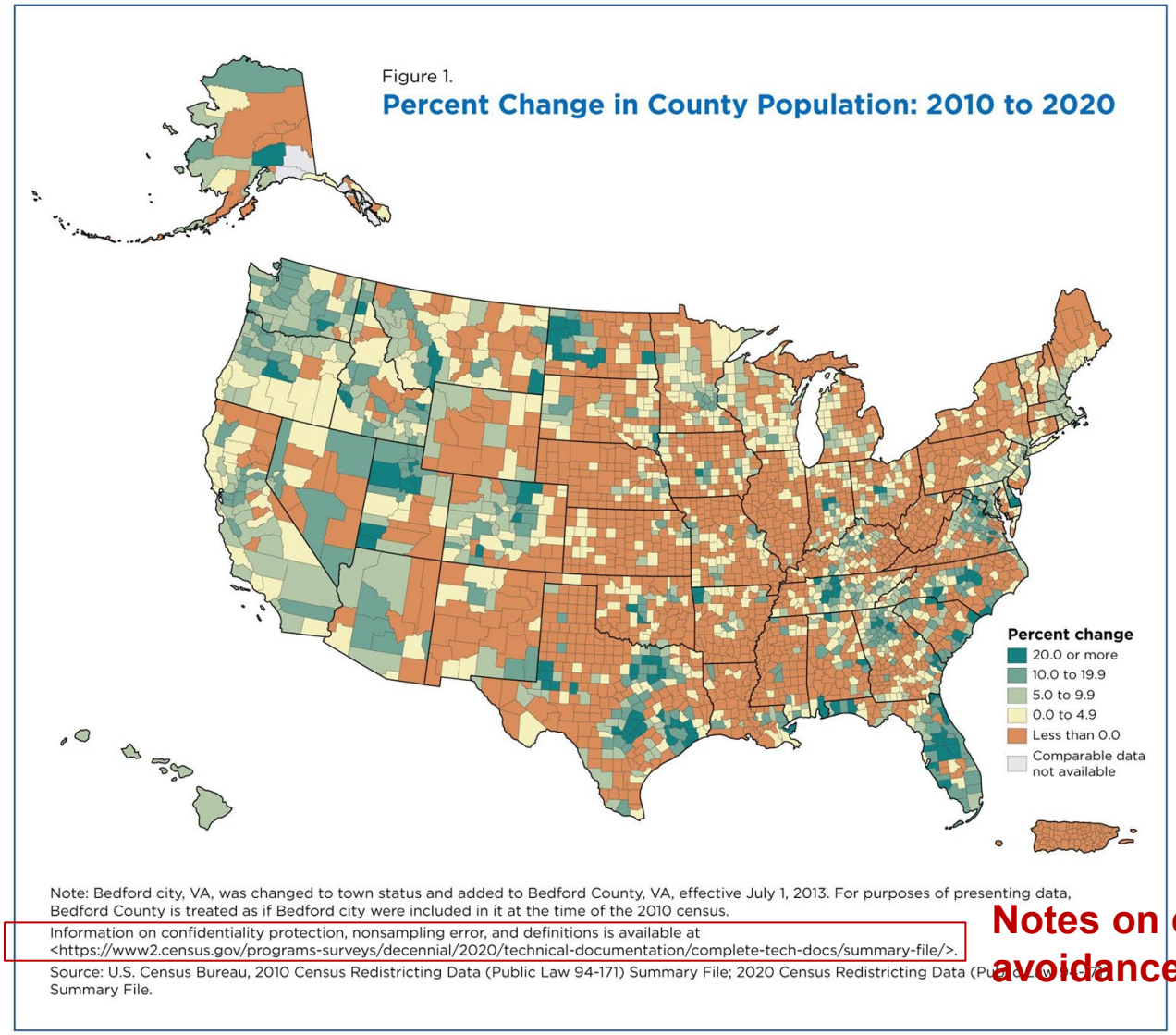
Give this article 37

The Census Bureau is bounded by Title 13 of the United States Code[1] not to "make any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

Statistical noise is added to census data prior to mapping to comply with privacy laws and regulations.

[1] 13 U.S.C. § 9

Impacts of Differential Privacy on Census Maps?



Notes on disclosure avoidance

Contextualizing Mapmaking Through Transparency

Transparency enables reproducibility, which allows map readers to fully explore how maps may distort the realities they present.

Key components to be transparent:

- Design
- Labeling
- **Data selection**
- Data slicing

However, many raw data are sensitive and are rarely made public, especially those that contain private information such as the **individual-level population data** (e.g., census microdata).

Bloomberg
US Edition

Sign In Subscribe

Live Now | Markets | Economics | Industries | Technology | Politics | Wealth | Pursuits | Opinion | Businessweek | Equality | Green | CityLab | Crypto | More

CityLab
Design

How to Detect the Distortions of Maps

All maps have biases. A new online exhibit explores the history of map distortions, from intentional propaganda to basic data literacy.

A map of population density in Tokyo, circa 1926, shows how maps splice and dice demographic data. *Leventhal Map Center, Boston Public Library*

By Laura Bliss
May 28, 2020 at 9:10 AM EDT

LIVE ON BLOOMBERG
Watch Live TV
Listen to Live Radio
Bloomberg Television

Research Objectives

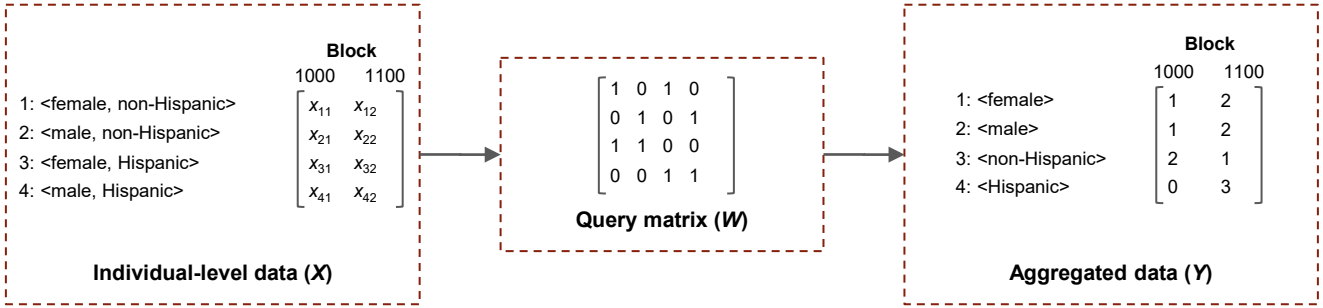
The objectives of this research are:

- To develop a realistic synthetic population dataset that is suitable for public use.
- To illustrate the use of synthetic population data for understanding cartographic processes and contextualizing cartographic artifacts.

Methods

Step 1: Collecting publicly available census tables as aggregated data.

Step 2: Creating matrix representations of individual-level (**X**) and aggregated (**Y**) data, as well as queries (**W**) used in data aggregation.



Step 3: Formulating an optimization problem to determine each element in **X:**

$$\min \quad ||\mathbf{WX}' - \mathbf{Y}||^2 ,$$

Objective: minimizing the squared difference between synthetic and actual census tables.

$$\text{subject to} \quad x'_{kj} \in \mathbb{Z}^* \quad \forall k, j ,$$

Constraint: ensuring integer decision variables.

Computational Experiments

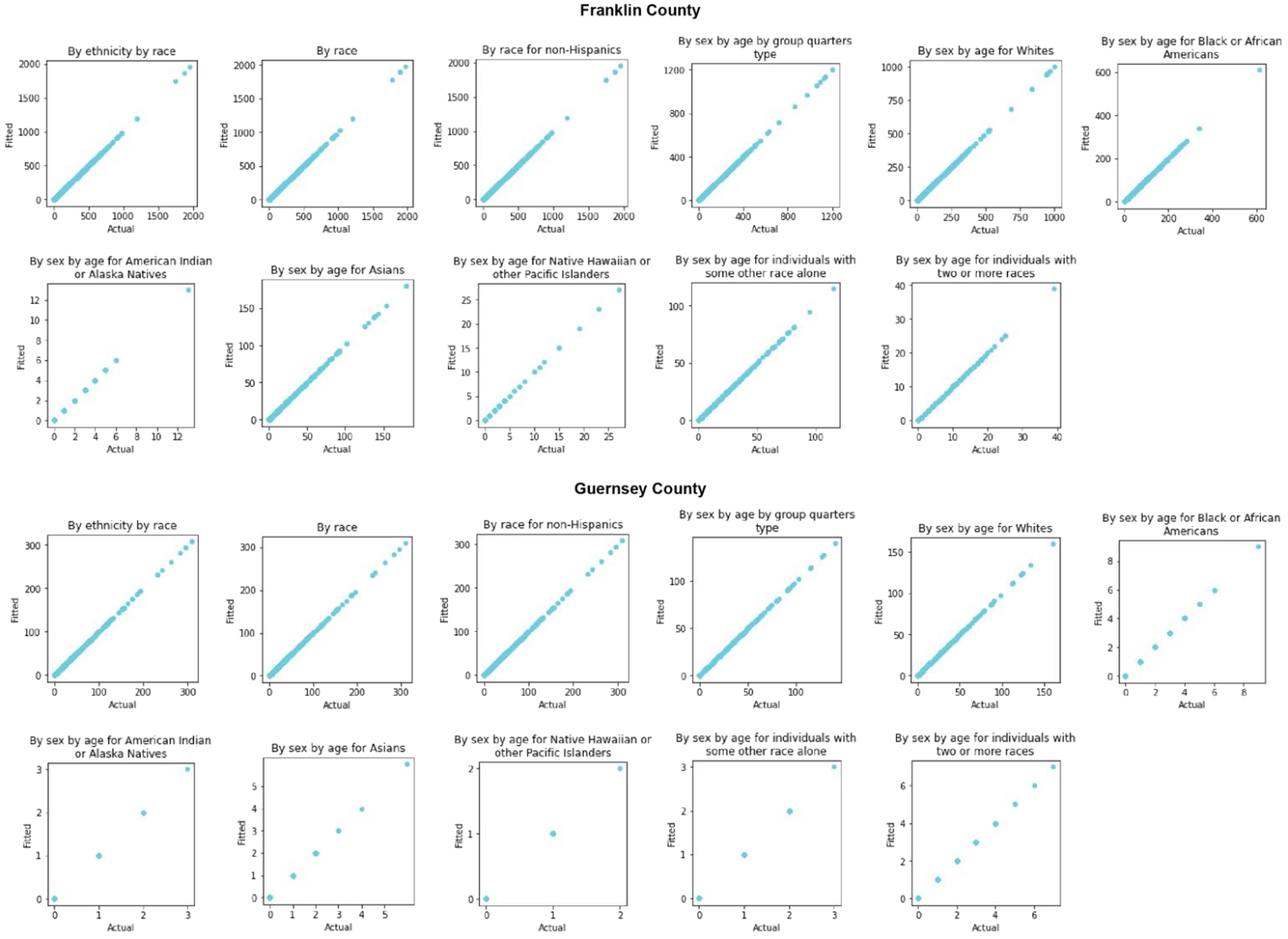
Data: Eleven census tables from the 2010 United States Census Summary File 1 (SF1).

- Two Ohio counties: Franklin and Guernsey.
- Five attributes: housing type, voting age, ethnicity, race, and sex.
- Population counts broken down by one or more of the five attributes at the census block level.

Table No.	Description	Number of columns
P5	Population counts broken down by ethnicity by race	126
P8	Population counts broken down by race	63
P9	Population counts broken down by race for non-Hispanics	63
P43	Population counts broken down by sex by age by group quarter types	28
P12A	Population counts broken down by sex by age for Whites	4
P12B	Population counts broken down by sex by age for Black or African Americans	4
P12C	Population counts broken down by sex by age for American Indian or Alaska Natives	4
P12D	Population counts broken down by sex by age for Asians	4
P12E	Population counts broken down by sex by age for Native Hawaiian or other Pacific Islanders	4
P12F	Population counts broken down by sex by age for individuals with some other race alone	4
P12G	Population counts broken down by sex by age for individuals with two or more races	4

Internal Validation

Compared with the original SF1 data.



Internal Validation

Compared with the original SF1 data.

Table P5

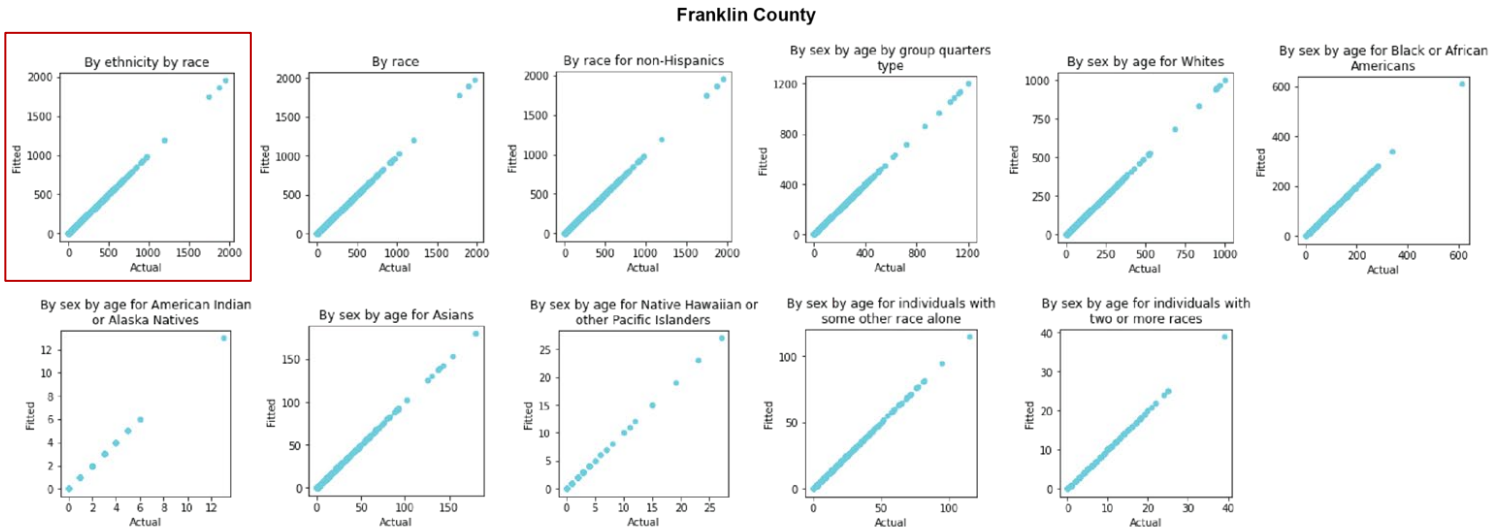
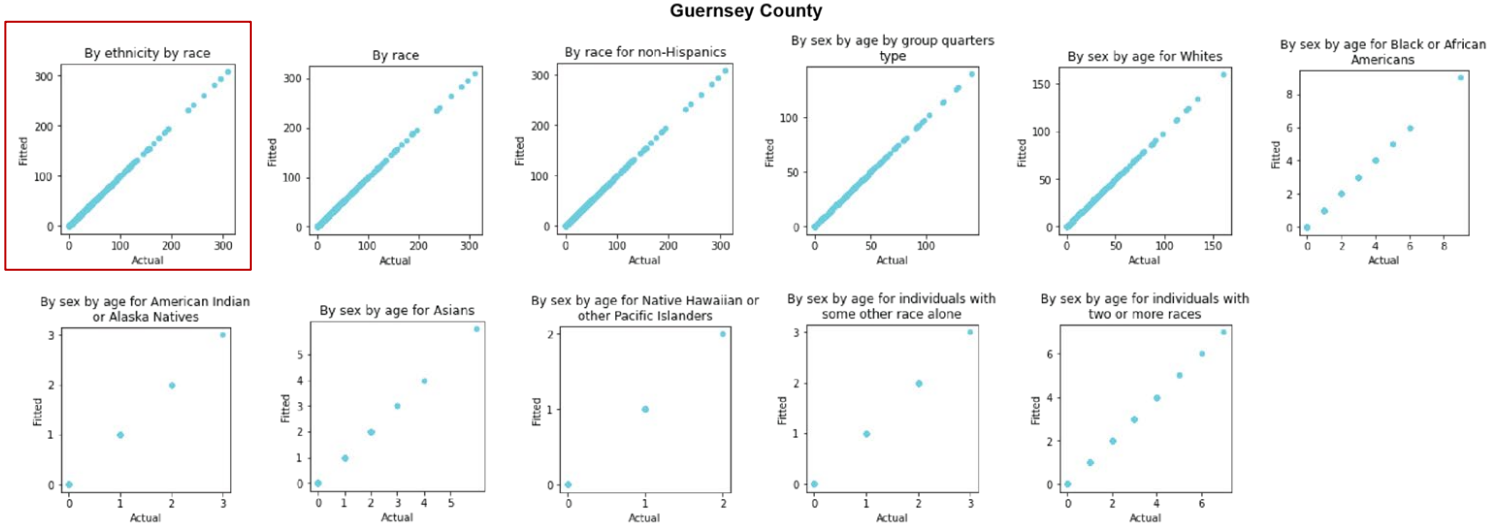
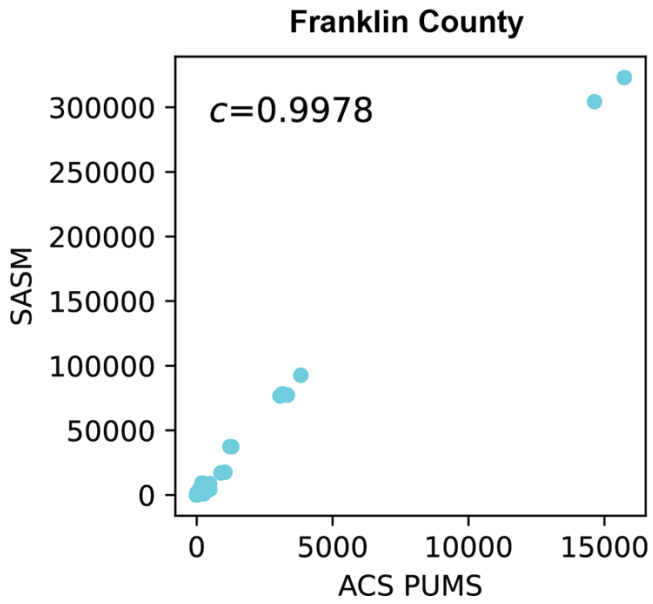


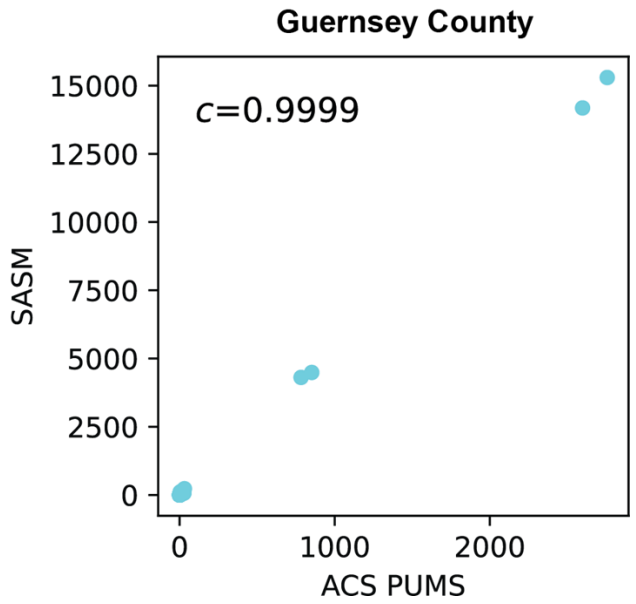
Table P5



External Validation

Compared with an external data source known as the American Community Survey Public Use Microdata Sample (ACS PUMS).



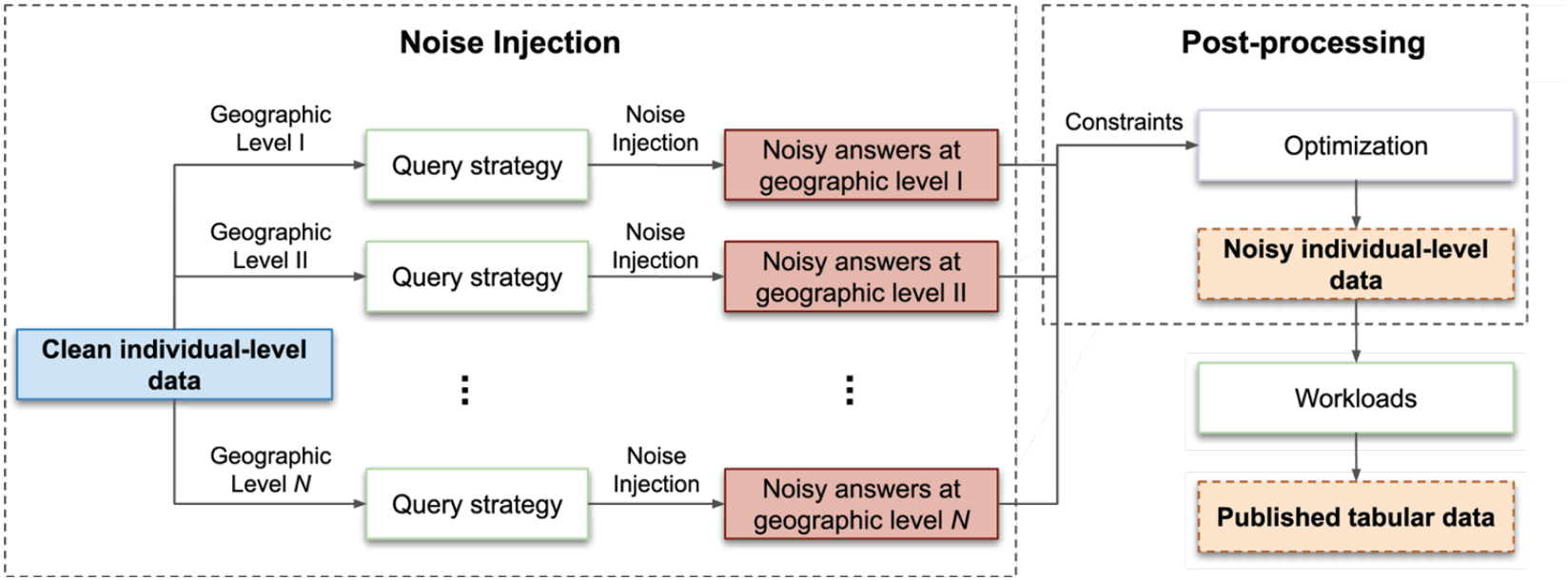


Case Study

Purpose: Contextualize census racial maps with the synthetic population data.

Methods:

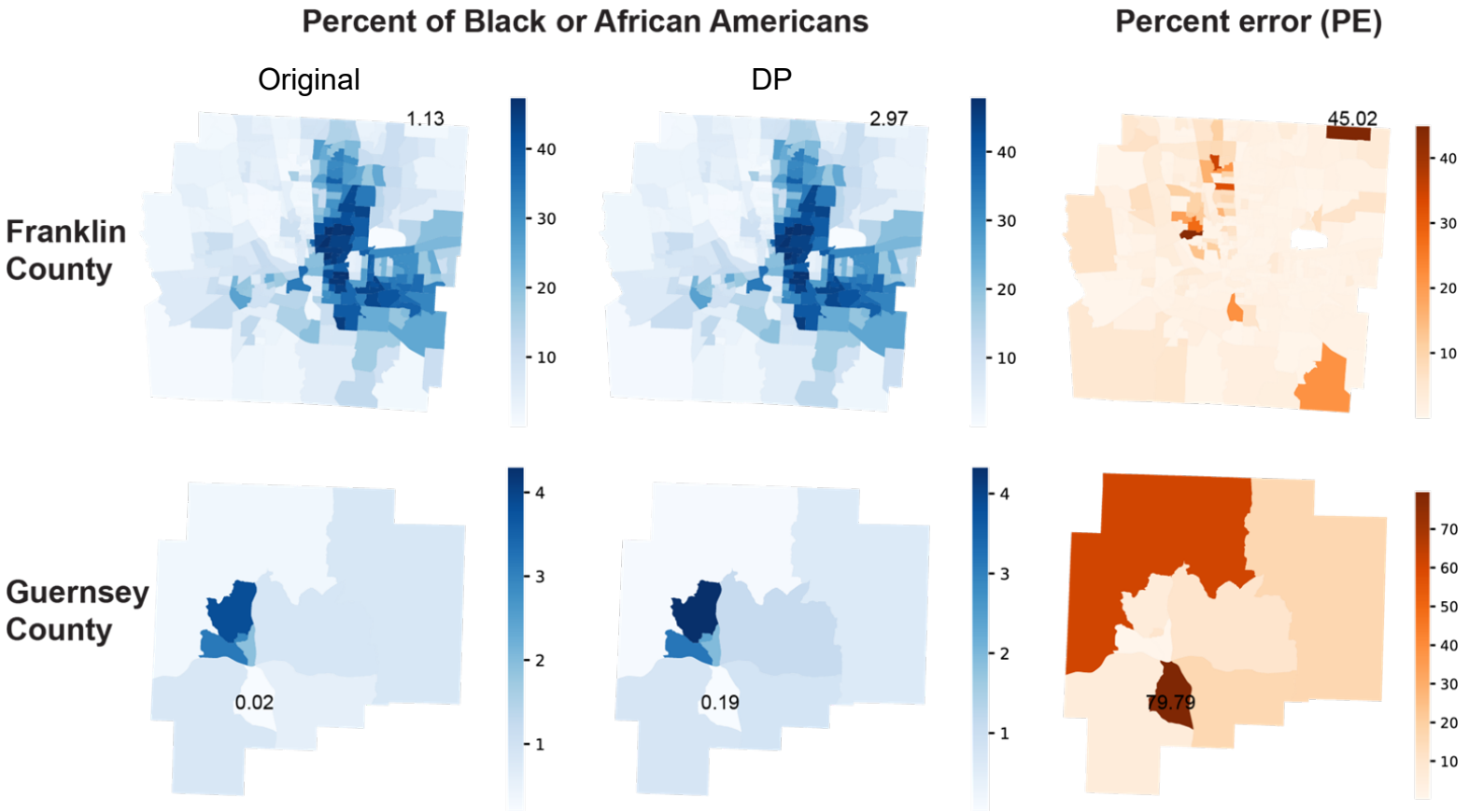
- Step 1: Implement the differential privacy (DP) mechanism on the synthetic population data using the computer source code released by United States Census Bureau (2020).
- Step 2: Compare the original and the differentially private synthetic population data.



Findings

How do we interpret census maps with privacy-preserving data?

- Areas with low percentages of Black or African Americans tend to have high values of PE and thus low data utility.
- Particularly in Guernsey County, tracts with percentage of Black or African Americans less than 1 percent have their PE higher than 60 percent.



Summary

We demonstrate in this paper how to generate and use an open and realistic synthetic population dataset to assist in the contextualization of maps.

- Synthetic population data are especially useful when true data are sensitive and not publicly available.
- Public use synthetic data facilitate transparency in mapmaking and help readers understand the contexts of a map.
- One of the future directions is to expand the scope of this dataset beyond the United States to other regions.

References

- Crampton, J. W. (2001). Maps as social constructions: power, communication and visualization. *Progress in Human Geography*, 25(2), 235-252.
- Harley, J. B. (1989). Deconstructing the map. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 26(2), 1-20.
- Harley, J. B. (1990). Cartography, ethics and social theory. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 27(2), 1-23.
- United States Census Bureau (2020). DAS 2010 demonstration data products disclosure avoidance system release. Retrieved October 7, 2022, from https://github.com/uscensusbureau/census2020-das-2010ddp