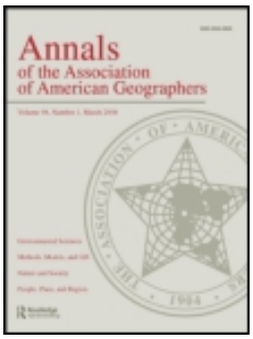


Routledge
Taylor & Francis Group

Annals of the Association of American Geographers

ISSN: 0004-5608 (Print) 1467-8306 (Online) Journal homepage: https://www.tandfonline.com/loi/raag20

Measuring Ethnic Clustering and Exposure with the *Q* Statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark

Antonio Páez , Manuel Ruiz , Fernando López & John Logan

To cite this article: Antonio Páez , Manuel Ruiz , Fernando López & John Logan (2012) Measuring Ethnic Clustering and Exposure with the *Q* Statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark, Annals of the Association of American Geographers, 102:1, 84-102, DOI: 10.1080/00045608.2011.620502

To link to this article: https://doi.org/10.1080/00045608.2011.620502

Measuring Ethnic Clustering and Exposure with the *Q* Statistic: An Exploratory Analysis of Irish, Germans, and Yankees in 1880 Newark

Antonio Páez,* Manuel Ruiz,† Fernando López,† and John Logan‡

*Centre for Spatial Analysis, School of Geography and Earth Sciences, McMaster University
†Department of Quantitative Methods and Information, Universidad Politécnica de Cartagena
‡Department of Sociology, Brown University

The study of population patterns has animated a large body of urban social research over the years. An important part of this literature is concerned with the identification and measurement of segregation patterns. Recently, emphatic calls have been made to develop measures that are better able to capture the geography of population patterns. The objective of this article is to demonstrate the application of the *Q* statistic, developed for the analysis of spatial association of qualitative variables, to the detection of ethnic clustering and exposure patterns. The application is to historical data from 1880 Newark in the United States, with individuals classified by ethnicity and geocoded by place of residence. Three ethnic groups, termed Irish, Germans, and Yankees, are considered. Exploratory analysis with the *Q* statistic identifies significant differences in the tendency of individuals and building occupancy to cluster by ethnicity. In particular, there is evidence of a strong affinity within ethnic clusters and some intermingling between Yankee and Irish residents. In contrast, the exposure of Germans to individuals of other groups is found to be more limited. *Key Words: clustering, exposure, Q statistic, segregation, spatial association.*

对人口模式的研究多年来已经掀起了城市社会研究的热潮。这些研究文献的一个重要组成部分是有关隔离模式的识别和测量。近期已产生有力的呼吁，以制定能够更好地捕捉人口模式地理的措施。这篇文章的目的是展示为分析定性变量的空间关联而开发的 *Q* 统计对于监测民族聚类和曝光模式的应用。该应用分析美国纽瓦克从 1880 年以来的历史数据，个人按种族分类和由居住地编码。考虑到三个民族，即爱尔兰，德国，和洋基人。用 *Q* 统计的探索分析，确定了在个人的倾向和建设居住种族集群上的显著性差异。特别是有证据显示，在民族集群内和在一些洋基人和爱尔兰的居民之间，有一个很强的亲和力。相比之下，德国人与其他群体的个人接触较为有限。*关鍵词：集群，曝光，Q 统计，隔离，空间关联。*

A través del tiempo, el estudio de los patrones de población ha construido un abultado cuerpo de investigación social urbana. Buena parte de esta literatura se ocupa de la identificación y medida de los patrones de segregación. Recientemente, son notables los llamados enfáticos que propenden por medidas más efectivas que capten la geografía de los patrones de población. El propósito de este artículo es demostrar la aplicación de la estadística *Q*, desarrollada para analizar la asociación espacial de variables cualitativas, con la cual detectar la agrupación étnica y patrones de exposición. La aplicación se hace a datos históricos de la Newark de 1880 en los Estados Unidos, clasificando los individuos por etnicidad y geocodificados por lugar de residencia. Se tomaron en cuenta tres grupos étnicos, denominados irlandeses, alemanes yanquis. El análisis exploratorio con la estadística *Q* identifica diferencias significativas en la tendencia de los individuos y ocupación de edificaciones a agruparse por etnicidad. En particular, existe evidencia de una fuerte afinidad al interior de los agrupamientos étnicos, lo mismo que de cierta mezcla de residentes yanquis e irlandeses. Por contraste, se ha encontrado que la exposición de los alemanes ante individuos de otros grupos es más limitada. *Palabras clave: agrupamiento, exposición, estadística Q, segregación, asociación espacial.*

Population segregation is not a new phenomenon. More than a century ago, in 1903, DuBois saw it as a barrier to comity between ethnic groups and lamented that it "caused each to see the worst in the other" (DuBois 1903, cited in Charles 2003, 167). In all probability, segregation was old even then.

There are a few modern accounts of historical segregation patterns and their effects. Kantrowitz (1979), for example, studied the segregation of minority populations in Boston from 1830 to 1970 as a way to inform current (at the time) debates on public school desegregation programs. Boyd (1998) investigated the

Annals of the Association of American Geographers, 102(1) 2012, pp. 84–102
Initial submission, September 2010; revised submission, January 2011; final acceptance, March 2011
Published by Taylor & Francis, LLC.

situation of black merchants in the early 1900s in the United States and suggested that, whatever other social effects it might have had, segregation seemed to have encouraged the emergence of a new class of black entrepreneurs. In an example of how segregation can emerge and be perpetuated, Gotham (2000) traced the origins of residential segregation in Kansas City in the first half of the twentieth century back to the racial attitudes of key players in the budding real estate market in that city. The use of covenants to codify segregation was not uncommon at various points in history (Weaver 1978). These modern studies and others (e.g., Hershberg et al. 1979; Spain 1979) provide valuable historical perspectives on the phenomenon. Besides some epochal accounts of segregation (e.g., by DuBois), it appears that the formal study of this social phenomenon only started with the studies of the Chicago School of Sociology that empirically described the social ecology of Chicago neighborhoods (Dawkins, Reibel, and Wong 2007). Population segregation research has since gained in scope and depth, and today it is a topic that animates a large body of urban social research from a number of different perspectives, including sociology (e.g., Logan and Zhang 2010), geography (e.g., Deurloo and de Vos 2008), urban studies (e.g., Harsman 2006), and economics (e.g., Cutler, Glaeser, and Vigdor 2008), to mention just a few.

One of the reasons segregation is of interest is that it remains a key to understanding inequality and social mobility issues and therefore is still of significant social science and policy interest (Pettigrew 1979; Charles 2003; Simpson 2004). The specific focus of segregation research varies by context—for instance, from ethnicity and race in the United States (Massey and Denton 1993), the United Kingdom (Peach 1996; Johnston, Forrest, and Poulsen 2002), and Australia (Poulsen and Johnston 2000) to religion in Northern Ireland (Lloyd 2010), income and race in Brazil (Feitosa et al. 2007), and age and income Canada (Smith 1998; Fong and Shibuya 2000). The general motivation, however, remains the same: trying to understand the processes and patterns of separation, whether willing or imposed, of members of a social group from others. Although interest in segregation seems to have ebbed and flowed in the past few decades (Charles 2003), judging from the number of articles, specially collected issues (e.g., Kaplan and Woodhouse 2004, 2005; Dawkins, Reibel, and Wong 2007; Wong, Reibel, and Dawkins 2007; Simpson and Peach 2009; Bolt, Ozuekren, and Phillips 2010), and the passion that animates some of the debates (e.g., Peach 2009), segregation research is currently at a high point, and work continues along numerous fronts.

One area of ongoing interest in the segregation literature is motivated by the need to produce reliable statistics to inform academic and policy discussions. Use of the Dissimilarity Index was long the standard approach used in segregation studies (Massey and Denton 1988). Especially after the systematic review of concepts and measures of Massey and Denton (1988), it became generally recognized that segregation is a concept that spans multiple dimensions and is not easily captured by any one single index. This prompted research that adopted or developed a number of indicators useful to capture the various dimensions of segregation, including Theil's Index, the delta index, the Gini Index, and so on.

One limitation of many early indicators used to measure the different dimensions of segregation is that they consider people in space but rarely their spatial relationships beyond propinquity in the same administrative division (e.g., the census tract). In other words, many of these indexes operate by aggregating areal population values and disregarding all other spatial structures (White 1983; Reardon and O'Sullivan 2004). In recent years, increasingly emphatic calls have been made to develop measures that are better able to capture geographic patterns of segregation (Wong 1997, 1999; Brown and Chung 2006; Johnston, Poulsen, and Forrest 2009). A geographical perspective adds depth to segregation analysis by allowing researchers to consider the spatial pattern of aspatial segregation measures (Wong 1997; Dawkins 2004), the local patterns of segregation (Wong 2002; Feitosa et al. 2007; Lloyd 2010), and, important from our perspective, by conceptualizing spatial association itself as a measure of segregation (Wong 1999; Brown and Chung 2006; Johnston, Poulsen, and Forrest 2009).

The objective of this article is to demonstrate the use for the analysis of population segregation of the newly developed *Q* statistic for spatial association of qualitative variables (Ruiz, López, and Páez 2010). As will be shown, the *Q* statistic can be used to assess patterns of clustering and exposure. It constitutes a valuable tool not only to explore these types of patterns but also to statistically test them against the hypothesis of spatial randomness, an old debate in the literature (see Reiner 1972; Zelder 1972). Unlike other approaches that are exclusively area-based and designed for continuous variables, *Q* is designed for qualitative variables and its support of them can also be the point.

This means that it can be applied to analysis at the personal level, with, say, ethnicity defining a qualitative attribute of the individual. Furthermore, it can also be scaled up to other levels of geography by categorizing higher level outcomes.

We demonstrate the proposed approach by means of historical data from 1880 Newark in the United States, with individuals classified by ethnicity and marital status, and geocoded by place of residence. Three ethnic groups, termed Yankees, Irish, and Germans, are considered. Application of the Q statistic identifies significant differences in the tendency of individuals to cluster by ethnicity and of buildings by dominant occupancy. In particular, there is evidence of a strong affinity for clustering within ethnic groups and some intermingling between Yankee and Irish residents. In contrast, exposure of German individuals to members of other groups is significantly more limited. The same is observed for predominantly German and other buildings.

The structure of the article is as follows: A review of previous work that has adopted a spatially explicit perspective to the measurement of population patterns precedes a technical section that describes the Q statistic. Next, we introduce the data set used in the application, followed by the results of the analysis. Finally, in the concluding section, we summarize our main points and sketch directions for future research.

Literature Review: Measuring Segregation Spatially

A number of recent articles discuss the state of the practice, the art, and challenges in segregation research (e.g., Kaplan and Woodhouse 2004, 2005; Dawkins, Reibel, and Wong 2007; Wong, Reibel, and Dawkins 2007; Simpson and Peach 2009; Bolt, Ozuekren, and Phillips 2010). Readers interested in a more extensive panoramic of the field are redirected to these sources. In our review we concentrate only on recent contributions that discuss the spatial aspects of measuring segregation. These studies tend to emphasize one or a combination of three major issues. First, traditional measures are notorious for their inability to appropriately incorporate the spatial relationships between units of analysis (i.e., the checkerboard problem of White 1983). The second issue is the presence of spatial pattern when mapping segregation measures. Finally, there are the questions of aggregation and scale, which could have an important impact on findings and recommendations. These issues are not necessarily independent but, as a number of articles reviewed here show, might in fact interact in various ways.

An early criticism of the standard tool of segregation research, the Dissimilarity Index, came from White (1983), who noted that paring geography out renders the measure insensitive to spatial pattern and incapable of distinguishing between residential clusters and ghettos. This issue, shared by most other segregation measures, is the so-called checkerboard problem. In an attempt to improve on this state of affairs, White proposed a proximity index that was directly based on distance between members of the population. White's index, unfortunately, is difficult to interpret, although several of its base components (i.e., the average distance between members of the same or different population groups) do indeed provide valuable spatial information. A notable aspect of White's proximity index is that it incorporated, perhaps for the first time in segregation research, a distance-decay function. This function is in essence a spatial kernel used to decrease the contribution to the proximity index of a given pair of individuals as the distance between them increases (Jakubs [1981] used the distances between all pairs of areal units for optimization). Kernel functions have since taken on a more prominent role in spatial segregation measures (see, for instance, Wong [1998], who used a rectangular kernel and alluded to distance-decay functions).

Kernel functions directly speak to the issues of spatial relationships and scale. Reardon and O'Sullivan (2004) proposed a formal framework to generate spatial segregation measures. At the core of this framework is the use of kernel functions to establish relations of proximity. Depending on data availability, the measures proposed are applicable at very high levels of granularity, potentially even the individual person. Despite this, the measures are based on population proportions and densities and are therefore inherently areal; however, the areas can be established by the analyst as part of defining the functional form and parameters of the kernel function. Reardon and O'Sullivan suggested that analysts can specify these elements of the framework based on theoretical notions of how space influences social interaction. In practice, it has been more common to use kernel-based approaches in an exploratory fashion, to investigate separate but related issues of scale and aggregation that form part of the modifiable areal unit problem (Wong, Lasus, and Falk 1999). Feitosa et al. (2007), for instance, demonstrated their global and local indicators of segregation

by exploring bandwidths ranging between 400 and 4,400 m. O'Sullivan and Wong (2007) proposed to use the union and intersection of two kernel functions for different population groups to measure segregation. The approach was applied in their article to the cities of Philadelphia and Washington, DC, using kernel bandwidths of 2.5 to 10 km, in 2.5-km increments. Reardon et al. (2008) implemented the principles outlined in Reardon and O'Sullivan (2004) to investigate the scale of segregation using bandwidths from 100 to 4,000 m. This produces so-called segregation profiles that track the degree of segregation at different scales. A similar idea was put to work in an article by Deurloo and de Vos (2008), who applied the *k*-function-inspired multiscale measure of Marcon and Puech (2003) to assess the concentration of various ethnic groups with respect to each other, at relatively small scales up to 560 m. More recently, Lloyd (2010) used geographically weighted descriptive statistics to investigate population concentration patterns by community background in Northern Ireland.

The ability to investigate segregation patterns at various scales begs the question of whether it makes a difference, and research has been conducted to clarify this, by comparing the results of conventional (census geography-based) to spatial measures of segregation. Kramer et al. (2010) investigated black and white segregation in 231 of the largest Metropolitan Statistical Areas in the United States using the diversity and exposure indexes and, after Reardon and O'Sullivan (2004), surface-based versions of the same. The results indicate that both types of measures, spatial and aspatial, are highly correlated, but the differences are not uniform, a fact that might mask potentially valuable information. In particular, these researchers report that the differences between census-tract-based and surface-based measurements are greater for smaller cities and at higher levels of resolution (i.e., when calculations are made based on smaller kernel bandwidths). The latter results stands in contrast to an earlier report by Wong (1997) indicating that the dissimilarity index tends to be deflated at lower levels of resolution (i.e., when calculations are based on geographically larger units of analysis). Although Wong (1997) did not report results by population size, he argued that the dissimilarity index is sensitive to the level of autocorrelation of the population values. These values, unfortunately, were not reported by Kramer et al. (2010). Other research by Dawkins (2004) did in fact confirm that spatial autocorrelation can in some cases account for a large part of the measured segregation.

That spatial autocorrelation provides a naturally spatial measure of segregation might seem evident. Already, Massey and Denton (1988) mentioned the use of measures of spatial autocorrelation to assess clustering patterns. Besides work by Wong (1999) that propounded the use of centrographic analysis as a way to assess segregation levels, however, until recently there were only a few examples of research where autocorrelation measures were used as measures of segregation. In part, this might have been due to the fact, noted by Dawkins (2004, 835), that autocorrelation measures, such as Moran's coefficient, are limited to the analysis of clustering and fail to capture unevenness. With the advent of local spatial analysis techniques, in particular the G_i and LISA statistics (Getis and Ord 1992; Anselin 1995), there are increased opportunities to investigate clustering and unevenness. A few recent articles adopt this approach. These include Logan, Alba, and Zhang (2002) and the use of LISA statistics to identify immigrant enclaves and ethnic communities in New York and Los Angeles. Brown and Chung (2006) advocated a geographical perspective in the analysis of segregation and supported their case with a study of Franklin County, Ohio. There, it was shown that blacks tend to be more clustered spatially than whites and that Asians and Hispanics, although displaying significant levels of clustering, do not reach the same levels as blacks and whites. In addition to the analysis of clustering using Moran's coefficient, the local version of the statistic detected patterns of unevenness, with black and white over- and underrepresentation in the central city, respectively, and the opposite pattern for the suburbs. The typical central city–suburban dichotomy, in contrast, did not hold for Asians, a group found to be overrepresented in the northwest and underrepresented in the southern part of the city. Three distinctive clusters of Hispanic residents were also found. Even more recently, Johnston, Poulsen, and Forrest (2009) also made a plea to "put more geography in" segregation analysis and showed the way by means of global and local autocorrelation analysis of population patterns in Auckland, New Zealand. The results of their analysis indicated not only high levels of population concentration for Europeans, Maori, Pacific Islanders, and Asians but also the regions of the city where these concentrations are particularly marked.

Together, the works reviewed here persuasively show the richness of detail that can be achieved by adopting a geographical perspective. In what follows, we describe an alternative analytical framework based on the use of the *Q* statistic.

Methods: Spatial Association of Qualitative Variables

Autocorrelation analysis of continuous variables is a time-honored practice in analytical geography (Getis 2008). More recently, interest in the analysis of variables of a qualitative and nominal nature has spurred renewed attention to techniques useful to explore and model spatial qualitative processes. One recent development is Q, a statistic designed to test the spatial association of qualitative variables (Ruiz, López, and Páez 2010). As we show here, Q provides an intuitive way to measure ethnic clustering and exposure patterns. In this section we briefly discuss the conceptual basis for the use of Q and describe the statistic.

Conceptual Basis

In their original review, Massey and Denton (1988) proposed five dimensions of segregation, namely, *evenness* (the over- or underrepresentation of a social group in specific areas), *exposure* (of members of one group to members of other groups), *concentration* (proximity of members of the same group in an area), *centralization* (concentration in the central city), and *clustering* (the extent to which members of a group adjoin one another). This typology of segregation has been revisited by later authors. It has been argued, for instance, that the relevance of centralization is much diminished in contemporary polynucleated cities. This has led to a reduction in the number of dimensions of segregation. Further, the remaining dimensions are seen to lie in a two-dimensional continuum: evenness-clustering and isolation-exposure in the case of Reardon and O'Sullivan (2004), and evenness-concentration and clustering-exposure in the case of Brown and Chung (2006). The original work of Massey and Denton (1988; see Tables 4 and 5) already indicated the interrelationships between some of these dimensions, in particular clustering and exposure, and also showed that evenness and clustering measures tend to retain a fairly distinctive character.

As should become clear in what follows, the Q statistic more naturally fits the clustering-exposure dimension of Brown and Chung (2006). According to these authors, the dimension of clustering refers to units close to others units of the same type, thus forming a contiguous grouping of likes. Further, the dimension of exposure is the degree to which units share a neighborhood with other units of different types. It follows then that high clustering is in fact a manifestation of low exposure and vice versa (Brown and Chung 2006, 126). The Q statistic is built around proximity relationships of spatial units that are classified according to their type. By designing neighborhoods of a specified size, it becomes possible to summarize the class membership of all units in a given neighborhood and therefore to investigate to what extent the members are of the same type (clustering) or of different types (exposure). This basic notion is formalized next.

Q Statistic

Q is developed for the analysis of a spatial variable, say $\mathbf{Y}$, that is the outcome of a discrete process. In other words, each realization y of the process can take one and only one of k different values, say $a_1, a_2, \ldots, a_k$, that are recorded at sites $i = 1, 2, \ldots, N$ with coordinates $\boldsymbol{s}_i$. In the simplest case, when $k = 2$, the process can be represented by a black-and-white map (i.e., $a_1 = w = 0$ and $a_2 = b = 1$). Borrowing a set of typical diagrams from the segregation literature, maps with different spatial configurations of the qualitative variable could be as shown in Figure 1. Note that for simplicity the diagrams use a regular distribution of cases; in actual practice, the statistic can be applied to an irregular distribution of cases as well. The diagrams represent two extreme cases

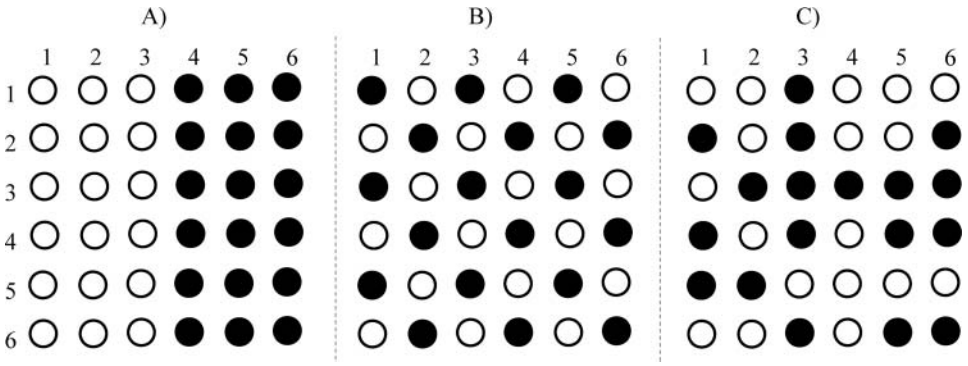
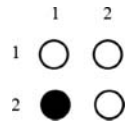


Figure 1. Examples of black-and-white spatial patterns: (A) nonrandom, (B) nonrandom (checkerboard), and (C) random.

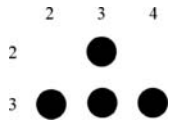
of nonrandom patterns (Figure 1A and 1B), as well as one random pattern (Figure 1C).

To capture relations of proximity between realizations of the spatial variable, we define for a location $\mathbf{s}_0$ a local neighborhood of size m, called an m-surrounding. The size of the m-surrounding is determined by the analyst (the analog of defining the pattern of contiguities in matrix $\mathbf{W}$ in autocorrelation analysis), but for the sake of the example consider m-surroundings of size 4 ($m = 4$), to give subsets of 4 cells. A rule must be defined to identify the $m - 1$ nearest neighbors that, together with site $\mathbf{s}_0$, form a neighborhood of size 4. Ruiz, López, and Páez (2010) take the $m - 1$ nearest neighbors based on distance and, in the case of ties, based on the smallest angle (counterclockwise) from the x axis to ensure the uniqueness of each member in the m-surrounding.

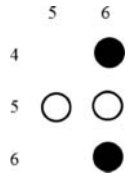
According to these rules, we can find the values of the spatial variable in the 4-surrounding for specific locations. For instance, for Figure 1C, the 4-surrounding of the cell in the first row and first column, or site (1,1), is



The neighbors have been selected based on distance and direction as follows: site (2,1) is the first neighbor, (2,1) is the second neighbor (same distance as the first neighbor, but greater angle from the x axis), and (2,2) is the third neighbor (more distant than the first and second neighbors). Clearly, the white unit at (1,1) displays high clustering, and low exposure to black. The 4-surrounding for site (3,3) is



The distance to all three neighbors is the same, so in order of direction they are (3,4) (2,3), and (3,2). The black unit in (3,3) is part of a cluster of black, and has zero exposure to white. As a final example, consider the 4-surrounding for site (5,6):



As it can be seen, in this case the neighborhood has equal clustering and exposure properties.

The shapes of the m-surroundings in these examples are different due to the rules used to select the $m - 1$ nearest neighbors. Because the arrangement of cases is regular, there are distance ties that are broken by making reference to the angle with respect to site $\mathbf{s}_0$. These rules are used for convenience but can be modified to incorporate anisotropy or other considerations. Distance ties are extremely rare when the distribution of cases is not regular, and the m-surroundings will in general be irregularly shaped.

The local configuration of values of y can be represented in a compact form by means of *symbols*. A symbol, denoted by σ, is a string that collects in a prespecified order the values of the variable in an m-surrounding. According to our rule, the symbol for site (1,1) in Figure 1C is {w,w,b,w} = {0,0,1,0}. The first element of the string is the value of y at $\mathbf{s}_0$, that is, at site (1,1). Cells (1,2) and (2,1) are equidistant from (1,1), and therefore the tie is broken by making reference to the angle from the x axis, so that (1,2) is picked first. Cell (2,2) is picked last because it is the farthest of $m - 1$ nearest neighbors. In similar fashion, the symbol for site (3,3) becomes {b,b,b,b} = {1,1,1,1}, and the symbol for site (5,6) becomes {w,b,w,b} = {0,1,0,1}. Every other cell in the diagram can be symbolized in the same way.

Now, because there are $k = 2$ classes and the m-surrounding is of size 4, it is straightforward to see that there are in fact $k^m = 16$ unique symbols (the number of permutations with repetition), as shown in Table 1. As per Table 1, we say that location (1,1) in the third diagram is of type σ_3, location (3,3) is of symbol σ_{16}, location (5,6) of symbol σ_9, and so on. After symbolizing the locations, it is possible to calculate the frequency of each symbol as the number of locations that are of type σ_j:

$$n_{\sigma_j} = \#\left(\mathbf{s} | \; s \text{ is of type } \sigma_j\right) \tag{1}$$

The relative frequency is simply the number of times that symbol σ_j ($j = 1, 2, \ldots, k^m$) is observed divided by the number of symbolized locations S. It should be clear that there will be some overlap between m-surroundings at different locations. This overlap could compromise some approximations required for developing a test of

Table 1. List of symbols for $k = 2$ and $m = 4$

$\sigma_1 = \{0,0,0,0\}$	$\sigma_5 = \{1,0,0,0\}$	$\sigma_9 = \{0,1,0,1\}$	$\sigma_{13} = \{0,1,1,1\}$
$\sigma_2 = \{0,0,0,1\}$	$\sigma_6 = \{0,0,1,1\}$	$\sigma_{10} = \{1,0,1,0\}$	$\sigma_{14} = \{1,1,1,0\}$
$\sigma_3 = \{0,0,1,0\}$	$\sigma_7 = \{0,1,1,0\}$	$\sigma_{11} = \{1,0,0,1\}$	$\sigma_{15} = \{1,1,0,1\}$
$\sigma_4 = \{0,1,0,0\}$	$\sigma_8 = \{1,1,0,0\}$	$\sigma_{12} = \{1,0,1,1\}$	$\sigma_{16} = \{1,1,1,1\}$

Table 2. Frequency and relative frequency of symbols in example

	Diagram 1		Diagram 2		Diagram 3	
	n_{σ_j}	p_{σ_j}	n_{σ_j}	p_{σ_j}	n_{σ_j}	p_{σ_j}
$\sigma_1 = \{0,0,0,0\}$	IIIII IIIII II	0.333		0.000	II	0.056
$\sigma_2 = \{0,0,0,1\}$		0.000		0.000	I	0.028
$\sigma_3 = \{0,0,1,0\}$		0.000		0.000	IIII	0.111
$\sigma_4 = \{0,1,0,0\}$	IIIII I	0.167		0.000	II	0.056
$\sigma_5 = \{1,0,0,0\}$		0.000	IIIII IIIII IIIII I	0.444	II	0.056
$\sigma_6 = \{0,0,1,1\}$		0.000		0.000	II	0.056
$\sigma_7 = \{0,1,1,0\}$		0.000	II	0.056	I	0.028
$\sigma_8 = \{1,1,0,0\}$		0.000		0.000	II	0.056
$\sigma_9 = \{0,1,0,1\}$		0.000		0.000	III	0.083
$\sigma_{10} = \{1,0,1,0\}$		0.000		0.000	III	0.083
$\sigma_{11} = \{1,0,0,1\}$		0.000	II	0.056	IIII	0.111
$\sigma_{12} = \{1,0,1,1\}$		0.000		0.000		0.000
$\sigma_{13} = \{0,1,1,1\}$		0.000	IIIII IIIII IIIII I	0.444	III	0.083
$\sigma_{14} = \{1,1,1,0\}$	IIIII	0.139		0.000	III	0.083
$\sigma_{15} = \{1,1,0,1\}$	I	0.028		0.000	II	0.056
$\sigma_{16} = \{1,1,1,1\}$	IIIII IIIII II	0.333		0.000	II	0.056

Note: In the frequency, each I indicates one occurrence of the symbol in the diagram.

hypothesis, so to reduce the overlap, the number of symbolized locations can be less than the number of observations N (more on this later). The relative frequency of each symbol is then:

$$p_{\sigma_j} = \frac{n_{\sigma_j}}{S} \tag{2}$$

The frequencies and relative frequencies of the symbols (ignoring the overlap condition) in each of the three examples in Figure 1 are shown in Table 2. It can be seen there that in general when the map is patterned, few symbols tend to dominate, whereas when the map is random no single symbol dominates. For a fixed $m \geq 2$, the relative frequency of symbols can be used to define the *symbolic entropy* of the spatial process as the Shanon's entropy of the distinct symbols:

$$h(m) = -\sum_j p_{\sigma_j} \ln(p_{\sigma_j}) \tag{3}$$

The entropy function is bounded between $0 < h(m) \leq \eta$. The lower bound is straightforward to derive. When a sequence of values is repeated in space, the information content of the map will in general be low because symbols become to some extent predictable (as is the case of maps with strong patterns of spatial association). The symbolic entropy tends to 0 because $p_{\sigma_i} \to 1$ and $p_{\sigma_j} \to 0$ for all $j \neq i$, which implies that $p_{\sigma_i} ln(p_{\sigma_i}) \to 0$ and $p_{\sigma_j} ln(p_{\sigma_j}) \to 0$. The upper bound depends on the frequency of outcomes of the categorical variable. In the particular case when the values of the variable appear with identical frequency (in the examples in Figure 1, black and white appear eighteen times each) the expected relative frequency for a random spatial process is $p_{\sigma_j} = 1/k^m$ for all j, and the upper bound is

$$\eta = \ln(k^m) \tag{4}$$

In general, the frequencies of the outcomes a_j will not be identical (e.g., some ethnic groups are minorities). In such cases, the upper bound of the entropy function (for a spatial random sequence) depends on the frequency of the various outcomes a_j and is given by (see Ruiz, López, and Páez 2010):

$$-\sum_{i=1}^{k^m} \frac{n_{\sigma_i}}{S} \sum_{j=1}^{k} \alpha_{ij} \ln(q_j) \tag{5}$$

where α_{ij} is the number of times that class a_j appears in symbol σ_i and $q_j = P(y = a_j)$.

The Q statistic is essentially a likelihood ratio test between the symbolic entropy of the observed pattern and the entropy of the system under the null hypothesis of a random spatial sequence:

$$Q(m) = 2S(\eta - h(m)) \tag{6}$$

The statistic is asymptotically χ^2 distributed with degrees of freedom equal to the number of symbols minus one. Let $0 \leq \alpha \leq 1$. A decision rule to reject the null hypothesis of spatial randomness at a confidence level of $100(1-\alpha)$ percent can be established as follows:

$$\begin{cases} \text{If } Q(m) > \chi^2_{k^m-1\alpha} \text{then reject } H_0 \\ \text{Otherwise do not reject } H_0 \end{cases} \quad (7)$$

where $\alpha = P(\chi^2_{k^m-1} > \chi^2_{k^m-1\alpha})$.

An in-depth exploration of the finite sample properties of the statistic is found in Ruiz, López, and Páez (2010). As noted earlier and discussed in detail in said reference, performance of the statistic can become compromised due to the overlap of m-surroundings. To meet all key approximations for testing, the overlap is controlled by letting the maximum number of symbolized locations S be less than the actual number of observations N, as follows:

$$S = \left[\frac{N-m}{m-r}\right] + 1 \quad (8)$$

where $[x]$ is the integer part of a real number x, and r is the overlap degree allowed between the m-surroundings of proximate locations. A procedure to select S locations that satisfy the designated overlap degree was introduced in Ruiz, López, and Páez (2010, 289) and briefly described in Appendix A. Simulation experiments reported in Ruiz, López, and Páez (2010) indicate that increasing the degree of overlap leads to a smaller size of the statistic (thereby reducing the risk of false positives) but at the cost of reduced power. Increasing the degree of overlap allows the analyst to retain more observations, which increases the power of the statistic but also slightly increases the risk of false positives.

Equivalent Symbols

The symbolization protocol proposed by Ruiz, López, and Páez (2010) and described previously—we call these *standard symbols*—contains a large amount of topological information regarding the units of analysis, including proximity and direction. In this sense, the protocol is fairly general. On the other hand, it is easy to see that the combinatorial possibilities can very quickly become unmanageable. For a process with $k = 3$ outcomes and $m = 5$, the number of symbols becomes $3^5 = 243$; for $k = 6$ and $m = 4$ it is $6^4 = 1{,}296$. Depending on the number of observations N, the explosion in the number of symbols can very rapidly consume degrees of freedom for hypothesis testing, because as a rule of thumb it is recommended that the number of symbolized locations be at least five times the number of symbols used (e.g., $S \geq 5k^m$), and S will usually be a fraction of N as per Equation (8). In addition, the large number of symbols might obfuscate the interpretation of results.

Table 3. Equivalent symbols for $k = 2$ and $m = 4$

Equivalent symbol	Standard symbols
$\sigma_1^* = \{4,0\}$	{0,0,0,0}
$\sigma_2^* = \{3,1\}$	{0,0,0,1}, {0,0,1,0}, {0,1,0,0}, {1,0,0,0}
$\sigma_3^* = \{2,2\}$	{0,0,1,1}, {0,1,1,0}, {1,1,0,0}, {1,0,0,1}, {0,1,0,1}, {1,0,1,0}, {1,0,0,1}
$\sigma_4^* = \{1,3\}$	{1,0,1,1}, {0,1,1,1}, {1,1,1,0}, {1,1,0,1}
$\sigma_5^* = \{0,4\}$	{1,1,1,1}

Note: Asterisk denotes equivalent symbols.

As an alternative, hereby we propose a symbolization protocol that sacrifices some amount of topological detail for conciseness. The alternative is based on the standard scheme; however, instead of retaining proximity and direction relationships, it maintains only the total number of occurrences of each outcome in an m-surrounding. We call these *equivalent symbols*. Because order in the sequence is not considered in this protocol, instead of a permutation with repetition, the number of symbols reflects a combination with repetition: $k(k+1)\ldots(k+m-1)/m!$. For example, the number of equivalent symbols for $k = 6$ and $m = 4$ is $6(6+1)(6+2)(6+3)/4! = 126$. Table 3 shows the equivalent symbols corresponding to the standard symbols for $k = 2$ and $m = 4$ (compare to Table 1). These equivalent symbols are read as follows: $\sigma^*{}_1$ is a location for which, in a neighborhood of four, there are no blacks; $\sigma^*{}_2$ is for locations where three out of four neighbors are white. Note the reduction in information: $\sigma^*{}_2$ includes the case where the nearest neighbor of a white is black (σ_4), as well as the case where the first two nearest neighbors are white (σ_2). In exchange, the number of symbols is greatly reduced, which relieves some pressure to work with smaller data sets. At least as important, interpretation of results also becomes more straightforward, something that facilitates the visual inspection of the frequency of classes.

Case Study: Data

Data used in the analyses to follow were compiled by the Urban Transition Historical GIS Project (UTP) at Brown University (Logan et al. 2011; see also

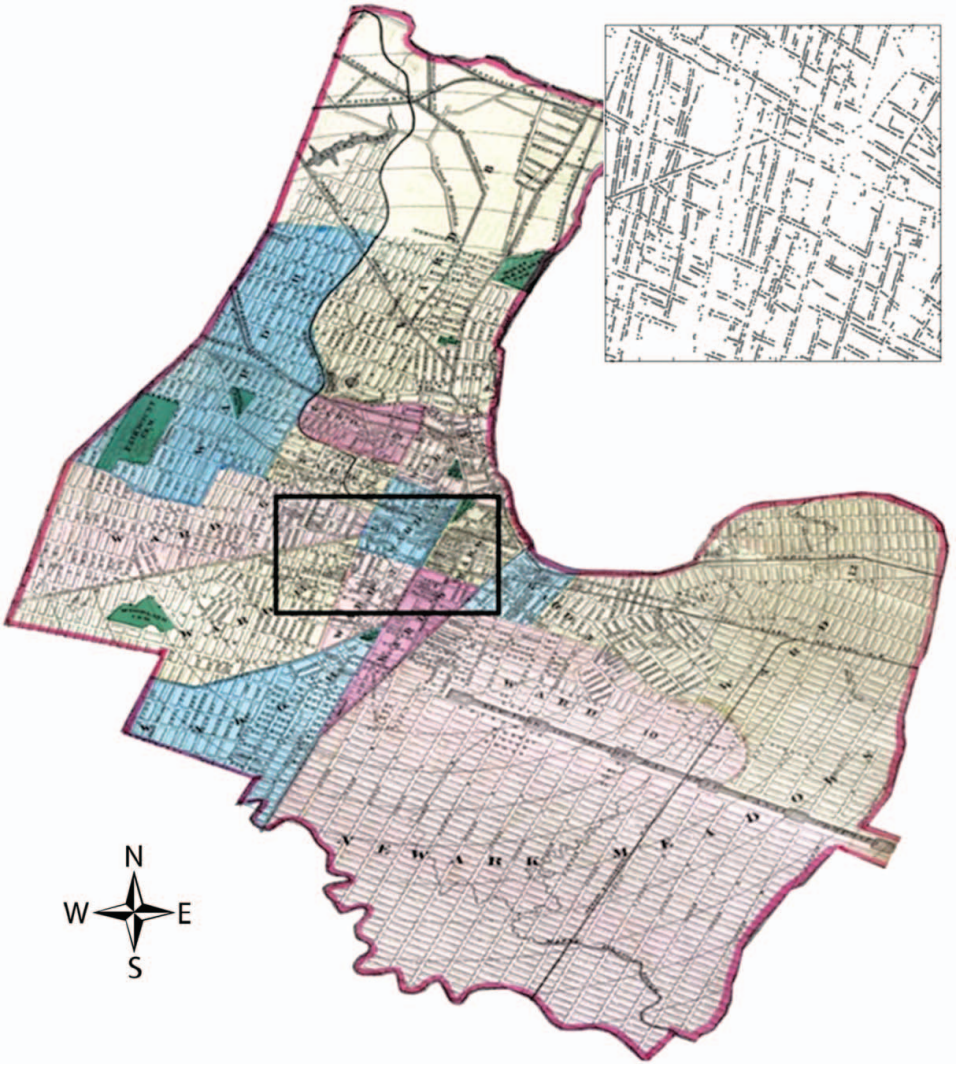


Figure 2. The City of Newark, New Jersey, in 1880, showing the approximate boundaries of the study area. The spatial distribution of cases appears in the inset. (Color figure available online.)

www.s4.brown.edu/utp). The project takes advantage of the 100 percent digital transcription of records from the 1880 Census that was organized by the Church of Jesus Christ of Latter Day Saints and prepared for scholarly use by the Minnesota Population Center. For thirty-nine major cities UTP has added addresses for all residents and is geocoding those addresses based on historical sources. Mapping begins with a contemporary geographic information systems (GIS) map of Essex County, which required considerable editing (deletion of new roads and other features, insertion of roads that had been eliminated, and correction of street names changed since 1880). In the case of Newark, key resources were a city directory from 1880 that includes address ranges for most streets and a detailed ward map circa 1872 showing the historical street grid. Nearly 97 percent of addresses have been successfully geocoded.

For the current application, only a portion of the data has been used. First, the main population groups in Newark in 1880 were Germans, Irish, and Yankees. Germans and Irish are persons who were born or had at least one parent who was born in Germany or Ireland, and Yankees are whites born in the United States with U.S.-born parents. These groups made up about 80 percent of the population, and for simplicity the analysis only considers the members of these groups. Further, only adults age eighteen and older are included. Second, the analysis is limited to the dense central portion of the city. As shown in Figure 2, the study area extends from the downtown area (near the river and including City Hall) and westward into Wards 6 and 13.

Out of a citywide total of 63,390 adult Germans, Irish, and Yankees with geocoded addresses, 21,520 lived in this portion of Newark. The ethnic

Table 4. Number and frequency of cases by class

Resolution	Class	Yankee	German	Irish	Mixed
Individual	Ethnicity	7,659(35.65%)	9,450(43.90%)	4,411 (20.50%)	
	Ethnicity and age < 30	2,667(12.40%)	3,545(16.57%)	1,682 (7.82%)	
Building	Dominant ethnicity	1,191(24.88%)	1,710(35.72%)	323 (6.75%)	1,563(32.61%)

composition of the study area was somewhat more German and less Irish than the city as a whole, but all three ethnic groups were well represented. In addition to discrete classification of individuals based on ethnicity, it is possible to introduce further classifications; for instance, by occupation, gender, age, marital status, and so on. For the sake of the following illustration, we also consider two age categories, namely, individuals younger than thirty and older than thirty years of age. The number and frequency of cases by class are shown in Table 4. In addition to data at the individual level, we also aggregate the information to obtain building occupancy. There are 4,787 unique locations (buildings) that we classify as follows: If the proportion of residents in any one building is greater than 50 percent, the building is classified as of that ethnic group. If no group is dominant at the 50 percent level, the building is classified as mixed. The number and frequency of cases appear in Table 4.

Analysis and Results[1]

Clustering and Exposure at the Individual Level

In this section we present the results of the ethnic clustering and exposure analysis. We begin with the more general case of clustering by ethnicity. The results of applying Q to the data are summarized in Table 5. The parameters used in the analysis appear there. Using an m-surrounding of 5 and overlap degree of 1 (so that any two proximate m-surroundings overlap at most in one observation), the number of symbolized locations[2] is 5,379. We calculate the statistic using standard and equivalent symbols. The results are highly significant and reject the null hypothesis of a random spatial sequence.

Lack of randomness, as illustrated by the examples in Figure 1, could take different forms. Because the cases are individuals, this could be separation of ethnic groups or intermingling in the case of the checkerboard pattern. Exploration of the relative frequency of symbols provides additional information about the characteristics of the spatial population pattern. The histogram of the relative frequency of symbols for individuals classified by ethnic group is shown in Figure 3. Figure 3 corresponds to the statistic for the twenty-one equivalent symbols in Table 5. Each bar in the histogram is accompanied by the expected relative frequency of the symbol under the null hypothesis of randomness and its respective 95 percent interval of confidence (see Appendix B for details on how these intervals are calculated). Recall that the expected value is calculated in consideration of the frequency of members of each ethnic group. Bars for symbols with frequencies that significantly depart from their expected value are color coded: Light gray indicates that the frequency exceeds the expectation, and dark gray indicates that the frequency is below the expectation under the null. Several bars in the figure are within the interval of confidence for the symbol. This implies that those symbols do not appear more or less frequently than what would be expected by chance.

Eight symbols appear with significantly more frequency than what would be expected under the null. This includes 5-surroundings composed exclusively of Germans (5G) or Yankees (5Y), which indicates clustering of these groups. In contrast, Irish do not display a similar tendency toward clustering. Irish, in fact, appear in a cluster as a majority only when in a neighborhood with Yankees. As a minority, they have a tendency to also appear more frequently in neighborhoods with Yankees (see symbols 4YI and 3Y2I) and to a lesser

Table 5. Clustering by ethnicity

Number of cases N	21,520		
Symbolized locations S	5,379		
Number of classes k	3		
Size of m-surrounding	5		
Degree of overlap r	1		
Number of standard symbols (σ)	243		
Number of equivalent symbols (σ^*)	21		
Frequency of classes	Y: 0.3559	I: 0.2050	G: 0.4391
Spatial association test	Statistic	Degrees of freedom	p Value
$Q(5)$ (standard symbols)	2,276.89	242	0.0000
$Q(5)$ (equivalent symbols)	2,050.01	20	0.0000

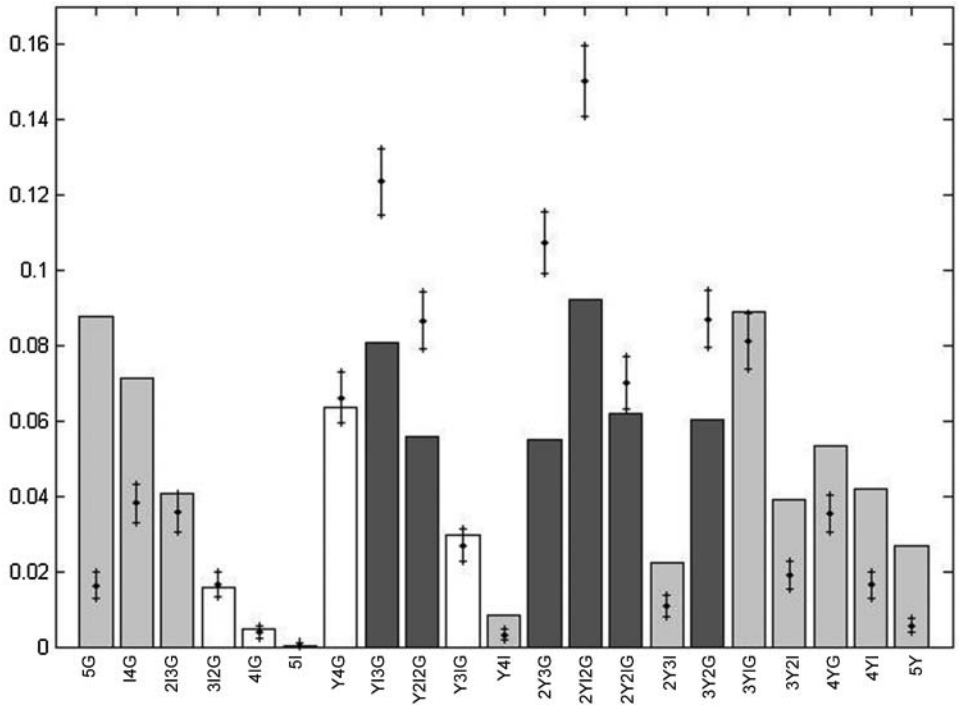


Figure 3. Relative symbol frequency for $Q(5)$ and ethnic classes Yankee, Irish, and German. The sequences on the x axis denote the composition of an m-surrounding of 5 (e.g., 5G is five Germans; 3YIG is three Yankees, one Irish, one German). Dark gray bars represent significantly lower frequencies, light gray bars are significantly higher frequencies, and white bars are not significant.

extent with Germans (see I4G). In contrast, mixed m-surroundings tend to be rare. All six symbols that appear significantly less frequently are for mixed neighborhoods and particularly mixed neighborhoods that include Germans (see YI3G, Y2I2G, 2Y3G, 2YI2G, 2Y2IG, and 3Y2G); the only exception is the case of a single German in a neighborhood that includes four Yankees (symbol 4YG). The overall pattern is one of ethnic clustering, especially for Germans and to a lesser extent Yankees. The spatial distribution of Irish individuals is reminiscent of the checkerboard pattern indicative of intermingling, in particular in combination with Yankees.

Patterns of exposure are more easily seen if we reclassify the cases. We illustrate two situations: exposure of Germans, an ethnic group that displays a significant tendency to cluster, to members of other ethnic groups; and exposure of Yankees, almost as numerous as Germans and with a tendency to cluster that does not match that of Germans. The results of running the statistic for these new classifications (Germans and Others, and Yankees and Others) are shown in Tables 6 and 7. After reclassification, $k = 2$, and the number of symbols is reduced. The statistic is calculated using both standard and equivalent symbols, and the results are, once again, highly significant for both cases.

Table 6. Exposure: Germans

Number of cases N	21,520		
Symbolized locations S	5,379		
Number of classes k	2		
Size of m-surrounding	5		
Degree of overlap r	1		
Number of standard symbols (σ)	32		
Number of equivalent symbols (σ^*)	6		
Frequency of classes	G: 0.4391	O: 0.5609	
Spatial association test	Statistic	Degrees of freedom	p Value
$Q(5)$ (standard symbols)	1,792.84	31	0.0000
$Q(5)$ (equivalent symbols)	1,774.23	5	0.0000

Table 7. Exposure: Yankees

Number of cases N	21,520		
Symbolized locations S	5,379		
Number of classes k	2		
Size of m-surrounding	5		
Degree of overlap r	1		
Number of standard symbols (σ)	32		
Number of equivalent symbols (σ^*)	6		
Frequency of classes	Y: 0.3559	O: 0.6441	
Spatial association test	Statistic	Degrees of freedom	p Value
$Q(5)$ (standard symbols)	1,138.00	31	0.0000
$Q(5)$ (equivalent symbols)	1,121.54	5	0.0000

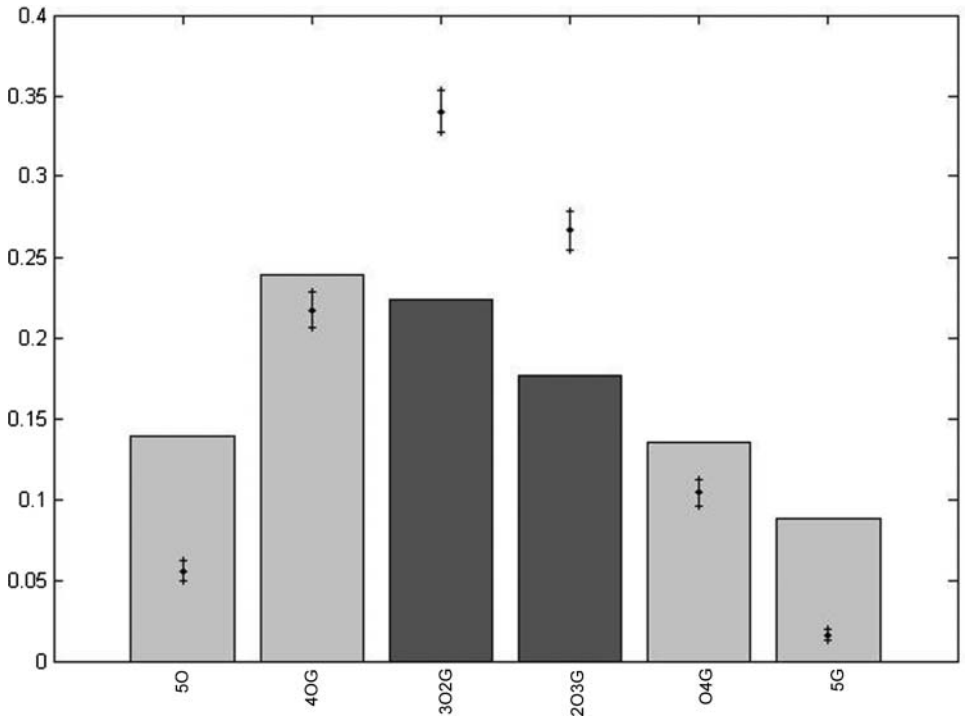


Figure 4. Relative symbol frequency for $Q(5)$ and classes German and other. The sequences on the x axis denote the number of others (i.e., non-German) and German in an m-surrounding of 5 (e.g., 5G is five Germans, 302G is three non-Germans and two Germans). Dark gray bars represent significantly lower frequencies, and light gray bars are significantly higher frequencies.

As before, it is possible to explore the characteristics of the nonrandom pattern by means of the frequency of symbols. Histograms for the relative frequency of symbols are shown in Figures 4 and 5. In the case of exposure of Germans, all symbols deviate significantly from their expected frequencies. It can be seen in Figure 4 that a frequent occurrence is, in a neighborhood of five, a given German exposed solely to other Germans or, at most, one single individual of a different ethnic group (symbols 5G and 4GO). Likewise, there are significantly more cases of members of other groups not exposed to Germans than what would be expected

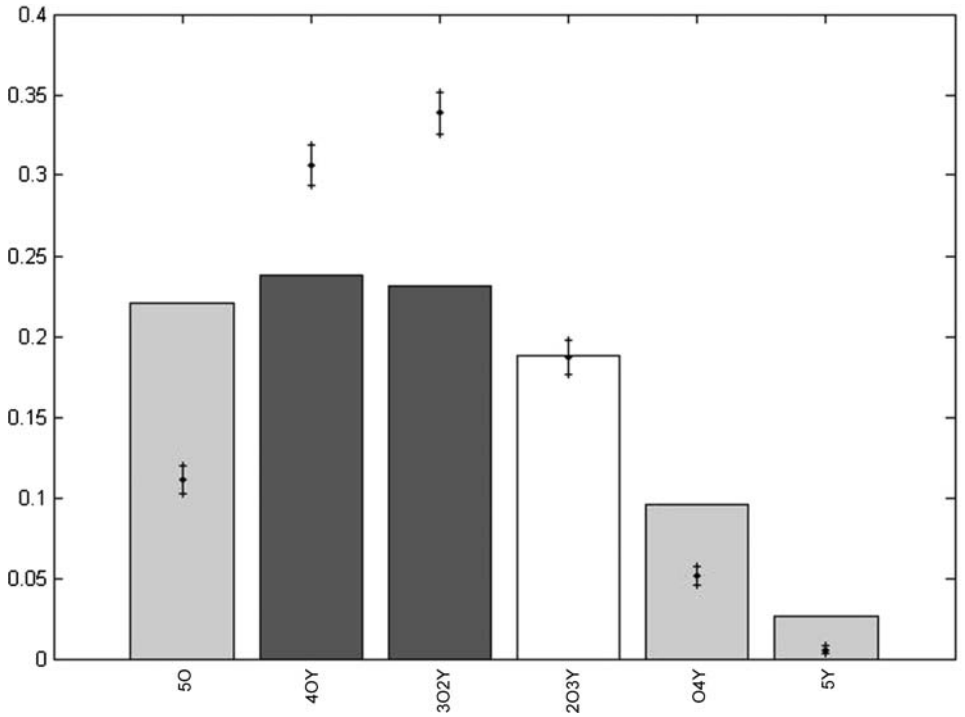


Figure 5. Relative symbol frequency for $Q(5)$ and classes Yankee and other. The sequences on the x axis denote the number of others (i.e., non-Yankee) and Yankees in an m-surrounding of 5 (e.g., 5Y is five Yankees, 302Y is three non-Yankees and two Yankees). Dark gray bars represent significantly lower frequencies, light gray bars are significantly higher frequencies, and the white bar is not significant.

by chance (symbol 50), although there are also more cases where members of other groups are exposed to a single German (symbol G40).

Clearly, as Germans and Yankees are the two most numerous groups, exposure relationships must be reciprocal between these two groups. Nonetheless, the exposure of Yankees to other Yankees is less marked, in relative terms, than was the case for Germans. And, although there are more cases than expected of members of other groups not exposed to Yankees, again the relative deviation from the expected value is less dramatic. Members of other groups also tend to be less exposed to a single Yankee; however, neighborhoods with three Yankees and two members of other groups occur as one would expect purely by chance.

Subclasses: Ethnicity and Age

Exploration of the spatial distribution of individuals of ethnic groups indicates a tendency toward intraethnic clustering, with some mixing between Yankees and Irish. The analysis could be refined by considering additional dimensions; for example, marital status, gender, or, as we illustrate in this subsection, age. A new classification scheme now subdivides each ethnic group according to age, those who are younger than thirty (represented in the symbol by lowercase: y, i, g, for Yankee, Irish, and German, respectively) and thirty or older (represented in the symbol by uppercase: Y, I, G, for Yankee, Irish, and German, respectively). The number of symbols increases with the number of classes ($k = 6$), and we adjust the size of the m-surrounding to reflect this. Application of the statistic with $m = 3$ and $r = 1$ indicates again that the distribution of individuals by ethnic group and age is not random (see Table 8).[3]

A number of symbols are within the 95 percent confidence intervals of their expected frequency under the null (Figure 6). Most are significantly more or less frequent than expected. Inspection of the histogram of relative symbol frequencies adds depth to the previous analysis by ethnicity only. For instance, although Germans of all ages tend in general to be in ethnic clusters, this tendency is relatively stronger for older Germans (≥30; see the magnitude of the deviation of symbols 3G and g2G from their expected values, compared to symbols 2gG and 3g). The evidence of clustering among Irish was, compared to the other ethnic groups, less strong. In particular, the cluster for five Irish individuals was not significant in the previous analysis, as seen in Figure 3. When exploring neighborhoods of three after classifying individuals by ethnicity and age, it turns out that older Irish do tend to cluster together (see symbol 3I) but younger Irish do not (see 3i). Of nineteen symbols that appear less frequently than expected, thirteen correspond to mixed neighborhoods, with one individual of each ethnic group. It is intriguing to note that although mixed neighborhoods of three older residents (i.e., YIG) are less frequent than expected, this significance is practically lost for mixed neighborhoods of younger residents (i.e., yig). Positive deviations from the expected value occur only for mixed group that include Yankees and Irish of different generations. In general, there is more interethnic and intergenerational clustering among Yankees and Irish (see positive deviations for 2Yi, 2YI, and Y2I), than among Germans and any other group (see negative deviations for 2Yg, Y2g, y2G, 2YG, y2g, and Y2G).

Clustering at the Building Level

In our final analysis we show how Q can be applied to a higher level of geography, in this case by aggregating

Table 8. Clustering by ethnicity and age

Number of cases N	21,520		
Number of symbolized locations S	10,759		
Number of classes k	6		
Size of m-surrounding	3		
Degree of overlap r	1		
Number of standard symbols (σ)	216		
Number of equivalent symbols (σ^*)	56		
Frequency of classes	*YL30*: 0.1239	*IL30*: 0.0782	*GL30*: 0.1647
	YG30: 0.2320	*IG30*: 0.1268	GG30: 0.2744
Spatial association test	Statistic	Degrees of freedom	p Value
$Q(3)$ (standard symbols)	1,667.45	215	0.0000
$Q(3)$ (equivalent symbols)	1,482.29	55	0.0000

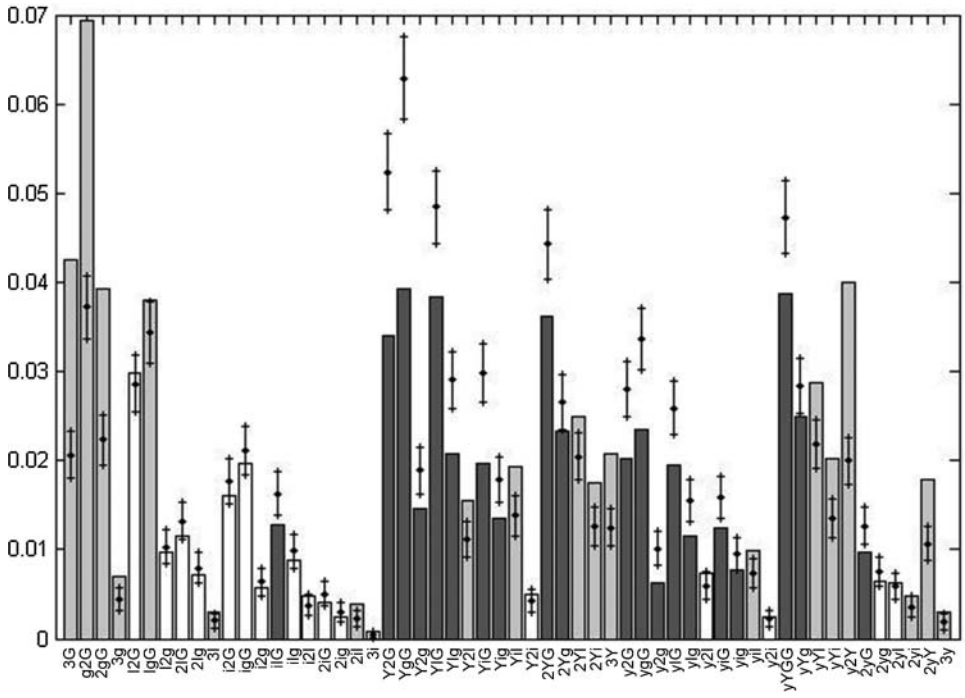


Figure 6. Relative symbol frequency for $Q(3)$ and classes Yankee (y for age <30 and Y for age >30), Irish (i for age <30 and I for age >30), and German (g for age <30 and G for age >30). The sequences on the x axis denote the composition of an m-surrounding of 3 (e.g., 2gG is two younger and one older German, y2I is one younger Yankee and two older Irish). Dark gray bars represent significantly lower frequencies, light gray bars are significantly higher frequencies, and white bars are not significant.

cases to buildings and classifying buildings as Yankee, Irish, and German according to the dominant ethnicity and mixed if no ethnic group is in the majority. The results of the analysis using $m = 4$ and $r = 1$ appear in Table 9. According to the decision rule, the null hypothesis of a random spatial sequence is rejected, using both standard and equivalent symbols.[4] Further, inspection of the relative symbol frequencies (Figure 7) indicates that mixed buildings tend to be more proximate than expected by chance (see 4M, Y3M, and I3M) except when there is a German building in the vicinity (symbol G3M). Although mixed buildings tend to cluster, mixed clusters of various building types are significantly less common than expected (YIGM). German buildings

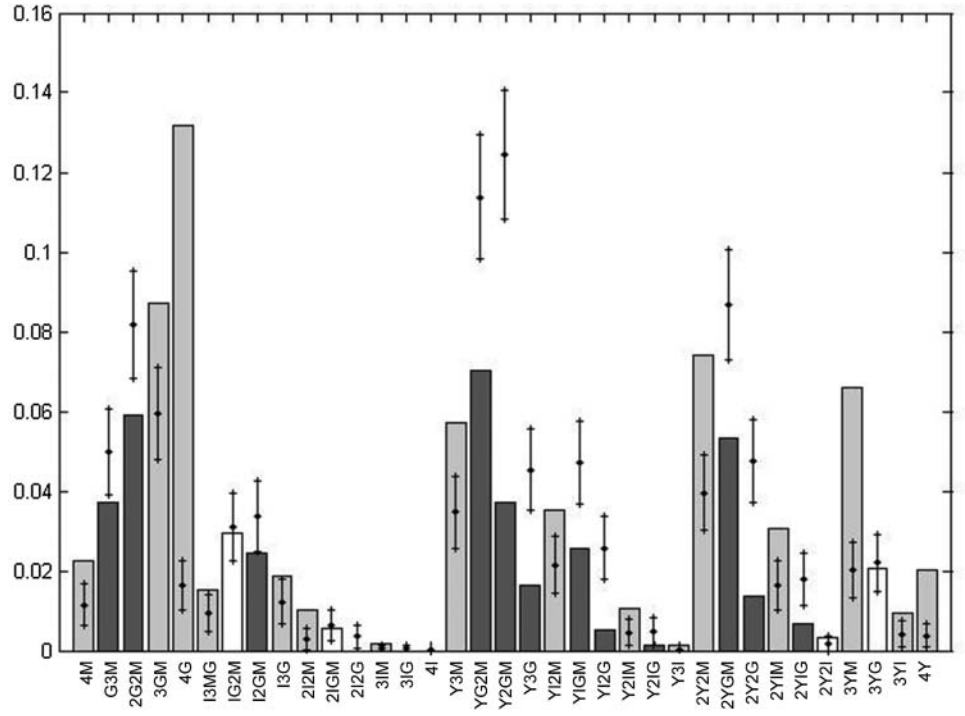


Figure 7. Relative symbol frequency for $Q(4)$ and building classes Yankee, Irish, German, and mixed. The sequences on the x axis denote the number of buildings of each class (e.g., 4M is four mixed buildings; 2YIG is two Yankee buildings, one Irish, and one German). Dark gray bars represent significantly lower frequencies, light gray bars are significantly higher frequencies, and white bars are not significant.

Table 9. Clustering by building ethnicity

Number of cases N	4,787		
Number of symbolized locations S	1,595		
Number of classes k	4		
Size of m-surrounding	4		
Degree of overlap r	1		
Number of standard symbols (σ)	256		
Number of equivalent symbols (σ^*)	35		
Frequency of classes	Y: 0.2488	I: 0.0675	
	G: 0.3572	M: 0.3265	
Spatial association test	Statistic	Degrees of freedom	p Value
$Q(4)$ (standard symbols)	1,503.82	255	0.0000
$Q(4)$ (equivalent symbols)	1,231.25	34	0.0000

are disproportionally more likely to cluster (4G, I3G, 3GM) except when there is a Yankee building in the vicinity (Y3G). This changes when German buildings are not the majority, as clusters of this type are rare (G3M, 2G2M, YG2M, and Y2GM).

Irish buildings, like Irish individuals, display less clustering and spatial association. When they do, even accounting for their smaller numbers, they tend to be in more integrated neighborhoods (see I3M and 2I2M) or embedded in other ethnic neighborhoods (see I3G and 3YI). Finally, we also find that Yankee buildings also tend to colocate (see 4Y) and, when in the majority, they appear with more frequency in the company of Irish or mixed buildings (see 3YI and 3YM).

Further Opportunities for Spatial Analysis

Symbolization, in addition to forming the basis for statistical analysis as detailed in the preceding sections, also provides the basis for further opportunities for spatial analysis. Having already determined, for instance, that a certain symbol (e.g., four German buildings in m-surroundings of size 4) appears more (or less) than what would be expected by chance, a question of interest is whether these clusters display a spatial pattern. Figure 8 illustrates this possible use of the symbolized cases. Figure 8 shows in three panels the symbols corresponding to clusters of four Yankee, four German, and four mixed buildings in m-surroundings of size 4. Clusters of four Irish buildings were not found with greater or lesser frequency than under the null hypothesis of randomness, and are therefore of limited interest. As seen in Figure 8A, clusters of four German buildings display a coherent spatial pattern of concentration along the center and especially north of the study area. In contrast, few clusters of other ethnic buildings are found in the region. Clusters of four mixed buildings are mostly located to the east, mainly along two or three parallel streets (Figure 8B). Clusters of Yankee buildings are mostly in the east and west of the study region, with a tendency toward the southern edge (Figure 8C).

Concluding Remarks

In a recent survey of the state of research on ethnic segregation, Kaplan and Woodhouse (2005) reflected on a number of problems that affect traditional

Figure 8. Spatial distribution of symbolized cases ($m = 4$): (A) four German buildings in m-surrounding (○); (B) four mixed buildings in m-surrounding (□); (C) four Yankee buildings in m-surrounding (◊).

approaches to measure segregation and noted progress along different fronts. These issues include the measurement of segregation in situations where multiple groups are present, the fact that many measures do not consider the spatial relationships between units of analysis, and the question of geographical scale. A number of developments in the past few years contribute toward addressing some of these issues. In the case of scale, the use of distance-based kernel approaches now allows for the measurement of segregation at multiple scales (e.g., Reardon and O'Sullivan 2004; Wong 2004; Feitosa et al. 2007; O'Sullivan and Wong 2007; Kramer et al. 2010). In terms of spatial relationships between units of analysis, several examples exist of studies that explicitly incorporate them by casting different autocorrelation measures as indicators of segregation (e.g., Brown and Chung 2006; Johnston, Poulsen, and Forrest 2009; Poulsen, Johnston, and Forrest 2010). Lastly, there has been progress in the development of measures of multigroup segregation (e.g., Wong 1998; Reardon and Firebaugh 2002) and the definition of typologies that seek to refocus analysis on the mix of the population (Johnston, Poulsen, and Forrest 2010).

Use of the Q statistic, demonstrated in this article in an application to historical data, follows on the heels of some of these advances and augments the analytical possibilities in segregation research. Our approach has a number of qualities to recommend it, indeed, some that contribute positively to several of the issues identified by Kaplan and Woodhouse (2005) in their survey. First, the Q statistic can naturally accommodate multiple groups, as illustrated in our analysis of three different population classes (Irish, Germans, and Yankees) and even subclasses (e.g., age classifications). Second, Q is inherently about spatial relationships between units of analysis, which enter the statistic by means of the definition of m-surroundings or neighborhoods of size m. This leads to the first noteworthy aspect of our approach with respect to scale. Q detects association of a spatial qualitative variable at the level of m, and therefore selection of m allows the analyst to explore spatial association at different scales. It must be noted, though, that interpretability of the results might become an issue at larger m-surrounding sizes as the number of symbols increases. The second issue related to scale is that, unlike other approaches that are based on proportions or densities and that are therefore inherently areal, the Q statistic can be applied at the most basic level of analysis (the individual) and can be scaled up as desired, as illustrated in our analysis of building occupancy by ethnicity. An intriguing possibility in terms of scaling up the analysis to administrative areas is to combine the spatial dimension of Q with the (inherently categorical) typology of population mixes of Johnston, Poulsen, and Forrest (2007). This, we suggest, is a worthy avenue for future research. As well, comparing Q to other existing measures of clustering and exposure might provide additional insights into the appropriateness of various measures in different application contexts.

On a final note, our approach does not aspire to be general (Reardon and O'Sullivan 2004; Wong 2005). Rather, we would argue that its value resides precisely in its specificity, as it unambiguously deals with only one dimension of segregation, on the clustering–exposure continuum. That being said, we suggest that a more complete spatial analytical framework to explore segregation could combine our approach to measuring clustering exposure and the use of local indicators of spatial association (Anselin 1995) or concentration (Getis and Ord 1992) as measures of concentration-evenness. This would provide a fully spatial picture of the two major dimensions of segregation.

Acknowledgments

The authors express their appreciation to three anonymous reviewers for their valuable comments on previous versions of this article. The work of Manuel Ruiz was partially supported by MCI (Ministerio de Ciencia e Innovación) grant MTM2009–07373, Fundación Séneca of Región de Murcia and by grant 861–2009–2010 from the Social Sciences and Humanities Research Council of Canada. Fernando López received financial support from the project ECO-2009–10534-ECON of Ministerio de Ciencia y Tecnología and from the project 11897/PHCS/09 of Fundación Séneca de la Región de Murcia. John Logan was supported by research grants from the National Science Foundation (0647584) and the National Institutes of Health (1R01HD049493–01A2). The usual disclaimer applies: The authors alone are responsible for the contents of this article.

Notes

1. MATLAB code to calculate and test Q is available as supplementary material that accompanies Ruiz, López, and Páez (2010). The code can also be downloaded at http://www.science.mcmaster.ca/geo/faculty/paez/publications.html#journals
2. As noted earlier, decreasing overlap degree reduces the risk of false positives but also the power of the statistic. The application is therefore very conservative. For

thoroughness, we calculated the statistic using $r = 2$, 3, and 4. The statistic is highly significant and rejects the null hypothesis of randomness in every case. As well, the relative frequency of symbols, and their significance, does not display undue variations. Detailed results for this sensitivity analysis are available from the authors.

3. We also calculated the statistic using $r = 2$. The results hold.
4. We also calculated the statistic using $r = 2$ and 3. The results hold.

References

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical Analysis* 27:93–115.

Bolt, G., A. S. Ozuekren, and D. Phillips. 2010. Linking integration and residential segregation. *Journal of Ethnic and Migration Studies* 36 (2): 169–86.

Boyd, R. L. 1998. Residential segregation by race and the Black merchants of northern cities during the early twentieth century. *Sociological Forum* 13 (4): 595–609.

Brown, L. A., and S. Y. Chung. 2006. Spatial segregation, segregation indices and the geographical perspective. *Population Space and Place* 12 (2): 125–43.

Charles, C. Z. 2003. The dynamics of racial residential segregation. *Annual Review of Sociology* 29:167–207.

Cutler, D. M., E. L. Glaeser, and J. L. Vigdor. 2008. Is the melting pot still hot? Explaining the resurgence of immigrant segregation. *Review of Economics and Statistics* 90 (3): 478–97.

Dawkins, C. J. 2004. Measuring the spatial pattern of residential segregation. *Urban Studies* 41 (4): 833–51.

Dawkins, C. J., M. Reibel, and D. W. Wong. 2007. Introduction—Further innovations in segregation and neighborhood change research. *Urban Geography* 28 (6): 513–15.

Deurloo, M. C., and S. de Vos. 2008. Measuring segregation at the micro level: An application of the M measure to multi-ethnic residential neighbourhoods in Amsterdam. *Tijdschrift Voor Economische en Sociale Geografie* 99 (3): 329–47.

DuBois, W. E. B. 1903. *The souls of black folk: Essays and sketches*. New York: Vintage Books.

Feitosa, F. F., G. Camara, A. M. V. Monteiro, T. Koschitzki, and M. P. S. Silva. 2007. Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science* 21 (3): 299–323.

Fong, E., and K. Shibuya. 2000. The spatial separation of the poor in Canadian cities. *Demography* 37 (4): 449–59.

Getis, A. 2008. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis* 40 (3): 297–309.

Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24 (3): 189–206.

Gotham, K. F. 2000. Urban space, restrictive covenants and the origins of racial residential segregation in a US city, 1900–50. *International Journal of Urban and Regional Research* 24 (3): 616–33.

Harsman, B. 2006. Ethnic diversity and spatial segregation in the Stockholm region. *Urban Studies* 43 (8): 1341–64.

Hershberg, T., A. N. Burstein, E. P. Ericksen, S. Greenberg, and W. L. Yancey. 1979. Tale of 3 cities—Blacks and immigrants in Philadelphia—1850–1880, 1930 and 1970. *Annals of the American Academy of Political and Social Science* 441 (1): 55–81.

Jakubs, J. F. 1981. A distance-based segregation index. *Socio-Economic Planning Sciences* 15 (3): 129–36.

Johnston, R., J. Forrest, and M. Poulsen. 2002. Are there ethnic enclaves/ghettos in English cities? *Urban Studies* 39 (4): 591–618.

Johnston, R., M. Poulsen, and J. Forrest. 2007. The geography of ethnic residential segregation: A comparative study of five countries. *Annals of the Association of American Geographers* 97 (4): 713–38.

———. 2009. Measuring ethnic residential segregation: Putting some more geography in. *Urban Geography* 30 (1): 91–109.

———. 2010. Moving on from indices, refocusing on mix: On measuring and understanding ethnic patterns of residential segregation. *Journal of Ethnic and Migration Studies* 36 (4): 697–706.

Kantrowitz, N. 1979. Racial and ethnic residential segregation in Boston 1830–1970. *Annals of the American Academy of Political and Social Science* 441 (1): 41–54.

Kaplan, D. H., and K. Woodhouse. 2004. Research in ethnic segregation I: Causal factors. *Urban Geography* 25 (6): 579–85.

———. 2005. Research in ethnic segregation II: Measurements, categories and meanings. *Urban Geography* 26 (8): 737–45.

Kramer, M. R., H. L. Cooper, C. D. Drews-Botsch, L. A. Waller, and C. R. Hogue. 2010. Do measures matter? Comparing surface-density-derived and census-tract-derived measures of racial residential segregation. *International Journal of Health Geographics* 9 (29): 1–15.

Lloyd, C. D. 2010. Exploring population spatial concentrations in Northern Ireland by community background and other characteristics: An application of geographically weighted spatial statistics. *International Journal of Geographical Information Science* 24 (8): 1193–1221.

Logan, J. R., R. D. Alba, and W. Q. Zhang. 2002. Immigrant enclaves and ethnic communities in New York and Los Angeles. *American Sociological Review* 67 (2): 299–322.

Logan, J. R., J. Jindrich, H. Shin, and W. Zhang. 2011. Mapping America in 1880: The Urban Transition Historical GIS Project. *Historical Methods* 44 (1): 49–60.

Logan, J. R., and C. Zhang. 2010. Global neighborhoods: New pathways to diversity and separation. *American Journal of Sociology* 115 (4): 1069–1109.

Marcon, E., and F. Puech. 2003. Evaluating the geographic concentration of industries using distance-based methods. *Journal of Economic Geography* 3 (4): 409–28.

Massey, D. S., and N. A. Denton. 1988. The dimensions of residential segregation. *Social Forces* 67 (2): 281–315.

———. 1993. *American apartheid: Segregation and the making of the underclass*. Cambridge, MA: Harvard University Press.

O'Sullivan, D., and D. W. S. Wong. 2007. A surface-based approach to measuring spatial segregation. *Geographical Analysis* 39 (2): 147–68.

Peach, C. 1996. Does Britain have ghettos? *Transactions of the Institute of British Geographers* 21 (1): 216–35.

———. 2009. Slippery segregation: Discovering or manufacturing ghettos? *Journal of Ethnic and Migration Studies* 35 (9): 1381–95.
Pettigrew, T. F. 1979. Racial change and social policy. *Annals of the American Academy of Political and Social Science* 441 (1): 114–31.
Poulsen, M. F., and R. J. Johnston. 2000. The ghetto model and ethnic concentration in Australian cities. *Urban Geography* 21 (1): 26–44.
Poulsen, M., R. Johnston, and J. Forrest. 2010. The intensity of ethnic residential clustering: Exploring scale effects using local indicators of spatial association. *Environment and Planning A* 42 (4): 874–94.
Reardon, S. F., and G. Firebaugh. 2002. Measures of multigroup segregation. *Sociological Methodology* 32:33–67.
Reardon, S. F., S. A. Matthews, D. O'Sullivan, B. A. Lee, G. Firebaugh, C. R. Farrell, and K. Bischoff. 2008. The geographic scale of metropolitan racial segregation. *Demography* 45 (3): 489–514.
Reardon, S. F., and D. O'Sullivan. 2004. Measures of spatial segregation. *Sociological Methodology* 34 (1): 121–62.
Reiner, T. A. 1972. Racial segregation: A comment. *Journal of Regional Science* 12 (1): 137.
Ruiz, M., F. López, and A. Páez. 2010. Testing for spatial association of qualitative data using symbolic dynamics. *Journal of Geographical Systems* 12 (3): 281–309.
Simpson, L. 2004. Statistics of racial segregation: Measures, evidence and policy. *Urban Studies* 41 (3): 661–81.
Simpson, L., and C. Peach. 2009. Measurement and analysis of segregation, integration and diversity: Editorial introduction. *Journal of Ethnic and Migration Studies* 35 (9): 1377–80.
Smith, G. C. 1998. Change in elderly residential segregation in Canadian metropolitan areas, 1981–91. *Canadian Journal on Aging-Revue Canadienne du Vieillissement* 17 (1): 59–82.
Spain, D. 1979. Race-relations and residential segregation in New Orleans—2 centuries of paradox. *Annals of the American Academy of Political and Social Science* 441 (1): 82–96.
Weaver, J. C. 1978. From Land Assembly to Social Maturity—Suburban Life of Westdale (Hamilton), Ontario, 1911–1951. *Histoire Sociale-Social History* 11 (22): 411–40.
White, M. J. 1983. The measurement of spatial segregation. *American Journal of Sociology* 88 (5): 1008–18.
Wong, D. W. S. 1997. Spatial dependency of segregation indices. *Canadian Geographer-Geographe Canadien* 41 (2): 128–36.
———. 1998. Measuring multiethnic spatial segregation. *Urban Geography* 19 (1): 77–87.
———. 1999. Geostatistics as measures of spatial segregation. *Urban Geography* 20 (7): 635–47.
———. 2002. Modeling local segregation: A spatial interaction approach. *Geographical and Environmental Modelling* 6 (1): 81–97.
———. 2004. Comparing traditional and spatial segregation measures: A spatial scale perspective. *Urban Geography* 25 (1): 66–82.
———. 2005. Formulating a general spatial segregation measure. *The Professional Geographer* 57 (2): 285–94.
Wong, D. W. S., H. Lasus, and R. F. Falk. 1999. Exploring the variability of segregation index D with scale and zonal systems: An analysis of thirty US cities. *Environment and Planning A* 31 (3): 507–22.
Wong, D. W., M. Reibel, and C. J. Dawkins. 2007. Introduction-segregation and neighborhood change: Where are we after more than a half-century of formal analysis. *Urban Geography* 28 (4): 305–11.
Zelder, R. E. 1972. Racial segregation: A reply. *Journal of Regional Science* 12 (1): 149–53.

Appendix A: Procedure for Selecting Sites with Controlled Degree of Overlap

To select S locations for the analysis, coordinates are selected such that for any two coordinates s_i, s_j the number of overlapping nearest neighbors of s_i and s_j are at most r. The set S, which is a subset of all the observations N, is defined recursively as follows. First choose a location $\boldsymbol{s}_0$ at random and fix an integer r with $0 \leq r < m$. The integer r is the degree of overlap, the maximum number of observations that contiguous m-surroundings are allowed to have in common. Let $\{s_1^0, s_2^0, \ldots, s_{m-1}^0\}$ be the set of nearest neighbors to $\boldsymbol{s}_0$, where the s_i^0s are ordered by distance to $\boldsymbol{s}_0$, or angle in the case of ties. Let us call $s_1 = s_{m-r-1}^0$ and define $A_0 = \{s_0, s_1^0, \ldots, s_{m-r-2}^0\}$. Take the set of nearest neighbors to $\boldsymbol{s}_1$, namely, $\{s_1^1, s_2^1, \ldots, s_{m-1}^1\}$, in the set of locations $S \setminus A_0$ and define $s_2 = s_{m-r-1}^1$. Now for $i > 1$ we define $s_i = s_{m-r-1}^{i-1}$ where s_{m-r-1}^{i-1} is in the set of nearest neighbors to s_{i-1}, $\{s_1^{i-1}, s_2^{i-1}, \ldots, s_{m-1}^{i-1}\}$, of the set $S \setminus \{\cup_{j=0}^{i-1} A_j\}$. Continue this process while there are locations to symbolize.

Appendix B: Intervals of Confidence for Histogram

In addition to providing a decision rule to reject the null hypothesis of spatial randomness, the Q statistic can also be used to explore in more detail the characteristics of pattern. This is done by preparing a visual representation of the frequency or relative frequency of classes. This can be a leaf-and-stem plot or a histogram. A question of interest is whether a specific symbol appears more or less frequently than what would be expected by chance. This question can be addressed by including as part of the histogram the intervals of confidence with respect to the expected (relative) frequency under the null hypothesis. These intervals of confidence can be calculated in the following way.

Fix a symbol σ. Then the number of times that a symbolized location is of σ-type, namely, Ψ_σ, can be approximated to a binomial distribution:

$$\Psi_\sigma \approx B(S, p_\sigma)$$

where S is the total number of symbolized locations. When S is large enough, the binomial distribution can be approximated to a normal distribution with the following parameters:

$$\Psi_\sigma \approx B\,(S, p_\sigma) \approx N\left(Sp_\sigma, \sqrt{Sp_\sigma\,(1-p_\sigma)}\right)$$

And therefore we get that:

$$\frac{\Psi_\sigma - Sp_\sigma}{\sqrt{Sp_\sigma\,(1-p_\sigma)}} \approx N\,(0,1)$$

Let $0 \leq \alpha \leq 1$. Let $z_{\alpha/2}$be the real number satisfying that $P(N(0,1) \geq z_{\alpha/2}) = \alpha/2$. Then, because the normal standard distribution N(0,1) is symmetric with respect to $x = 0$ axis, we have that:

$$\begin{aligned}\alpha &= P\left(-z_{\alpha/2} \leq \frac{\Psi_\sigma - Sp_\sigma}{\sqrt{Sp_\sigma\,(1-p_\sigma)}} \leq z_{\alpha/2}\right)\\ &= P\left(Sp_\sigma - z_{\alpha/2}\sqrt{Sp_\sigma\,(1-p_\sigma)} \leq \Psi_\sigma \leq Sp_\sigma\right.\\ &\quad \left. + \; z_{\alpha/2}\sqrt{Sp_\sigma\,(1-p_\sigma)}\right)\end{aligned}$$

and therefore we get that:

$$\left(p_\sigma - z_{\alpha/2}\sqrt{\frac{p_\sigma\,(1-p_\sigma)}{S}}, p_\sigma + z_{\alpha/2}\sqrt{\frac{p_\sigma\,(1-p_\sigma)}{S}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for the relative frequency of a symbol to occur $\frac{\Psi_\sigma}{S}$.

Correspondence: School of Geography and Earth Science, McMaster University, Hamilton, ON, L8S 3Z9, Canada, e-mail: paezha@mcmaster.ca (Páez); Departamento de Métodos Cuantitativos e Informáticos, Universidad Politécnica de Cartagena, Cartagena, 30201, Spain, e-mail: manuel.ruiz@upct.es (Ruiz); fernando.lopez@upct.es (López); Spatial Structures in the Social Sciences, Brown University, Providence, RI 02912, e-mail: John_Logan@brown.edu (Logan).