Published in final edited form as:

Wiley Interdiscip Rev Comput Stat. 2012 ; 4(1): . doi:10.1002/wics.196.

Clusters in irregular areas and lattices

William F. Wieczorek^{1,*}, Alan M. Delmerico¹, Peter A. Rogerson², and David W.S. Wong³

¹Center for Health and Social Research, State University of New York College at Buffalo, CLAS A203 Buffalo, NY, USA

²Department of Geography, Wilkeson Hall, University at Buffalo, Buffalo, NY 14261, USA

³Geography and GeoInformation Science, College of Science, George Mason University, MS 6A2 Fairfax, VA, USA

Abstract

Geographic areas of different sizes and shapes of polygons that represent counts or rate data are often encountered in social, economic, health, and other information. Often political or census boundaries are used to define these areas because the information is available only for those geographies. Therefore, these types of boundaries are frequently used to define neighborhoods in spatial analyses using geographic information systems and related approaches such as multilevel models. When point data can be geocoded, it is possible to examine the impact of polygon shape on spatial statistical properties, such as clustering. We utilized point data (alcohol outlets) to examine the issue of polygon shape and size on visualization and statistical properties. The point data were allocated to regular lattices (hexagons and squares) and census areas for zip-code tabulation areas and tracts. The number of units in the lattices was set to be similar to the number of tract and zip-code areas. A spatial clustering statistic and visualization were used to assess the impact of polygon shape for zip- and tract-sized units. Results showed substantial similarities and notable differences across shape and size. The specific circumstances of a spatial analysis that aggregates points to polygons will determine the size and shape of the areal units to be used. The irregular polygons of census units may reflect underlying characteristics that could be missed by large regular lattices. Future research to examine the potential for using a combination of irregular polygons and regular lattices would be useful.

Keywords

cluster analysis; geographic information systems; geostatistics; neighborhoods

INTRODUCTION

Geographic areas are polygons that may represent social constructs, such as neighborhoods, or tabulation areas for economic, population, or other variables. The definition and measurement of geographic areas has been an enduring issue for social science and health research because there is no generally agreed upon definition of a neighborhood so that political boundaries or the geographic units of the census are often used by necessity. Geographic areas are crucial measurement units because all social, economic, and health interactions occur in a spatial context. Areal units are often used in multilevel models such as hierarchical linear models¹ or in geographic cluster analysis. The size and shape of the geographic areas are likely to influence the type and quality of the activities that occur there,

as well as the statistical properties of the areas. The modifiable areal unit problem² is a generalized conceptualization of the ecological fallacy in which statistical association may vary across geographical areas of differing scales. This combination of issues highlights some of the challenges of utilizing geographical areas for statistical analysis.

Assessing the impact of using geographic areas of various sizes and shapes is not a simple process because much of the data (e.g., US Census) are available only for specifically defined areas such as census block groups, tracts, zip-code areas, and political boundaries (e.g., cities, counties, states). However, when point data with a geographic reference (e.g., an address) are available, it is possible to allocate these points to polygons of various shapes and sizes. This approach allows for a systematic examination of how the visualization and statistical properties (e.g., clustering) vary across areas of differing shapes and sizes. In this analysis, alcohol outlets were allocated to areas of various shapes (irregular polygons defined by census geography and regular lattices of various shapes) and sizes (census tract, zip-code areas, and lattices of roughly similar size). Census tracts and zip-code areas were selected because these are commonly used in social science, planning, and health applications. A spatial clustering statistic and visualization were used to assess the impact of neighborhood shape and size on the geostatistical clusters and appearance of the same underlying data set.

BACKGROUND

Although areal units can be used in statistical analysis of almost anything that can be measured in a geographic area, the most relevant application for the current research is for social science and health applications. In these fields of research there is a major interest in measuring the effect of neighborhoods on health and socio-behavioral outcomes (e.g., crime). Many neighborhood effects in epidemiological and social science research have been found primarily through the utilization of multilevel models that incorporate census-based measures of neighborhood characteristics (e.g., poverty, income, vacant housing).^{3–6} A few illustrations are neighborhood effects on childhood aggression⁷ and neighborhood crime exposure as a factor in adverse birth outcomes.⁸ Of relevance to the current study is the broad literature on the role of alcohol outlets on alcohol consumption, problem drinking, and related problems (e.g., violence).⁹ The review by Livingston et al.⁹ shows that the findings for the impact of alcohol outlets are not always consistent and that one issue is the measurement of alcohol availability.

Researchers in these fields are aware of some of the current limitations of neighborhood-level measures and recommend further theoretical and empirical research to improve neighborhood-level measures.^{3–6} Further, other research has identified some key aspects of measuring neighborhood areas. Coulton et al.¹⁰ examined the perceptions of residents' definition of neighborhoods and found that the boundaries did not match census boundaries. However, they also found that the size of the residents' neighborhoods were similar to the area covered by census units such as tracts. Similarly, Kruger¹¹ found that 0.25 mile buffer areas that assessed structural deterioration were associated with crime and quality of life. Thus, neighborhood scale (particularly for urban and suburban locales) appears to be captured with areas similar to tracts. Also, zip-code areas are often used in multilevel research because many data are reported at that level or the only available spatial reference is a zip code. However, zip codes are designed for facilitating mail delivery and are known to have a number of problems such as shifting boundaries due to changes in the mail delivery system.¹²

For research that utilizes census data only, there are few options available for neighborhood-level units outside of traditional, small area census geographies. In some areas of research,

point data of interest are available such as the locations of crimes (crime rate), alcohol outlets (alcohol availability), or cases of a specific disease (asthma). In many instances though, census geographies are still used to define the neighborhood unit. However, in these cases other options for neighborhood area definitions are available but poorly understood as alternatives due to limited research exploring these and their associated geostatistical properties.

RESEARCH APPROACH

This project explores the spatial and statistical properties of various irregular polygons (i.e., census) and regular lattice areal units. Specifically, we examine the spatial clustering properties of census and lattice neighborhood units when aggregating the same set of data points to these areas. To better assess the impact of shape and size, the cluster bandwidth (i.e., the anticipated size of the cluster) is systematically varied. The units were selected to match the total number of some of the most commonly used neighborhood units (tracts and zip-codes areas). The rationale for this research is to gain a better understanding of potential alternative, noncensus neighborhood units for use in geospatial analyses, as well as for multilevel models of alcohol availability and environmental alcohol exposure.

METHODS

The current inquiry uses ‘real world’ point data to compare the statistical properties of neighborhood polygons. The neighborhood units to be evaluated are typical census units used as neighborhood measures (zip-codes areas [also known as ZCTAs or zip-code tabulation areas] and census tracts) and alternative neighborhood polygon areas (lattices of regular hexagons and squares). ZCTAs are boundaries based on zip codes that are used by the Census Bureau as a defined boundary; they may not fully align with zip codes used for mail delivery. The point data used are alcohol outlet locations for Erie County, NY for 2007. The data were obtained from the New York State Liquor Authority as addresses that were subsequently geocoded to point locations in ArcGIS. Census 2000 data were obtained for Erie County for census tract units and for ZCTAs. Regular grids of square and hexagonal polygons were developed for the same location (Erie County, NY). The number of squares and hexagons that comprise this area was chosen to reflect a similar number of units as compared to the two types of census polygons. Alcohol outlets were aggregated based on their spatial location to produce counts per neighborhood polygon. These counts were then denominated by the population in each unit; these population data are from the 2000 Census for census geographies but must be calculated for the lattice polygons. For the lattice polygons, census block centroids with associated population counts were spatially aggregated to develop a total population count for each lattice unit. The resulting measure for each census and lattice polygon is: Count of All Alcohol Outlets/10,000 Population. These alcohol outlet density rates were then used as the measure to evaluate the census and lattice neighborhood units’ spatial clustering properties.

One approach to finding geographic clusters, when their location is unknown, is described by Rogerson¹³; it avoids the use of simulation and requires that a bandwidth corresponding to an estimate of the cluster size be specified. When an estimate for the cluster size is unknown, this method can be used in a more exploratory mode; the significance associated with varying bandwidths can be compared and evaluated. This method comprises a global statistic and associated local statistics. The local statistic is a summation of differences between the observed and expected values for an individual polygon unit and its neighbors with a kernel-based weighting scheme and an adjustable bandwidth. The global statistic is the sum of squares of the local statistics for the study area. The individual local statistics can

be compared against a calculated, Bonferroni-adjusted critical value to test for significance (0.05 for all cases here).

Specifically, for a study area consisting of a set of regions, the method first constructs local statistics z_i for each region using original observations (denoted y_i) that have been standardized by subtracting their mean, and then dividing the result by the standard deviation. Weights are defined as $w_{ij} = (\sqrt{\pi}\sigma)^{-1} \exp(-d_{ij}^2/2\sigma^2)$, where d_{ij} is the distance from region i to region j and the bandwidth (σ) is chosen as the standard deviation of a normal distribution that matches the size of the hypothesized cluster. Here, a bandwidth of one ($\sigma = 1$) is defined as the standard deviation of the kernel being equivalent to the average distance between neighborhood region centroids. The local statistic for region i is defined as $z_i = \sum_j w_{ij} y_j$. Our statistic of interest is the maximum among these z_i values: $U = \max_i z_i$. The local statistics constructed at or near edges will not have as many regional neighbors as other

regions. Consequently, the sum of the squared weights ($\sum_j w_{ij}^2$), and the variance of the local statistic (which is based on the sum of the squared weights), will be smaller for regions near edges than for other regions. To help address these edge effects, as well as similar effects induced by regional configurations other than a square grid, an improvement can be made by using modified, scaled weights instead of the original weights. These are defined in terms of the original weights as follows:

$$w_{ij}^* = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}}.$$

The local statistics, z_i , all have normal distributions with mean zero and variance one.^{13,14}

The next step is to find the critical value U^* such that $Pr(U = \max_i z_i > U^*) = \alpha$ by using that value of U^* that leaves probability $(1 + 0.81 \sigma^2) / A$ in the tail of the standard normal distribution, for the total area of the region (A) that has been subdivided into a grid of m square cells, each having side of unit length. Alternatively, U^* may be approximated by

$$U_{\alpha}^* = \sqrt{-\sqrt{\pi} \ln \left(\frac{4\alpha(1+0.81\sigma^2)}{m} \right)}.$$

This method was employed for five bandwidths for each set of neighborhood units, specifically, bandwidths of 0, 0.5, 1, 1.5, and 2, with the nonzero bandwidths ranging from one half to twice the average distance between neighborhood region centroids. The results from this analysis were then mapped by polygon unit for each of the five bandwidths. The number, spatial location, and relationship of statistically significant neighborhood clusters were then evaluated across census and lattice neighborhood units.

RESULTS

Table 1 shows the global and local clustering statistics for both the zip-code-level and tract-level neighborhood units. All of the areas regardless of size, shape, or bandwidth show significant clustering at the global level. The largest global χ^2 values are found for the 0.05 bandwidth. Table 1 also shows the number of areal units that approach significance ($U_i >$

1.96) and exceed the critical value for U . Larger bandwidth leads to a greater number of areal units that exceed the local critical value. Irregular census polygons have more areas exceeding the critical value than do either square or hexagonal lattices.

The zip-code level cluster maps have substantial spatial similarities as well as differences (see Figure 1). The areal extent of the main cluster (which is the urban area of the county as shown by the small size of the zip-code areas) gets larger through the addition of neighboring areas as bandwidth increases. For the regular lattices in Figure 1, the main cluster becomes notably larger at higher bandwidths, eventually extending to about half of the total area. The square and hexagonal lattices had a main cluster that was much larger in spatial extent than for the irregular polygons at bandwidths of 1.5 and 2.0. However, outlying clusters tend to become nonsignificant as bandwidth increases, with no significant outlying clusters above a bandwidth of 0.5 for zip-code size areas. Significant small clusters to the south of the main cluster are present at small bandwidths (i.e., 0.0, 0.1). The greatest visual similarity is for the maps at the 0.05 bandwidth (the level with the maximum χ^2).

The tract-sized clusters in Figure 2 show substantially more detailed clusters than for the larger zip-code areas. There is a main, urban cluster (recognized by the small size of the tracts) for all tract-level polygons regardless of shape and bandwidth. Similar to the zip-code areas, this main cluster tends to increase in size as bandwidth increases; however, the overall geographic area covered at the highest bandwidths is notably smaller than for the zip-code results. Both the hexagonal and square lattices identify several outlying clusters only at low bandwidths (i.e., 0.0, 0.1). The irregularly shaped tracts, however, have at least one significant outlying cluster across all but the largest bandwidth.

In both figures, the light red polygons show areas with positive clustering that fails to reach significance. These nonsignificant positive clustering areas are often present at higher bandwidths and tend to coincide spatially with the significant clusters at the lower bandwidth. This result indicates that these outlying clusters tend to be of limited spatial extent, which is why they do not reach significance as the number of neighboring polygons increases as the kernel distance increases.

CONCLUSION

We examined the clustering of the same underlying data set of geographic points allocated to regular lattices (hexagons and squares) and irregular polygons (tracts and zip-code areas) to explore the potential for using neighborhood definitions other than those defined by political/census boundaries. This topic is an important issue in spatial analysis and multilevel modeling because few options are available and there is a limited literature on the topic, especially in health research. The results indicate that the specific circumstances of a spatial or multilevel analysis that uses points aggregated to polygons will drive the decision on the size and shape of the polygons.

There does not appear to be any clear overall preference for the type or shape of the polygons when assessing global clustering. There are recognizable differences in the local patterns of clustering; however, there is no specific polygon shape or size that appears to be preferable in all situations. There are nuanced differences between the choice of polygon shape, size, and clustering bandwidth. There is a strong tendency at higher bandwidths for clusters to become larger and more homogeneous, while outlying and smaller clusters are no longer significant. Larger bandwidths are more appropriate when the focus of the analysis is to identify large areas, and small, localized variation is considered unimportant. Smaller bandwidths (1.0 or less) are useful for identifying smaller and more outlying clusters. The smaller sized tract polygons were better suited to the recognition of smaller, localized

clusters, than are the zip-code level areas (as would be expected). A noteworthy implication for irregular census tract polygons is that these boundaries may be better able to capture meaningful clustering in relatively isolated areas, especially across a relatively wide range of bandwidths (see Figure 2).

The results suggest that an exploratory analysis should utilize more than one initial approach in terms of polygon shape to provide an improved understanding of spatial clustering across the entire region. Future research is needed to examine the feasibility of utilizing a stepped approach to polygon shape and size, particularly for regions that include a range of urban, suburban and rural areas. This approach might utilize a political or census boundary for relatively populated areas (urban and suburban) while using a square or hexagonal overlay for rural areas. The potential advantage of this approach would be to retain the utility of the more urban census boundaries while adding more spatial resolution in rural areas by substituting smaller area lattices for the relatively large census units typically found in rural locations.

Acknowledgments

This project was supported by grant R01 AA016161 from the National Institute on Alcohol Abuse and Alcoholism.

References

1. Raudenbush, SW.; Bryk, AS. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage Publications, Inc; 2002.
2. Openshaw, S. The Modifiable Areal Unit Problem: Concepts and Techniques in Modern Geography. Vol. 38. Norwich: Geo books; 1984.
3. Dietz RD. The estimation of neighborhood effects in the social sciences: an interdisciplinary approach 1. Social Sci Res. 2002; 31:539–575.
4. O'Campo P. Invited commentary: advancing theory and methods for multilevel models of residential neighborhoods and health. American Journal of Epidemiology. 2003; 157:9. [PubMed: 12505885]
5. Roux D, Ana V. The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. Epidemiol Rev. 2004; 26:104. [PubMed: 15234951]
6. Galea, S.; Osypuk, TL. What level macro? Choosing appropriate levels to assess how place influences population health. In: Galea, S., editor. Macrosocial Determinants of Population Health. New York: Springer; 2007. p. 399-435.
7. Romano E, Tremblay RE, Boulerice B, Swisher R. Multilevel correlates of childhood physical aggression and prosocial behavior. J Abnormal Child Psychol. 2005; 33:565–578.
8. Messer LC, Kaufman JS, Dole N, Savitz DA, Laraia BA. Neighborhood crime, deprivation, and preterm birth. Ann Epidemiol. 2006; 16:455–462. [PubMed: 16290179]
9. Livingston M, Chikritzhs T, Room R. Changing the density of alcohol outlets to reduce alcohol-related problems. Drug Alcohol Rev. 2007; 26:557–566. [PubMed: 17701520]
10. Coulton CJ, Korbin J, Chan T, Su M. Mapping residents' perceptions of neighborhood boundaries: a methodological note. Am J Comm Psychol. 2001; 29:371–383.
11. Kruger DJ. Verifying the operational definition of neighborhood for the psychosocial impact of structural deterioration. J Comm Psychol. 2008; 36:53–60.
12. Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas(ZCTAs) for the spatial analysis of epidemiological data. Int J Health Geogr. 2006; 5:58. [PubMed: 17166283]
13. Rogerson PA. A statistical method for the detection of geographic clustering. Geogr Anal. 2001; 33:215–227.
14. Siegmund DO, Worsley KJ. Testing for a signal with unknown location and scale in a stationary Gaussian random field. Ann Stat. 1995; 23:608–639.

FURTHER READING

15. Getis, A.; Fischer, MM., editors. Handbook of Applied Spatial Analysis. New York: Springer; 2010.
16. Gelfand, AE.; Diggle, PJ.; Fuentes, M.; Guttorp, P., editors. Handbook of Spatial Statistics. New York: CRC Press; 2010.
17. Rogerson, P.; Yamada, I. Statistical Detection and Surveillance of Geographic Clusters. New York: CRC Press; 2009.
18. Wieczorek WF, Delmerico AM. Geographic information systems. WIREs Comp Stat. 2009; 1:167–186.

FIGURE 1.
Clusters for zip-code-sized polygons.

FIGURE 2.
Clusters for tract-sized polygons.

Global and Local Clustering Statistics

TABLE 1

Bandwidth ()																									
		0				0.5				1				1.5				2							
		df	U ²	2	df	U ²	2	df	U ²	2	df	U ²	2	df	U ²	2									
Global	N	69	26.349	2728.794	102.168 ²	5.718	8701.758	108.100 ²	3.085	8941.692	74.159 ²	2.450	5792.202	49.845 ²	1.980	3553.932	33.697 ²								
ZCTA		69	26.349	2728.794	102.168 ²	5.718	8701.758	108.100 ²	3.085	8941.692	74.159 ²	2.450	5792.202	49.845 ²	1.980	3553.932	33.697 ²								
HexZCTA		67	6.484	11832.081	156.760 ²	6.565	12815.856	180.524 ²	3.560	12968.999	125.319 ²	2.730	8484.704	81.026 ²	2.164	4970.384	50.605 ²								
SqZCTA		68	7.320	5362.821	74.524 ²	6.918	6659.800	95.061 ²	3.378	8159.254	76.597 ²	2.573	5474.272	51.124 ²	2.117	3292.255	34.063 ²								
Tracts		249	151.575	1260.740	262.537 ²	11.977	3477.785	93.780 ²	4.840	4663.581	56.169 ²	3.352	4599.302	42.606 ²	2.760	4045.270	35.330 ²								
HexTract		243	20.350	3319.066	125.452 ²	15.610	4511.912	146.447 ²	5.137	8467.334	103.850 ²	3.759	8838.950	85.243 ²	3.090	7276.083	64.772 ²								
SqTract		245	19.878	3052.884	113.259 ²	15.022	4292.389	134.684 ²	5.125	8381.700	102.976 ²	3.789	8792.017	85.882 ²	3.140	7331.651	66.607 ²								

1															1.5															2														
		0				0.5				1				1.5				2						1				1.5				2												
		Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}	Max U _i	U _{crit}	U _i > 1.96	U _i > U _{crit}											
Local	N	69	27.16	3.2182	13	12	25.68	3.1671	23	22	22.26	3.0505	28	27	17.44	2.9186	28	27	13.37	2.7923	33	30																						
ZCTA		69	27.16	3.2182	13	12	25.68	3.1671	23	22	22.26	3.0505	28	27	17.44	2.9186	28	27	13.37	2.7923	33	30																						
HexZCTA		67	94.76	3.2101	10	6	94.53	3.1589	12	8	60.6	3.0419	11	11	36.92	2.9097	20	18	25.91	2.7830	26	26																						
SqZCTA		68	46.19	3.2143	9	7	50.71	3.1630	12	9	44.23	3.0463	12	10	29.88	2.9142	18	17	19.75	2.7877	24	22																						
Tracts		249	7.85	3.5542	37	13	12.81	3.4683	77	51	10.21	3.4030	105	81	8.09	3.3500	129	105	6.81	3.3053	143	122																						
HexTract		243	36.02	3.5481	23	14	40.99	3.4621	22	14	39.9	3.3966	25	16	32.26	3.3435	24	22	25.49	3.2987	31	26																						
SqTract		245	27.33	3.5501	27	17	32.78	3.4642	22	19	35.9	3.3988	22	16	32.89	3.3457	26	22	28.17	3.3009	32	29																						

*I*_p < 0.001.