

1998

A Perimeter-based Clustering Index for Measuring Spatial Segregation: A Cognitive GIS Approach

Dennis P Culhane, *University of Pennsylvania*

Chang Moo Lee, *University of Pennsylvania*

A perimeter-based clustering index for measuring spatial segregation: a cognitive GIS approach

C-M Lee

Wharton Real Estate Center, University of Pennsylvania, 3600 Market Street, Philadelphia, PA 19104-2648, USA; e-mail: leecm@wharton.upenn.edu

D P Culhane

School of Social Work, University of Pennsylvania, PA 19104-2648, USA;
e-mail: dennis@cmhpsr.upenn.edu

Received 25 November 1996; in revised form 2 May 1997

Abstract. Many efforts have been made to develop segregation indices that incorporate spatial interaction based on the contiguity concept. Contiguity refers to how similar the concentration of the subject of interest in one areal unit is to that in adjacent areal units. However, highly segregated situations are typically considered to be isolated sections or enclaves rather than smoothly formed peaks of concentration in space. Therefore, highly segregated enclaves may not exhibit contiguity. In this paper, a new index to measure the degree of clustering is developed and it is compared with the existing indices of concentration or segregation. The proposed clustering index (I^c) tends to give more weight to 'enclaveness' rather than contiguity alone. This may be a good property for those cases in which the primary concern of an investigator is the formation of enclaves of a socioeconomic subject, including minority populations, poverty, crime, epidemics, and mortgage red-lining. Additionally, its property of robustness to the citywide rate allows us to perform properly an intercity comparison of a given subject by index score even when the citywide rate varies significantly, unlike the other measures.

1 Introduction

Numerous efforts have been made to develop a proper index to measure the spatial segregation of a population group.⁽¹⁾ Though each index characterizes somewhat different aspects of a spatial distribution, one can distinguish two types of indices: measures ignoring spatial interaction between areal units; and measures incorporating spatial interaction.

The problem with measures of segregation that lack spatial interaction components, including the dissimilarity index, the Gini coefficient, and the entropy index (Theil, 1972), is well illustrated in the case of the 'checkerboard problem', described by White (1983). Several efforts have been made to develop segregation indices that incorporate spatial interaction, including the index of spatial proximity (White, 1986) and the distance-based index of dissimilarity (Morgan, 1982). In general, these measures include spatial interaction by distance or binary adjacency between two areal units. Recently Wong (1993) formulated a new segregation index, which uses the length of the common boundary of two areas as an indicator of the degree of social interaction between the residents of the two areas.

Spatial-interaction measures in geography, and segregation measures incorporating spatial interaction in sociology are similar in concept. However, spatial-interaction measures in geography are based only on distribution in physical space, whereas the segregation measures take account in population distribution overlaid on physical space along with the distribution of physical space itself. The spatial-interaction segregation indices in sociology are derived from Dacey (1968) and Geary (1954), where they have been labeled 'contiguity' measures.

⁽¹⁾ See Massey and Denton (1988) for the existing measures.

Contiguity refers to how similar the concentration of the subject of interest in one areal unit is to that in adjacent areal units. If the figures for adjoining areal units are generally closer than those for the areal units not adjoining, this condition yields a contiguous distribution of the subject of interest (Dacey, 1968; Geary, 1954). This contiguity aspect of spatial distribution has been well developed into a field of spatial statistics known as spatial autocorrelation. In the last twenty years, a number of instruments for testing for and measuring spatial autocorrelation have appeared (Anselin, 1988). To geographers, the best-known statistics are Moran's I , and, to a lesser extent, Geary's c (Cliff and Ord, 1973).

In some cases (Massey and Denton, 1988), the contiguity measures in geography are interpreted as clustering indices, with some modifications. However, a high degree of clustering does not always represent a high degree of contiguity. For example, one can imagine a spatial distribution pattern in which one subject of interest forms isolated enclaves which have visible boundaries. That distribution is not supposed to yield a high degree of contiguity, as the difference at the boundaries of the enclaves reduces the overall degree of contiguity. In the real world, one is generally concerned about isolated enclaves of a population group which are recognized by both high concentration and separateness, rather than the spatial contiguity of their distribution alone. For point data in the natural space, there are some measures which use nearest-neighbor methods to describe the degree of clustering (Ripley, 1981). However, for areal data in urban space, overlaid with population, one needs to have a different measure of clustering rather than the existing contiguity measure of segregation.

In this paper, a new index to measure the degree of clustering is developed and then compared with the existing indices of segregation. In section 2, clustering is defined in an operational way, and in sections 3 and 4 a method for calculating the new clustering index and its properties are discussed. In section 5, four existing indices to be compared with the new clustering index are discussed briefly, and the clustering index and the four other indices are compared in two hypothetical settings including binary distribution in a regular lattice, and semicontiguous distribution in a regular lattice. In section 6, the five indices are compared in a real-world application, the five boroughs of New York City.

2 Operational definition of clustering

When the spatial distribution of a subject of interest on a map is examined, viewers tend to draw arbitrary boundaries of clusters and define a set of clusters cognitively, whether it is a point distribution or an areal data distribution. This cognition could be said to have three attributes: the total size of the clusters, their shape, and the closeness between them. Here, a clustering index is derived based on these three attributes.

In order to draw the boundaries of clusters, an objective way to define clusters is needed. In an urban setting, the probability of occurrence of a subject in an areal unit depends on population rather than the size of the areal unit. For example, all other factors being equal, the expected number of the poor in a census tract depends on the number of people residing in the tract rather than the physical size of the tract. Once the rate of an object group to population in each tract is determined, the next issue is how to define concentration of the object group. One popular way of defining concentration is the location quotient (Q^1).

The location quotient is a device frequently used to identify specialization, concentration, or the potential of an area for selected employment, industry, or output indicators (Bendavid-Val, 1983; Chen, 1994). It refers to a ratio of the fractional share of the subject of interest at the local level to the ratio at the regional level. When a local Q^1 in a region is greater than 1, the locality has a higher concentration of the subject of interest relative to the other localities of the region combined. For example, a census

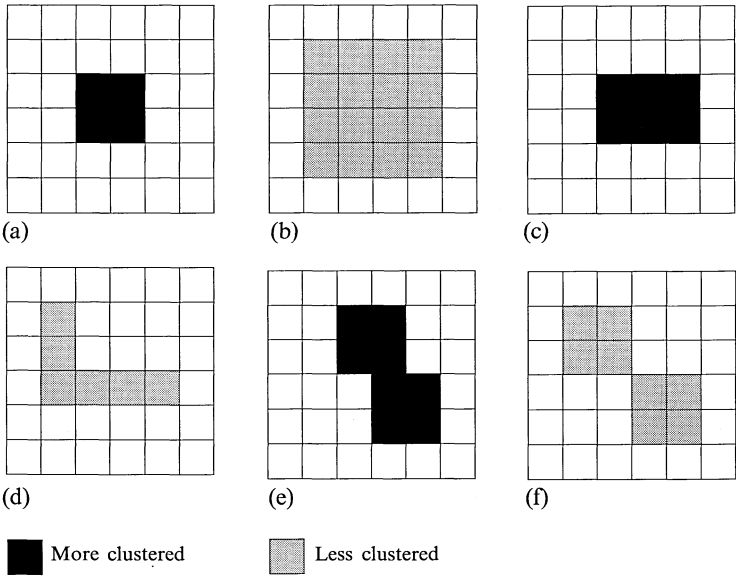


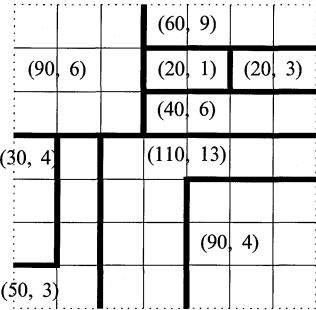
Figure 1. Size, shape, and adjacency of subclusters: (a) small size, $p = 8$; (b) large size, $p = 16$; (c) regular shape, $p = 10$; (d) irregular shape, $p = 14$; (e) adjacent, $p = 14$; (f) separated, $p = 16$.

tract or a block group may be equivalent to a locality, and a city to a region. Thus, Q^1 may be used to identify census tracts that contain a higher percentage share of a subject of interest than a city as a whole, and which have a Q^1 value greater than 1. Here, adjacent areal units showing a high concentration of the subject form a few clusters on the map.

Once clusters are obtained, one needs to quantify the size, shape, and closeness of the clusters. A measure that combines these three factors is the total perimeter of the clusters. When shape and adjacency of the clusters are the same, the total perimeter (P) of the clusters is a proper measure of the total size [see figures 1(a) and 1(b)]. When the size and adjacency of the clusters are constant, circular shapes have the minimum possible values [see figures 1(c) and 1(d)]. When the size and shape of the clusters are the same, two adjoining clusters have a smaller total perimeter than two separated clusters [see figures 1(e) and 1(f)]. Therefore, one can measure the degree of clustering by assessing how small the total perimeter of the clusters (the concentrated areas of a subject of interest) is, where the concentrated areas are selected by Q^1 .

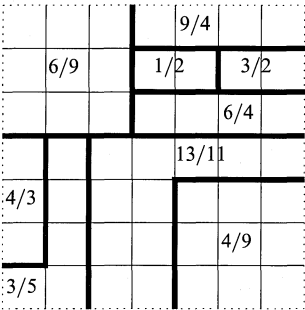
3 Calculation of a clustering index

Based on the operational definition of clustering, one can develop a clustering index. In order to illustrate the process for calculating the clustering index, a hypothetical city space is assumed as in figure 2(a) (see over), where P is the population of each census tract, and x is the number of an object group in the tract. As a first step, every census tract is divided into two groups: highly concentrated census tracts, and less concentrated census tracts based on Q^1 [see figure 2(b)]. When two highly concentrated tracts are adjacent to each other, the common boundary lines are deleted and the two polygons of the tracts are merged to form one polygon [see figure 2(c)]. This merging process continues and finally a few polygons result, which represent highly concentrated areas or clusters [see figure 2(d)]. The more adjacent the highly concentrated tracts are, the more common boundaries are erased, and the smaller the ratio of the sum of the perimeters

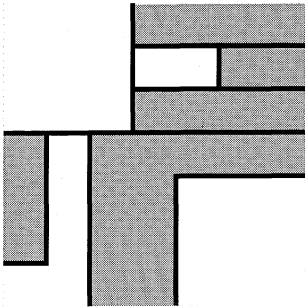


Total perimeter: 33 (61 – 28)
(excluding the boundary of the study area)
Total population: 490
Total object group: 49
The numbers in parentheses are (P, x)

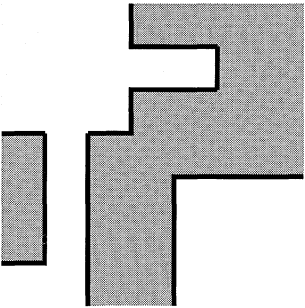
(a)



(b)



(c)



(d)

Total perimeter of merged polygons: 23
(excluding the boundary of the study area)
Clustering index = 1 – 23/33 = 0.30

Figure 2. The merging process used to calculate the clustering index: (a) population and object group; (b) calculating the location quotient; (c) identifying concentrated tracts; (d) merging concentrated tracts.

of the merged polygons to the sum of the perimeters of the original tracts will be. In our case, the boundaries of the study area are not included in the calculation.

In this concept, the clustering index can be denoted as follows:

$$I^c = 1 - \frac{\sum_i \sum_j |l_i - l_j| b_{ij}}{\sum_i \sum_j b_{ij}},$$

where l_i is a binary value for tract i (1 if $Q^1 \geq 1$; 0 if $Q^1 < 1$); and b_{ij} is the length of the common boundary between census tracts i and j (0 if tracts i and j are not connected or $i = j$). If a pair of adjacent tracts have the same l value (either 1 or 0), $|l_i - l_j|$ becomes 0, and their common boundary b_{ij} does not count in the numerator in the equation. Only the boundary between a pair of adjacent tracts which have different l values (high concentration versus low concentration) remains.

One advantage of this measure for an irregular polygon layout is that the degree of proximity between polygons is automatically taken into account during the merging

process. If the shared boundary between two tracts is longer than the boundaries between the others, the intensity of interaction between the two tracts is higher than that between the others. If the index is calculated manually with a map, it should be problematic. However, the use of an arc-node typology table such as Arc Attribute Table (AAT) in Arc/Info (ESRI Inc., Redlands, CA) in conjunction with a polygon topology table such as Polygon Attribute Table (PAT) in Arc/Info can simplify the calculation process.⁽²⁾

4 Properties of the clustering index

In a regular lattice, the clustering index proposed can achieve a minimum value of 0, when the subject of interest is distributed perfectly like a checkerboard, and its maximum value approaches 1, when all members of the subject of interest reside in an areal unit. If this subject is distributed randomly, the mean expected value of the index will be 0.5, as the probability that two adjacent areal units have different categories of Q^1 (greater than 1 or equal to and less than 1) is 0.5 and that the common boundary between the two areal units remains after the polygon merging process.

The statistical distribution of the index cannot be obtained through an analytic derivation, because Q^1 and its binary categorization into the high-concentration area and the low-concentration area are not easily incorporated into the analytic derivation process. In this case, we apply the Monte Carlo method to establish the distribution of

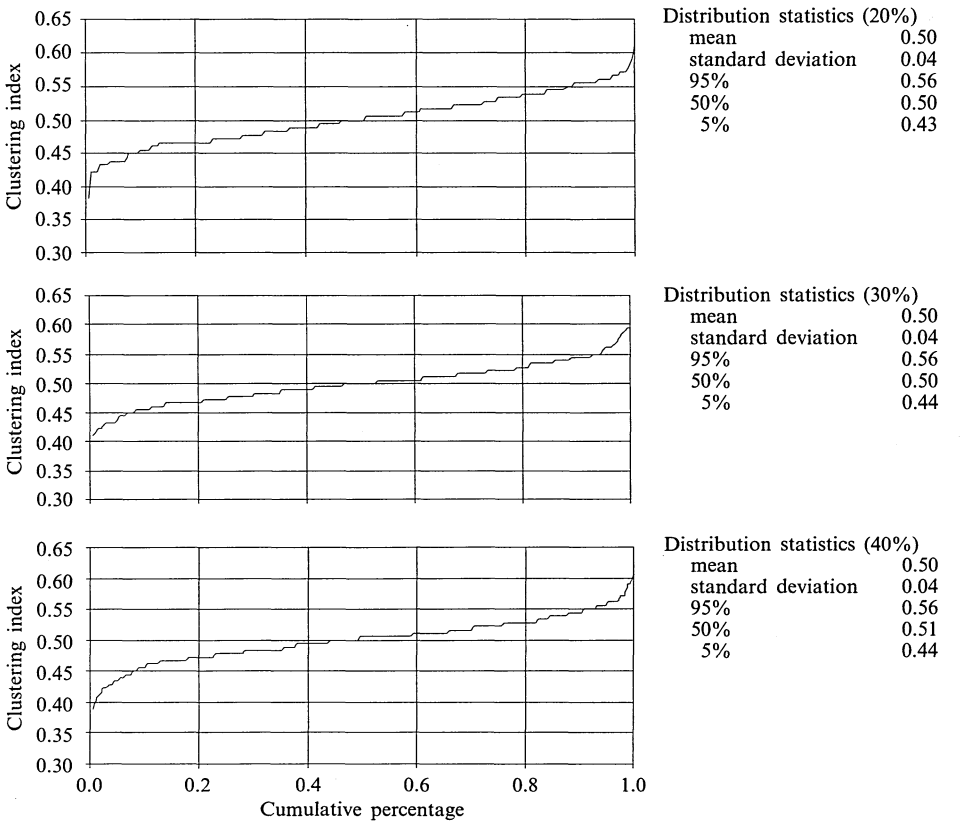


Figure 3. Distribution of the clustering index for three citywide rates of the object group: (a) 20%; (b) 30%; and (c) 40%.

⁽²⁾ SAS codes for the calculation of the clustering index can be obtained from the authors.

the index based on the assumption of a stochastic process, as the distribution is not obtainable analytically. Based on the assumption that each member of an object group can be located freely, the probability that a member can be placed in a certain areal unit is proportional to the ratio of the total number of the object group in the tract to the citywide total.

The hypothetical city forms a 10×10 regular lattice, each tract of which contains 100 people. Therefore the total population of the city is 10 000. Three different citywide rates of the object group are chosen for a sensitivity analysis: 20%, 30%, and 40%. The resulting distributions of the index based on the assumption of the stochastic process are given in figure 3. As expected, the means (standard deviations) of all three simulations are 0.50 (0.04).

In most cases, the segregation indices are used to differentiate highly segregated situations. Therefore the statistical distribution of the index obtained by the assumption of randomness may not be so informative in real-world applications of the index. In the following section, some special examples of the segregated distributions are chosen to compare the clustering index with other existing segregation indices.

5 Comparisons in hypothetical space

In this section, the clustering index (I^c) is compared with the other four indices: the dissimilarity index (I^d) among nonspatial-interaction segregation measures; White's spatial proximity index (I^{sp}) among spatial-interaction segregation measures; Moran's index (I^M) among spatial autocorrelation measures in geography; and Wong's modified dissimilarity index (I^{md}) as a recently developed perimeter-based segregation index. The mathematical expressions of the four indices are given in appendix A.

For a diagrammatic comparison of the five indices, we assume a hypothetical city space, which has 100 square-shaped tracts forming a 10×10 grid pattern. The total population of the city is 1000, distributed evenly in each tract, so that each tract contains 10 people.

In the first setting, only binary distribution is allowed on the regular lattice. Therefore all 10 people in a tract are in an object group or none are in the object group. In the second setting, contiguous distribution is allowed, while the total number in an object group remains constant. Therefore each tract can contain any number of people (not exceeding 10) in the object group. However, the total number of people in the object group in the city should remain constant. These two settings are chosen to analyze the effects of the marginal change of spatial setting.

5.1 Binary distribution in a regular lattice

In the distributional patterns given in figure 3,⁽³⁾ I^d values are set to 1, as only binary distributions are allowed. Interestingly, I^{sp} produces exactly the same scores as $I^M + 1$, and I^{md} produces the same values as I^c . One reason for these similarities is the regular lattice and binary distribution adopted in this hypothetical space. Although all four are spatial-interaction indices, the two sets of indices produce quite different values in ranking. I^{sp} and I^M yield the highest score for figure 4(a) whereas I^{md} and I^c yield the highest score for figure 4(d), which has a visually smaller cluster than figure 4(a). Also, I^{sp} and I^M produce a higher score for figure 4(b), which has four separate clusters, than for figure 4(f), which has a single linear cluster; I^{md} and I^c produce a higher value for figure 4(f) than for figure 4(b). It is interesting that I^{sp} and I^M report higher scores for the distributional patterns forming a larger cluster and separated clusters, whereas I^c and I^{md} do so for a single cluster and for smaller clusters.

⁽³⁾ This hypothetical setting is identical to the lattice used in Wong (1993) except figure 4(c).

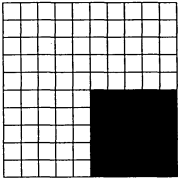
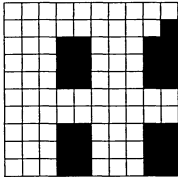
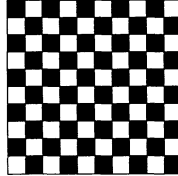
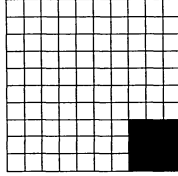
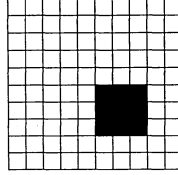
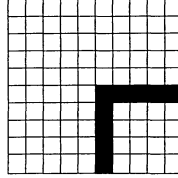
	I^d	I^{sp}	I^M	I^{md}	I^c
<div></div> <div>(a)</div>	1	1.852 (1)	0.852 (1)	0.944 (2)	0.944 (2)
<div></div> <div>(b)</div>	1	1.533 (4)	0.533 (4)	0.828 (5)	0.828 (5)
<div></div> <div>(c)</div>	1	0.000 (6)	-1.000 (6)	0.000 (6)	0.000 (6)
<div></div> <div>(d)</div>	1	1.730 (2)	0.730 (2)	0.967 (1)	0.967 (1)
<div></div> <div>(e)</div>	1	1.693 (3)	0.693 (3)	0.933 (3)	0.933 (3)
<div></div> <div>(f)</div>	1	1.434 (5)	0.434 (5)	0.900 (4)	0.900 (4)
<div>Percentage</div> <div><div></div>0</div> <div><div></div>100</div>					

Figure 4. Six configurations of binary distribution. (Note: the numbers in parentheses are the rankings.)

As discussed in the introduction, there are some differences between contiguity and clustering. Contiguity and clustering may correspond to a certain degree. However, a highly compact case of clustering may not exhibit a corresponding degree of contiguity, as is evident in these examples.

In this hypothetical setting, the total number in an object group varies in each distribution. It may be misleading, because an index obtained in a city having, for example, a 20% black population may not be quite comparable with that in another city with a 40% black population. As an example, one can assume that two cities have identical geographical settings; however, one has 9 minority people, whereas the other has 4 minority people. If they can choose their locations freely, the likelihood of all minority people choosing to live in a single tract is lower in the city populated with 9 people than in the city populated with 4 people.

5.2 Semicontinuous distribution in a regular lattice

The citywide proportion of an object group varies across cities, and segregation indices are used primarily for intercity comparisons. However, the impact of varying citywide group rates on the comparability of segregation indices has not been fully examined. Here, the relative impact of the varying group rates on various segregation measures in the setting of semicontinuous distributions is examined. We also examine the relative impact of separated clusters on various segregation measures.

In figure 5, three spatial distribution patterns are displayed. Figures 5(a), 5(c), and 5(e) contain 160 people of an object group out of a total of 1000, and figures 5(b), 5(d), and 5(f) have 320 people of the group out of a total of 1000. Hence figures 5(b), 5(d), and 5(f) have a citywide group rate of 32%, whereas figures 5(a), 5(c), and 5(e) have a value of 16%. Thus each areal unit in figures 5(b), 5(d), and 5(f) has twice the rate of group members in figures 5(a), 5(c), and 5(e), respectively. In terms of distributional pattern, figures 5(a) and 5(b) exhibit a smoothly peaked concentration at the center; figures 5(c) and 5(d) have one homogeneous cluster; and figures 5(e) and 5(f) have two clusters.

I^M yields the highest values for figures 5(a) and 5(b), whereas I^d , I^{sp} , I^{md} , and I^c produce the highest value for figure 5(d). However, I^{sp} produces a higher value in figure 5(b) than in figures 5(c) and 5(e). As discussed in the previous section, this indicates that I^M and I^{sp} are more sensitive to contiguity in object group values than enclaveness, which is the operational definition of clustering proposed here.

I^M and I^c generate the same scores regardless of the variation in overall rate, whereas I^d , I^{sp} , and I^{md} produce a lower score for the distributions with the low group rate [figures 5(a), 5(c), and 5(e)] than for the distributions with the high group rate [figures 5(b), 5(d), and 5(f)]. These indicate that I^M and I^c are robust with respect to an overall rate change, and I^d , I^{sp} , and I^{md} are sensitive to an overall rate change. The index most vulnerable to an overall rate change is I^{sp} ; it yields the 3 highest values for the distributions with the high group rate [figures 5(b), 5(d), and 5(f)] and the 3 lowest values for the distributions with the low group rate [figures 5(a), 5(c), and 5(e)].

I^{md} produces the two lowest values for figures 5(a) and 5(b), and I^c produces its two lowest values for figures 5(e) and 5(f). This indicates that I^{md} is less sensitive to a division of clusters than I^c . In its functional form, I^{md} is I^d minus its spatial interaction component. For example, I^{md} (0.699) in figure 5(a) is calculated as I^d (0.739) minus a spatial interaction component (0.04), which is relatively small. In such a case, I^{md} generates a score similar to I^d , which does not account for spatial interaction between areal units.

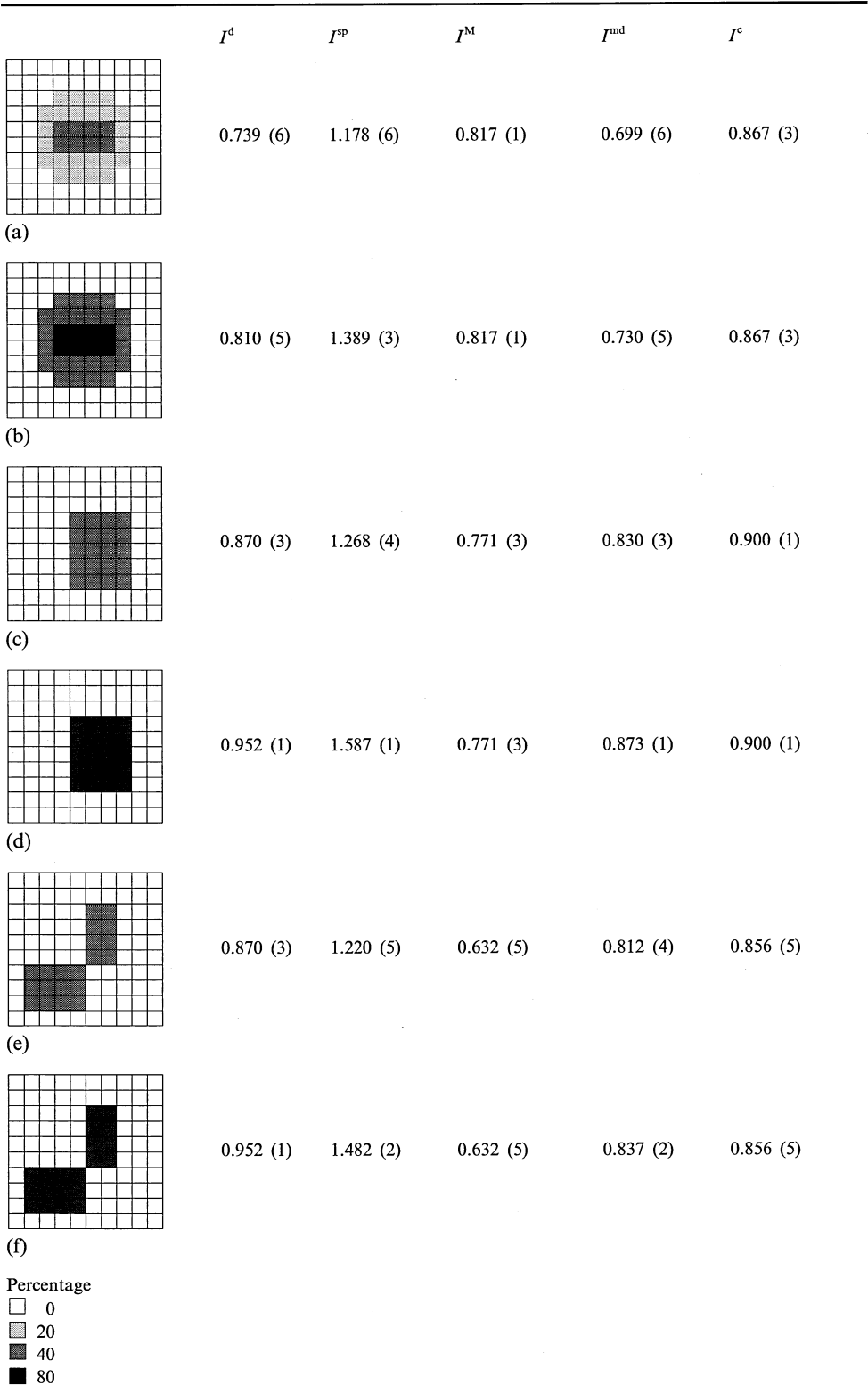


Figure 5. Six configurations of continuous distribution. (Note: the numbers in parentheses are the rankings.)

6 Comparisons in a real-world setting

In this section, the five indices are compared in an actual setting, as the hypothetical settings in the last section were chosen arbitrarily by the authors and may give rise to some selection bias. We examine whether the properties of the indices discussed in the hypothetical settings are found consistently in a real-world setting.

The five boroughs in New York City are chosen for the study area. Among the five boroughs, Manhattan and Brooklyn have relatively regular and similarly sized tract formations, whereas the Bronx, Queens, and Staten Island have comparatively irregular tract formations of varying sizes.

Segregation indices have mostly been used for intercity comparison of segregation of an object group. However, in some cases, a research question would be which object group is more segregated than the others. Thus, in addition to the interborough comparison, three different subjects of interest are examined: the proportion of black people, rates of poverty (people below the poverty level), and the origins of the homeless (prior addresses of the users of family shelters).⁽⁴⁾

Race and poverty data are obtained from the 1990 Census and the data on users of homeless shelters are obtained from data collected by the New York City Department of Homeless Services. Each index is calculated at the census tract level. Sample statistics of the subjects for each borough are given in table 1, and the calculated indices and corresponding Q^1 maps are shown in figures 6, 7, and 8. The proportion of black people varies between 8% in Staten Island and 38% in Brooklyn, and the poverty rate ranges from 8% in Staten Island to 27% in the Bronx. Alternatively, the rate of origins of the homeless is much smaller, ranging from 0.06% in Staten Island to 0.54% in the Bronx.

In sections 6.1 and 6.2, the indices are compared in two different formats: first, the indices of each subject are compared by borough; and, second, the indices of each borough are compared by subject.

Table 1. Sample statistics.

	Manhattan	Bronx	Brooklyn	Queens	Staten Island
Total population	1 496 861	1 247 530	2 300 664	1 996 710	386 855
Black population	330 278 (22.06)	462 918 (37.11)	873 620 (37.97)	424 314 (21.25)	31 949 (8.26)
The homeless	4 841 (0.32)	6 769 (0.54)	7 052 (0.31)	2 004 (0.10)	249 (0.06)
The poor	299 228 (19.99)	337 023 (27.02)	514 163 (22.35)	212 092 (10.62)	29 343 (7.59)

Note: Numbers in parentheses are percentages in the total population.

6.1 Interborough comparison

Instead of using the actual index scores, we compare the measures by means of the relative rank of each measure across boroughs, subject by subject. The ranks are given in table 2 (see page 340) (see table A1 for the actual index scores).

In terms of the black population, Brooklyn obtains the highest score for every index, and Staten Island gets the lowest score for I^{sp} and I^{M} , whereas the Bronx gets the lowest score for I^{d} , I^{md} , and I^{c} . As shown in figure 6, Brooklyn exhibits only one obvious enclave in the distribution of the black population. A visual examination reveals

⁽⁴⁾ See Culhane et al (1996) for a more detailed description about the data on users of shelters for the homeless. Note that the prior addresses of such users from 1989 to 1992 are used.

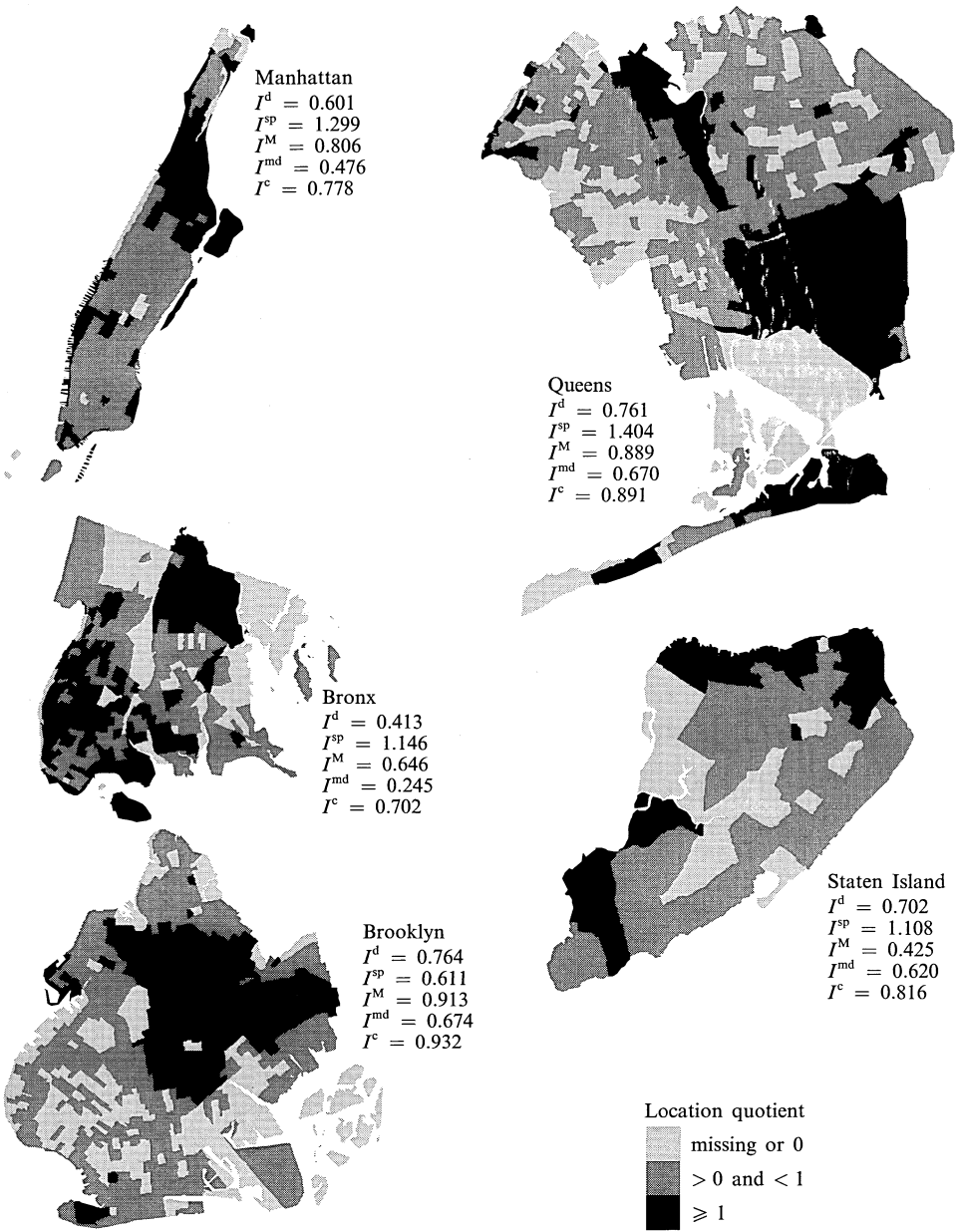


Figure 6. Census tract map of the black population distribution.

that Staten Island exhibits concentrations in a relatively smaller number of tracts than the Bronx, where the distribution of the black population is most broadly dispersed.

For the poverty distribution, the Bronx obtains the highest score for all the indices except I^{md} , which produces its highest score for Staten Island. Queens gets the lowest score among all indices except I^c , which yields the lowest score for Staten Island. As shown in figure 8, Staten Island, generating the highest score for I^{md} and the lowest score for I^c , exhibits several separated clusters in a smaller number of tracts. As discussed in the previous section, this shows that I^c is more sensitive to a subdivision of clusters than I^{md} .

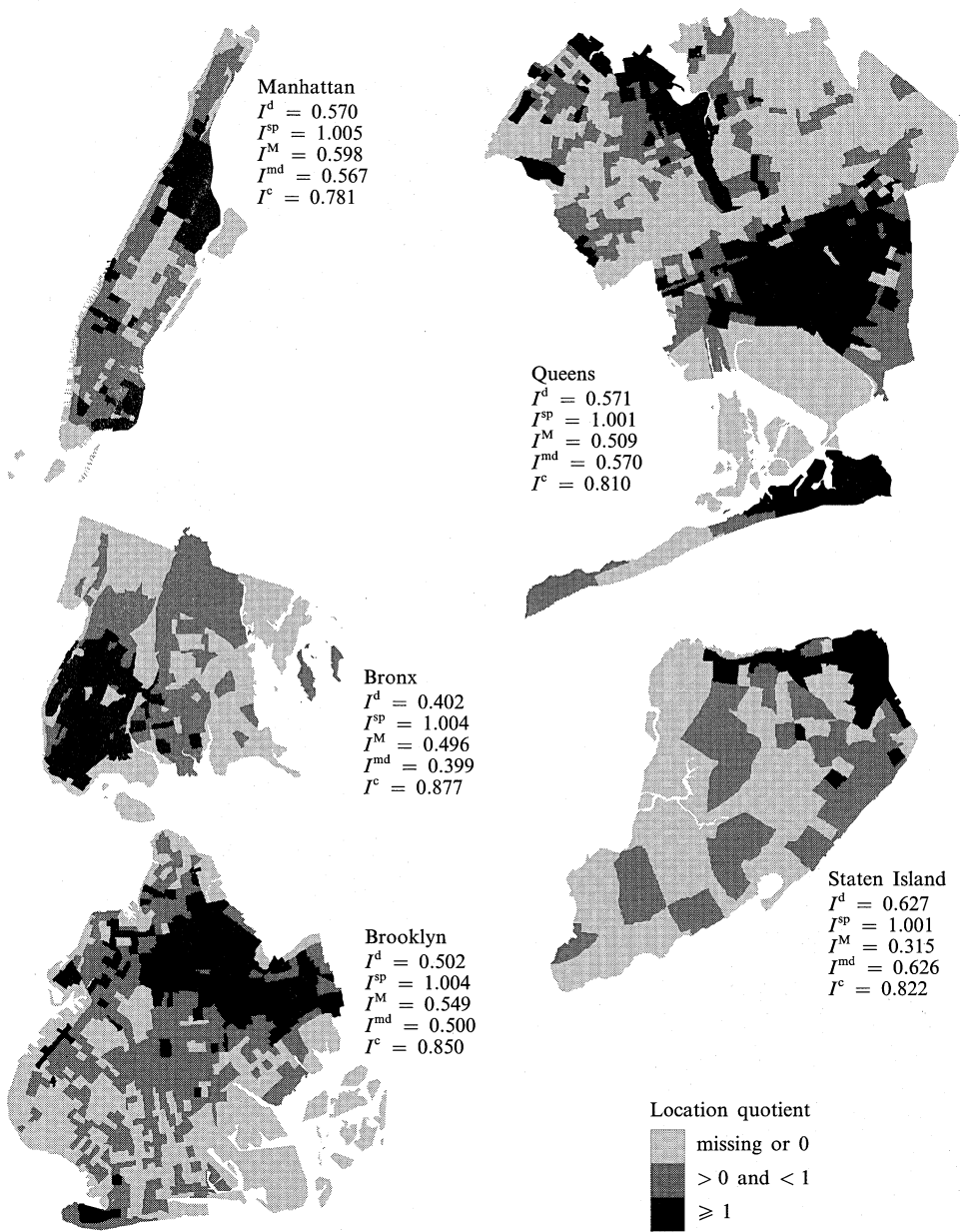


Figure 7. Census tract map of the origins of the homeless.

The distribution of the origins of the homeless generates a mixed result. I^{sp} and I^M produce the highest scores for the Bronx, whereas I^d and I^{md} generate the highest scores for Staten Island, and I^c is highest for Brooklyn. In contrast, I^d and I^{md} generate the lowest scores for the Bronx, whereas I^{sp} does the same for Queens and Staten Island, I^M for Staten Island, and I^c for Manhattan. As shown in figure 7, Bronx, Brooklyn, and Staten Island exhibit one well-formed cluster, and the three boroughs are ranked as the top three in I^c .

Overall, the distribution of the black population generates a fairly consistent ranking of each borough, and the distribution of the origins generates the most mixed rankings

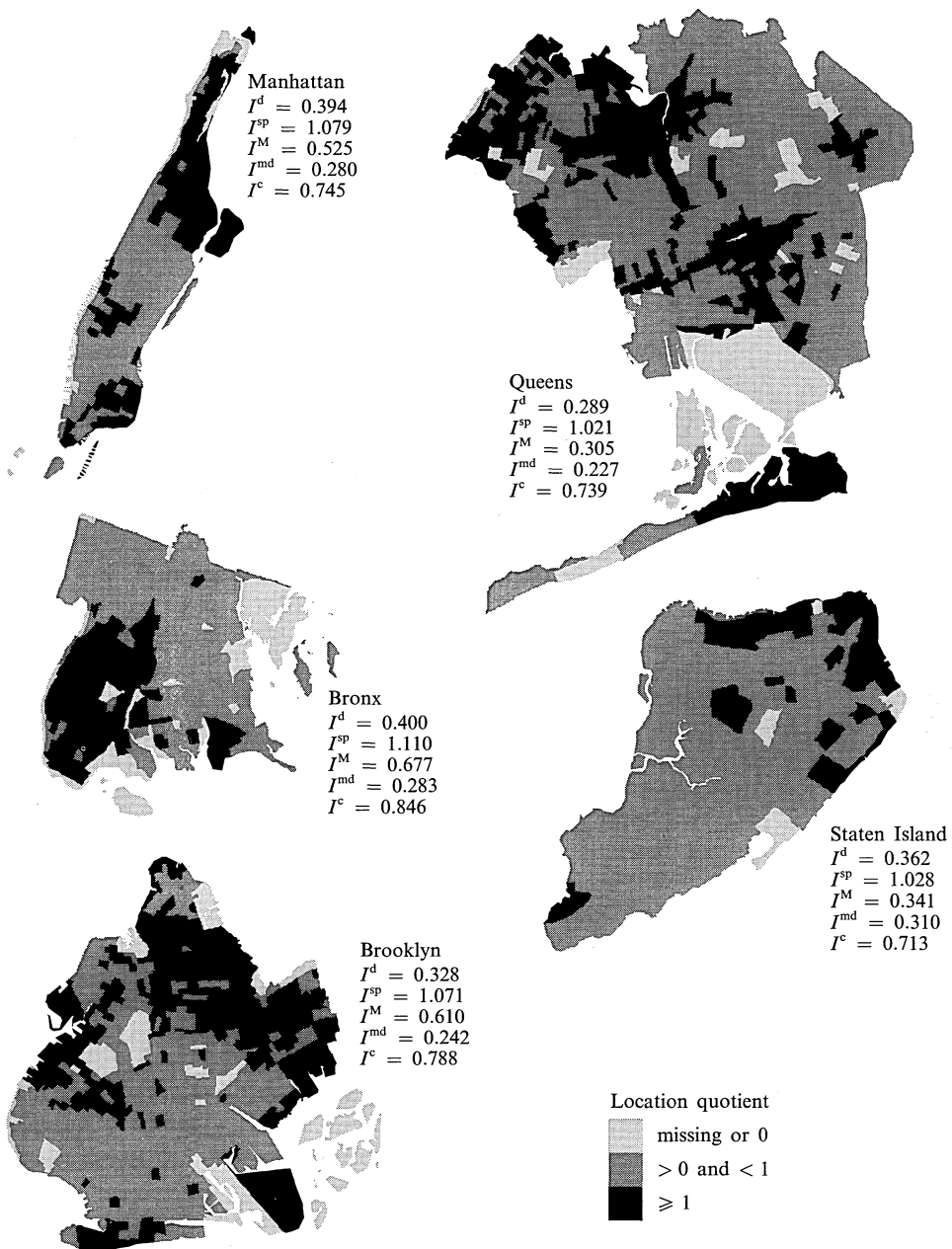


Figure 8. Census tract map of poverty distribution.

over the indices. As observed in the hypothetical settings, I^{md} produces a similar ranking to I^d and is sensitive to the relative number of highly concentrated tracts. I^{sp} and I^M produce similar rankings to each other and underestimate the highly concentrated enclaves such as the origins in Staten Island. I^c produces distinctive rankings and is very sensitive to the separation of clusters such as the three clusters of origins in Manhattan, which produce the lowest ranking for I^c among the boroughs.

Table 2. Comparison of the relative ranking of each measure across boroughs.

Index	Black population					Homelessness					Poverty				
	MN	BX	BK	QN	SI	MN	BX	BK	QN	SI	MN	BX	BK	QN	SI
I^d	4	5	1	2	3	3	5	4	2	1	2	1	4	5	3
I^{sp}	3	4	1	2	5	1	2	2	4	4	2	1	3	5	4
I^M	3	4	1	2	5	1	4	2	3	5	3	1	2	5	4
I^{md}	4	5	1	2	3	3	5	4	2	1	3	2	4	5	1
I^c	4	5	1	2	3	5	1	2	4	3	3	1	2	4	5

Note: see text for an explanation of the variables; MN, Manhattan; BX, Bronx; BK, Brooklyn; QN, Queens; SI, Staten Island.

6.2 Intersubject comparison

The three subjects, the black population, poverty, and origins of the homeless, exhibit similar areas of concentration. However, their detailed distribution patterns and degrees of clustering differ. The ranks of each subject by borough are given in table 3.

According to the old dissimilarity index (I^d), the black population is the most segregated, and poverty is the least segregated in all five boroughs. However, when the spatial interaction is accounted for, the figure is changed. Based on I^{sp} , the origins are the least segregated, whereas the black population is still the most segregated. I^M produces mixed rankings for origins and poverty, whereas the black population generates the highest scores in all boroughs except the Bronx. However, I^{md} and I^c produce the highest scores for homelessness in three boroughs—Manhattan, Bronx, and Staten Island—and black population does the same in the other two boroughs.

The Q^1 maps of the Bronx and Brooklyn show visible differences among the subjects and their indices. Here, the ranking statistics are compared along with visual analyses of the Q^1 maps for the two boroughs. In the Bronx, the black population distribution shows a more scattered distribution of high-concentration tracts; the distribution of homelessness produces a well-formed single cluster, and poverty shows a slightly larger cluster than that of homelessness. In terms of index rankings, the black population distribution gets the highest score for I^d and I^{sp} , poverty gets the highest for I^M , and homelessness obtains the highest for I^{md} and I^c . Based on our clustering index I^c , the homeless are more clustered than the black population and the poor in the Bronx.

In Brooklyn, the black population distribution exhibits a relatively large tight cluster whereas homelessness has a fuzzy cluster that also occupies a smaller number of tracts than the black population. Poverty exhibits the most scattered distribution of all three subjects. In terms of ranking, the black population obtains the highest score for every index, and poverty gets the second highest for I^{sp} and I^M , whereas homelessness gets the

Table 3. Comparison of the relative ranking of each measure by subject.

Index	Manhattan			Bronx			Brooklyn			Queens			Staten Island		
	BL	HL	PV	BL	HL	PV	BL	HL	PV	BL	HL	PV	BL	HL	PV
I^d	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
I^{sp}	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2
I^M	1	2	3	2	3	1	1	3	2	1	2	3	1	3	2
I^{md}	2	1	3	3	1	2	1	2	3	1	2	3	2	1	3
I^c	2	1	3	3	1	2	1	2	3	1	2	3	2	1	3

Note: see text for an explanation of the variables; BL, black population; HL, homelessness; PV, poverty.

second highest for I^d , I^{md} , and I^c . Again, I^{sp} and I^M fail to report a high score for a tight cluster forming a visible enclave.

As shown in this intersubject comparison, each index shows quite a different degree of segregation for each subject. This result implies that the choice of an index is a critical issue when the degrees of segregation of different subjects of interest are compared. As an example, a policymaker may need to choose between a population-based program (targeting people under the poverty level) and a neighborhood-based program (targeting the neighborhoods which generate more homeless people) for the prevention of homelessness. Based on the clustering index, origins of the homeless form tighter enclaves than poverty in New York City, in contrast to other indices. This result indicates that a neighborhood-targeted program may be an alternative or a supplement to a population-only-targeted strategy.

7 Conclusion

As observed in both the hypothetical and the real-world applications, the proposed clustering index (I^c) appears to capture a different aspect of spatial distribution, which is, we believe, clustering as enclaveness (both contiguity and separateness). As expected, White's spatial proximity index (I^{sp}) captures contiguity dimensions similar to Moran's index (I^M). These two indices tend to underestimate concentration (ignoring spatial interaction), which is well captured by the dissimilarity index (I^d), and enclaveness; the indices also tend to produce a high score for a more dispersed clustering. Wong's modified dissimilarity index (I^{md}) is a good measure to combine both concentration and the two heterogeneous aspects of clustering as enclaveness. However, the final score of this index is usually overwhelmed by its original dissimilarity index component. As a result, I^{md} tends to produce a similar score to the original dissimilarity index.

The proposed clustering index tends to give more weight to enclaveness than contiguity alone. This may be a good property for those cases in which the primary concern of an investigator is the formation of enclaves of a socioeconomic subject, including minority populations, poverty, crime, epidemics, and mortgage red-lining. Additionally, its property of robustness to the citywide rate allows us to perform properly an intercity comparison of a given subject by index score, even when the citywide rate varies significantly, unlike in the case of the other measures. However, we admit that the index may be less sensitive to different levels of concentration in the enclaves, as the attribute variable becomes a binary variable ($Q^1 > 1$ versus $Q^1 \leq 1$). Further research is therefore required for detailed refinements of the clustering index to better capture enclaveness of a subject of interest.

Acknowledgement. An earlier version of this paper was presented at the URISA '95 Conference, San Antonio, TX on 20 July 1990.

References

- Anselin L, 1988 *Spatial Econometrics: Methods and Models* (Kluwer, Dordrecht)
- Bendavid-Val A, 1983 *Regional and Local Economic Analysis for Practitioners* (Praeger, New York)
- Chen L, 1994, "Modeling housing and demographic diversity at census tract versus block group levels of aggregation" *Journal of the Urban and Regional Information Systems Association* 6 11–20
- Cliff A D, Ord L K, 1973 *Spatial Autocorrelation* (Pion, London)
- Culhane D P, Lee C-M, Wachter S M, 1996, "Where the homeless come from: a study of the prior address distribution of families admitted to public shelters in New York City and Philadelphia" *Housing Policy Debate* 7 327–365
- Dacey M F, 1968, "A review on measures of contiguity for two and K-color maps", in *Spatial Analysis: A Reader in Statistical Geography* Eds B J L Berry, D F Marble (Prentice-Hall, Englewood Cliffs, NJ) pp 479–495
- Geary R C, 1954, "The contiguity ratio and statistical mapping" *Incorporated Statistician* 5 115–141
- James D R, Taeuber K E, 1985, "Measures of segregation", in *Sociological Methodology* Ed. N Tuma (Jossey-Bass, San Francisco, CA) pp 1–32

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

Morgan B S, 1982, "The properties of a distance-based segregation index" *Journal of Socio-economic Planning Sciences* **16** 167 – 171

Odland J, 1988 *Spatial Autocorrelation* (Sage, New York)

Ripley B D, 1981 *Spatial Statistics* (John Wiley, New York)

Shen Q, 1994, "An application of GIS to the measurement of spatial autocorrelation" *Computers, Environment, and Urban Systems* **18** 167 – 191

Theil H, 1972 *Statistical Decomposition Analysis* (North Holland, Amsterdam)

White M J, 1983, "The measurement of spatial segregation" *American Journal of Sociology* **88** 1008 – 1018

White M J, 1986, "Segregation and diversity: measures in population distribution" *Population Index* **52** 198 – 221

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

Appendix A: four indices for comparison

(1) The dissimilarity index I^d ,

$$I^d = \sum_{i=1}^n \left[\frac{t_i |r_i - R|}{2PR(1 - R)} \right],$$

where t_i is the population of areal unit i , r_i is the object group proportion of areal unit i , P and R are the total population and the object group proportion in the whole region, which consists of n areal units.

(2) White's spatial proximity index I^{sp} ,

$$I^{sp} = \frac{1}{TB_{tt}} (XB_{xx} + YB_{yy})$$

and

$$B^X = \sum_{i,j=1}^n \frac{1}{X^2} x_i x_j c_{ij},$$

where B_{xx} , B_{yy} , and B_{tt} are the average proximities between X members, between Y members, and among all members (T) of the population, respectively; X , Y , and T are the total number of X , Y , and all members (T) of the population, respectively; and x_i and x_j are the numbers of X members at areal units i and j , and c_{ij} is the proximity between areal units i and j .

(3) Moran's index I^M ,

$$I^M = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (p_i - \bar{p})(p_j - \bar{p})}{\sum_i (p_i - \bar{p})^2},$$

where n is the number of census tracts and the double summation indicates summation over all pairs of tracts; p_i is the ratio of an object group of tract i to the population of tract i ; \bar{p} is the mean of p_i ; and w_{ij} is a proximity weight for the pair of tracts i and j , which is 0 when $i = j$.

(4) Wong's modified dissimilarity index I^{md} ,

$$I^{md} = I^d - \frac{1}{2} \sum_i \sum_j w_{ij} |z_i - z_j|$$

and

$$w_{ij} = \frac{d_{ij}}{\sum_i \sum_j d_{ij}},$$

where z_i is the proportion of an object group in areal unit i ; and d_{ij} is the length of the common boundary of areal units i and j .

Table A1. Calculated indices.

Index	Manhattan			Bronx		
	black population	homelessness	poverty	black population	homelessness	poverty
I^d	0.601	0.570	0.394	0.413	0.402	0.400
I^{sp}	1.299	1.005	1.079	1.146	1.004	1.110
I^M	0.806	0.598	0.525	0.649	0.496	0.677
I^{md}	0.476	0.567	0.280	0.245	0.399	0.283
I^c	0.778	0.781	0.745	0.702	0.877	0.846
	Brooklyn			Queens		
I^d	0.764	0.502	0.328	0.761	0.571	0.289
I^{sp}	1.611	1.004	1.071	1.404	1.001	1.021
I^M	0.913	0.549	0.610	0.889	0.509	0.305
I^{md}	0.674	0.500	0.242	0.670	0.570	0.227
I^c	0.932	0.850	0.788	0.891	0.810	0.739
	Staten Island			Mean		
I^d	0.702	0.627	0.362	0.512		
I^{sp}	1.108	1.001	1.028	1.126		
I^M	0.425	0.315	0.341	0.574		
I^{md}	0.620	0.626	0.310	0.446		
I^c	0.816	0.822	0.713	0.806		

Note: see text for an explanation of the variables.