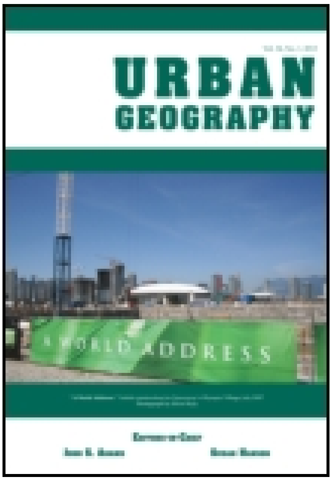

Urban Geography

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rurb20

USING SPATIAL SEGREGATION MEASURES IN GIS AND STATISTICAL MODELING PACKAGES

David W. S. Wong [a] & Wing K. Chong [a]

[a] Geography and Earth Systems Science, George Mason University, Fairfax, Virginia 22030 Tel: 703-993-1212 Fax: 703-993-1216 dwong2@gmu.edu

Published online: 16 May 2013.

RESEARCH NOTE

USING SPATIAL SEGREGATION MEASURES IN GIS AND STATISTICAL MODELING PACKAGES[1]

David W. S. Wong and *Wing K. Chong*
Geography and Earth Systems Science
George Mason University
Fairfax, Virginia 22030
Tel: 703-993-1212
Fax: 703-993-1216
dwong2@gmu.edu

Abstract: It is generally known that many traditional measures of segregation cannot distinguish different spatial population patterns. Several spatial measures of segregation have been proposed to overcome this problem, but these spatial measures are difficult to use because they require explicit spatial information and complex computation. This paper shows that Geographic Information Systems (GIS) can provide the spatial information required by these indices and that statistical packages can offer the complex computation functions needed to calculate the indices. We use ARC/INFO as the GIS package and S-Plus as the statistical package to demonstrate how spatial segregation indices can be calculated by combining the capabilities of these two types of systems.

The purpose of this paper is to demonstrate how Geographic Information Systems (GIS) can help in calculating a family of spatial indices of segregation. These spatial segregation indices utilize spatial information explicitly and require complex mathematical manipulations. GIS is very effective in capturing spatial information, but is generally not efficient in spatial analysis and modeling. Thus, the spatial components stored in GIS have to be extracted and used by other tools to perform the complicated calculations. By combining the capabilities of GIS (we use ARC/INFO) and powerful statistical packages (we use S-Plus), calculation of spatial segregation indices is within our reach due to advanced computing technology.

Urban geographers and sociologists have studied segregation for more than half a century. A significant development in this area of research is the use of the index of dissimilarity or segregation index D (Duncan and Duncan, 1955) to measure segregation objectively in a two-group setting. The index has been used extensively in studying different forms of segregation (e.g., Taeuber and Taeuber, 1965; Farley and Taeuber, 1974; White, 1987). The index is generally defined as

$$D = 0.5 \times \sum_i \left| \frac{b_i}{B} - \frac{w_i}{W} \right| . \qquad (1)$$

Urban Geography, 1998, **19**, 5, pp. 477–485.

The traditional setting for studying segregation is between White and Black populations, where b_i and w_i are denoted respectively as the Black and White population counts in areal unit i. B and W are the total Black and White population counts of the entire study area. This index of segregation is popular partly because it can be calculated easily using census data, and partly because it possesses other desirable properties. For instance, D falls between 0 and 1, where 0 indicates no segregation and 1 refers to a perfectly segregated situation.

The index of segregation has a major deficiency that constrains its usefulness. Morrill (1991) and Wong (1993) demonstrated that the D index is incapable of differentiating various patterns of ethnic group distribution. They used two contrasting spatial configurations: one with a pattern similar to a checkerboard, with different ethnic groups dominating adjacent cells, and another with two regions, each dominated by one group. The D index for these two configurations is 1, indicating perfect segregation, because D is a measure of internal homogeneity of areal units in a region. As long as each areal unit is dominated by one group (i.e., internally homogeneous), the D index will be 1. But most people would perceive that the configuration similar to the checkerboard should have a lower level of segregation than the two-region configuration. Our perception of segregation level is partly affected by the notion that people can interact across areal units; thus, two ethnic groups can have a high level of interaction in the checkerboard configuration, and people in the two-region configuration will have a lower chance to interact with members of another group. This perception was summarized by Newby (1982), who argued that segregation implies the separation of ethnic groups and that separation in space therefore can inhibit interaction among groups.

Based on Newby's (1982) notion of segregation, Morrill (1991) introduced a modified D that includes spatial adjacency information to reflect spatial interaction potential across areal boundaries. As a result, the modified index can differentiate various spatial patterns. The index is defined as

$$D(adj) = D - \frac{\sum_i \sum_j |c_{ij}(z_i - z_j)|}{\sum_i \sum_j c_{ij}} \tag{2}$$

where c_{ij} is the i-jth element in a binary connectivity matrix, 1 reflects adjacency and 0 otherwise, and z_i is the proportion of Blacks in areal unit i. Formally, z_i is defined as

$$z_i = \frac{b_i}{b_i + w_i}. \tag{3}$$

Thus, $z_i - z_j$ is the difference in the proportion of Blacks across zones, or the difference in the spatial concentration of Blacks.

D(adj) assumes that as long as areal units i and j are next to each other, the population in these areal units can interact. Wong (1993) suggested that the interaction can be modeled more realistically if the intensity of interaction is affected by the length of the common boundary. Thus, he proposed

$$D(w) = D - \frac{\sum_{i}\sum_{j} w_{ij}\left|z_i - z_j\right|}{\sum_{i}\sum_{j} w_{ij}} \tag{4}$$

and

$$w_{ij} = \frac{d_{ij}}{\sum_{j} d_{ij}} \tag{5}$$

where d_{ij} is the length of the common boundary of areal units i and j.

Wong (1993) further argued that the ease of crossing the areal boundary to interact with people in neighboring units is determined by proximity to or accessibility of the boundary. This factor is in turn affected by the geometric characteristics of areal units. These geometric characteristics include size of regions and length of zonal boundary, suggesting another spatial index. It is defined as

$$D(s) = D - \sum_{i}\sum_{j}\left\{\frac{w_{ij}\left|z_i - z_j\right|}{\sum_{i}\sum_{j} w_{ij}} \times \frac{\frac{1}{2}\left(\frac{P_i}{A_i} + \frac{P_j}{A_j}\right)}{MAX\left(\frac{P}{A}\right)}\right\} \tag{6}$$

where P_i and A_i are the perimeter and area, respectively, of unit i. The perimeter-area ratio reflects the perimeter shared by each unit of area, which can be a proxy of population size. MAX (P/A) is the maximum perimeter-area ratio in the study region. The numerator in the last term is the average compactness of two neighboring areal units. The average will be smaller than the MAX (P/A), which reflects the least compact geometric shape that is ideal to facilitate interzonal interaction between population groups. Thus, the interaction across zonal boundary facilitated by the length of common boundary is once more constrained by the compactness of the two areal units.

The approach adopted by Morrill and Wong is basically a spatial interaction approach suggested by Newby (1982). The spatial indices introduced by Morrill and Wong incorporate spatial elements capturing the opportunity of interaction among groups over areal unit boundaries, an approach that extends the work of Jakubs (1979) regarding distance as a major factor affecting segregation. However, Massey and Denton (1988) criticized spatial approaches using neighborhood or adjacency information as impractical and almost impossible. They believed that the information has to be extracted by visual inspection of census maps. Moreover, the D(w) and D(s) indices require geometric information such as perimeter and area of areal units, which seem to be difficult to obtain. Massey and Denton's perceptions of this approach are also shared by many researchers. This is a major reason why these spatial indices have not been adopted.

It is true that these spatial measures cannot be used easily because calculations are more complicated than what popular programs such as spreadsheets can handle. The complexity of these spatial measures lies partly in the spatial components incorporated into the indices. Though these spatial components, which include neighborhood information and geometric characteristics of areal units, seem difficult to obtain, they are readily available through GIS. They either can be used to calculate the spatial indices inside GIS or can be extracted from GIS as inputs to other calculation tools, such as mathematical and statistical modeling packages, to derive these indices. This paper demonstrates how ARC/INFO, a GIS package, and S-Plus, a statistical package, can work together to derive the family of spatial segregation measures. ARC/INFO is one of the most popular GIS packages. We use S-Plus to handle the complex computations of the spatial indices because it has a bridge (GisLink) to ARC/INFO that transfers data seamlessly between the two packages. In addition, S-Plus has the powerful data and matrix manipulation capabilities that are essential in computing the indices.

IMPLEMENTING SPATIAL SEGREGATION INDICES IN GIS

Given today's computation technology, it is feasible to implement all of the above spatial indices of segregation to analyze real-world situations. There are several ways to implement these spatial indices in a GIS environment. Probably the most straightforward method is to utilize GIS to extract the pertinent geographic and geometric information. Many GIS packages store all the geographic and geometric variables required to compute the above spatial indices. These variables include membership of neighboring units (c_{ij}), length of common boundary (d_{ij}), perimeter (P_i), and area (A_i). Different GIS packages store these variables in different ways; therefore, the detailed methods to extract these spatial and geometric variables are system-dependent. This paper will focus on the ARC/INFO system environment. Spatial and geometric variables are extracted from ARC/INFO and then accessed through GisLink by S-Plus, which in turn performs the mathematical modeling.

ARC/INFO is not only one of the most popular commercial GIS pakages, but also one of several that store in a relational database structure all the pertinent geographic variables required by this project. S-Plus is a powerful statistical modeling and visualization tool with robust matrix manipulation capabilities. The GisLink module allows one to move data back and forth between ARC/INFO and S-Plus, and to execute S-Plus commands in ARC/INFO. In other words, users can examine maps displayed in ARC/INFO while simultaneously visualizing the data using statistical graphics in S-Plus.

ARC/INFO utilizes the node-arc topological data structure to store cartographic data. This type of topological data is extremely important for various spatial modeling techniques, including spatial statistics (Griffith, 1989). In the ARC/INFO data structure, the Arc Attribute Table (AAT) stores the internal number for the left polygon (LPOLY#) and the right polygon (RPOLY#) of an arc (ESRI, 1994). Based on these two items, we can create the connectivity or adjacency matrix (i.e., c_{ij} in Equation 2). The AAT file also

ARC/INFO

Process TIGER (TIGER92.AML)
Import STF population data into PAT (ADDITEM, JOINITEM)
Construct the sparse spatial weights matrices (CONWEIGH.AML)
Convert the sparse matrices into full C matrices (a C program)
Calculate the modified perimeter measure (PERI.AML and TAB.AML)
Merge the measure back to PAT

S-Plus

Remove outer polygon record in the PAT
Remove unnecessary items in the PAT
Scan in each population group as vectors
Scan in the modified perimeter and area as vectors
Scan in the vectors by row and by column to form the Z and Z' matrices
Multiply the corresponding value in C by (Z - Z')
Calculate the perimeter-area ratio for each areal unit
Derive D(adj), D(w), and D(s) using matrix operation

Fig. 1. Operations inside ARC/INFO and S-Plus environments.

indicates the length of an arc, which in many cases can be a common boundary between two areal units (an arc will not be a common boundary if one side of the arc is the outer or universal polygon). Therefore, we can find out the length of the common boundaries (i.e., d_{ij} in Equations 4 and 5) from the AAT. The Polygon Attribute Table (PAT) file stores the characteristics of polygons. The standard items in a PAT file include area (A_i) and perimeter (P_i) of each polygon.

The major steps in extracting variables from ARC/INFO are summarized in Figure 1. We use an Arc Macro Language (AML) script (TIGER92.AML)[2] to import the 1992 Census TIGER/Line files. Polygon coverages at the census tract and block group levels are built. We use two other AML scripts (PERI.AML and TAB.AML) created by the authors to compute a modified perimeter measure. The perimeter used in the spatial indices excludes the sides of a polygon that are shared with the universal polygon. It is assumed that the universal polygon contains only the study region. Because the area of each areal unit is already stored in the PAT file, the perimeter-area ratio is calculated from the PAT using the modified perimeter derived from the two AML scripts. Then population attribute data are imported from Census Summary Tape Files (STF) 3A to the Polygon Attribute Table (PAT) in ARC/INFO. Another AML script, CONWEIGH.AML, written by Dodson (Anselin et al., 1992), is used to create sparse spatial weights matrices. This AML script extracts information from AAT and PAT files to construct the sparse matrices, which record all neighboring units of a given areal unit. If a stochastic spatial weights (i.e., row standardized) matrix based on length of common boundary is requested, the length instead of the binary adjacency information will be used as the weight. A program written in the programming language C and provided by Anselin et al. (1992) is used to convert the sparse matrices into full matrices with dimension n by n, where n is the number of areal units in the entire study region.

Several data sets are created in the ARC/INFO environment. They include the PAT files with population count and the perimeter-area ratio for each polygon except the universal polygon. The binary connectivity matrix and the spatial weights matrix based upon common boundary are also available. These matrices are accessed and manipulated by S-Plus for analysis.

COMPUTATION AND MANIPULATION IN S-PLUS

The major operations in S-Plus are summarized also in Figure 1. S-Plus is an object-oriented statistical package. All elements are treated as objects with defined behavior or operations. When S-Plus accesses the PAT file, the file becomes an object. Because the outer or universal polygon is not used in the analysis, its record in the PAT is removed. Other non-pertinent items or variables in the PAT object are also dropped, so that only the population data and the perimeter-area ratios are retained. For further manipulation, we treat population counts for each group as a vector. The perimeter-area ratio is also read as a vector. Using some simple vector manipulations, we derive the vector z, which is the proportion of Blacks in every areal unit in Equations (2), (3), (4), and (6). The vector is rescanned by row to derive a matrix Z_i such that it has n identical z rows. The matrix is then transposed to derive Z_j .

Other components are also scanned in as matrix objects. They include the binary connectivity matrix (C) and the spatial weights matrix based upon length of common boundary (W). Instead of using ordinary matrix functions, we utilize the algebraic operations for corresponding cells in S-Plus for different matrices. $Z_i - Z_j$ results in the difference in minority concentration for all possible pairs of areal units. But not all i-j pairs are adjacent to each other. Thus, $|Z_i - Z_j|$ is multiplied by the corresponding elements of either the C matrix or the W matrix. Other calculations are relatively straightforward matrix algebraic operations or arithmetic operations and it is not necessary to describe them in detail here.

CASE STUDIES

We used STF 3A data and TIGER/Line files for Washington, DC; we also implemented the above procedures for the entire state of Connecticut, treating each of its eight counties as a study region. In both cases, we calculated D, D(adj), D(w), and D(s) at the census tract level and the census block group level. The results at the tract level are reported in Table 1, and those at the block group level are reported in Table 2.

As previous literature had predicted, D was the highest in all situations because it does not take into account the fact that potential opportunity for interaction among populations in different areal units may lower the level of segregation. In most counties in Connecticut, D(adj) and D(w) were very similar, indicating that the length of common boundary may not be a significant factor. As expected, D(s) was higher than D(adj) and D(w), but

TABLE 1.—THE FAMILY OF SEGREGATION MEASURES AT THE TRACT LEVEL, CONNECTICUT COUNTIES AND WASHINGTON, DC[a]

Counties/city	D	D(adj)	D(w)	D(s)
Fairfield	0.67	0.57	0.57	0.63
Hartford	0.70	0.63	0.63	0.68
Litchfield	0.44	0.43	0.43	0.43
Middlesex	0.53	0.47	0.48	0.51
New Haven	0.67	0.59	0.59	0.67
New London	0.49	0.42	0.42	0.48
Tolland	0.46	0.44	0.43	0.45
Windham	0.48	0.47	0.47	0.47
Washington	0.77	0.62	0.63	0.70

[a]All results are rounded to two decimal places.

TABLE 2.—THE FAMILY OF SEGREGATION MEASURES AT THE BLOCK GROUP LEVEL, CONNECTICUT COUNTIES AND WASHINGTON, DC[a]

Counties/city	D	D(adj)	D(w)	D(s)
Fairfield	0.70	0.62	0.62	0.68
Hartford	0.73	0.67	0.68	0.72
Litchfield	0.63	0.61	0.61	0.62
Middlesex	0.58	0.54	0.54	0.56
New Haven	0.71	0.62	0.63	0.71
New London	0.57	0.51	0.51	0.57
Tolland	0.57	0.55	0.54	0.56
Windham	0.67	0.65	0.65	0.66
Washington	0.79	0.65	0.66	0.75

[a]All results are rounded to two decimal places.

lower than D in most cases, because in D(s) the potential for interaction among populations in different areal units is factored by the shape of the areal units. Thus, the level of segregation indicated by D(s) was not diminished from D by as great a magnitude as the levels indicated by D(adj) and D(w). One interesting observation is that as we moved the analysis from the tract level to the block group level, the differences between D and D(s) became even smaller in Connecticut counties. In fact, at the block group level, D and D(s) were similar in several cases. For Washington, DC, the differences among the segregation indices were more apparent.

TABLE 3.—THE FAMILY OF SEGREGATION MEASURES AT THE TRACT AND BLOCK GROUP LEVELS, THE METROPOLITAN REGIONS OF HARTFORD AND NEW HAVEN[a]

Cities	N	D	D(adj)	D(w)	D(s)
Tract level					
Hartford	49	0.65	0.49	0.49	0.58
New Haven	30	0.50	0.26	0.28	0.49
Block group level					
Hartford	106	0.66	0.50	0.51	0.59
New Haven	130	0.62	0.43	0.43	0.62

[a] All results are rounded to two decimal places.

DISCUSSION

We must bear in mind two very important and related aspects of the Connecticut study. First, segregation indices were calculated for each county. Within each county, the enumeration units (tracts or block groups) might vary tremendously in size. Units closer to cities were usually smaller, whereas rural units were larger. Second, all segregation measures we used were global measures. These measures provided a summary of the overall condition of the study regions (each county). Because our study regions in Connecticut were relatively large and some regions were highly heterogeneous, it was possible that the impact of small enumeration units on the indices might have been counterbalanced by the impact of large enumeration units. Thus, including highly heterogeneous spatial components in those spatial indices may not provide spatial measures significantly different from the non-spatial D index.

To demonstrate that a large and heterogeneous study region may nullify the incorporation of spatial factors in measuring segregation, we created two spatial subsets surrounding the cities of Hartford and New Haven in Connecticut. We calculated the family of spatial segregation indices for these two regions (Table 3). The segregation measures were quite different among these indices in the Hartford area, but they varied less in the New Haven area. When we examined the geometric variations of enumeration units in the two city regions, we found that the standard deviations in area and perimeter measures for New Haven were almost twice as large as the standard deviations of the corresponding measures for Hartford. In other words, the enumeration units in the New Haven region were so heterogeneous that the impacts from spatial factors canceled out each other, and thus the D and its spatial variants were not that different. On the other hand, the Hartford region consisted of enumeration units of similar geometric characteristics; thus including the spatial factors in measuring segregation became important.

CONCLUSION

We have revisited existing literature, arguing that D is not very useful in measuring spatial segregation, and reviewed several spatial measures of segregation. These spatial measures are conceptually and theoretically sound, but have not been used partly because of the difficulty of implementing them in real-world analysis. Though the process requires some clever matrix manipulations and computations, we have demonstrated that these spatial segregation measures that were virtually impossible to implement in the past can be implemented easily with GIS and other mathematical modeling tools in the current "GIS era."

Although the relative performance of different spatial segregation measures is beyond the scope of this paper, the general approach we have adopted here can be extended to other spatial modeling methods that rely heavily on spatial information. If readers are interested in implementing the spatial segregation measures in GIS and S-Plus, they may contact the authors to obtain necessary AML scripts and S-Plus codes for the calculation.

NOTES

[1] We would like to thank the College of Arts and Sciences at George Mason University for providing a Graduate Research Assistantship to Wing Chong, who was David W. S. Wong's graduate student. We would also like to give thanks to Joseph S. Wood of George Mason University and Professor James O. Wheeler, Co-Editor of *Urban Geography*, for their comments.

[2] This AML script is available from ESRI, Inc., the vendor of ARC/INFO.

LITERATURE CITED

Anselin, L., Hudak, S., and Dodson, R., 1992, *Spatial Data Analysis and GIS: Interfacing GIS and Econometrics Software*. Santa Barbara, CA: National Center for Geographic Information Analysis, University of California.

Duncan, D. and Duncan, B., 1955, A methodological analysis of segregation indexes. *American Sociological Review*, Vol. 20, 210–217.

ESRI, 1994, *Understanding GIS: The ARC/INFO Method*. Redlands, CA: Environmental Systems Research Institute.

Farley, R. and Taeuber, A. F., 1974, Racial segregation in the public schools, 1967 to 1972: Assessing the effect of governmental policies. *Sociological Focus*, Vol. 8, 3–26.

Griffith, D. A., 1989, *Advanced Spatial Statistics*. New York, NY: Kluwer.

Jakubs, J. F., 1979, A consistent conceptual definition of the index of dissimilarity. *Geographical Analysis*, Vol. 11, 315–321.

Massey, D. S. and Denton, N. A., 1988, The dimensions of residential segregation. *Social Forces*, Vol. 67, 281–315.

Morrill, R. L., 1991, On the measure of geographical segregation. *Geography Research Forum*, Vol. 11, 25–36.

Newby, R. G., 1982, Segregation, desegregation, and racial balance: Status implications of these concepts. *The Urban Review*, Vol. 14, 17–24.

Taeuber, K. E. and Taeuber, A. F., 1965, *Negroes in Cities: Residential Segregation and Neighborhood Change*. Chicago, IL: Aldine.

White, M. J., 1987, *American Neighborhoods and Residential Differentiation*. New York, NY: Russell Sage.

Wong, D. W. S., 1993, Spatial indices of segregation. *Urban Studies*, Vol. 30, 559–572.