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 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?

Note: Bedford city, VA, was changed to town status and added to Bedford County, VA, effective July 1, 2013. For purposes of presenting data, Bedford County is treated as if Bedford city were included in it at the time of the 2010 census.

Information on confidentiality protection, nonsampling error, and definitions is available at <https://www2.census.gov/programs-surveys/decennial/2020/technical-documentation/complete-tech-docs/summary-file/>.

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).
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 $X$:

$$\min \|WX' - Y\|^2,$$

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

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

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

Compared with the original SF1 data.
Internal Validation

Compared with the original SF1 data.
External Validation

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

Franklin County

\[
c = 0.9978
\]

Guernsey County

\[
c = 0.9999
\]
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