Investigating MAUP Effects on Census Data Using Approximately Equal-Population Aggregations

Yue Lin
Department of Geography, The Ohio State University, Columbus, OH, USA

Ningchuan Xiao
Department of Geography, The Ohio State University, Columbus, OH, USA

Abstract

The modifiable areal unit problem (MAUP) can significantly impact the use of census data as different choices in aggregating geographic zones can lead to varying outcomes. Previous research studied the effects using random aggregations, which, however, may lead to the use of impractical and unrealistic zones that deviate from recommended census geography criteria (e.g., equal population). To address this issue, this study proposes the use of approximately equal-population aggregations (AEPAs) for exploring MAUP effects on various statistical properties of census data, including Moran coefficients, correlation coefficients, and regression statistics. A multistart and recombination algorithm (MSRA) is used to generate multiple sets of high-quality AEPAs for testing MAUP effects. The results of our computational experiments highlight the need for more well-defined census geographies and realistic alternative zones to fully understand MAUP effects on census data.

2012 ACM Subject Classification Computing methodologies → Modeling and simulation

Keywords and phrases Census, heuristics, modifiable areal unit problem, spatial aggregation, spatial autocorrelation

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.47

Category Short Paper

1 Introduction

The U.S. Census Bureau reports data using a nested hierarchy of geographic zones, beginning with census blocks and progressing to block groups, tracts, counties, and states. The boundaries of these zones are typically created before the digital computer era and are often arbitrary [4], resulting in significant variations in size, population, and demographic makeup [8]. When data is aggregated into different geographic zones and at different scales, statistical properties of the data, such as spatial autocorrelation and correlation coefficients, often demonstrate significant differences from the officially defined zones. This problem is referred to as the modifiable areal unit problem, or MAUP [6], which often causes uncertain and potentially biased census data [5].

A spatial aggregation is a particular way of grouping low-level units (e.g., census blocks) into contiguous high-level units (e.g., census block groups or tracts). Figure 1 illustrates such an aggregation where census block groups (light grey lines) are aggregated into tracts (dark grey lines). During this process, the aggregated data is expected to have different, often reduced, spatial autocorrelation compared to the original data. When using the aggregated data in correlation and regression analysis, the coefficients may also differ from those obtained using the original data.

To understand MAUP effects on the statistical properties of spatial data, algorithms have been developed to generate alternative random aggregations [5, 2]. However, these algorithms often cannot yield spatial aggregations that satisfy some prerequisites of census geography. For example, census tracts in the United States are designed to have a relatively uniform


© Yue Lin and Ningchuan Xiao;
licensed under Creative Commons License CC-BY 4.0

Figure 1 The official 2010 census block groups (light grey) and tracts (dark grey).

population size of around 4,000 people [9], a criterion that random aggregations typically fail to meet. Using random aggregations, therefore, may only partially reveal MAUP effects on the statistical properties of spatial data.

In the meantime, past research has also advocated the use of geographic zones that are standardized in terms of size, population, and other socioeconomic components [4, 8]. To achieve this, automated zone design algorithms have been developed to generate alternative aggregations that meet these criteria [4, 1, 7]. However, these algorithms often focus on producing one of the many possible aggregations that align with these criteria and may not allow us to fully understand MAUP effects on census data.

This paper reports our work in progress where we utilize a heuristic search method called multistart and recombination algorithm (MSRA) [10] to generate multiple approximately equal-population aggregations (AEPAs). We compare the analysis of MAUP effects derived from AEPAs with that from random aggregations. Specifically, we investigate whether AEPAs significantly differ from random aggregations in (1) spatial autocorrelation, (2) correlation coefficients, and (3) other regression statistics. In the following sections, we first describe the MSRA and then illustrate how AEPAs generated by the MSRA can be used to explore the impact of spatial aggregation on the statistical properties of census data.

2 The Multistart and Recombination Algorithm

The MSRA is a heuristic search method that identifies a diverse set of high-quality spatial aggregations where the populations in the zones are approximately equal [10]. The algorithm consists of two phases. In the first phase, a multistart process generates a pool of independent aggregations. Each aggregation is randomly created and then improved using an efficient method called the give-and-take algorithm to reduce population differences between zones by swapping units [3]. The second phase is an iterative process where in each iteration two aggregations from the pool are randomly selected and then combined to create a new aggregation. If the new aggregation is new and superior to the worst in the pool, it is added to the pool by replacing the worst aggregation.

3 Computational Experiments: Design and Results

We chose Franklin County, Ohio as our study area due to its diverse social, economic, and demographic characteristics, as well as its mix of densely populated urban areas and extensive rural areas. The county is composed of 887 block groups that are combined to form 284 tracts (Figure 1). To explore potential alternatives to these census tracts, we use the MSRA to generate 500 AEPAs, each combining 887 block groups into 284 zones with approximately equal population (Figure 2a–b). We also employ the algorithm proposed in [7]

to generate 10,000 random aggregations for comparison (Figure 2c–d). Figure 3 illustrates the population distribution among official census tracts, as well as zones in AEPAs and random aggregations. The results suggest that random aggregations have the greatest variation in population distribution, followed by the official census tracts, and then the AEPAs generated by the MSRA. Using the AEPAs, we investigate MAUP effects on three variables related to the Franklin population: the number of people who work from home (x_1), the number of non-Hispanics (x_2), and the number of people with a Bachelor's degree or higher (y).

Figure 2 Two AEPAs (a, b), and two random aggregations (c, d).

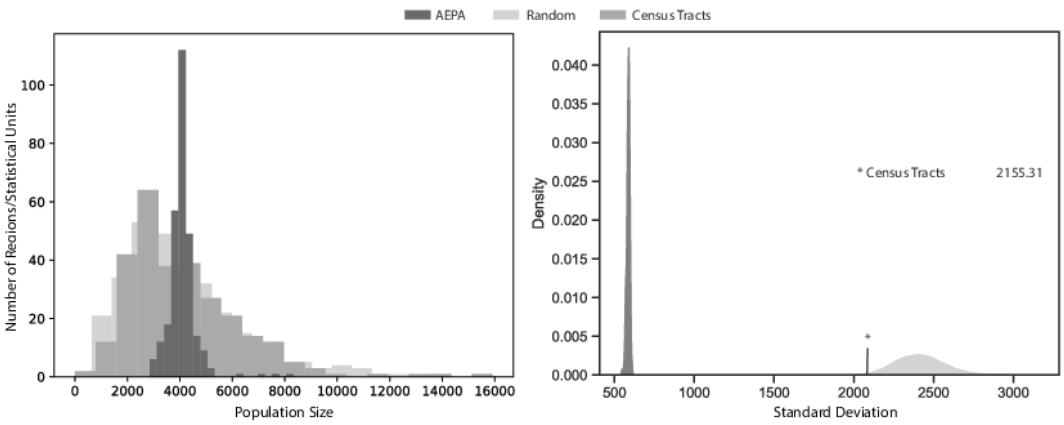

Figure 3 Population distribution and comparison. The left panel displays the population distribution for official census tracts, zones in an AEPA, and zones in a random aggregation. The right panel shows the standard deviation of population size for zones in all AEPAs and random aggregations.

3.1 MAUP effects on Moran coefficients

The Moran coefficient (MC) is a statistical measure that determines the degree of spatial autocorrelation of a variable. Figure 4 demonstrates MAUP effects on the MC using AEPAs and random aggregations, revealing three significant findings. First, the MC resulting from AEPAs can be quite different from the MC at the block group level. For instance, for variable x_2, the MC at the block group level (0.341) suggests moderate spatial autocorrelation, whereas the average MC resulting from AEPAs (0.136) indicates weak spatial autocorrelation. This finding implies that aggregation can affect the statistical properties of census data. Second, the MC obtained by aggregating data using the official census tracts may not always provide a reliable representation of the MCs that can be obtained through AEPAs. For example, the average MC for variable x_2 obtained through AEPAs is 0.136, while the tract-level MC

is noticeably higher at 0.354. This may seem surprising, but can be explained by the fact that although equal population served as a general principle when the Census Bureau first designed the boundaries of census tracts, these boundaries have not been updated in decades while the population within them has changed substantially. As a result, the population distribution in existing census tracts deviates from equal population, and the statistical properties of census data can also differ from those that can be obtained through AEPAs. Third, it is observed that the distributions of MC under equal population and random aggregations may differ significantly. For example, there is minimal overlap between the distribution of MC under random and equal population aggregations for variables x_2 and y.

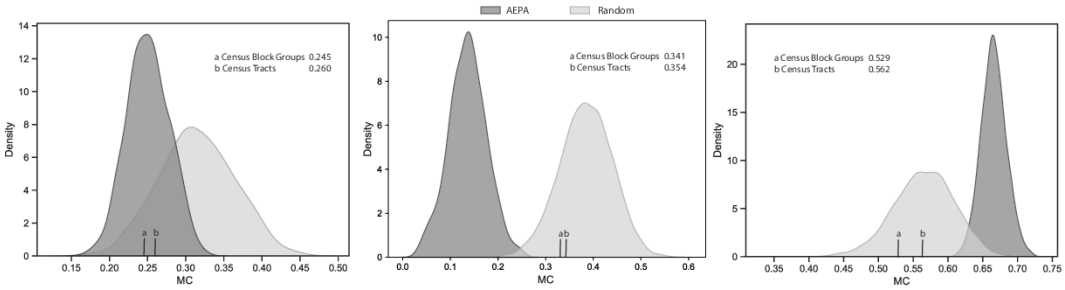

Figure 4 MAUP effects on the MC of variables x_1 (work from home, left), x_2 (non-Hispanics, middle), and y (Bachelor's degree or higher, right).

3.2 MAUP effects on correlation coefficients

Correlation coefficients play a critical role in statistical analysis by quantifying the degree and direction of the relationship between two variables. Figure 5 presents MAUP effects on correlation coefficients between variables x_1 and y and between x_2 and y. The findings are similar to those observed for the MC. First, the correlation coefficients derived using AEPAs can differ substantially from the block group-level coefficient, as demonstrated by the correlation coefficient between x_2 and y being 0.666 at the block group level, whereas the average resulting from AEPAs is substantially lower at 0.325. Second, the correlation coefficients obtained through AEPAs can differ considerably from the tract-level correlation coefficient. For instance, the average correlation coefficient between x_2 and y for AEPAs is 0.325, suggesting a slightly weak association, while the correlation coefficient at the tract level is 0.669, indicating a strong association. Third, the correlation coefficients resulting from AEPAs and random aggregations have little overlap, as previously observed for the MC. This finding reinforces the idea that random aggregations may not yield representative results that reflect MAUP effects when the equal population criterion is applied to modify census geography.

3.3 MAUP effects on regression statistics

Regression analysis is a widely used statistical tool to explore the relationship between a dependent variable and independent variables. In the case of two variables, it is important to consider MAUP effects on the regression slope, which indicates the direction and strength of the association. In addition, if the regression residuals exhibit spatial autocorrelation, the assumption of independent residuals is violated, compromising the validity of linear regression analysis. It is therefore crucial to investigate the spatial autocorrelation of the residuals under different aggregations to determine if they exacerbate or mitigate the issue.

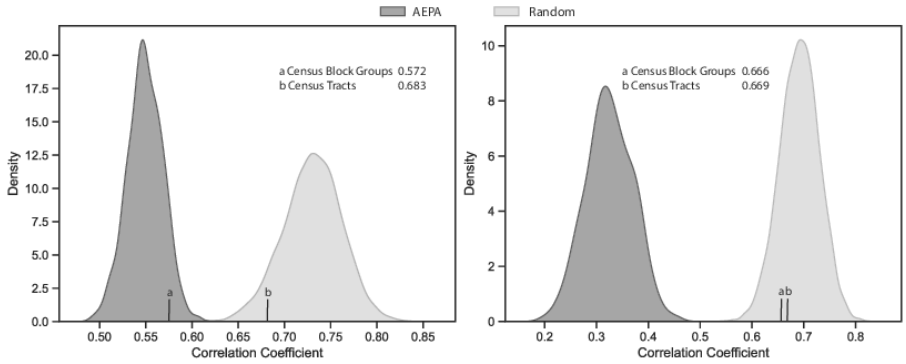

Figure 5 MAUP effects on the correlation coefficients between x_1 (work from home) and y (Bachelor's degree or higher) on the left, and between x_2 (non-Hispanics) and y on the right.

Here, we examine two regression models of the form $y = \alpha + \beta x_1$ and $y = \alpha + \beta x_2$, where α represents the regression intercept and β the regression slope. Figure 6 presents MAUP effects on the regression slope β and the Moran coefficient of the regression residuals (MCRR). It is observed that both β and MCRR exhibit differences using AEPAs compared to block group-level regression statistics. In addition, there are differences between β and MCRR obtained through AEPAs and existing census tracts. Consistent with our findings for MC and correlation coefficient, distributions of β and MCRR generated using AEPAs and random aggregations can have minimal overlap.

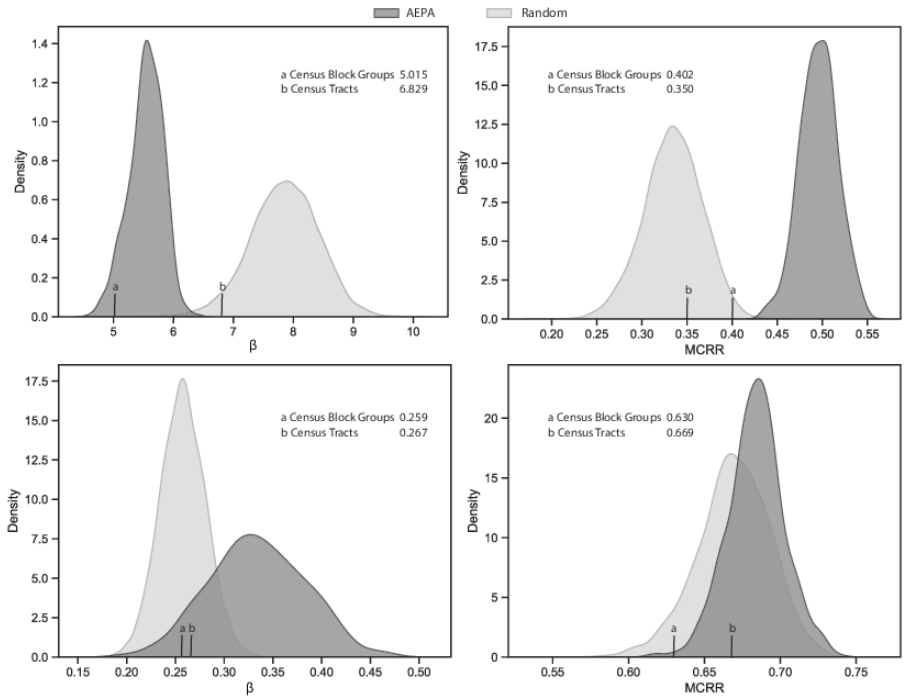

Figure 6 MAUP effects on the regression slope β and the MCRR for two models: $y = \alpha + \beta x_1$ (top) and $y = \alpha + \beta x_2$ (bottom).

4 Conclusions

We present a renewed exploration of MAUP effects on univariate and bivariate statistics of census data using multiple approximately equal-population aggregations (AEPAs). Our study yields three key findings. The first highlights the significance of MAUP effects when aggregating census data from low-level units, which can greatly impact the interpretation of statistical properties such as Moran coefficients, correlation coefficients, and regression statistics. Second, the current census geography deviates from the principles that guided its design decades ago, which poses a challenge for understanding and addressing MAUP effects. Our analyses show how existing census tracts, established since 1790 and evolved over time, barely adhere to the equal population criterion today, resulting in statistical properties that differ from what we would expect under equal population aggregation. This finding underscores the need to re-examine the existing census geography and to develop more well-defined geographic zones to help better understand MAUP effects. Finally, our analyses reveal that random aggregations and AEPAs can yield vastly different results regarding MAUP effects. While it is recognized that not all randomly generated sets of zones are suitable for use as census geography, they are still commonly used to study MAUP effects. Our study demonstrates the limitations of such an approach and emphasizes the importance of employing realistic zones with approximately equal population to better understand MAUP effects on census data. Future research can be directed to generalize these findings to other variables and explore the impact of AEPAs in multivariate situations.

References

1 Samantha Cockings, Andrew Harfoot, David Martin, and Duncan Hornby. Maintaining existing zoning systems using automated zone-design techniques: methods for creating the 2011 census output geographies for england and wales. *Environment and Planning A*, 43(10):2399–2418, 2011.

2 A Stewart Fotheringham and David WS Wong. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A*, 23(7):1025–1044, 1991.

3 Myung Jin Kim. Give-and-take heuristic model to political redistricting problems. *Spatial Information Research*, 27(5):539–552, 2019.

4 David Martin. Optimizing census geography: The separation of collection and output geographies. *International Journal of Geographical Information Science*, 12(7):673–685, 1998.

5 Stan Openshaw. A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In *Statistical Applications in the Spatial Science*, pages 127–144. Pion, 1979.

6 Stan Openshaw. *The Modifiable Areal Unit Problem*. Geo Books, Norwich, 1983.

7 Stan Openshaw and RS Baxter. Algorithm 3: A procedure to generate pseudo-random aggregations of n zones into m zones, where m is less than n. *Environment and Planning A*, 9(12):1423–1428, 1977.

8 Stan Openshaw and Liang Rao. Algorithms for reengineering 1991 census geography. *Environment and Planning A*, 27(3):425–446, 1995.

9 United States Census Bureau. Glossary: Census tract, 2022. URL: `https://www.census.gov/programs-surveys/geography/about/glossary.html#par_textimage_13`.

10 Ningchuan Xiao, Peixuan Jiang, Myung Jin Kim, and Anuj Gadhave. A multistart heuristic approach to spatial aggregation problems. In *International Conference on GIScience Short Paper Proceedings*, volume 1, pages 349–351, 2016.