J Geograph Syst (2003) 5:139–160
DOI: 10.1007/s10109-003-0110-3

Journal of Geographical Systems

Developing local measures of spatial association for categorical data

Barry Boots

Department of Geography and Environmental Studies, Wilfrid Laurier University, Waterloo, Ontario N2L 3C5, Canada (e-mail: bboots@wlu.ca)

Received: 5 January 2003 / Accepted: 1 May 2003

Abstract. This paper describes a procedure for extending local statistics to categorical spatial data. The approach is based on the notion that there are two fundamental characteristics of categorical spatial data; composition and configuration. Further, it is argued that, when considered locally, the latter should be measured conditionally with respect to the former. These ideas are developed for binary, gridded data. Local composition is measured by counting the numbers of cells of a particular type, while local configuration is measured by join counts. The approach is illustrated using a small, empirical data set and an *ad hoc* procedure is developed to deal with the impact of global spatial autocorrelation on the local statistics.

Key words: categorical spatial data, local spatial statistics, spatial autocorrelation

JEL classification: C0, C15, C49

1 Introduction

A set of spatial data consists of measurements (*data values*) taken at specific locations (*data sites*) in a geographic space (*study region*). Numerous measures of spatial association have been developed to examine the nature and extent of spatial dependence in such data. Global measures of spatial association do this by using the complete data set to derive a single value for the entire study region. As such these measures emphasize average or typical characteristics of the complete data set. Inherent in such an approach is the implicit assumption that the single value holds throughout the study region, i.e., the study region is

The author gratefully acknowledges financial support from the GEOIDE Network of Centres of Excellence (ENV #4) and the helpful comments of three anonymous reviewers.

environmentally homogeneous (Unwin 1996a; Fotheringham and Brunsdon 1999). Statistically, this is equivalent to assuming that the process(es) which give rise to the data values are stationary over the study region (Unwin 1996b). Unless stationarity holds, the global value will not be universally applicable throughout the study region, and may not even apply in any part of it (Fotheringham 1997; Fotheringham and Brunsdon 1999). In contrast to global measures, local measures of spatial association examine spatial dependence in subsets (variously known as neighbourhoods, windows or kernels) defined with respect to each data site i in the complete data set. Such measures focus on the identification of variations, rather than regularities, in the nature of spatial association within the study region (Fotheringham 1997, 1999; Fotheringham and Brunsdon 1999). Thus, while global approaches yield a single value for the entire data set, local approaches are capable of generating a local value for each data site in the data set.

Getis and Ord (1996) suggest that local measures of spatial association can be used for several purposes, including:

1. assessing the assumption of stationarity for a given study region;
2. identifying the existence of pockets of distinctive data values (hot and cold spots);
3. identifying the scale (spatial extent) at which there is no discernible association of data values.

In addition, if the measure satisfies the two requirements of a local indicator of spatial association (LISA) (Anselin 1995), it can also be used to decompose the corresponding global measure into the contributions of individual data sites, thus revealing which sites have the most impact on the global measure.

While a number of local measures of spatial dependence have been developed for quantitative data, most notably local Moran's I, local Geary's c, and the G and O statistics of Getis and Ord (for a review of the first three see Boots 2002, for the last see Ord and Getis 2001), research aimed at developing measures for use with categorical data is just beginning (Wilhelm and Sander 1998; Gebhardt 1999; Brunsdon et al. 2002). There are, however, some related antecedents in image analysis (Woodcock and Strahler 1987; Musick and Grover 1991), and landscape analysis (Baker and Cai 1992; LaGro 1991; Mead et al. 1981; Murphy 1985; Perera and Baldwin 2000; Riitters and Wickham 1995; Riitters et al. 1997). However, in such work, local measures are used primarily as smoothing or low pass filtering devices aimed at reducing or removing local variation with the intention of making the data values more homogeneous (Moore 2000) and thus the use of global measures more appropriate.

Since categorical data is often encountered in environmental and ecological data, the main purpose of this paper is to develop some local measures of spatial association for such data. We refer to such measures as local indicators for categorical data (LICD). In the next section, we examine a number of issues related to the development of LICDs and propose two such measures. At present, these measures are envisaged as operating in an exploratory spatial data analysis (ESDA) role, in the spirit

of a geographical analysis machine (GAM) (Openshaw et al. 1987; Fotheringham and Zhan 1996), with the primary aim of addressing the second purpose of Getis and Ord (1996) given above. The use of these LICDs is demonstrated in Section 3 using an empirical example. This illustration raises a number of additional considerations, and one of these, the effect of significant global spatial dependence in the data values, is examined more fully in Section 4. Here we present an *ad hoc* procedure for dealing with such situations. The paper concludes with a discussion of a number of open problems together with some suggestions for further work.

2 Local indicators for categorical data (LICDs)

In general, spatial categorical data with k classes can be represented as a k-colour map ($k \geq 2$). We will limit our attention to binary maps ($k = 2$; black/white). We do this, in part, because, when k is large and the number of observations in the subregion is small, it is unlikely that any meaningful questions can be posed or answered using statistical methods. This is even the case when $k = 2$ if the probability of one of the classes is small. Also even when $k > 2$, a single type may be of particular interest; e.g., the binary model of habitat (suitable/unsuitable) in island-biogeography theory (Gustafson 1998, p. 150) or error/non-error in remotely sensed data (Congalton 1988). When $k > 2$, k-colour maps can be converted to binary maps in two ways; a single colour may be tested for its relationship with all other colours or any partition of the k colours into two groups may be tested against each other. To further simplify the presentation, we confine our attention to gridded (raster) lattice data.

In some respects, categorical data is less straightforward than its quantitative counterpart. Following Li and Reynolds (1993, 1994, 1995) and Gustafson (1998), it is possible to identify two sets of characteristics of categorical maps: *composition* which relates to aspatial characteristics of the different classes and *configuration* which refers to characteristics of the spatial distribution of the classes.

For binary data, global composition can be measured by the number of black (or white) cells in the map. Let p_b be the proportion of black cells in the entire study region. Under the assumption of no global spatial dependence (i.e., the black cells are distributed at random throughout the study region), the probability of a cell being black is equal for every cell and the colour in one cell is independent of the colours in all other cells. Then the probability ($\Pr(X = x)$) of finding x black cells in a subregion of r cells is given by the binomial distribution (Turner et al. 1989)

$$\Pr(X = x) = \binom{r}{x} p_b^x (1 - p_b)^{r-x} \tag{1}$$

Using Eq. (1) we can test if there is significant presence or absence of black cells in a specified subregion. We count the number of black cells X in $(m \times m)$ $(m = 3;\ 5;\ 7; \ldots)$ windows centred on each cell in the map and evaluate the probability of finding x using Eq. (1). Cells for which $\Pr(X \geq x) < p$ or $\Pr(X \leq x) < p$ represent those for which there

is significant presence or absence, respectively, at the p significance level.[1] Note that the window will be truncated for those cells around the edge of the study region and allowance must be made for this. This measure, which we refer to as the *local composition* for a given cell i, parallels directly the approach taken by the Getis family of statistics for quantitative data (Getis and Ord 1992, 1996; Ord and Getis 1995, 2001).

Numerous measures of configuration have been proposed in landscape ecology. Many of these, including various contagion and fragmentation indices designed to measure the degree of interspersion of patch types, incorporate join count information. Given that global join count statistics have been employed extensively, especially in geography, as a measure of the nature and extent of spatial autocorrelation in nominal data, it seems reasonable to consider local versions.

How might we use join counts locally? We could express the number of black/white (or black/black or white/white) joins in a subregion as a proportion of all joins in that subregion, and then test if this proportion differs significantly from the corresponding proportion in the entire study region. Providing that both the number of cells in the subregion and p_b were sufficiently large, a one-sample difference-of-proportions test could be employed. However, we do not feel that this would be an appropriate approach. This is because, as landscape ecologists have recognized in both theoretical (Gustafson 1998; Haines-Young and Chopping 1996; Lavorel et al. 1993) and empirical (Hulshoff 1995) studies, configuration is not independent of composition. Thus, this approach would only appear to be valid when the composition of the subregion does not differ significantly from that of the entire study region. Whenever the local composition differs significantly from the global one, we should not be surprised to find that the local configuration also differs significantly from the global one. In view of this, we suggest that the appropriate hypothesis for examining local configuration should be conditional in form. Thus, we might ask, given the number of black cells (composition) in a subregion, do the number of joins of a specified type differ from what would be expected if the black cells were located by chance in the subregion? In order to evaluate such an hypothesis, we need to know the sampling distributions of the join counts. It appears that the normal approximation is only reasonable for the join-counts when none of n, np_b, or $n(1-p_b)$ is small, where n is the number of cells and p_b is the proportion of them that are black (Upton and Fingleton 1985, p. 163). In this context, small values can be considered to be $n < 30$, p_b or $(1-p_b) < 0.2$ (Cliff and Ord 1981, Chapt. 2). This would include all 3×3 and 5×5 windows, regardless of the number of black cells, those 7×7 windows when $np_b \leq 9$ or $n(1-p_b) \geq 40$, those 9×9 windows when $np_b \leq 16$ or $n(1-p_b) \geq 65$, and so on. In such situations, the counts can either be enumerated completely or approximated by a random sample of all possible outcomes. We have enumerated all of these cases for

[1] Strictly speaking, since the population is not infinite, the hypergeometric distribution should be used instead of the binomial distribution. However, the binomial usually provides an adequate approximation of the hypergeometric when $r < 0.1n$, where n is the number of observations. (Johnson and Kotz 1969, p. 148).

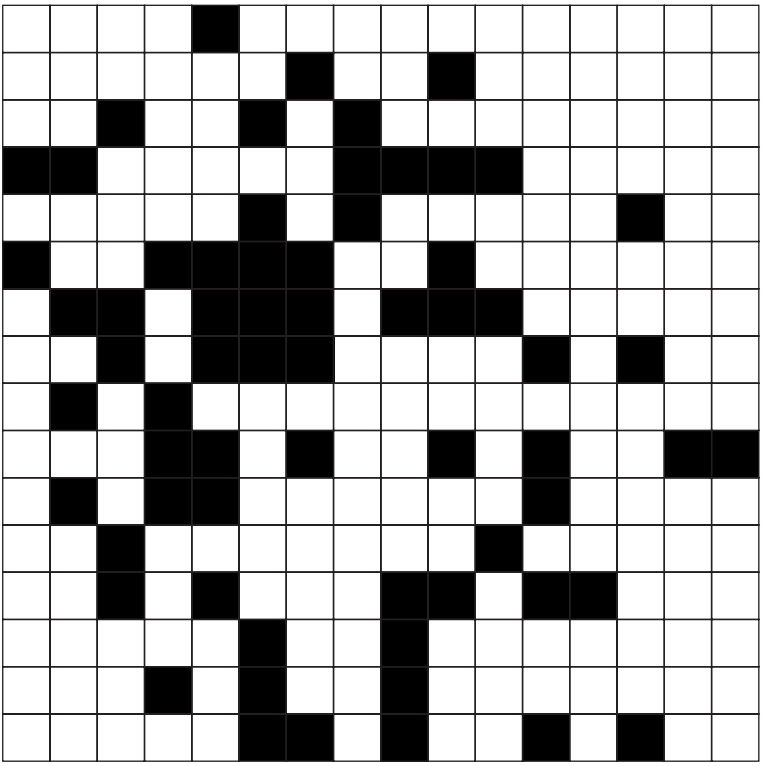

Fig. 1. Location map of *Atriplex hymenelytra* (re-drawn from Upton and Fingleton, 1985, Fig. 3.3)

windows of sizes up to 7×7. Thus, we are able to test the conditional hypothesis.[2] We refer to these local join counts as measures of *local configuration*.

3 Illustration

To illustrate the use of the measures of local composition and local configuration described in the previous section, consider Fig. 1 which is re-drawn from Fig. 3.3 in Upton and Fingleton (1985). In this figure, the cells coded black/white correspond to quadrats where the perennial shrub *Atriplex hymenelytra* is present/absent in a sample area in Death Valley, California. p_b is equal to $\frac{65}{256} = 0.254$. Using a global join count, Upton and Fingleton (1985, Example 3.2, pp. 159–160) find there is no reason to reject the null hypothesis of the random distribution of plants in the study region. To explore local composition, we count the number of black cells X in

[2] This data is available from the author. Note also that Tinkler (1977) has computed the means and standard deviations of join counts, for non-free sampling, for all square grid lattices in the size range 2×3 to 16×16.

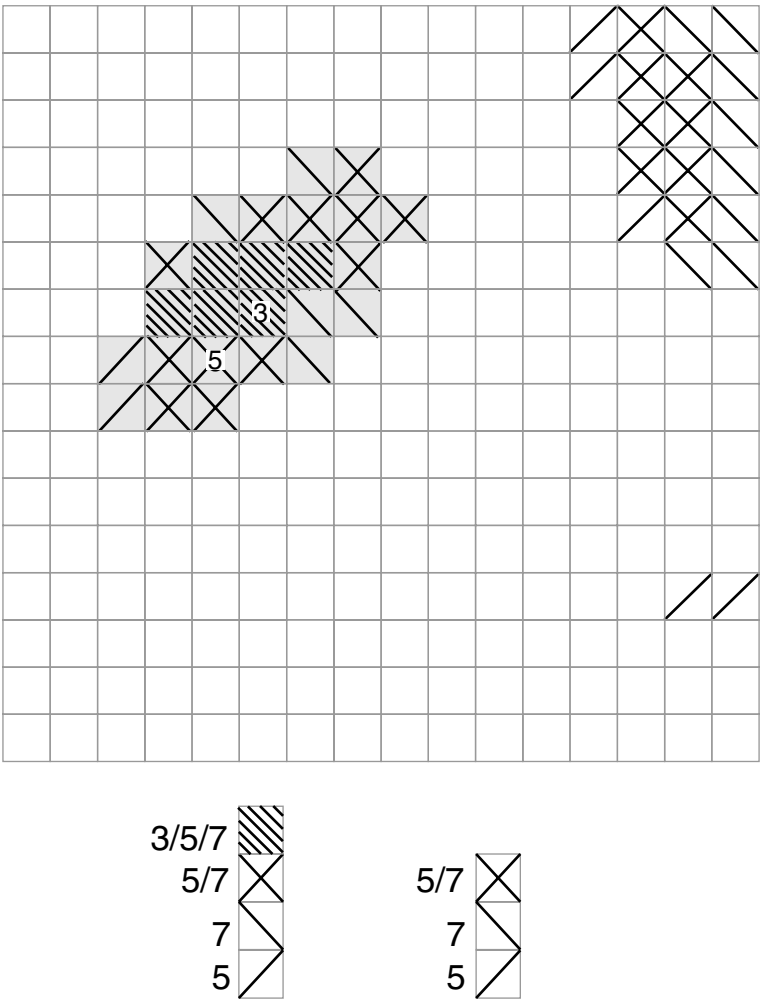

Fig. 2. Location of cells for which the number of black cells x in a $(m \times m)$ window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation. Cells for which $\Pr(X \geq x) < 0.05$ are indicated by underlying grey shading. (3, 5 indicates that the probability is significant for a 3×3 and a 5×5 window, respectively, after applying the Sidak correction for multiple tests)

$(m \times m)$ $(m = 3, 5, 7)$ windows centred on each cell in the map and evaluate the probability of finding x using Eq. (1). Note that the window is truncated for those cells around the edge of the study region. Figure 2 shows those cells for which $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$. Following the lead of Ord and Getis (2001, p. 423), we use this conventional single-test cutoff value because we are employing the LICD in an ESDA role. However, it is acknowledged that a statistical evaluation of the value for any given cell will suffer from the effects of correlation between counts in overlapping windows and multiple testing.

Figure 2 suggests that there are clear variations in the spatial distribution of the shrub within the study region. In particular, there is a zone, aligned north-east to south-west, in the north-west quadrant, where there is a marked presence of black cells. This zone consists of a core of six cells where there is a marked presence at all three window sizes, surrounded by a further nineteen cells where there is marked presence for at least one window size.[3] In contrast, there is a zone of nineteen cells in the north-east corner where there is a marked absence of black cells for the two larger window sizes. There are also a pair of cells along the eastern edge of the study region with a marked absence of black cells for the 5×5 window.[4]

To explore local configuration, we count the number of black/black (b/b), white/white (w/w), and black/white (b/w) joins in $(m \times m)$ $(m = 3, 5, 7)$ windows centred on each cell in the map, using the rook's count. If the black cells cluster in the window, this will be reflected in an abundance of b/b and w/w joins and a deficit of b/w joins. Conversely, dispersion of black cells results in an abundance of b/w joins and a deficit of b/b and w/w ones. Using the enumerations described in Sect. 2, we identify those cells for which the probability of at least (or at most) the observed number of any of the three join counts is <0.05 under the assumption of no spatial dependence. Depending on the magnitudes of the counts, such cells are considered to represent the foci of local clustering or dispersion (see Fig. 3). As Fig. 3 shows, for the three window sizes examined, this procedure identifies 10 cells, all but one of which is indicative of local clustering of black cells. All of the cells exhibiting local clustering are located either within or adjacent to the zone of marked presence of black cells identified by the local composition measure.

The two LICD measures may be combined to create a joint classification with eight potential classes, although only five of these are realized for the data in Fig. 1 (see Fig. 4). There are only four cells in Fig. 4 which display both distinctive local composition and local configuration and in each case these are where larger numbers of black cells cluster in space. Given that these four cells are part of a single cluster in Fig. 4, it is conceivable that they

---

[3] If we apply the Sidák correction, $\alpha^* = 1 - (1 - \alpha)^{\frac{1}{n}}$, where α^* is the adjusted significance level, we find that there are two cells, one each for a (3×3) and a (5×5) window (see Fig. 2), where we can be confident that the clustering of black cells in the subregion centred on the cell is statistically significant.

[4] The cells with "significantly high" black counts form single clumps of sizes 6, 20, and 23, for window sizes 3×3, 5×5, and 7×7, respectively. The cells with "significantly low" black counts form two clumps of size 11 and 2 for the 5×5 window and a single clump of size 16 for the 7×7 window. Here a clump is defined using edge contiguity (i.e., rook's count or von Neumann neighbourhood). Based on estimates from 10,000 simulated random patterns with the same proportion black ($p_b = 0.254$), the probability of getting 6 or more significant cells for a 3×3 window in a random pattern is $\simeq 0.078$ but the probability of clump of size 6 or greater $\simeq 0.022$. The corresponding probabilities for the observed values for a 5×5 and a 7×7 window are $\simeq 0.0031$ and $\simeq 0.0017$, and $\simeq 0.0070$ and $\simeq 0.0085$, respectively. For the significantly low counts, the probability of finding 13 or more significant cells for window size 5×5 is $\simeq 0.0110$, while the probabilities of clumps of size 2 or greater, and 11 or greater are $\simeq 0.5309$ and $\simeq 0.0190$, respectively. The corresponding probabilities for the observed values for a 7×7 window are $\simeq 0.0224$ and $\simeq 0.0227$.

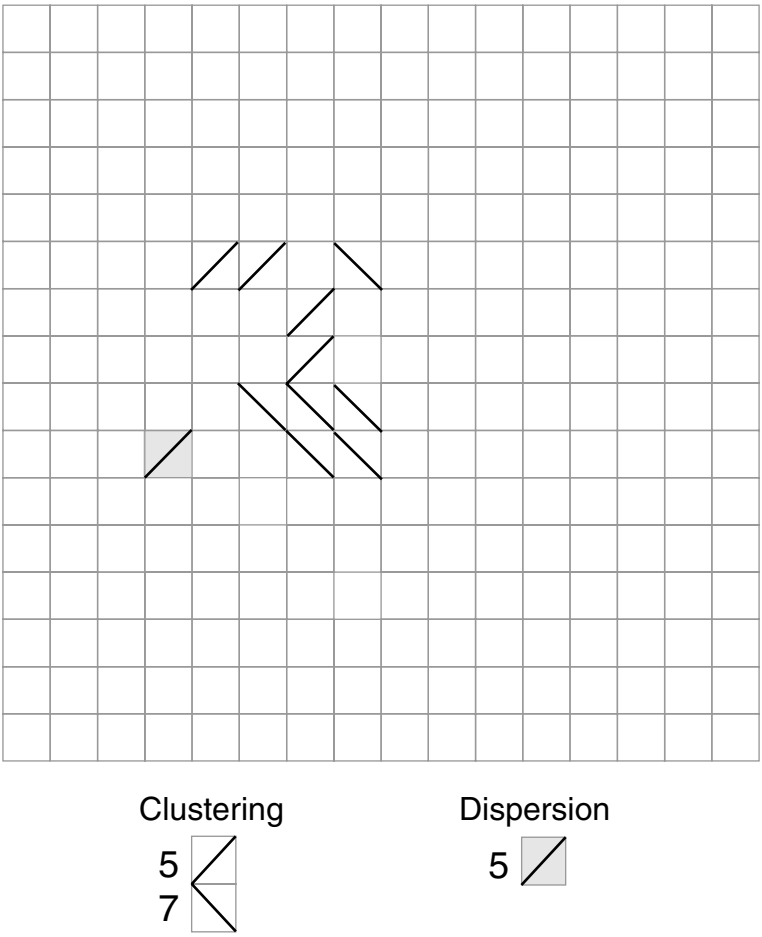

Fig. 3. Location of cells for which the probability of the observed number or greater (the observed number or less) of any of the three join counts (b/b, w/w, b/w) is < 0.05 under the null hypothesis of no global spatial autocorrelation

represent locations which are particularly favourable to the presence of *Atriplex hymenelytra*.

4 Dealing with global spatial dependence

The pattern of *Atriplex hymenelytra* analyzed in the previous section did not display significant global spatial autocorrelation. However, when there is significant global spatial autocorrelation, we can anticipate that local statistics will be too liberal, identifying excessively locations with "significant" local spatial dependence. This will certainly be the case for the local measure of composition in this paper. However, the behaviour of the local measure of configuration will not be affected since the significance of this measure is evaluated conditionally on the composition of the window used to

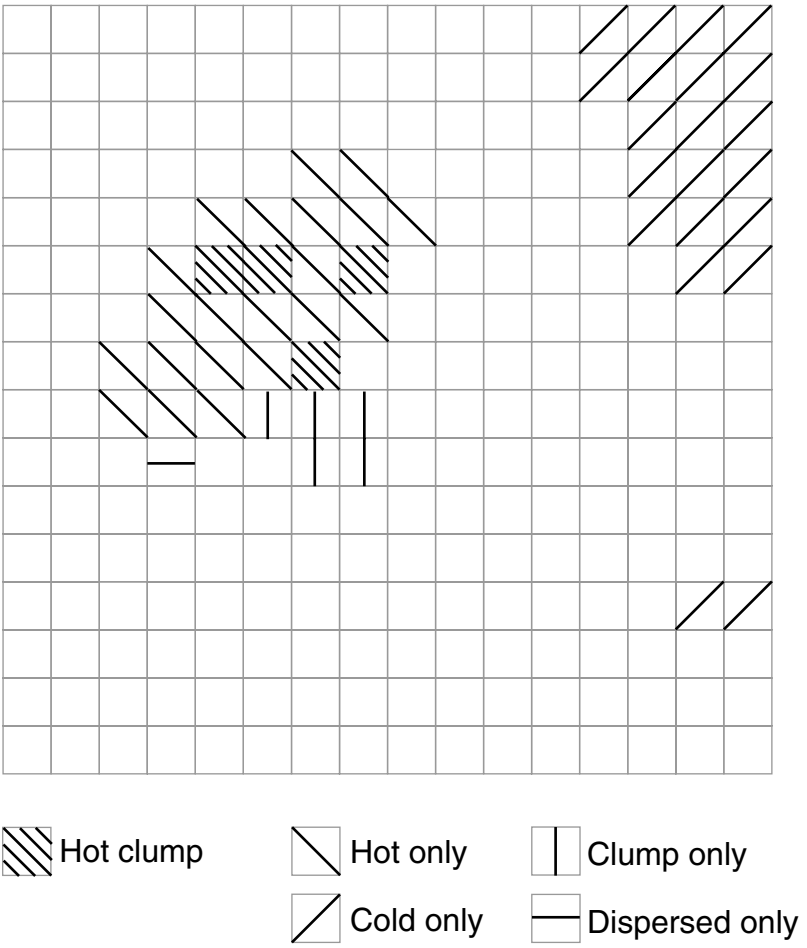

Fig. 4. Location of cells for which the observed values of local composition and/or local configuration are significant under the null hypothesis of no global spatial autocorrelation ($\alpha = 0.05$)

compute the measure. Hence, in this section, we concentrate on the local measure of composition and propose an *ad hoc* procedure for accounting for the effects of global spatial autocorrelation.

Currently, there are three general strategies for attempting to deal with the effects of global spatial autocorrelation on local statistics. The first of these is to attempt to remove global spatial autocorrelation before undertaking local analysis (Griffith and Layne 1999). This can be done by fitting a global model, computing residuals, and analyzing the residuals rather than the original data values. However, this will likely be viable only when the global spatial dependence takes the form of a trend surface or some other simple form. Its success is also dependent on the global model being specified correctly. In practice, this approach would also have the effect of transforming the original binary data into a non-categorical form. In

addition, edge effects are difficult to handle with measures derived from counts.

A variant on this approach is to obtain the conditional distribution of the local statistic for a specified global process. For example, Tiefelsdorf (1998, 2000) presents a method for deriving the conditional distribution of I_i when the global process is assumed to be either a Gaussian autoregressive or a Gaussian moving average spatial process. The conditional distribution measures the stochastic variation around the global process and is thus an indicator of local heterogeneities in the global process.

A second strategy is to retain the original data values but to evaluate the local measures using some form of restricted randomization, rather than the complete or conditional forms of randomization which are invalid under global spatial autocorrelation. For example, Fortin and Jacquez (2000) suggest performing randomizations that maintain the same degree of global spatial autocorrelation. Although they do not specify a procedure for doing this, one possibility might be the conditional pixel swapping procedure described by Liebisch et al. (2002). However, it is anticipated that there might be situations where it would be difficult to obtain sufficient independent realizations with the same global spatial autocorrelation.

The third strategy is to attempt to partition the study area into relatively homogeneous subareas. This approach is based on the assumption that any local anomalies will be at a different spatial scale to the global spatial autocorrelation. Ord and Getis (2001) adopt this strategy in the development of their test statistic, O_i, based on their G statistics. It seems particularly appropriate when we may not have any clear idea about the number, location, shape or size of the local anomalies. Our approach uses the same strategy and also shares much in common with that used by Rogerson (2002). The approach is as follows.

First, test for global spatial autocorrelation using global join-count statistics. If these indicate that there is no significant global spatial autocorrelation, apply the LICDs as illustrated in Sect. 3.[5] However, if the global join-counts reveal significant global spatial autocorrelation, perform the following steps.

1. Use equation 1 to identify cells with excessively high and low counts for a specified window size and nominal significance level α, under the assumption of no global spatial autocorrelation (henceforth referred to as "significant cells"). Potentially, significant cells may be of two kinds; those that occur simply because they are in subareas where the proportion of black cells is much greater/smaller than the global proportion (i.e., they are artifacts of the global spatial autocorrelation) and those that represent real local anomalies (independent of the existence of global spatial autocorrelation). Regardless, all cells in the pattern now fall into one of three classes, significant cells with high black counts, significant cells with low black counts, and cells with typical (expected) black counts.
2. Examine the number and spatial distribution of the two types of significant cells. Because of the multiple testing involved, even in a

[5] If the window size is sufficiently large, the binomial test may be replaced with a one-sample difference of proportions test.

pattern with no global spatial autocorrelation (henceforth referred to as a "random pattern"), we might expect to find a total of approximately $n\alpha$ significant cells. While it is impossible to say anything about the spatial distribution of such cells, this knowledge can be obtained by simulating random patterns with the same number of black cells as the pattern being analyzed.[6]

3. Identify clumps of x significant cells for which the probability of observing a clump of size at least x cells is <0.05 in a random pattern (henceforth referred to as "significant clumps").
4. Count the number of black cells in each significant clump and in the remainder of the study area not covered by significant clumps and compute the probability of finding a black cell in each of these subregions.
5. Evaluate the local composition using the additive binomial rather than the simple binomial in Eq. (1). The additive binomial is given by

$$\Pr(X = x) = \sum_{i=1}^{m} a_i \binom{r}{x} p_{b_i}^{x} (1 - p_{b_i})^{r-x} \tag{2}$$

where r and x are the same as in Eq. (1), m is the number of subregions covering the window, a_i is the proportion of the window covered by subregion i, and p_{b_i} is the proportion of black cells in subregion i.

We illustrate this procedure using two patterns, one in which there is global autocorrelation but no local anomalies, the other with global spatial autocorrelation and an anomalous subregion. The former situation is shown in Fig. 5. While this pattern has the same number of black cells as the pattern in Fig. 1, the black cells are located according to an inhomogeneous planar Poisson process in which the probability of receiving a black cell declines with distance from the left edge of the study region. As expected, global join-count statistics identify significant positive global spatial autocorrelation associated with a clustering of black cells ($z(b/b) = 2.473$; $z(w/w) = 1.912$; $z(b/w) = -2.327$). Thus, we undertake the five steps described above.

Figures 6, 7, and 8, show the significant cells, as defined in step 1, for 3×3, 5×5, and 7×7 windows, respectively, using a nominal significance level of $\alpha = 0.05$. For the 3×3 windows (see Fig. 6), step 1 identifies 19 cells with significantly high black counts. To implement step 2, we ran 10,000 simulations of random patterns with the same proportion of black cells as the pattern in Fig. 5 (hereafter referred to as the "random simulations"). No occurrences with a total of 19 or more significant cells were found in these random simulations. Further, the 19 significant cells form two clumps of size five and fourteen. In the random simulations, given that there is a at least one significant cell[7], the probabilities of finding a clump of 5 or more, or 14 or more significant cells are 0.044185 and 0.000083, respectively. Following step 3, we thus conclude that the 19 cells represent two significant clumps. The

[6] However, Ahuja and Schacter (1983, Chap. 2) do report some limited results for the number of clumps in a random pattern on a 100×100 square grid.

[7] In the random simulations, for a 3×3 window, the probability of observing no significant cells is $\simeq 0.2432$. For a 5×5 window the probability of finding no significant cells with excessively high and low black counts is $\simeq 0.3371$ and $\simeq 0.3884$, respectively. The corresponding probabilities for a 7×7 window are $\simeq 0.5128$ and $\simeq 0.4653$, respectively.

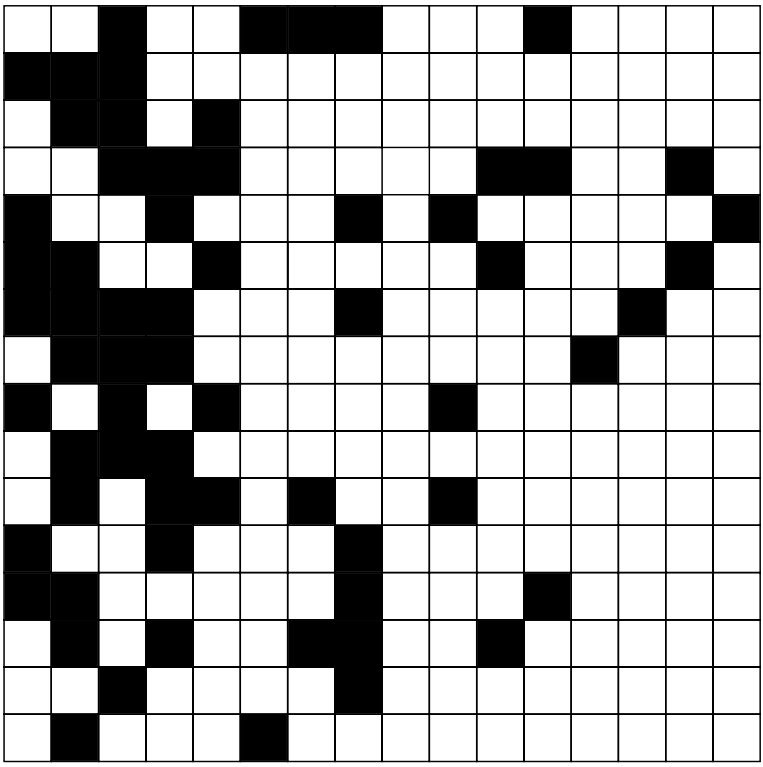

Fig. 5. Pattern created by an inhomogeneous planar Poisson process (probability (black) = 0.254)

original pattern has now been subdivided into three subregions, two clumps (A and B) of significantly high black counts, and the remainder (C). In accordance with step 4, we calculate the probability of a black cell in regions A, B, and C. These probabilities are 0.800, 0.786, and 0.211, respectively (recall the global probability of finding a black cell is 0.254). Finally, in step 5, we use these probabilities in Eq. (2) to evaluate the local composition. The results indicate that none of the counts for the cells in the pattern are significant at a nominal significance level of $\alpha = 0.05$.

For the 5×5 windows (see Fig. 7), step 1 identifies 38 cells with significantly high black counts and 20 with significantly low counts. In the random simulations, no occurrence of 38 or more cells with significantly high counts was found, while the probability of finding 20 or more cells with significantly low counts is 0.001300. Both the cells with significantly high black counts and those with significantly low counts form single clumps. No clump of 38 significantly high counts was found in the random simulations, while the probability of finding a clump of significantly low counts of at least 20 cells is 0.001298. We conclude that each of these may be considered a significant clump, so that the pattern is now subdivided into three regions, clump A, clump B, and the remainder C. The probability of finding a black cell in each of these three regions is 0.553, 0, and 0.222, respectively. Using

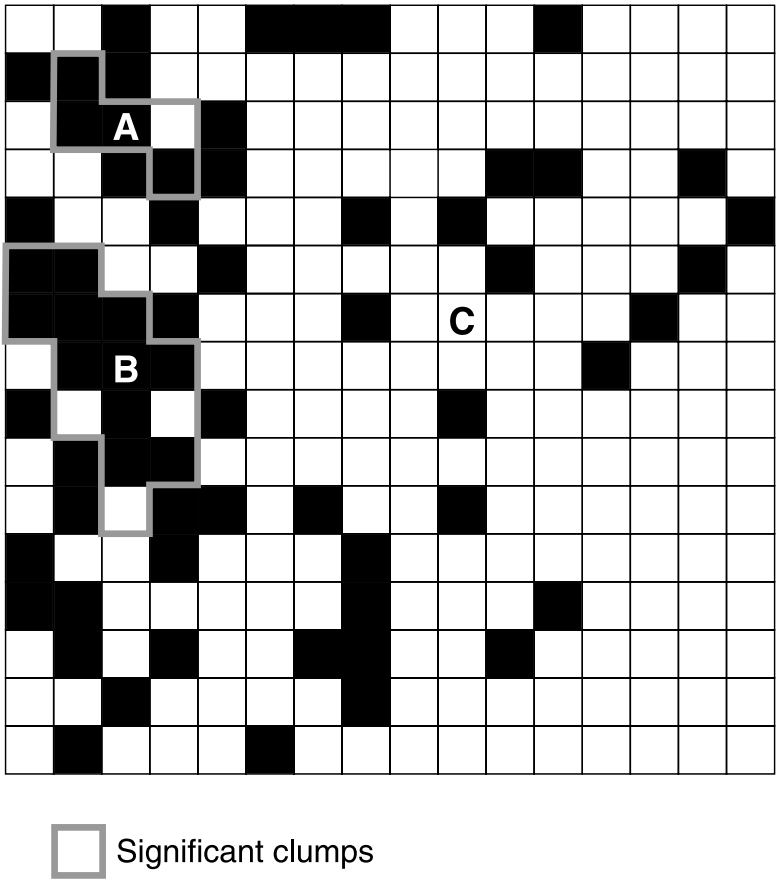

Fig. 6. Location of cells for which the number of black cells x in a 3×3 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

these values in Eq. (2) reveals that none of the counts can be considered either significantly high or low.

For the 7×7 windows (see Fig. 8), there are 49 and 40 cells with significantly high and low black counts, respectively. The 49 significantly high black counts form a single clump. No occurrence of this kind was found in the random simulations and thus this is considered a significant clump (clump A). The 40 significantly low counts form a clump of 38 cells and two isolated cells. No occurrence of 40 or more significant cells (and thus a clump of size 40 or more) occurred in the random simulations. However, the probability of at least an isolated significant cell with a low count is 0.469075. Consequently, clump B is considered significant while cells C and D are not. This results in three sub areas, clumps A and B, and the remainder E, with the associated probabilities of a black cell being 0.633, 0.026, and 0.195, respectively. Using these probabilities none of the cells are significant. Note that cells C and D with counts of 2 black cells, which were significant using

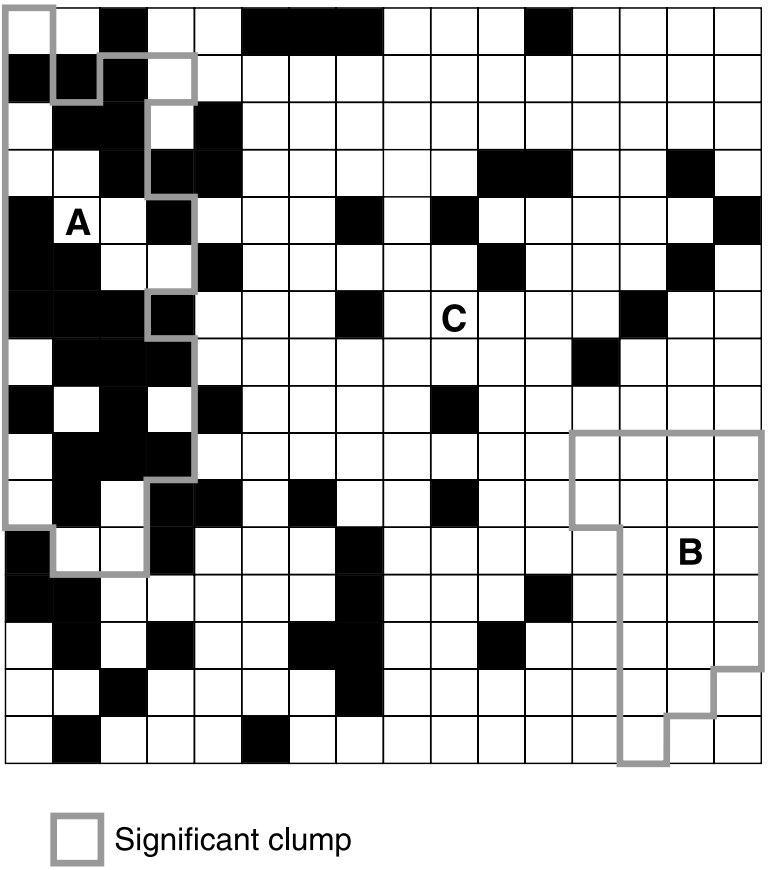

Fig. 7. Location of cells for which the number of black cells x in a 5×5 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

the global probability ($p_b = 0.254$), are no longer significant because of the lower values of p_b ($= 0.195$ or 0.026) in their vicinity when the pattern is partitioned into three subregions. Thus, our procedure indicates that the pattern in Fig. 5 shows no local anomalies at any of the three window sizes examined, which is consistent with our knowledge of how the pattern was generated.

In the second example, we adjust the pattern in Fig. 5 to create an anomaly. We do this by moving six black cells from location A in the higher density (left hand side) to location B in the lower density (right hand side) as shown in Fig. 9. This has the effect of creating a small anomalous region of black cells in the vicinity of B. It also reduces the clustering of black cells so that the join count statistics are now $z(b/b) = 1.9976$; $z(w/w) = 1.4948$; $z(b/w) = -1.8514$, although the first of these is still significant at $\alpha = 0.05$.

Figures 10, 11, and 12, show the significant cells, as defined in step 1 above, for 3×3, 5×5, and 7×7 windows, respectively, using a nominal significance level of $\alpha = 0.05$. For the 3×3 windows (see Fig. 10), step 1 identifies

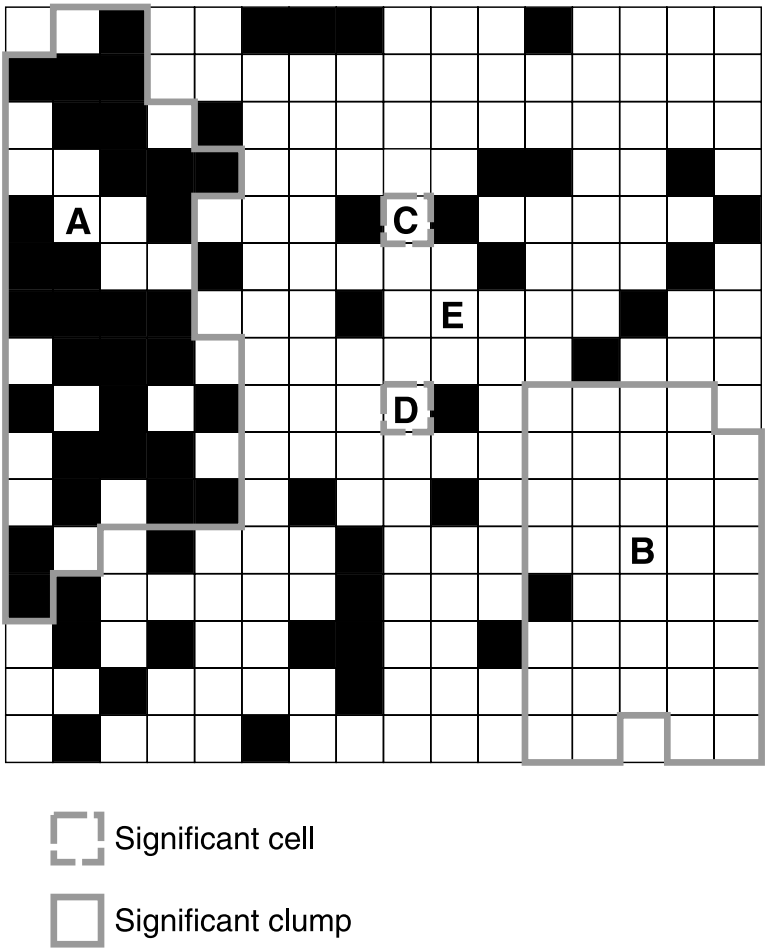

Fig. 8. Location of cells for which the number of black cells x in a 7×7 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

two clumps of cells with significantly high black counts, A, in the higher density area, with 14 cells and B, focused on the anomaly, with two cells. In the random simulations, given that there is at least one significant cell, the probabilities of finding a clump of 2 or more, or 14 or more significant cells are 0.445019 and 0.000083, respectively. In view of this, we retain only A as a significant clump, so that the pattern is divided into two subregions, A with $p_b = 0.786$ and the remainder with $p_b = 0.223$. Using these probabilities in Eq. (2), reveals that only the two cells in region B are now significant.

For the 5×5 windows (see Fig. 11), step 1 identifies a single clump of 23 cells with significantly high black counts and a single clump of 9 cells with significantly low counts. In the random simulations, the probabilities of finding these occurrences were 0.000797 and 0.045800, respectively and so we conclude that each of these may be considered a significant clump. Thus, the

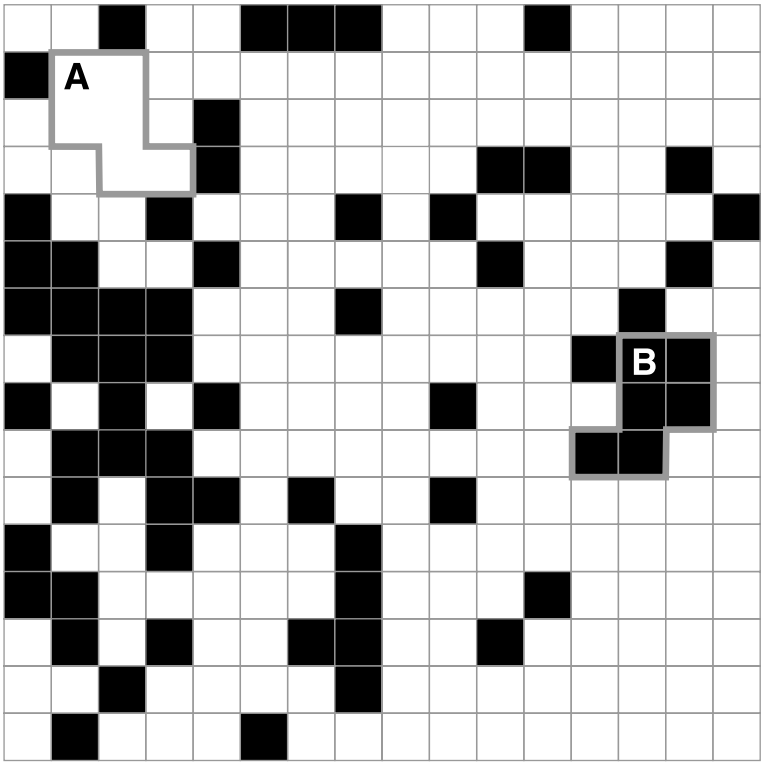

Fig. 9. Pattern created by a modified inhomogeneous planar Poisson process (probability (*black*) = 0.254)

pattern is subdivided into three regions, clump A, clump B, and the remainder C. The probability of finding a black cell in each of these three regions is 0.609, 0, and 0.228, respectively. Using these values in Eq. (2) reveals that none of the counts can be considered either significantly high or low.

For the 7×7 windows (see Fig. 12), there is a single clump of 30 cells with significantly high counts and three clumps, of sizes 19, 1 and 1, with significantly low black counts. Since the associated probabilities for these events in the random simulations are 0.001737, 0, 012433, and 0.469075, respectively, only clumps A and B are considered significant clumps. This results in three subareas, clumps A and B, and the remainder E, with the associated probabilities of a black cell being 0.633, 0.053, and 0.217, respectively. Using these probabilities, none of the cells are significant.

Collectively, the results from the three window sizes suggest that our procedure is capable of identifying the local anomaly at the appropriate scale (i.e., only for the 3×3 window).

5 Conclusions

This paper has described a procedure for extending local statistics to categorical data. Our approach is based on the notion that there are two

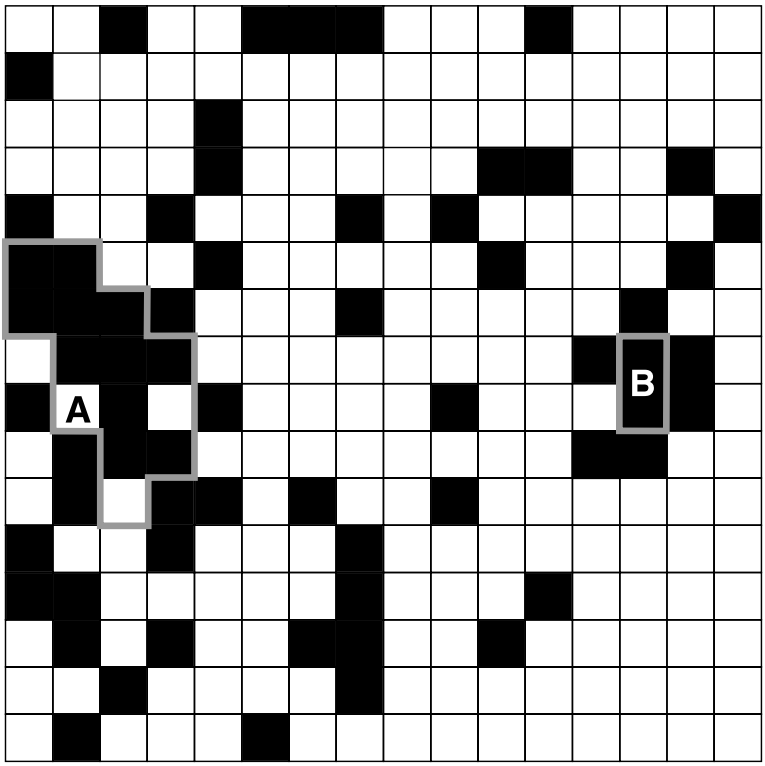

Fig. 10. Location of cells for which the number of black cells x in a 3×3 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

fundamental characteristics of categorical data, composition and configuration. Further, it is argued that, when considered locally, the latter should be measured conditionally with respect to the former. The approach was illustrated using a small, empirical data set and an ad hoc procedure was developed to deal with the impact of global spatial autocorrelation on the local statistics. However, a number of issues remain for further consideration. These are of two kinds, those related to the specific ad hoc test for dealing with global spatial autocorrelation and those related to tests for categorical spatial data in general. We consider the former first.

The implementation of step 1 for the ad hoc procedure requires the selection of a nominal significance level, which, in turn, influences the subsequent steps of the test. In this paper, a value of 0.05 was selected and so it would be interesting to examine how the choice of other values relates to the power of the test. Such investigations should also consider the influence of the size of the study area.

A more subtle issue arises in step 5 where the evaluation requires the specification of the parameters for the additive binomial equation. In order to implement this step, these are estimated using the observed values. However, since the empirical pattern results from a stochastic process, the

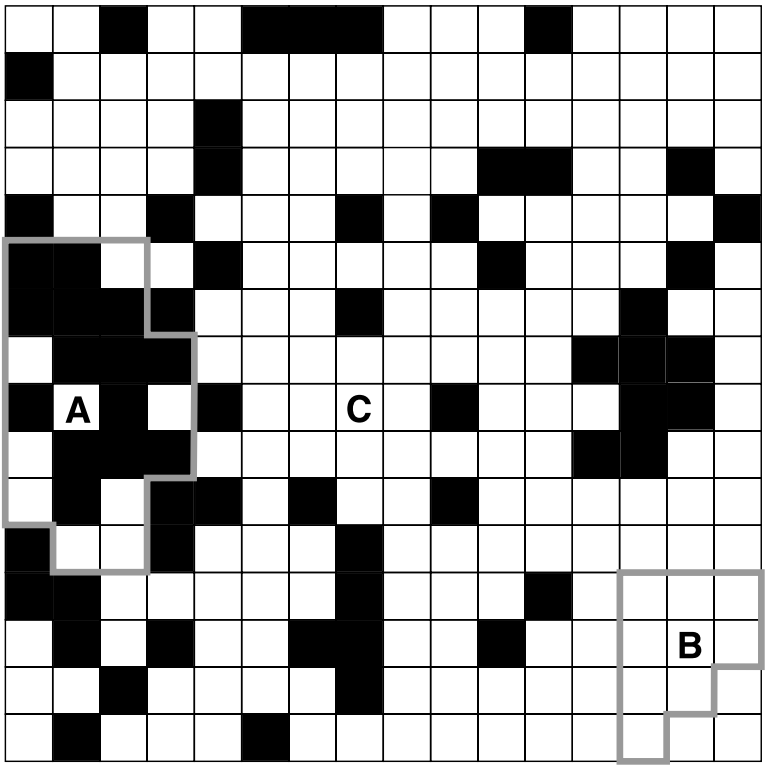

Fig. 11. Location of cells for which the number of black cells x in a 5×5 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

true values are actually variable rather than fixed quantities. Thus, any inference made in step 5 should be considered conditional on the observed values.

With regard to more general issues, first of all, we have only considered binary data on regular lattices. In theory, there is no obstacle to extending the approach to irregular lattices, although "windows" will need to be defined in terms of spatial adjacency lags or distance, which means that they will be of variable size. Further, if they are defined in terms of lags, first order windows will typically be smaller in size than the smallest regular window (3×3) used in this paper, containing on average about six cells. This will inhibit examining both local composition and configuration at this scale.[8]

Consideration of multinomial data means that we need to replace the binomial model in Eq. (1) with the multinomial one. This is probably not worthwhile for small data sets since the increasing number of categories makes it less likely to identify distinctive events. However, it does seem

[8] For example, for a 3×3 window in a regular lattice, of the 30 possible outcomes for local configuration (0-9 black cells, b/b, w/w, b/w counts) only 4 yield significant events (i.e., $p < 0.05$).

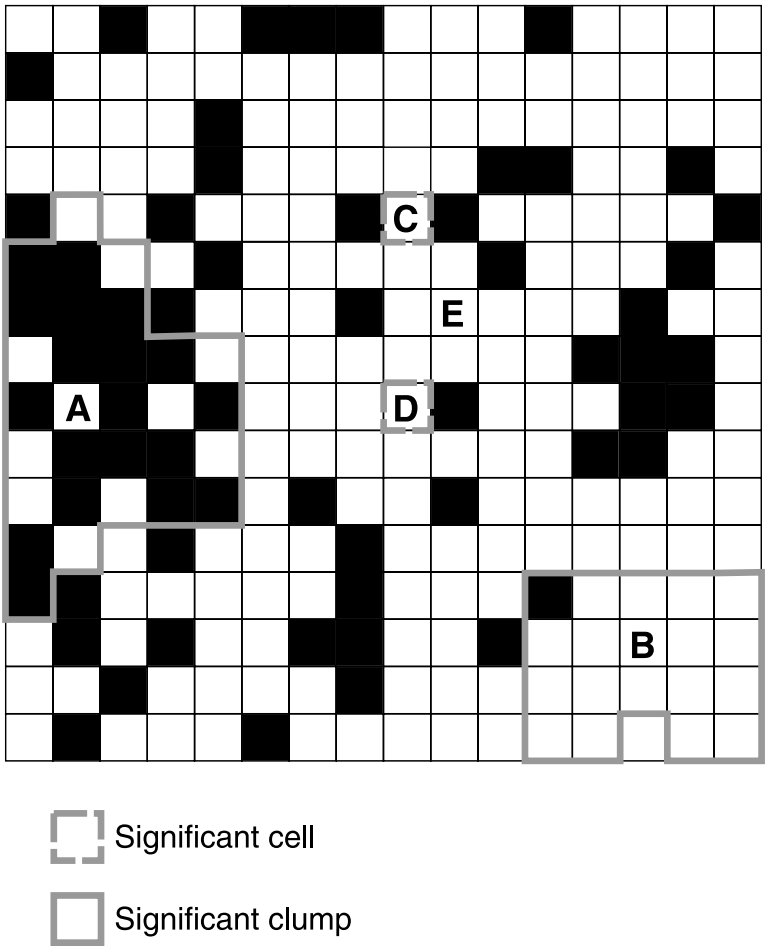

Fig. 12. Location of cells for which the number of black cells x in a 7×7 window centred on the cell have probabilities $\Pr(X \leq x)$ or $\Pr(X \geq x) < 0.05$ under the null hypothesis of no global spatial autocorrelation

appropriate if the approach is applied to large data sets such as remotely sensed imagery (Griffith 2002). Such large data sets may also allow for the use of alternative ways of partitioning the study region into quasi-homogeneous subareas when global spatial autocorrelation is present. Possible candidates include multifractal analysis (Milne 1991) or local semivariance analysis (Aldstadt and Getis 2002).

Although there seems little choice in terms of how local composition is measured, there are many other possibilities for measuring local connectivity besides the local join-counts used here. These include various measures of complexity developed in cartography (Monmonier 1974; Muller 1976; MacEachern 1985; Bregt and Wopereis 1990; Johnsson 1995) and measures of configuration developed in landscape ecology (Haines-Young and Chopping 1996; Frohn 1998; He et al. 2000). Indeed, the computation of

multiple measures for a single data set may provide additional insight in the context of exploratory spatial data analysis.

Consideration also needs to be given to the development of local measures for multivariate categorical data sets. However, this is currently inhibited by the lack of effective global measures for such data, although the work by Wartenberg (1985) and Lee (2001) are important steps in this direction. In the meantime, an interim solution has been proposed by Sokal et al. (1998) who suggest computing local measures for each variable, ranking these, and then determining a summary rank for each data site.

References

Ahuja N, Schacter BJ (1983) *Pattern models*. John Wiley, New York

Aldstadt J, Getis A (2002) Partitioning heterogeneous spaces. *GIScience 2002 The Second International Conference on Geographic Information Science, Abstracts* 9–11

Anselin L (1995) Local indicators of spatial autocorrelation – LISA. *Geographical Analysis* 27: 93–115

Baker WL, Cai Y (1992) The r.le programs for multiscale analysis of landscape structure using the GRASS geographical information system. *Landscape Ecology* 7: 291–302

Boots B (2002) Local measures of spatial association. *Ecoscience* 9: 168–176

Bregt AK, Wopereis MCS (1990) Comparison of complexity measures for choropleth maps. *The Cartographic Journal* 27: 85–91

Brunsdon C, Fotheringham S, Charlton M (2002) Geographically weighted local statistics applied to binary data. In: Egenhofer MJ, Mark DM (eds) *Geographic Information Science: Second International Conference, GIScience 2002, Boulder CO, USA, September 2002 Proceedings* pp. 38–50. (Lecture Notes in Computer Science 2478) Springer Verlag, Berlin

Cliff AD, Ord JK (1981) *Spatial processes: models and applications*. Pion, London

Congalton RG (1988) Using spatial autocorrelation analysis to explore the errors in maps generated by remotely sensed data. *Photogrammetric Engineering and Remote Sensing* 54: 387–392

Fortin M-J, Jacquez GM (2000) Randomization tests and spatially autocorrelated data. *Bulletin of the Ecological Society of America* 81: 201–205

Fotheringham AS (1997) Trends in quantitative methods I: stressing the local. *Progress in Human Geography* 21: 88–96

Fotheringham AS (1999) Guest editorial: local modelling. *Geographical & Environmental Modelling* 3: 5–7

Fotheringham AS, Brunsdon C (1999) Local forms of spatial analysis. *Geographical Analysis* 31: 340–358

Fotheringham AS, Zhan F (1996) A comparison of three exploratory methods for cluster detection in point patterns. *Geographical Analysis* 28: 200–218

Frohn RC (1998) *Remote sensing for landscape ecology: new metric indicators for monitoring, modeling, and assessment of ecosystems*. Lewis Publishers, Boca Raton, FL

Gebhardt F (1999) Cluster tests for geographical areas with binary data. *Computational Statistics and Data Analysis* 31: 39–58

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

Getis A, Ord JK (1996) Local spatial statistics: an overview. In: Longley P, Batty M (eds) *Spatial analysis: modelling in a GIS environment* pp. 261–277. GeoInformation International, Cambridge

Griffith DA (2002) Modeling spatial dependence in high spatial resolution hyperspectral data sets. *Journal of Geographical Systems* 4: 43–51

Griffith DA, Layne LJ (1999) *A casebook for spatial statistical data analysis: a compilation of analyses of different thematic data sets*. Oxford University Press, New York

Gustafson EJ (1998) Quantifying landscape spatial pattern: What is the state of the art? *Ecosystems* 1: 143–156

Haines-Young R, Chopping M (1987) Quantifying landscape structure: a review of landscape indices and their application to forested landscapes. *Progress in Physical Geography* 20: 418–445

He HS, DeZonia BE, Mladenoff DJ (2000) An aggregation index (*AI*) to quantify spatial patterns on landscapes. *Landscape Ecology* 15: 591–601

Hulshoff RM (1995) Landscape indices describing a Dutch landscape. *Landscape Ecology* 10: 101–111

Johnson NL, Kotz S (1969) *Distributions in statistics: discrete distributions*. Houghton Mifflin, Boston

Johnsson K (1995) Fragmentation index as a region based GIS operator. *International Journal of Geographical Information Systems* 9: 211–220

LaGro J (1991) Assessing patch shape in landscape metrics. *Photogrammetric Engineering and Remote Sensing* 57: 285–293

Lavorel S, Gardner RH, O'Neill RV (1993) Analysis of patterns in hierarchically structured landscapes. *Oikos* 67: 521–528

Lee S-I (2001) Developing a bivariate spatial association measure: an integration of Pearson's *r* and Moran's *I*. *Journal of Geographical Systems* 3: 369–385

Li H, Reynolds JF (1993) A new contagion index to quantify spatial patterns of landscapes. *Landscape Ecology* 8: 155–162

Li H, Reynolds JF (1994) A simulation experiment to quantify spatial heterogeneity in categorical maps. *Ecology* 75: 2446–2455

Li H, Reynolds JF (1995) On definition and quantification of heterogeneity. *Oikos* 73: 280–284

Liebisch N, Jacquez G, Goovaerts P, Kaufmann A (2002) New methods to generate neutral images for spatial pattern recognition. In: Egenhofer MJ, Mark DM (eds) *Geographic Information Science: Second International Conference, GIScience 2002, Boulder CO, USA, September 2002 Proceedings* pp. 181–195. (Lecture Notes in Computer Science 2478) Springer Verlag, Berlin

MacEachern AM (1982) Map complexity: comparison and measurement. *The American Cartographer* 9: 31–46

Mead RA, Sharik TL, Prisley SP, Heinen JT (1981) A computerized spatial analysis system for assessing wildlife habitat from vegetation maps. *Canadian Journal of Remote Sensing* 7: 34–40

Milne BT (1991) Lessons from applying fractal models to landscape patterns. In: Turner MG, Gardner RH (eds) *Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity* pp. 199–235. Springer-Verlag Inc., New York

Monmonier MS (1974) Measures of pattern complexity for choroplethic maps. *American Cartographer* 1: 159–169

Moore K (2000) Resel filtering to aid visualisation within an exploratory data analysis system. *Journal of Geographical Systems* 2: 375–398

Muller JC (1976) Objective and subjective comparison in choroplethic mapping. *The Cartographic Journal* 13: 156–166

Murphy DL (1985) Estimating neighborhood variability with a binary comparison matrix. *Photogrammetric Engineering and Remote Sensing* 51: 667–674

Musick HB, Grover HD (1991) Image textural measures as indices of landscape patterns. In: Turner MG, Gardner RH (eds) *Quantitative methods in landscape ecology: the analysis and interpretation of landscape heterogeneity* pp. 77–103. Springer-Verlag Inc., New York

Openshaw S, Charlton ME, Wymer C, Craft AW (1987) A Mark I Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1: 359–377

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis* 27: 286–306

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science* 41: 411–432

Perera AH, Baldwin DJB (2000) Spatial patterns in the managed forest landscape of Ontario. In: Perera AH, Euler DL, Thompson ID (eds) *Ecology of a managed terrestial landscape: patterns and processes of forest landscapes in Ontario*. UBC Press, Vancouver

Riitters KH, O'Neill RV, Hunsaker CT, Wickham J, Yankee DH, Timmins SP, Jones KB, Jackson BL (1995) A factor analysis of landscape pattern and structure metrics. *Landscape Ecology* 10: 23–39

Riitters KH, Wickham DJ (1995) *A landscape atlas of the Chesapeake Bay watershed*. SERDP, Arlington, Va.

Rogerson P (2002) Change detection thresholds for remotely sensed images. *Journal of Geographical Systems* 4: 85–97

Sokal RR, Oden NL, Thomson BA (1998) Local spatial autocorrelation in biological variables. *Biological Journal of the Linnean Society* 65: 41–62

Tiefelsdorf M (1998) Some practical applications of Moran's *I*'s exact conditional distribution. *Papers of the Regional Science Assoication* 77: 101–129

Tiefelsdorf M (2000) *Modelling spatial processes: the identification and analysis of spatial relationships in regression residuals by means of Moran's I*. (Lecture Notes in Earth Sciences 87) Springer-Verlag, Berlin, Heidelberg

Tinkler KJ (1977) *Joint count statistics for 2-colour lattices up to* 16×16. Department of Geography, Brock University

Turner MG, O'Neill RV, Gardner RH, Milne BT (1989) Effects of changing scale on the analysis of landscape pattern. *Landscape Ecology* 3: 153–162

Unwin A (1996a) Exploratory spatial analysis and local statistics. *Computational Statistics* 11: 387–400

Unwin D (1996b) GIS, spatial analysis and spatial statistics. *Progress in Human Geography* 20: 540–541

Upton G, Fingleton B (1985) *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*. John Wiley & Sons, Chichester

Wartenberg D (1985) Multivariate spatial correlation: a method for exploratory geographical analysis. *Geographical Analysis* 17: 263–283

Wilhelm A, Sander M (1998) Interactive statistical analysis of dialect features. *Journal of the Royal Statistical Society: Series D (The Statistician)* 47: 445–455

Woodcock C, Strahler A (1987) The factor of scale in remote sensing. *Remote Sensing of Environment* 21: 311–322