Visualizing Scalar Effects of Urban Data Aggregation

Jonathan K. Nelson*

University of Wisconsin-Madison

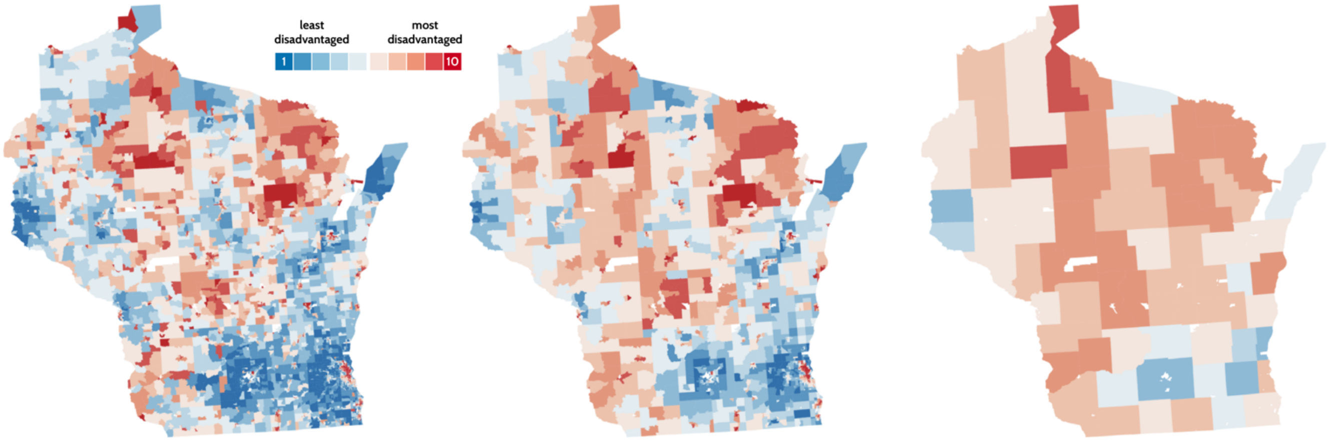


Figure 1: Mapping Area Deprivation Indices at the census block group (left), tract (middle), and county (right) levels of aggregation.

Abstract

The process of geospatial data aggregation provides a means for abstracting the complexity of urban systems to not just better understand them, but also protect the privacy of the individuals within them. However, level of aggregation and the arbitrary sizes, shapes, and arrangements of areal units may lead to statistical and visual bias that affects the reliability and validity of findings derived from the analysis of areally aggregated urban data. This bias and resulting analytical uncertainty – *known as the Modifiable Areal Unit Problem (MAUP)* – has implications for public policy implementation and allocation of critical resources in both urban and rural areas. Despite a wealth of geographic research on MAUP and development of advanced statistical approaches to quantifying its effects, many of these insights and techniques remain largely inaccessible and subsequently unadopted by GIS professionals working on city planning applications. This paper introduces a simple vector-to-raster choropleth mapping workflow that enables a broad range of urban analysts to visually assess the scalar effects of the modifiable areal unit problem.

Keywords: Urban data aggregation, modifiable areal unit problem, choropleth map, spatial decision support tool.

1 Introduction and Background

Making sense of urban systems oftentimes requires the analysis and visualization of areally aggregated data. The aggregation process makes spatial data more comprehensible and helps to maintain the privacy of individuals reflected in the data. Geographic boundaries, however, are commonly defined by historical or political processes that may lead to inaccurate and misleading analytical results. Both level of aggregation (i.e., *the scale effect*) and the arbitrary sizes, shapes, and arrangements of zones (i.e., *the zoning effect*) contribute to statistical and visual bias that affects the reliability and validity of findings derived from the analysis of areally aggregated data [4, 14]. This bias and resulting analytical uncertainty, known as the Modifiable Areal Unit Problem (MAUP), has implications for public policy implementation and the extent to which economic, health, and other resources are distributed (or not) to the individuals in both urban and rural areas [17].

Researchers have employed a wide range of global and local spatial statistical modeling techniques, such as geographically weighted regression [2], spatial autocorrelation analysis [11], and spatial aggregation entropy [18] to simulate, quantify, and better understand the effects of the longstanding and unresolved MAUP. Complex interactive visualization frameworks [5] and bivariate choropleth mapping techniques [13] have also been devised to help make sense of the dynamic interrelation of geographic phenomena across scale. Moreover, MAUP has been explored in many urban contexts, including socioeconomics and equity [7, 10], health [9, 16, 17], and a variety of transportation applications [3, 12, 15]. The advancement of local statistical frameworks has further enabled new insights into why data are distributed over space in certain ways, reflecting a fundamental shift in examining MAUP through the lens of not just data properties but also process properties [1].

Despite an active and rich body of geographic research on MAUP and development of more advanced analytical approaches to quantifying and visualizing its effects, many of these insights and technical frameworks remain largely inaccessible and subsequently unadopted by GIS professionals working in the city planning space. This is problematic given the vital role these individuals play in translating urban data into insights for decision makers. This work presents a case study, introducing a simple vector-to-raster choropleth mapping workflow that enables a broad range of urban analysts to visually assess the scalar effects of MAUP in the context of socioeconomic disadvantage.

2 Approach and Case Study

The proposed vector-to-raster choropleth mapping workflow is designed to support GIS professionals in making more informed decisions on which scales to aggregate, analyze, and represent

* email: jknelson3@wisc.edu

urban data for problem-specific city planning applications. In addition to serving as a spatial decision support tool at earlier stages in the analytical process, the workflow also aims to enable GIS professionals to recognize and more effectively communicate to stakeholders the underlying reasons for conflicting results when spatial analysis is conducted at different levels of aggregation. The workflow can be performed using common desktop GIS mapping software applications (e.g., QGIS and Esri ArcGIS Pro) and consists of four major steps:

1. Aggregate variable-of-interest (VoI) to all levels of aggregation deemed relevant for comparison.
2. Create choropleth maps depicting VoI values at selected levels of aggregation.
3. Rasterize choropleth maps.
4. Generate difference maps depicting the magnitude and direction of change in VoI values between selected levels of aggregation.

The following subsections expand on each major step using a case study to illustrate the approach.

2.1 Data Selection and Aggregation

Data used to illustrate this workflow come from the Neighborhood Atlas Project [8]. The VoI is the area deprivation index (ADI) aggregated to census block groups across Wisconsin State (USA). ADI is a metric used to rank "neighborhoods" based on socioeconomic disadvantage at the state or national level. ADI considers factors such as income, education level, employment status, and housing quality and aims to inform healthcare delivery and policy. This dataset was selected to demonstrate this methodology because [8] explicitly state that ADI was constructed at the census block group level and is not valid if aggregated to other geographic units, thus offering a helpful framework for comparatively visualizing mapped depictions of ADI at a valid level of aggregation versus other common but invalid levels of aggregation (e.g., census tract and county). The decision to demonstrate this methodology using Wisconsin as the area-of-interest (opposed to other US States) was arbitrary and inspired solely by the Neighborhood Atlas Project originating from the University of Wisconsin School of Medicine and Public Health. Tract and county level enumeration units were selected for comparative analysis due to prevalence of use in a wide range of urban analytical applications and because census block groups nest neatly within tracts and tracts nest neatly within counties, making the aggregation of ADI values across these enumeration units straightforward to conduct, visualize, and interpret.

2.2 Create Choropleth Maps

Next, choropleth maps were created to visualize ADI values at the census block group, tract, and county levels of aggregation (Fig. 1). ADI scores range from 1 (least disadvantaged area) to 10 (most disadvantaged area), thus a divergent color scheme was employed with darker shades of blue representing areas least disadvantaged and darker shades of red representing areas most disadvantaged. An Albers Conic projection was applied to ensure equal area representation of all enumeration units and facilitate a more honest interpretation and comparison of mapped values. A visual comparison of the three choropleth maps quickly conveys the diminishing variance as level of aggregation increases and that local patterns appear more pronounced and similar at the block group and tract levels of aggregation in comparison to patterns reflected at the county level.

2.3 Rasterize Choropleth Maps

The third step in the workflow is to convert the choropleth maps from vector to raster format in preparation for creating difference maps (see following subsection). While difference maps can be created using vector data, the process is more convoluted, involving a series of spatial joins and field calculations. Another advantage of raster conversion is it enables the creation of choropleth difference maps between both nested and non-nested aggregation structures, the latter being very challenging to create in vector format.

As part of the raster conversion process, it is critical to ensure that the VoI is selected as the value to burn into the raster output (in this case ADI) and the output raster size aligns with the resolution of the geometrically most detailed enumeration unit considered in the analysis (in this case census block groups). The goal is to select a raster resolution that achieves visual consistency with the detail present in vector choropleth map input. Figure 2 provides a side-by-side comparison of vector (left) and raster (right) map outputs to illustrate a geometrically complex urban area of Wisconsin at a map scale of 1:50,000 and raster resolution of 100,000 x 100,000 pixels.

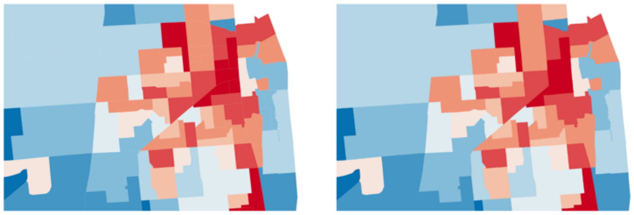


Figure 2: Establishing visual consistency between vector (left) and raster (right) choropleth map depictions.

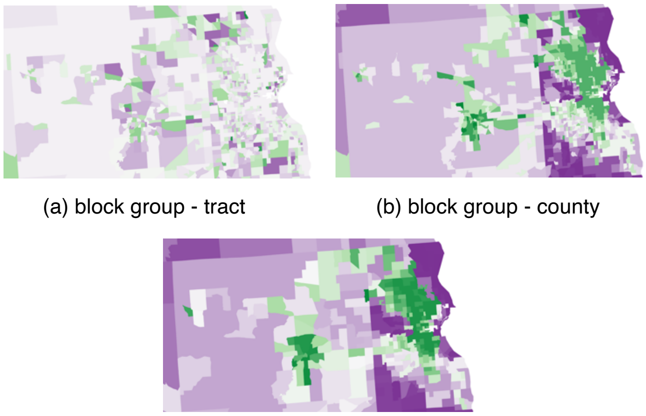
2.4 Generate Difference Maps

The final step in the workflow is to use a raster calculator to generate difference maps that visualize the change in magnitude and direction of ADI values between the three levels of aggregation. Tract- and county-level rasters can be subtracted from the block group raster (Fig. 3a and b); the county raster can also be subtracted from the tract raster (Fig. 3c). Difference values are depicted using unique value, unclassed choropleth maps that communicate both subtle and more pronounced changes in ADI values across aggregation levels. A diverging purple-green color scheme is employed to emphasize areas where there is a negative (purple) or positive (green) difference in values between two different levels of aggregation. Darker shades of purple signify areas where ADI tended to be greater at the coarser level of aggregation, whereas darker shades of green signify areas where ADI tended to be lower at the coarser level of aggregation. For example, a dark purple polygon in Figure 3b reflects an area that would be deemed more disadvantaged (higher ADI value) at the county level compared to the census block group level. Conversely, a dark green polygon in Figure 3c reflects an area that would be deemed less disadvantaged (lower ADI) at the county level compared to the census tract level. Lighter shades of these colors reflect areas where ADI values were more stable between the different levels of aggregation.

3 Discussion, Limitations and Outlook

The maps in Figure 3 communicate two important patterns. First, ADI values are relatively stable when aggregated to census block groups *or* tracts, but areas of local variation still exist (Fig. 3a). Second, the similarity in patterns shown in Figures 3b and 3c suggest that block group *and* tract ADI values differ from county

ADI values in a similar way. These insights provide GIS professionals with a better understanding of the local (in)stability of ADI across scale and enable them to combine this understanding with domain knowledge to make more informed decisions on which level(s) of aggregation are most appropriate to use for analysis.



(a) block group - tract

(b) block group - county

(c) tract - county

Figure 3: Visualizing **negative** and **positive** differences in ADI values between levels of aggregation in the Milwaukee metropolitan area.

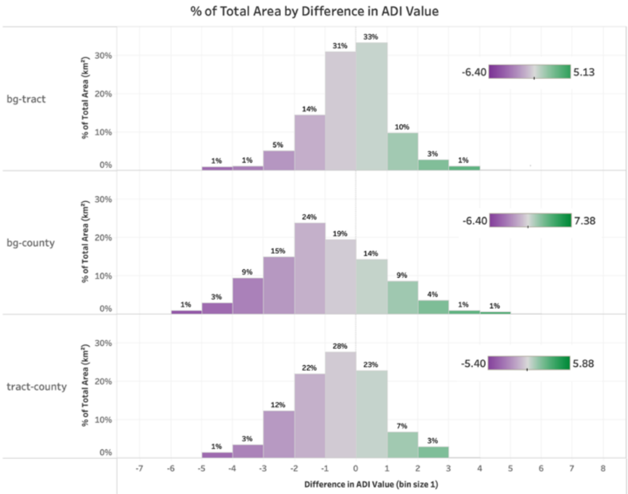


Figure 4: Percent of total area by difference in ADI value for entire State of Wisconsin.

A limitation of this workflow is that the resulting insights primarily support visual assessment rather than quantification. Raster summary statistics can be calculated to derive the distribution of area by difference in ADI value between any two levels of aggregation (Fig. 4); however, the workflow does not currently include any spatial statistical modelling techniques to more substantively quantify local stability across scale. Moreover, the workflow only enables analysts to visualize the scalar effects of urban data aggregation one variable at a time, thus does not shed light on how the dynamic interrelation among two or more urban phenomena may vary across scale. Thus, there is a trade-off between the accessibility and simplicity of the workflow versus its ability to support more advanced multivariate statistical analysis.

In the context of the human-centric *CityVis* design space [6], this geovisualization workflow can help to inform data quality and reliability across different levels of aggregation (C1); prompt discussion about how to design data aggregation structures that more effectively reflect both the individual's sense-of-place and the overall complexity of urban infrastructure (C2, C3); and equip a broad range of urban analysts with a simple approach to visually documenting the urban data aggregation process and assessing the appropriateness and validity of data representation at multiple scales and contexts (C4, C5, C6, C9). In summary, this vector-to-raster choropleth mapping workflow aims to exemplify how geovisualization techniques can support urban data governance processes by broadening exposure to, and understanding of, MAUP effects, thus promoting data/visual literacy among analysts and decisionmakers and facilitating a more informed use of aggregated urban data. Piloting this workflow with GIS professionals working across a variety of city planning projects is a critical next step in assessing the usability and utility of this workflow and identifying opportunities to increase its relevancy and value.

Acknowledgements

The author wishes to thank the anonymous reviewers for their insightful feedback and Justin Cole for sharing his perspective on the workflow through the lens of a GIS practitioner and educator. Support for this research was provided by the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

References

[1] A. S. Fotheringham and M. Sachdeva. Scale and local modeling: new perspectives on the modifiable areal unit problem and Simpson's paradox. J Geogr Syst, 24: 475–499, 2022. doi: 10.1007/s10109-021-00371-5

[2] A. S. Fotheringham, C. Brunsdon, and M. Charlton. Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons, 2003.

[3] F. Gao, S. Li, Z. Tan, Z. Wu, X. Zhang, G. Huang, and Z. Huang. Understanding the modifiable areal unit problem in dockless bike sharing usage and exploring the interactive effects of built environment factors. International Journal of Geographical Information Science, 35(9): 1905–1925, 2021. doi: 10.1080/13658816.2020.1863410

[4] C. E. Gehlke and K. Biehl. Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association, 29(185A):169–170, 1934.

[5] S. Goodwin, J. Dykes, A. Slingsby, and C. Turkay. Visualizing multiple variables across scale and geography. IEEE Transactions on Visualization and Computer Graphics 22(1): 599–608, 2015. doi: 10.1109/TVCG.2015.2467199

[6] S. Goodwin, S. Meier, L. Bartram, A. Godwin, T. Nagel, and M. Dörk. Unravelling the human perspective and considerations for urban data visualization. 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), 126–130, 2021. doi: 10.1109/PacificVis52677.2021.00024

[7] R. Javanmard, J. Lee, J. Kim, L. Liu, and E. Diab. The impacts of the modifiable areal unit problem (MAUP) on social equity analysis of public transit reliability. Journal of Transport Geography, 106, 2023. doi: 10.1016/j.jtrangeo.2022.103500

[8] A. J. H. Kind and W. Buckingham. Making Neighborhood Disadvantage Metrics Accessible: The Neighborhood Atlas. New England Journal of Medicine, 378: 2456–2458, 2018. doi: 10.1056/NEJMp1802313. PMCID: PMC6051533

[9] D. Lee, C. Robertson, C. Ramsay, and K. Pyper. Quantifying the impact of the modifiable areal unit problem when estimating the

health effects of air pollution. Environmetrics, 31(8): e2643, 2020. doi: 10.1002/env.2643

[10] G. Lee, D. Cho, and K. Kim. The modifiable areal unit problem in hedonic house-price models. Urban Geography, 37(2):223–245, 2016. doi: 10.1080/02723638.2015.1057397

[11] S. I. Lee, M. Lee, Y. Chun, and D. A. Griffith. Uncertainty in the effects of the modifiable areal unit problem under different levels of spatial autocorrelation: A simulation study. International Journal of Geographical Information Science, 33(6):1135–1154, 2019. doi: 10.1080/13658816.2018.1542699

[12] T. Li, M. Zhang, H. Jiang, and P. Jing. Understanding the Modifiable Areal Unit Problem and Identifying Appropriate Spatial Units while Studying the Influence of the Built Environment on the Traffic System State. Journal of Advanced Transportation, 2022. doi: 10.1155/2022/8288248

[13] J. K. Nelson and C. A. Brewer. Evaluating data stability in aggregation structures across spatial scales: revisiting the modifiable areal unit problem. Cartography and Geographic Information Science, 44(1):35–50, 2017. doi: 10.1080/15230406.2015.1093431

[14] S. Openshaw. The Modifiable Unit Areal Problem. Concepts and Techniques in Modern Geography. Geo Books, Norwich, UK, 38, 1984.

[15] A. Pani, P. K. Sahu, A. Chandra, and A. K. Sarkar. Assessing the extent of modifiable areal unit problem in modelling freight (trip) generation: Relationship between zone design and model estimation results. Journal of Transport Geography, 80: 102524, 2019. doi: 10.1016/j.jtrangeo.2019.102524

[16] M. P. Parenteau and M. C. Sawada. The modifiable areal unit problem (MAUP) in the relationship between exposure to NO 2 and respiratory health. International journal of health geographics, 10(58):1-15, 2011. doi: 10.1186/1476-072X-10-58

[17] E. D. Root. Moving Neighborhoods and Health Research Forward: Using Geographic Methods to Examine the Role of Spatial Scale in Neighborhood Effects on Health. Annals of the Association of American Geographers, 102(5):986–995, 2012. doi: 10.1080/00045608.2012.659621

[18] J. Xiao. Spatial aggregation entropy: a heterogeneity and uncertainty metric of spatial aggregation. Annals of the American Association of Geographers, 111(4):1236–1252, 2021. doi: 10.1080/24694452.2020.1807309