[FC-07-026] Problems of Scale and Zoning

Abstract

Spatial data are often encoded within a set of spatial units that exhaustively partition a region, where individual level data are aggregated, or continuous data are summarized, over a set of spatial units. Such is the case with census data aggregated to enumeration units for public dissemination. Partitioning schemes can vary by scale, where one partitioning scheme spatially nests within another, or by zoning, where two partitioning schemes have the same number of units but the unit shapes and boundaries differ. The Modifiable Areal Unit Problem (MAUP) refers to the fact the nature of spatial partitioning can affect the interpretation and results of visualization and statistical analysis. Generally, coarser scales of data aggregation tend to have stronger observed statistical associations among variables. The ecological fallacy refers to the assumption that an individual has the same attributes as the aggregate group to which it belongs. Combining spatial data with different partitioning schemes to facilitate analysis is often problematic. Areal interpolation may be used to estimate data over small areas or ecological inference may be used to infer individual behaviors from aggregate data. Researchers may also perform analyses at multiple scales as a point of comparison.

Keywords: aggregation, areal interpolation, dasymetric mapping, ecological fallacy, ecological inference, MAUP, multi-scale analysis, Uncertainty

Author & citation

Mennis, J. (2019). Problems of Scale and Zoning. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2019 Edition), John P. Wilson (Ed.).
DOI: 10.22224/gistbok/2019.1.2

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Problems of Scale and Zoning. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Definitions
2. Introduction
3. Scale vs. Zoning
4. The Modifiable Areal Unit Problem (MAUP)
5. Combining Spatial Data with Different Partitioning Schemes
6. Addressing Problems of Scale and Zoning

1. Definitions

Modifiable Areal Unit Problem (MAUP): An issue related to the analysis of spatially aggregated data where the results of mapping or statistical analysis may differ when using

different spatial units of aggregation.

Ecological fallacy: The incorrect assumption that individuals have the same characteristics or properties of the group to which they belong, in the case of spatial data, to the spatial unit within which the individual resides.

Simpson's paradox: An issue related to data analysis where statistical results may differ between an analysis conducted on an entire data set and subsets of that data set.

Areal interpolation: The estimation of data values over areas which have not been sampled or for which data are only available over a coarser resolution.

Dasysmetric mapping: A type of map which displays a continuous variable according to a set of exhaustive and non-overlapping polygons, where the polygon boundaries occur at the steepest escarpments of the surface; Also, an analytical technique used to disaggregate spatial data over small areas using ancillary data sources.

2. Introduction

Problems of scale and zoning in Geographic Information Science generally refer to how an area or region is exhaustively partitioned into a set of discrete, non-overlapping spatial units for purposes of mapping or statistical analysis. The "problem" concerns the fact that the nature of the partitioning can affect the interpretation and results of visualization and statistical analysis. Oftentimes, a researcher does not know a priori the spatial scale at which a geographic process under investigation operates, or the optimal spatial partitioning scheme to capture a particular geographic phenomenon. Researchers are also often constrained by technical or practical limitations to the capability to collect spatial data or make observations, as with the resolution constraints of remotely sensed imagery. Further, many spatial data sets can only be acquired in an aggregated form, such that the original individual observations are not available. Such is the case, for instance, with many socioeconomic and health data sets disseminated by governmental census and health agencies, where publicly accessible spatial data products are typically aggregated to spatial units in order to preserve the privacy of individuals.

Consequently, it is often unclear if an observed spatial pattern or statistical result is truly indicative of the geographic phenomenon under investigation or if it is an artifact of the partitioning scheme, or whether alternative spatial partitioning schemes would yield different mapped visual patterns or statistical results. The issue is present in spatial data that represent geographic phenomena conceptualized as continuous surfaces, i.e. considered to vary continuously over space, such as air temperature. Here, the issue concerns partitioning the area into a set of spatial units that best captures the natural breaks in the surface – i.e. the locations of the steepest surface escarpments or gradients. The problem of partitioning comes into play here when the surface is exhaustively tessellated into regions using a set of spatial units that do not optimally capture the surface variation, for instance where air temperature might be mapped using a set of political units, such as US counties. Such a map might visually suggest that air temperature changes abruptly at county boundaries, whereas air temperature is obviously not typically constrained by political demarcations (unless those demarcations also happen to coincide with a physical barrier like a mountain range).

More commonly, however, problems of partitioning are considered in count, or punctiform, data such as population counts or characteristics, which may also be expressed as a rate (e.g. the percentage of a population identifying as a certain race or ethnicity) or density (e.g. population density or people per square mile). Such count data are often made available as data products which aggregate individual observations to sets of discretely bounded enumeration units, such as census boundaries, neighborhoods, or postal codes. Oftentimes, the genesis of the enumeration unit is based on some convenience of enumeration, or based on a political or administrative reason, that has little to do with the underlying geographic phenomenon being represented. In this case, where the spatial unit of enumeration is divorced from the geographic phenomena being represented, problems associated with partitioning may be particularly pronounced.

3. Scale vs. Zoning

The issue of partitioning is typically considered a consequence of both scale and zoning. Scale refers to the number, and, relatedly, the size, of spatial units used to partition an area. For instance, a difference in scale of partitioning occurs between the use of a lesser number of larger units versus a greater number of smaller units, for example in the use of US Census Bureau tracts versus block groups, where one or more block groups nest within each tract. Zoning refers to the shapes and boundaries of the spatial units. In this case, two different partitioning schemes for an area may have the same number of spatial units, but the shapes and boundaries of the units may differ. For instance, consider a region partitioned using a set of postal codes (e.g. a US zip code) versus a set of legislative districts of approximately the same size and number but with different zone shapes.

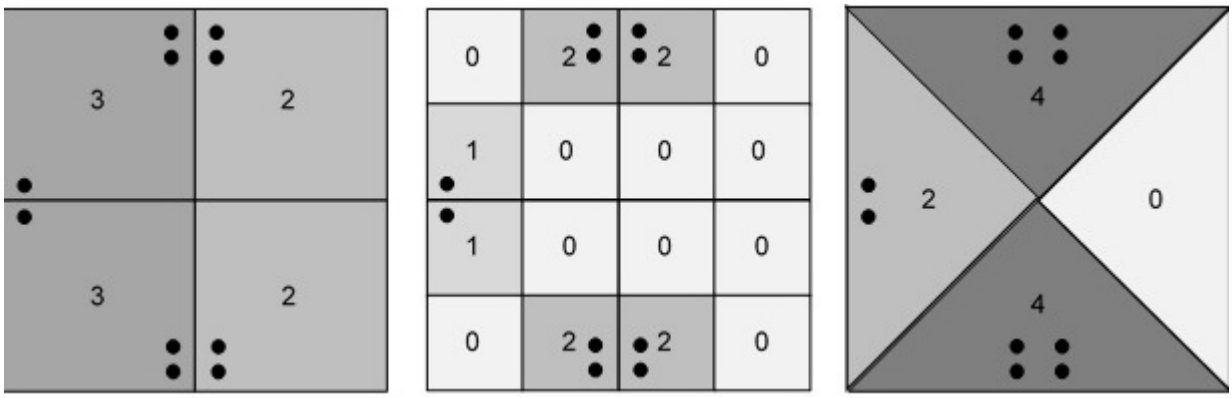


Figure 1. Abstract illustration of the impact on aggregating data for a set of points to an original partitioning scheme (left panel) due to differences in scale (middle panel) and zoning (right panel). Numbers and grayscale indicate the number of points within each spatial unit for that partitioning scheme (darker shade of gray indicates a higher number of points). See text for explanation. Source: author.

As an illustration of how changes in scale and zoning can impact data aggregation, Figure 1 shows a point process overlain with three different partitioning schemes. Each scheme represents an aggregation of the point data to a set of spatial units, where the numeric

label within each spatial units represents the number of points within that unit for that partitioning scheme. In the original partitioning scheme (left panel), two of the four spatial units contain two points and two units contain three points. A change in scale, in which each of the four original spatial units is partitioned into four equal size spatial units to yield 16 total units (middle panel), contains eight spatial units with zero points in each, four spatial units with two points in each, and two spatial units with 1 point in each. A change from the original partitioning scheme due to zoning (right panel) yields the same number of original spatial units, i.e. four, but here two of the spatial units contain four points each, one unit contains two points, and the other unit contains zero points. Thus, the number of points per unit obviously will vary depending on the nature of the partitioning scheme.

4. The Modifiable Aerial Unit Problem (MAUP)

Problems of scale and zoning in Geographic Information Science are often couched in an analytical issue referred to as the **modifiable areal unit problem (MAUP)**. The MAUP occurs when statistical results or visual patterns embedded in maps differ according to changes in the scale or zoning of the partitioning scheme used to aggregate spatial data (Openshaw, 1984). The problem commonly occurs when data on individual objects, such as people, are disseminated in data sets in aggregated form, such that the actual spatial distribution of the original individual objects is unobtainable. This is typically a concern in the use of census, health, or voting data, where privacy protection precludes the release of data which identify individual people or households.

One consequence of the MAUP is that spatial patterns that may be observed in a map using one partitioning scheme may not be visible using another partitioning scheme where the scale or zoning differs. For instance, consider the spatial patterns illustrated in Figure 1; Here, a grayscale color scheme is used to represent the number of points within each spatial unit, where a lighter gray is used to represent a lesser number of points and a darker gray is used to represent a greater number of points. Note that because of the differences in the nature of data aggregation in the different partitioning schemes, the visual patterns expressed within the three maps differ substantially in the depiction of both the spatial distribution and magnitude of the data, even though all three maps are derived from the same original point data. The left panel suggests a high number, or density, of points to the west (left), whereas a simple change of partitioning by zoning (right panel) suggests a higher number in the north (top) and south (bottom). An altogether different visual pattern of the distribution of points is depicted in the middle panel.

Analogously, data aggregations using different spatial partitioning schemes can produce different statistical results, even for the calculation of basic descriptive statistics. Again, consider Figure 1. The minimum and maximum number of points per unit vary from 2 (minimum) and 3 (maximum) in the left panel, to 0 and 2 in the middle panel, to 0 and 4 in the right panel, respectively. Similarly, values of the mean and standard deviation also vary among the maps – from 2.5 (mean) and 0.5 (standard deviation) in the left panel, to 0.6 and 0.9 in the middle panel, to 2.0 and 2.0 in the right panel. Notably, the standard deviation is much smaller than the mean in the left panel, but is identical to the standard deviation in the right panel. In the middle panel the standard deviation is greater than the mean. Thus, a change in partitioning scheme can produce substantially different statistical characterizations of the data.

Particularly in the case of the scale effect for data calculated as rates or percentages, the use of a set of fewer and larger spatial units tends to inflate correlation values, as compared to the use of a set of a greater number of smaller spatial units (Guelke and Biehl, 1934). This is due to the smoothing that occurs when aggregating over larger spatial units, where variation over smaller areas is effectively hidden. The concept can also be considered an example of **Simpson's paradox**, where the results of statistical analyses may vary, and indeed, be of opposite sign, when data are disaggregated into subgroups (Simpson, 1951). In the case of the MAUP, the subgroups occur as subdivisions over space. The effect of the MAUP in multivariate statistical analysis, such as multiple regression, has been shown to be pervasive, unpredictable, and prone to manipulation (Fotheringham and Wong, 1991). The MAUP has been widely recognized as problematic in a variety of types of social and health science data (Maantay, 2007; Nelson and Brewer, 2017; Root, 2012)

A related common error in the use of aggregated data sets is the **ecological fallacy**, which occurs when one assumes that an individual has the same attributes as the aggregate group to which it belongs, i.e., in the case of spatial data, the spatial unit within which it resides. Consider, for example, ethnicity data from the US Census Bureau. Figure 2 shows a map of the percentage of the population identifying as Hispanic according to the 2010 US Census by county in southeastern Pennsylvania, including Philadelphia, Montgomery, Bucks, Delaware, and Chester counties. The county level map implies a homogeneous distribution of Hispanics throughout each individual county in southeast Pennsylvania, with a maximum value of 12% Hispanic. The ecological fallacy would be to assume that each individual location within each county takes on the same percent Hispanic value as the county within which it resides. This is, of course, unlikely to be the case, as ethnicity tends to cluster in neighborhoods smaller than a county.

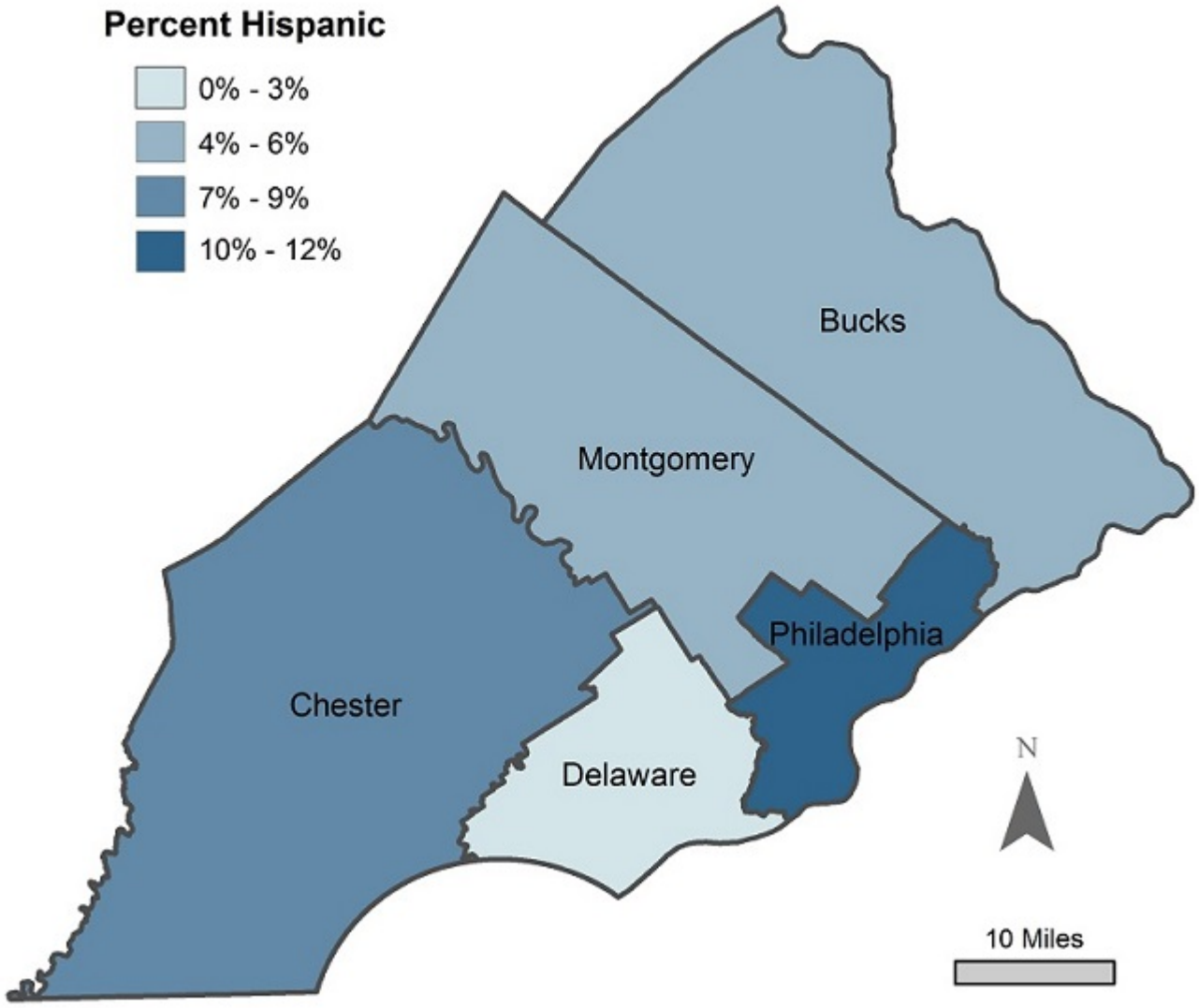


Figure 2. Percentage of the population identifying as Hispanic by county in southeastern Pennsylvania, US. Source: author.

Figure 3 shows a map of percent Hispanic over the same southeastern Pennsylvania region but instead of mapping the variable by county it is mapped by census tract, a much smaller spatial unit where multiple tracts nest within each county. This map clearly shows that the Hispanic population is concentrated in certain smaller areas within counties, whereas in other parts of each county there are few Hispanic residents. For instance, the tract level map shows that the Hispanic population is concentrated in the southern portion of Chester county and in the northeastern neighborhoods of Philadelphia county.

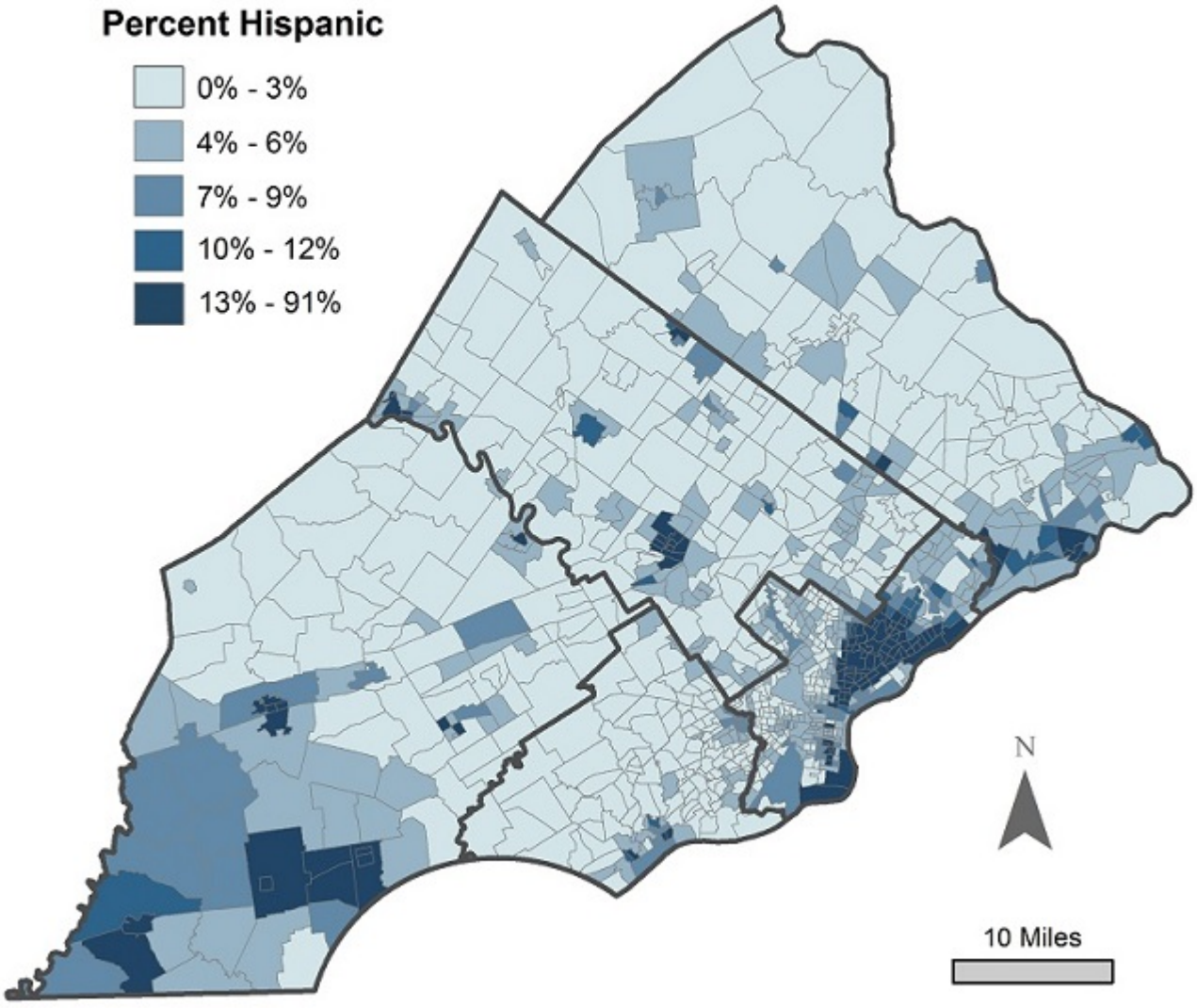


Figure 3. Percentage of the population identifying as Hispanic by US Census Bureau tract in southeastern Pennsylvania, US, with county boundaries overlain. Note the substantial within-county variation in percent Hispanic illustrated in the tract-level map. Source: author.

5. Combining Spatial Data with Different Partitioning Schemes

Another challenge concerns combining multiple spatial data sets together (referred to as data integration, fusion, or synthesis) when they use different partitioning schemes, whether in terms of differences in scale or zoning. This problem is sometimes referred to as integrating data with incompatible spatial units or, alternatively, a change of support in geostatistics. This would occur, for example, if one were to combine spatial data sets encoded using any combination of US Census tracts, zip codes, legislative districts, watershed boundaries, and so on, where spatial unit boundaries are not coincident across different data sets, nor are they spatially nested, but rather overlap. The problem typically

arises when data from a variety of sources are combined in a Geographic Information Systems (GIS)-based site suitability analysis, or other types of spatial analysis investigating the association or relationship among spatial variables, as with a statistical regression. Such analytical frameworks typically assume a domain of variable values mapped onto a consistent set of spatial unit observations.

The problem of integrating data with incompatible spatial units also emerges when bespoke regions are created using measurements of distance or accessibility to certain features. For example, a GIS operation may be used to generate a spatial unit that captures an area within a certain drive time of a hospital as a proxy for access to health care. Or, a GIS operation may be used to generate a spatial unit that captures the area within a certain distance of a hazardous feature as a proxy for an area of risk exposure. The problem of incompatible spatial units emerges when such a spatial unit capturing an area of access or exposure is integrated with, say, demographic data encoded in census enumeration units in order to assess the population with access to the amenity or exposed to the hazard. Because the spatial unit representing the area of access or exposure and the census data are encoded using different partitioning schemes whose boundaries are not consistent but rather overlap, the exact population within the area of access or exposure cannot be precisely calculated.

6. Addressing Problems of Scale and Zoning

There is no "solution" to problems of scale and zoning, as the issue is simply endemic to generalizing and aggregating data over space. However, there are approaches that attempt to mitigate the impacts of problems of scale and zoning on spatial analysis, or aid in the interpretation of analytical results that may be impacted by problems of scale and zoning. One approach is to attempt to disaggregate aggregated spatial data so as to estimate the underlying spatial variation that may be suppressed by the use of coarser resolution spatial units. **Areal interpolation** refers to a group of methods that aim to facilitate data integration with incompatible spatial units by estimating data values over small areas which are not directly observable, for which sampling has not occurred, or for which data are simply not readily available. The most basic areal interpolation method, referred to as areal weighting, simply assumes a homogeneous distribution within each spatial unit, thus allowing for data disaggregation to any spatial partitioning scheme. More sophisticated areal interpolation methods incorporate statistical approaches such as regression, kriging, or the expectation-maximization (EM) algorithm (Liu, Kyriakidis, and Goodchild, 2008). **Dasymetric mapping** methods for areal interpolation incorporate ancillary data, external data such as remotely sensed imagery that may be used to model data estimates over small areas, and which may be combined with other statistical areal interpolation approaches (Mennis, 2009).

Another approach to problems of scale and zoning seeks to infer characteristics or behaviors of individuals from aggregate data. Such methods, referred to as ‘ecological inference,’ attempt to exploit the association among variables captured at an aggregated level, for instance between aggregated race and voting behavior within a legislative district. The method of bounds focuses on the limits placed on one aggregated variable based on its association with another aggregated variable. For instance, given electoral data for a district with 400 voters, of whom 100 are white and 300 Hispanic, if 10% of the vote went

to a certain candidate, then the maximum number of whites who could have possibly voted for that candidate is 40 (i.e. 10% of the 400 total votes); Thus, the likelihood of any individual white (or Hispanic) person's voting behavior in the district is constrained. Statistical regression has also been employed for ecological inference, as well as the combination of the two approaches and their extension using Bayesian and hierarchical modeling methods (King, Rosen, and Tanner, 2004).

Researchers also sometimes conduct an analysis using a series of different spatial partitioning schemes to investigate the impact of the MAUP on mapping or statistical results. Typically, a researcher would examine the difference in results among different scales of data aggregation, for example in comparing the results of analyses of US Census block groups, tracts, and counties. Notably, the aim is not to establish that some statistical result persists over multiple scale of data aggregation; It is not necessarily the case that a statistical result that occurs using one partitioning scheme or scale of analysis but not another is somehow invalid (though it is certainly possible that statistical significance may be observed under random conditions at some scales of analysis and not others). Rather, such multiscale analytical results may be useful as an indication of the scale at which a geographic process or phenomenon under investigation operates or occurs.

References

Fotheringham, S. and Wong, D.S. (1991). The modifiable areal unit problem in multivariate statistics. Environment and Planning A, 23, 1025-1044.

Gehlke, C.E., and Biehl, K. (1934). Certain effects of grouping upon the size of the correlation coefficient in census tract material. Journal of the American Statistical Association, 29, 169-170.

King, G., Rosen, O., and Tanner, M.A. (Eds.) (2004). Ecological Inference: New Methodological Strategies. Cambridge, UK: Cambridge University Press.

Liu, X.H., Kyriakidis, P.C., and Goodchild, M.F. (2008). Population-density estimation using regression and area-to-point residual kriging. International Journal of Geographical Information Science, 22, 431-447.

Maantay, J. (2007). Asthma and air pollution in the Bronx: Methodological and data considerations in using GIS for environmental justice and health research. Health and Place, 13, 32-56.

Mennis, J. (2009). Dasymetric Mapping for Estimating Population in Small Areas. Geography Compass, 3, 727-745.

Nelson, J. K., and Brewer, C. A. (2017). Evaluating data stability in aggregation structures across spatial scales: Revisiting the modifiable areal unit problem. Cartography and Geographic Information Science, 44, 35-50.

Openshaw, S. (1984). The Modifiable Areal Unit Problem. Concepts and Techniques in Modern Geography No. 38. Norwich, England: Geo Books, Regency House.

Root, E.D. (2012). 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, 986-995. doi: 10.1080/00045608.2012.659621

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. Journal of the Royal Statistical Society, Series B, 13, 238-241.