B.4 Spatial Clustering

Jared Aldstadt

B.4.1 Introduction

Spatial clustering analysis has become common in many fields of research, and is most commonly used in epidemiology and criminology applications. Knox (1989, p.17) defines a spatial cluster as, '*a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance.*' This is a useful operational definition, but there are very few situations when phenomena are expected to be distributed randomly in space. In most cases an implicit assumption in spatial cluster analysis is that the researcher has accounted for all the factors known to influence the variable of study. This would lead to an examination of residual spatial variation in a spatial modeling exercise. Spatial clustering analysis is carried out on raw variables or rates when there are no *a priori* hypotheses regarding the process.

There are an ever increasing number of methods available for the analysis of spatial clustering. These techniques can be divided into two categories: those that are used to determine if clustering is present in the study region, and those that attempt to identify the location of clusters. The first category of tests is called global clustering techniques and these methods provide a single statistic that summarizes the spatial pattern of the region. These will be discussed in the section that follows. The second type of methodology is called local clustering. Local methods examine specific sub-regions or neighborhoods within the study to determine if that area represents a cluster of high values (a hot spot) or low values (a cold spot). These methods can be further differentiated as either focused or non-focused tests. Focused tests examine one or a small set of pre-defined foci of interest. Non-focused tests are designed to find clusters that exist throughout the entire region of analysis. Local clustering methods will be discussed in Section B.4.3. Considerations for choosing a spatial clustering method and some concluding remarks are provided in Section B.4.4.

M.M. Fischer and A. Getis (eds.), *Handbook of Applied Spatial Analysis: Software Tools, Methods and Application*s, DOI 10.1007/978-3-642-03647-7_15,

B.4.2 Global measures of spatial clustering

The methods developed to detect global clustering are also called general tests of clustering. In most cases, the null hypothesis is one of spatial randomness. These methods provide a single summary statistic which describes the degree of clustering present in the mapped pattern. The value of the statistic indicates whether the pattern is clustered, random, or dispersed. In contrast to a clustered pattern, a dispersed pattern is one where high values and low values are nearby each other more often than would be expected in a random pattern. Clustered and dispersed patterns may also be labeled positive and negative spatial autocorrelation respectively.

Areal data methods

The first set of methods deal with areal data, or the attributes of units that are mapped as polygons. These attributes are most often aggregate data such as a density or a rate per unit of population. It does not usually make sense to carry out spatial analysis with a raw count of events within a spatial unit. Much of the variation in the attribute is likely to be a function of the size of the unit or the population at risk within the unit. The use of rates may also confound cluster analysis when there is substantial variation in the size of the denominator to be used to calculate rates. Consequently, variants of general tests have been developed that account for this variation in population size and examine the spatial pattern of the excess or deficiency of events occurring in each spatial unit. These analyses are not limited to scale data, and a method that examines clustering in a map with two classes will also be discussed.

Global clustering statistics take a common form that compares the similarity of values at locations to the spatial proximity of the locations. This type of statistic is called a general cross-product statistic, and it was introduced by Mantel (1967) for computing the similarity between two matrices. The spatial proximity between each pair of locations i and j is denoted W_{ij} and entered into an n-by-n matrix called the spatial weights matrix. The spatial weights matrix is most often denoted as $\boldsymbol{W}$, and is discussed further below. The similarity of two data values x_i and x_j is denoted S_{ij} and can be entered into an n-by-n matrix that is labeled $\boldsymbol{S}$. Clustering is indicated when spatial proximity and similarity are positively related. In summation notation, the general form of the statistic is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\, S_{ij}. \tag{B.4.1}$$

Each of the techniques presented in this section are a variation of this form, with the distinguishing variant being the measure of similarity between values. Often the indices are normalized by global measures of similarity and spatial connectivity.

The spatial weights matrix defines the structure of spatial relationships in the study region. It delimits the extent of clustering that the clustering technique is able to detect. The choice of ***W***, therefore, should be considered carefully in clustering analysis. The simplest and perhaps most commonly used set of spatial weights is the binary contiguity matrix. Here, W_{ij} is equal to one if units i and j share a common boundary and zero otherwise. There are two variants of the binary contiguity matrix. The Rook case requires that neighbors share a common edge. A common vertex or point is all that is required for contiguity in the Queen case. Other binary weights matrices include a number of nearest neighbors and the complete set of neighbors with a given distance. Spatial relationships may also be defined as a function of the distance between units. Most commonly elements are defined as

$$W_{ij} = d_{ij}^{-\alpha} \tag{B.4.2}$$

where d_{ij} is the distance between units i and j *and* α is larger than zero. It should also be noted that the diagonal of the weights matrix, the values W_{ii}, are usually set to zero.

The weights matrix used in cluster analysis is often standardized so that the elements of each row sum to one (row standardization). This procedure serves to equalize the weight given each observation in the analysis with respect to its number of neighbors. The elements of this standardized matrix are calculated as

$$\widetilde{W}_{ij} = \frac{W_{ij}}{\sum_{j=1}^{N} W_{ij}}. \tag{B.4.3}$$

Standardization should not be carried out in cases when the weights have meaningful interpretation with regards to the analysis (Anselin 1988). For example, standardizing inverse distance matrices will distort the relative spatial relationships between units and cloud interpretation of the clustering index. The effects of standardization are examined and an alternative to row standardization is provided by Tiefelsdorf et al. (1999). A more complete examination of the spatial weights matrix with references to many alternative forms and several reviews is given by Getis and Aldstadt (2004).

Join-count statistic. The join count statistic is a measure of clustering for a binary classification of data. These values could be visualized as a two-category choropleth map. The two classes are usually referred to as black (*B*) and white (*W*). A join is another name for the contiguity relationship of two areas sharing a boundary. The statistic value is the number of joins of a given type. Each boundary may connect two black units (*BB*), two white units (*WW*) or one unit of each type (*BW*). Cliff and Ord (1973) define the number of *BW* joins as the general cross product statistic

$$BW = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - x_j)^2 \qquad \text{(B.4.4)}$$

where x_i equal to one corresponds to *B* and x_i equal to zero corresponds to *W*. Following from the definition of join, the weights, W_{ij} are usually restricted to a binary contiguity representation. Under a free sampling assumption, the expected number of *BW* joins in a random spatial distribution is

$$E[BW] = 2\,Jpq \qquad \text{(B.4.5)}$$

where *J* is the total number of joins. *p* is the probability that a unit is coded *B* and is often estimated as the proportion of units that are in the class *B*. *q* is the probability that a unit is coded *W* and is equal to *one minus p*. The number of joins may be calculated from the binary contiguity weights as

$$J = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}. \qquad \text{(B.4.6)}$$

If the classes are clustered together, there would be fewer observed *BW* joins than expected. Likewise, if the pattern is dispersed or similar to a checkerboard pattern, there would be more *BW* joins than expected in a spatially random pattern. The variance of the *BW* statistic under both free and non-free sampling are derived in Cliff and Ord (1973) along with an extension to the case when there are more than two classes.

Moran's I. Moran's *I* is a well known test for spatial autocorrelation (Moran 1950). The index is similar to covariance and correlation statistics. The measure of similarity between values at two locations *i* and *j* is the product of the deviation between the value at each location and the estimate of the global mean $\bar{x}$. This product is weighted by the spatial proximity of the two locations, and the sum of the resulting values for all pairs of locations is the spatial autocovariance. The standardized index is given as

$$I = \frac{N}{S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad i \neq j \tag{B.4.7}$$

where

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}. \tag{B.4.8}$$

The expected value for a spatially random distribution is minus one over (n–1). This quantity tends towards zero as the sample size increases. Values greater than this indicate clustering of units with high and or low values. Values that are smaller than the expected value indicate negative association between proximate locations. Unlike the Pearson's correlation coefficient, Moran's I is not bounded between negative one and one, but usually falls within this interval (Bailey and Gatrell 1995). A correlogram displays the Moran's I value calculated for a number of increasing distances. The distances are most often mutually exclusive distance bands or orders of contiguity. The correlogram can be used to determine the extent of spatial autocorrelation and at what distance spatial autocorrelation is maximized.

Cliff and Ord (1973) derive the distribution of Moran's I under the null hypothesis for two different sampling assumptions. Under the randomization assumption the n observed values are fixed, but they are relocated randomly among the locations in a random fashion. The normality assumption assumes that the values at each location are drawn from independent and identical normal distributions. Underlying both of these assumptions is the additional assumption of stationarity. In the spatial context, stationarity implies that the mean and variance of the variable of interest is constant throughout the study region. Cliff and Ord (1973) prove that under both the randomization and normality assumptions Moran's I is asymptotically normally distributed. When n is large, a reliable significance value can be computed based on this distribution. Tiefelsdorf and Boots (1995) show that the rate of convergence to normality is a function of the spatial weights matrix and the distribution of the data values as well as sample size. A Monte Carlo approach, as outlined by Besag and Newell (1991), is often used to generate significance values under either the randomization or normality assumptions.

Adjusting for heterogeneous variance. When the spatial units vary significantly in size, the assumption of constant variance is violated. Specifically, units with large populations are less likely to deviate from the global mean with respect to units with small populations (Haining 2003). Walter (1992) demonstrates that variation in size of population at risk can result in incorrectly rejecting the null hypothesis. Several methods have been proposed to test the spatial randomness hypothesis when the background population is heterogeneous (Waller and Gotway 2004). Oden (1995) proposed a version of Moran's I, I_{pop} , that is based on individual level data. Inference is again based on the randomization assumption. However, the randomization refers to the status of individuals. This is most often applied in studies of disease clustering where cases are denoted as one and the remaining individuals are denoted zero. Tango (1995) proposed the excess events test (*EET*) that is defined as

$$EET = \sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\left(c_i - n_i \tfrac{C}{n}\right)\left(c_j - n_j \tfrac{C}{n}\right) \tag{B.4.9}$$

where c_i is the number of cases in unit i, n_i is the population of unit i, and C is the total number of cases in the study region. Like I_{pop} a large variation from the expected number of cases within a region contribute to large statistics, and I_{pop} is an affine transformation of *EET* (Oden et al. 1998; Tango 1998). Tango suggested an exponentially decreasing function of distance as the weight between units exp $(-d_{ij}/\lambda)$, where d_{ij} is the distance between locations i and j, and λ is a measure of the spatial scale of clustering. The maximized excess events test (*MEET)* searches over a plausible range of λ for the minimum *p-value* (Tango 2000). This methodology examines clustering at a number of scales while accounting for multiple testing. Assunção and Reis (1999) propose an Empirical Bayes method for standardizing rates when variances are not stable. In this approach x_i is

$$x_i^{adj} = \frac{x_i - E[x_i]}{\sqrt{Var(x_i)}}. \tag{B.4.10}$$

In the accompanying simulation study, the authors determine that the standardized index is more powerful than the traditional Moran's I. Assunção and Reis (1999) also compare their method to Oden's I_{pop} which is powerful in detecting rate heterogeneity within units, but is not as useful for detecting spatial correlation of rates.

Geary's c. Geary's c is an alternative measure of spatial clustering that takes the familiar cross-product form (Geary 1954). The similarity of two locations is quantified as the difference between the values at each location squared. This leads to the statistic

$$c = \frac{(n-1)}{2S_0} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(x_i - x_j)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{B.4.11}$$

Two values that are similar will have a small contribution to the global value, therefore low values of c are indicative of a clustered pattern. The expected value of a random pattern is one, and c ranges between zero and two. Cliff and Ord (1973) derived the variance under the randomization and normalization assumptions.

Getis-Ord G. The Getis-Ord G statistic quantifies the relationship between two locations as the product of the values at the locations (Getis and Ord 1992). The statistic is

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}\, x_i\, x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i\, x_j}. \tag{B.4.12}$$

Use of the general G requires that the variable of analysis is positive valued with a natural origin. The expected value under a random pattern is

$$E[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}}{n(n-1)}. \tag{B.4.13}$$

G values greater than the expected value result from a pattern that is dominated by concentrations of high values because the product of neighboring units is large. A low G value results from a pattern dominated by clusters of low values. Acceptance of the null does not necessarily imply a random pattern, but may result in the case that clusters of both high and low values exist in the study region. The G statistic differs from the other indexes discussed in this section in that it is not strictly a measure of clustering, but provides an indication of the type of clustering that is present in the study region.

Point data methods

A second set of methods is used to analyze phenomena that are mapped as points. These could be the location of a set of objects or the locations of a set of events. Complete spatial randomness (CSR) describes the pattern of points that would occur by chance in a completely undifferentiated environment. The process that generates this pattern is called the homogeneous planar Poisson point process. In this process points are generated in a study are under the conditions: (a) each location in the study area has an equal probability of receiving a point; and (b) the selection of a location for a point is independent of the location of existing points. As with areal data, patterns may deviate from CSR by being either clustered or dispersed. In a clustered pattern, points are on average closer than expected in CSR. In a dispersed pattern, points are uniformly distributed throughout the study area.

The CSR hypothesis is limiting and rejection of this null may not be meaningful. There are few instances when the homogeneous and independent probability of occurrence is plausible. To avoid this limiting assumption, comparative analysis of two or more point patterns is conducted. This allows for examination of clustering above and beyond what would be expected due to spatial variation in the probability of occurrence. The aim is often to determine whether some attribute is clustered in a population given its heterogeneous distribution. When analyzing one or more types of events or objects, the point patterns are often referred to as marked point patterns.

Quadrat analysis. Quadrat analysis is one of the first techniques used to test the CSR hypothesis. Quadrat analysis involves partitioning the study area into a number of scattered or contiguous equal sized quadrats and was originally developed in the plant ecology literature (Greig-Smith 1952). The number of events in each cell is tabulated and a frequency table of these cell counts is computed. A goodness-of-fit test is then performed to determine if the frequencies are significantly different from those expected under a Poisson process. An excess number of low and high cell counts indicate a clustered pattern. An excess number of cells with average density indicate a dispersed pattern. The results are dependent on the size of the quadrats, and often the analysis is repeated for a range of quadrat sizes (Boots and Getis 1988). The general clustering methods described above are also used to analyze the pattern of events aggregated into quadrats.

Nearest neighbor analysis. Nearest neighbor analysis also has it origins in the plant ecology literature. These methods are based on the distance between each point and its closest neighbor. Clark and Evans (1954) derived the expected value and variance of the average nearest neighbor distance in a CSR pattern. The use of the mean nearest neighbor distance provides an easy to interpret summary statistic, but is a crude representation of a point pattern. For instance, a few very large nearest neighbor distances associated with isolated points could obscure an otherwise clustered pattern. Refined nearest neighbor analysis overcomes this issue by examining the entire distribution of nearest neighbor distances. The test statistic is the maximum difference between the observed nearest neighbor distance frequency distribution and the distribution expected under the null hypothesis (Diggle 1990). A rigorous analysis of a point data set can also include the analysis of higher order neighbors.

Ripley's K function. One problem with quadrat analysis and nearest neighbor analysis is that they examine only one scale of interaction at a time (Bailey and Gatrell 1995). Most commonly these techniques detect clustering at short distances. Advances in computational capabilities have enabled the examination of all inter point distances. Ripley's *K* function can be computed over a range of distances and be used to identify the scales over which clustering occurs (Ripley 1976). The estimator is defined as

$$\hat{K}(d) = \frac{R}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} \qquad \text{for } i \neq j \tag{B.4.14}$$

where *R* is the size of the study area. The weights matrix is binary and equal to one when points *i* and *j* are within distance *d*, and zero otherwise. A standardized measure that simplifies interpretation is given as

$$\hat{L}(d) = \sqrt{\frac{\hat{K}(d)}{\pi}}\,. \tag{B.4.15}$$

The expected value of $\hat{L}(d)$ under CSR is *d*. A value greater than *d* indicates clustering and a value less than *d* indicates dispersion. The statistical significance of the results is determined through Monte Carlo simulations under an appropriate null hypothesis (Besag and Diggle 1977).

The points outside the study region are unobserved and cannot be included in the summation. In order to correct for this edge effect, points near the boundary may be given a larger weight in the analysis. Ripley (1976) provided one such correction for rectangular study areas (see Chapter B.3). The boundary problem is also overcome by transforming or duplicating the existing dataset to create points outside the boundary. A comparison of the various edge correction methods is provided by Yamada and Rogerson (2003).

Ripley's *K* function is a form of second order analysis because it is examining the interaction or dependence between points. This is in contrast to the intensity of points, which are termed first-order effects. There is an implicit assumption that the density of points is uniform within the study area (Diggle 2003). When the density of points is heterogeneous within the study area, this first-order effect may be captured in the *K* function. To avoid this ambiguity the distances of analysis should be limited so that they are small relative to the size of the study area. One rule of thumb is to limit the maximum distance of analysis to no longer than one-half the length of the shorter side of a rectangular study area.

Bivariate point patterns. The methods above have only considered points of a single type. Bivariate point pattern methods may be used to answer questions concerning the spatial dependence of two types of events. One set of points may also be used as a control group to correct for the variations in density within the study area. This type of analysis is especially relevant to epidemiological studies where inhomogeneous populations at risk are the norm.

The cross *K* function is a useful tool for examining the relationship between two sets of events (Bailey and Gatrell 1995). The estimator is given as

$$\hat{K}_{12}(d) = \frac{R}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} W_{ij} \tag{B.4.16}$$

where n_1 and n_2 are the number of each type of points. The result can be standardized in the same manner as above (see Eq. (B.4.15)). In this case, a value greater than *d* indicates that attraction between the two types of events and a value lower than *d* indicates repulsion between the two types of events. Significance is calculated through randomization. In this case, the patterns are preserved in their original form, but they are shifted relative to one another. These shifts may be performed using a toroidal transformation of the study area.

Spatial randomness may not always be an important hypothesis to test. Very often the potential locations of an event are limited within the study area. Examples include crimes which are geocoded to the nearest available street address or cases of disease which are distributed among the population at risk. This type of heterogeneity can be accounted for using bivariate point pattern analysis. Cuzick and Edwards (1990) presented a method based on the number of nearest neighbors of each type of point. The method depends on a scale parameter, *k*, that indicates the extent of analysis in terms of the number of nearest neighbors. The method was designed to detect clusters in epidemiological datasets, and the events of interest are usually cases of disease. The second set of events is called controls and is selected as being representative of the population at risk. The statistic is given as

$$T_k = \sum_{i=1}^{n_1} m_i(k) \tag{B.4.17}$$

where n_1 is the number of cases, and $m_i(k)$ is the number of cases among the k nearest neighbors. When cases are clustered, the resulting statistic will be large. T_k will be small when the cases are dispersed and therefore, surrounded by controls. Jacquez (1994) developed a modification to the Cuzick and Edwards' test that can be used to evaluate aggregate data as well.

A form of the K function can be employed in the same situation (Diggle and Chetwynd 1991). The statistic becomes the difference between the two univariate K functions,

$$Diff(d) = \hat{K}_1(d) - \hat{K}_2(d) \tag{B.4.18}$$

where $\hat{K}_1(d)$ and $\hat{K}_2(d)$ are the K functions for each set of points. If the events of type one are distributed randomly in relation to the remaining points, the difference will be approximately zero. A positive difference indicates that points of type one are more clustered than points of type two. A negative value indicates that points of type one are more dispersed than points of type two. The significance of both T_k and $Diff(d)$ can be examined under the random labeling null hypothesis. The designation of event type is randomly permuted or shuffled among the points for each realization in a Monte Carlo procedure.

B.4.3 Local measures of spatial clustering

When the null hypothesis of spatial randomness is rejected by a general test for spatial clustering two additional questions are raised: where are the clusters and what is their spatial extent. Local clustering statistics are used to answer these questions. It should be noted, however, that there may be significant local clustering even in the case that the general test results in acceptance of the null hypothesis. Local measures can be either tests of clustering or focused tests.

Areal data methods

As with global clustering statistics, the local tests take a general form. A local clustering statistic is the product of a spatial weights vector and a similarity vector. It is represented in summation notation as

$$\sum_{j=1}^{n} W_{ij}\ S_{ij}\ . \tag{B.4.19}$$

Several of the global methods presented in Section B.4.2 have a local equivalent that is the *i*th unit's contribution to the global statistic.

Getis-Ord G_i and G_i^*. Getis and Ord (1992) present a local clustering test that is based on the concentration of values in the neighborhood of a unit. The original statistic was given as

$$G_i = \frac{\sum_{j=1}^{n} W_{ij} x_j}{\sum_{j=1}^{n} x_j} \qquad \text{for } i \neq j. \tag{B.4.20}$$

The authors derive the expected value and variance of G_i when W_{ij} are elements of a binary spatial weights matrix. Most often the weights are based on proximity with the value at all units within a given distance being summed in the numerator. The G_i^* statistic includes the contribution of the *i*th unit in the calculation of local concentration. This amounts to adding the value x_i to both the numerator and denominator in Eq. (B.4.20). The G_i^* matches the usual definition of cluster as a contiguous and non-perforated set of units. In this original formulation, the statistics are intended for use with variables that possess a natural origin.

Modified versions of the G_i and G_i^* statistics are presented by Ord and Getis (1995). The newer formulation standardizes the statistic by subtracting the expected value and dividing the difference by the standard error. This eases interpretation as the result can be interpreted as approximately following a standard normal distribution. A positive value indicates clustering of high values and a negative value indicates a cluster of low values. This update also allows for the use of non-binary weights matrices and variables without a natural origin. The standardized G_i^* statistic is given in Chapter B.3.

The Moran scatter plot and local Moran's I. The Moran scatter plot was introduced by Anselin (1996) as an exploratory spatial data analysis (ESDA) tool for assessing local patterns of spatial association (see also Chapter B.1). This bivariate scatter plot places the unit values (x_i) on the horizontal axis and the spatial lag (lag_i) for the same variable on the vertical axis (see Fig. B.4.1). The spatial lag is the spatially weighted average of the values at neighbouring units, and is calculated as

$$lag_i = \frac{\sum_{j=1}^{n} W_{ij}\, x_j}{\sum_{j=1}^{n} W_{ij}}. \tag{B.4.21}$$

The axes of the plot are drawn so that they cross at the average value of x_i and lag_i, respectively. The four quadrants of the plot separate the spatial association into four components. The first letter in the quadrant labels indicates whether the value of x_i is higher (H) or lower (L) than the average of all values. Correspondingly, the second letter in the quadrant labels indicates whether the value of lag_i is higher (H) or lower (L) than the average of all the spatial lags. Units that fall into the quadrants labelled ‘HH’ and ‘LL’ represent clustering of high and low values respectively. The remaining quadrants contain units that have negative association with their neighbours and can be considered as spatial outliers. A spatial outlier may arise from a cluster consisting of just one unit. The Moran scatter plot is a useful visualization tool for assessing spatial pattern and spatial clustering.

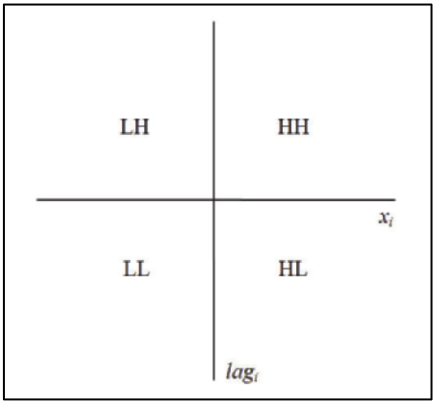

Fig. B.4.1. The Moran scatter plot

The significance of extreme points in the Moran scatter plot can be assessed using local Moran’s I or I_i (Anselin 1995). For each region, I_i is calculated as

$$I_i = \frac{\sum_{j=1}^{n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})} . \tag{B.4.22}$$

As discussed in Chapter B.3, I_i represents a decomposition of the global Moran’s I. This form of local method is called a Local Indicator of Spatial Association (LISA). Anselin (1995) also presents the formulation of the local Geary’s c or c_i. Statistical significance can be determined through the provided expected value and variance or by Monte Carlo procedure. A positive I_i indicates clustering of high or low values. A negative I_i indicates a spatial outlier. Several results are, therefore, reported for each unit. These include the statistic value, the significance value, and the label of corresponding quadrant of the Moran scatter plot.

Local clustering of categorical data. In the case of global clustering statistics, the methods for categorical data preceded the methods for metric data. This was not the case for local methods of pattern analysis. Boots (2003, 2006) details the issues in this research area and presents ESDA methods for describing and understanding patterns of categorical data.

Accounting for multiple and dependent testing

Local spatial statistics are often used in an exploratory mode to test for clustering at each location in the study area simultaneously. In this case, the issue of multiple and dependent testing is a concern when assessing the significance of clustering. Multiple testing problems arise whenever more than one hypothesis test is carried out using the same dataset. The probability of rejecting the null hypothesis at least once when it is true in all cases is much higher than the nominal type I error rate, α. The dependence part of the problem is a result of nearby local tests relying on many of the same data values. The results of these tests are, therefore, correlated. Failure to account for these effects results in over identification of clusters by local spatial statistics (Anselin 1995; Ord and Getis 1995).

The Bonferroni correction is commonly used to account for multiple testing (Warner 2007). In this approach, a new critical value is calculated for the individual tests by dividing the overall level of type I error by the number of tests. For example, if an overall significance level of 0.05 is desired for 20 simultaneous tests, a significance level of 0.0025 is used in each separate test. Caldas de Castro and Singer (2006) demonstrate the usefulness of a less conservative approach called the false discovery rate (FDR). FDR controls for the rate of false positives among the nominally significant results and was introduced by Benjamini and Hochberg (1995). It is based on the distribution of significance values for a set of tests, and is therefore adaptive to the characteristics of each dataset. Simulation studies performed by Caldas de Castro and Singer (2006) compared uncorrected local statistics with the Bonferroni and FDR corrected versions. FDR was superior in properly identifying the location and extent of spatial clusters. Another common approach to account for multiple testing is to examine just the most extreme value of all the individual tests (Baker 1996; Tango 2000). This approach provides a satisfying solution in a general test of clustering, but it does not address each local test individually.

Local spatial tests are most often evaluated under the assumption that there is no global spatial autocorrelation. Some attempts have been made to relax this assumption and evaluate clustering in the presence of spatial autocorrelation. One technique, that of Ord and Getis (2001), is described in Chapter B.3 of this handbook. Goovaerts and Jacquez (2004) present a geostatistical technique for generating datasets under a realistic null hypothesis. These models include both spatial autocorrelation and heterogeneous populations in the examinations of clustering.

Cluster detection algorithms

A second set of local methods are the automated search procedures and their associated test statistics. These computational techniques involve testing a large number of regions within the study area for spatial clustering. These methods have primarily been applied to spatial analysis of epidemiological data. They are flexible in that they can, for the most part, be applied to both point and aggregate data. In the case of aggregate data, the location associated with spatial units is most often taken as the centroid of the unit. They differ from the methods presented above in that they are not limited to a fixed definition of neighborhood, and thus cluster size, but are designed to detect clusters of varying sizes. It should be noted, however, that the test statistics discussed in the previous section could be used in conjunction with the search procedures outlined below.

Geographical Analysis Machine (*GAM*). The Geographical Analysis Machine (GAM) was the first automated approach to finding cluster locations in spatial patterns (Openshaw et al. 1987). The original GAM involves searching a large number of circles across the study area. The circles are centered on a grid, and the radius of these circles is allowed to vary over a suitable range of values. The number of cases in each circle is counted and the significance of the count is evaluated. A Monte Carlo procedure is used, and the circles that fall within a given threshold are retained. The resulting set of circles is then mapped to show cluster centers. One weakness of the GAM is the lack of control for multiple testing (Besag and Newell 1991). The GAM did, however, show the utility of a geocomputational approach to cluster detection and has inspired several modifications and improvements. Each of the methods described below has built on the foundation of the GAM. There have also been several improvements to the GAM procedure itself. One example is the method of Conley et al. (2005). This technique uses genetic algorithms to speed search times and reduce over-reporting of cluster sizes.

Besag and Newell's method. One additional shortcoming of the original GAM is that the circles examined are based on a distance only approach. If the population at risk varies, then circles of the same size contain different size populations. This variation in population at risk must be included in the analysis. The Besag and Newell (1991) method overcomes this difficulty by requiring the expected cluster size, say k, as a user input. Each unit with at least one case of disease is examined as a potential center of clustering. The circle is expanded in order of nearest neighbor distance until at least k cases are included within the circle. The inference is then based on the number of units, L_i, containing k cases. The significance of each potential cluster is evaluated using the Poisson cumulative distribution function under the uniform risk null hypothesis

$$P\left(L_i \leq l_i\right) = 1 - \sum_{j=1}^{k-1} \exp(-\mu)\,\frac{\mu^j}{j!} \tag{B.4.23}$$

where l_i is the observed number of units containing k cases, and μ is the expected number of cases within those units. μ is calculated as the product of the global risk and the population at risk within the set of units under examination. Fotheringham and Zhan (1996) compare GAM, Besag and Newell's method and their own modification of the GAM search algorithm. All methods are deemed successful at detecting clusters, but Besag and Newell's method is the least likely to result in false positives. Additionally, Fotheringham and Zhan (1996) provide a formulation of Besag and Newell's method for use with point data, as the original presentation was based on areal spatial units.

The SaTScan procedure. The SatScan procedure is another cluster finding procedure inspired by the GAM (Charlton 2006). Like the GAM, SaTScan searches a large number of circles and examines the number of cases in relation to the population at risk (Kulldorff 2004). Most analysts choose to examine clusters that are centered on cases or region centroids as in the Besag and Newell method, but any number of potential clusters could be examined. At each center, the size of the circle is increased until a user defined maximum cluster size is reached. The maximum cluster size could be given in terms of geographic area or population at risk. The minimum cluster size does not need to be specified. During the search procedure, the likelihood that each cluster has occurred by change is evaluated using the spatial scan statistic. Kulldorff (1997) derived the spatial scan statistic for count or marked point pattern data. Variants of the spatial scan statistic appropriate for other types of data have also been developed (Huang et al. 2007; Jung et al. 2007). The spatial scan statistic based on the Poisson distribution is employed for aggregate case data. A uniform risk null hypothesis is evaluated. $L(R)$ is the likelihood that there is a cluster in a region R, and L_0 is the likelihood under the null. A likelihood ratio test statistic is given by

$$\frac{L(R)}{L_o} = \left(\frac{c_R}{\mu_R}\right)^{c_R} \left(\frac{C - c_R}{C - \mu_R}\right)^{C - c_R} \tag{B.4.24}$$

if $c_R > \mu_R$, and one otherwise. Here C is the total number of cases for the population, c_R is the number of cases in region R, and μ_R is the expected number of cases in the region R. The most likely cluster or clusters are those with the largest likelihood ratio values. An exact p-value is calculated using a Monte Carlo procedure. A primary advantage of the spatial scan statistic is that it takes multiple testing into account. A version of the SaTScan procedure that examines elliptical regions as potential clusters is presented by Kulldorff et al. (2006).

Finding arbitrarily shaped clusters. To this point, each of the cluster detection methods discussed are limited to either a prespecified and fixed definition of neighborhood or the examination of a large number of circles or ellipses. In most cases there is little reason to expect that spatial clustering would take a regular shape. To overcome this limitation, a variety of tests have been developed to locate irregularly shaped clusters. Each of these approaches uses a definition of proximity equivalent to the binary contiguity matrix. Spatial units are treated as nodes on a connected graph. The resulting clusters are not limited to being regular shapes, but must be contiguous regions or connected sub-graphs.

Tango and Takahashi (2005) proposed an examination of all possible connected sub-graphs up to a pre-selected maximum cluster size. This approach works well for clusters containing a small number of units, but is not feasible for finding larger clusters. Two approaches use stochastic optimization techniques to overcome this shortcoming. Duczmal and Assunção (2004) employ simulated annealing, and Duczmal et al. (2007) a genetic algorithm. These techniques are not restricted to a maximum cluster size, but they require additional inputs, known as hyper parameters, that govern the search process.

Aldstadt and Getis (2006) proposed an iterative region growing approach to finding arbitrarily shaped clusters called AMOEBA. To begin this procedure a single unit is selected as the seed location. All possible combinations of contiguous units are examined and the set that maximizes the clustering statistic is retained. The algorithm then continues by examining the units at each order of contiguity until the addition of units no longer increases the test statistic. At this point a cluster based on the first seed location is delimited. The procedure can be repeated using every location as the seed location. The significance of each delimited cluster is evaluated using a Monte Carlo procedure. The iterative approach ensures that low value units will not be included in clusters of high values. This prohibits the linking of two or more disjoint clusters as one, which is possible in the other approaches.

Focused clustering methods

Focused clustering tests start with a predetermined set of foci, and examine the likelihood that each of these foci is the center of a cluster. Foci are most often represented as points, but they may also be linear or areal features. The most common application of these tests is the examination of disease clusters in proximity to a pollution source. The null hypothesis is that disease risk is not elevated in proximity to the foci. It bears repeating that the potential sources should be identified before the initiation of these focused tests. If potential foci are selected based on their proximity to areas of raised incidence identified through cluster detection procedures, the inference is biased toward rejection of the null hypothesis (Waller and Gotway 2004). This is known as the ‘Texas sharpshooter fallacy.’ The name comes from the Texan that shoots into the side of a barn and then paints a target centered on the hits so that it appears he is a sharpshooter.

The Lawson-Waller score test. Waller et al. (1992) and Lawson (1993) independently developed a score tests for focused clustering. The global risk can be estimated as the total number of cases, *C,* divided by the total population at risk, *n*. The resulting score statistic is a local version of Tango's EET statistic. The score statistic for a focus *i* is given as

$$T_i = \sum_{j=1}^{n} W_{ij}\left(c_j - n_j \tfrac{C}{n}\right) \tag{B.4.25}$$

where c_j is the number of cases in unit *j*, n_j is the population of unit *j*. Here, the spatial weight can take a variety of forms. A distance decay function depicts the setting where exposure decreases as distance to the foci increases. A binary weight may also be used to indicate that all units within a given distance are experiencing similar exposure. Under the constant risk null hypothesis, the expected value of the statistic is zero. The variance of T_i is

$$Var(T_i) = \tfrac{C}{n}\left(\sum_{j=1}^{n} W_{ij}^2 n_j\right) - \tfrac{C}{n^2}\left(\sum_{j=1}^{N} W_{ij}\, n_j\right)^2 . \tag{B.4.26}$$

The standardized statistic

$$Z(T_i) = \frac{T_i}{\sqrt{Var(T_i)}} \tag{B.4.27}$$

can then be compared to the standard normal distribution. Monte Carlo tests may be more appropriate when there is a small number of regions or for a vary rare disease (Waller et al. 1992). A method of determining the exact distribution of T_i is provided by Waller and Lawson (1995). Rogerson (2005) defines both a global test and a local clustering statistic based on the score test.

Other focused clustering tests. Stone (1988) developed a group of tests based on the first isotonic regression estimator. This method assumes that the relationship between exposure and risk is monotonic, but the relationship does not have to take a parametric form. This flexibility is unique among focused clustering tests. Bithell (1995) provided a set of tests that are called linear risk score tests. These tests are based on the notion of the relative risk function. Under this alternative hypothesis, relative risk of disease declines as distance to the focus increases. The test statistic is the sum of these estimated relative risk values. This test is commonly performed using the rank of distances to neighboring units. In this case the risk becomes a function of relative location as opposed to exact location. Tango (2002) provides an extended score test that allows for non-monotonic relative risk functions. The extended score test would be most useful in the situation where exposure is expected to peak at some distance form the putative source.

A focused clustering test for individual or point level data is provided by Diggle (1990) and refined by Diggle and Rowlingson (1994). This method can be applied to inhomogeneous point patterns when the locations of disease cases and a representative control group are known. The method is flexible in terms of the functional form of the spatial risk, but the type of model must be specified. The parameters of the kernel are estimated using non-linear binary regression. The regression framework allows for straightforward inclusion of covariates when they are available. If the kernel function is log-linear or a step function, the model reduces to logistic regression (Diggle and Rowlingson 1994).

B.4.4 Concluding remarks

The choice of clustering method depends on several factors. The first consideration is whether the method is appropriate for the available data type. Beyond this practical consideration it is of primary importance that the method evaluates an appropriate null and alternative hypotheses (Waller and Gotway 2004). Some null hypotheses that have been mentioned are spatial randomization, constant risk, and random labeling. Possible alternative hypotheses include variations of regional, local, or focused clustering. Beyond these criteria, an analyst might consider the power of the test in choosing between appropriate methods. In the case of spatial clustering, power refers to the probability of rejecting the null hypothesis given that the data have been generated under the alternative hypothesis. Monte Carlo methods are useful in this regard, and can be used to generate data under a variety of hypotheses. Kulldorff et al. (2003) developed a set of benchmark data, generated under a variety of alternative hypotheses, that can be used to evaluate and compare methods. A later paper compares a large set of methods using the benchmark data (Song and Kulldorff 2003). The power of a test can also be affected by the properties of the data and choice of parameters for clustering methods (Waller et al. 2006). For example, the power can vary widely based on the choice of spatial weights (Song and Kulldorff 2005). Takahashi and Tango (2006) provide a modified test for power that takes into account not only the ability to reject the null hypothesis but also whether the detected clusters are of the correct size and in the proper location. A discussion on method choice and statistical power can be found in Waller and Gotway (2004).

There was a time when, due to a lack of clustering methodologies, researchers could be excused for applying techniques without strict adherence to assumptions. For the most part, this is no longer the case. There are now tools available to handle most data types and a variety of hypotheses. The research in this field will progress by improving existing methods and developing new ones. These developments combined with the rapid innovation in software for spatial data analysis, as covered in Part A of this handbook, will increase the utility of spatial clustering analysis as a research tool.

References

Aldstadt J, Getis A (2006) Using AMOEBA to create a spatial weights matrix and identify spatial clusters. Geogr Anal 38 (4):327-343

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht

Anselin L (1995) Local indicators of spatial association – LISA. Geogr Anal 27(2):93-115

Anselin L (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer MM, Scholten HJ, Unwin D (eds) Spatial analytical perspectives on GIS. Taylor and Francis, London, pp.111–125

Assunção RM, Reis EA (1999) A new proposal to adjust Moran's *I* for population density. Stat Med 18(16):2147-2162

Bailey TC, Gatrell AC (1995) Interactive spatial data analysis. Longman, Harlow

Baker RD (1996) Testing for space-time clusters of unknown size. J Appl Stat 23(5):543-554

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B 57(1):289-289

Besag J, Diggle PJ (1977) Simple Monte Carlo tests for spatial pattern. J Appl Stat 26(3):327-333

Besag J, Newell J (1991) The detection of clusters in rare diseases. J Roy Stat Soc A 154(1):143-155

Bithell JF (1995) The choice of test for detecting raised disease risk near a point source. Stat Med 14:2309-2322

Boots BN (2003) Developing local measures of spatial association for categorical data. J Geogr Syst 5(2):139-160

Boots BN (2006) Local configuration measures for categorical spatial data: binary regular lattices. J Geogr Syst 8(1):1-24

Boots BN, Getis A (1988) Point pattern analysis. Sage, London

Caldas de Castro M, Singer BH (2006) Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. Geogr Anal 38(2):180-208

Charlton ME (2006) A mark 1 geographical analysis machine for the automated analysis of point data sets: twenty years on. In Fisher PF (ed) Classics from IJGIS: twenty years of the International Journal of Geographical Information Science and Systems. CRC Press (Taylor and Francis Group), Boca Raton [FL], London and New York, pp.35-40

Clark PJ, Evans FC (1954) Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35(4):445-453

Cliff AD, Ord JK (1973) Spatial autocorrelation. Pion, London

Conley J, Gahegan M, Macgill J (2005) A genetic approach to detecting clusters in point data sets. Geogr Anal 37(3):286-314

Cuzick J, Edwards R (1990) Spatial clustering for inhomogeneous populations. J Roy Stat Soc B 52(1):73-104

Diggle PJ (1990) A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. J Roy Stat Soc A 153(3):349-362

Diggle PJ (2003) Statistical analysis of spatial point patterns. Edward Arnold, New York

Diggle PJ, Chetwynd AG (1991) Second-order analysis of spatial clustering for inhomogeneous populations. Biometrics 47(3):1155-1163

Diggle PJ, Rowlingson BS (1994) A conditional approach to point process modelling of elevated risk. J Roy Stat Soc A157(3):433-440

Duczmal L, Assunção R (2004) A simulated annealing strategy for detection of arbitrarily shaped spatial clusters. Compu Stat Data Anal 45(2):269-286

Duczmal L, Cançado ALF, Takahashi RHC, Bessegato LF (2007) A genetic algorithm for irregularly shaped spatial scan statistics. Comp Stat Data Anal 52(1):43-52

Fotheringham AS, Zhan FB (1996) A comparison of three exploratory methods for cluster detection in spatial point patterns. Geogr Anal 28(3):200-218

Geary RC (1954) The contiguity ratio and statistical mapping. The Incorp Stat 5(3):115-145

Getis A (2009) Spatial autocorrelation. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.255-278

Getis A, Aldstadt J (2004) Constructing the spatial weights matrix using a local statistic. Geogr Anal 36(2):90-105

Getis A, Ord JK (1992) The analysis of spatial association by distance statistics. Geogr Anal 24(3):189-206

Goovaerts P, Jacquez GM (2004) Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. Int J Health Geogr 3(14), http://www.ij-healthgeographics.com/content/3/1/14

Greig-Smith P (1952) The use of random and contiguous quadrats in the study of the structure of plant communities. Ann Bot 16(2):293-316

Haining RP (2003) Spatial data analysis: theory and practice. Cambridge University Press, Cambridge

Haining RP (2009) The nature of geoferenced data. In Fischer MM, Getis A (eds) Handbook of applied spatial analysis. Springer, Berlin, Heidelberg and New York, pp.197-217

Huang L, Kulldorff M, Gregorio D (2007) A spatial scan statistic for survival data. Biometrics 63(1):109-118

Jacquez GM (1994) Cuzick and Edwards' test when exact locations are unknown. Am J Epidemiol 140(1):58-64

Jung I, Kulldorff M, Klassen AC (2007) A spatial scan statistic for ordinal data. Stat Med 26(7):1594

Knox EG (1989) Detection of clusters. In Elliott P (ed) Methodology of enquiries into disease clustering. Small Area Health Statistics Unit, London, pp.17-20

Kulldorff M (1997) A spatial scan statistic. Comm Stat Theor Meth 26(6):1481-1496

Kulldorff M (2004) SaTScan v4.0: Software for the spatial and space-time scan statistics. Information Management Services Inc.

Kulldorff M, Tango T, Park P (2003) Power comparisons for disease clustering tests. Comput Stat Data Anal 42(4):665-684

Kulldorff M, Huang L, Pickle L, Duczmal L (2006) An elliptic spatial scan statistic. Stat Med 25(22):3929-3943

Lawson AB (1993) On the analysis of mortality events associated with a prespecified fixed point. J Roy Stat Soc A156(3):363-377

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209-220

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(12):17-23

Oden N (1995) Adjusting Moran's *I* for population density. Stat Med 14(1):17-26

Oden N, Jacquez GM, Crimson R (1998) Authors reply. Stat Med 17:1058-1062

Openshaw S, Charlton ME, Wymer C, Craft A (1987) A mark 1 geographical analysis machine for the automated analysis of point data sets. Int J Geogr Inform Sci 1(4):335-358

Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27(4):286-306

Ord JK, Getis A (2001) Testing for local spatial autocorrelation in the presence of global autocorrelation. J Reg Sci 41(3):411-432

Ripley BD (1976) The second-order analysis of stationary point processes. J Appl Prob 13(2):255-266

Rogerson PA (2005) A set of associated statistical tests for spatial clustering. Environ Ecol Stat 12(3):275-288

Song C, Kulldorff M (2003) Power evaluation of disease clustering tests. Int J Health Geographics 2(1):9

Song C, Kulldorff M (2005) Tango's maximized excess events test with different weights. Int J Health Geographics 4(1):32

Stone R (1988) Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. Stat Med 7(6):649-660

Takahashi K, Tango T (2006) An extended power of cluster detection tests. Stat Med 25(5):841

Tango T (1995) A class of tests for detecting 'general' and 'focused' clustering of rare diseases. Stat Med 14:2323-2334

Tango T (1998) Adjusting Moran's *I* for population density by N. Oden. Stat Med 17(9):1055-1058

Tango T (2000) A test for spatial disease clustering adjusted for multiple testing. Stat Med 19(2):191-204

Tango T (2002) Score tests for detecting excess risks around putative sources. Stat Med 21 (4):497-514

Tango T, Takahashi K (2005) A flexibly shaped spatial scan statistic for detecting clusters. Int J Health Geographics 4(11), http://www.ij-healthgeographics.com/content/4/1/11

Tiefelsdorf M, Boots B (1995) The exact distribution of Moran's *I*. Environ Plann A 27(6): 985-999

Tiefelsdorf M, Griffith DA, Boots B (1999) A variance-stabilizing coding scheme for spatial link matrices. Environ Plann A 31(1):165-180

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley-Interscience, Hoboken [NJ]

Waller LA, Lawson AB (1995) The power of focused tests to detect disease clustering. Stat Med14:2291-2308

Waller LA, Hill EG, Rudd RA (2006) The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. Stat Medicine 25(5):853

Waller LA, Turnbull BW, Clark LC, Nasca P (1992) Chronic disease surveillance and testing of clustering of disease and exposure: application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. Environmetrics 3(3):281-300

Walter SD (1992) The analysis of regional patterns in health data – I. Distributional considerations. Am J Epidemiol 136(6):730-741

Warner RM (2007) Applied statistics: from bivariate through multivariate techniques. Sage, Thousand Oaks [CA]

Yamada I, Rogerson PA (2003) An empirical comparison of edge effect correction methods applied to *K* function analysis. Geogr Anal 35(2):97-110