4 Spatial Autocorrelation

THEORY

Learning Objectives

This chapter deals with

- Spatial autocorrelation and its importance to geographical problems
- Global and local spatial autocorrelation techniques like Moran's *I*, Getis-Ord G and Geary C
- Tracing spatial clusters of high values (hot spots) or low values (cold spots)
- Tracing spatial outliers
- Optimized hot spot analysis
- Interpreting the statistical significance of results
- Incremental spatial autocorrelation used to define the appropriate scale of analysis
- The multiple comparison problem and spatial dependence
- Introducing Bonferroni correction and the false discovery rate
- Spatiotemporal autocorrelation analysis using bivariate and differential Local Moran's *I* index
- Presenting step-by-step examples using ArcGIS and GeoDa

After a thorough study of the theory and lab sections, you will be able to

- Distinguish between global and local spatial autocorrelation
- Understand why spatial autocorrelation analysis is relevant to geographical analysis
- Apply local and global indices of spatial autocorrelation like local Moran's, Getis-Ord G_i and G_i^*
- Use Moran's *I* scatter plot to identify patterns
- Identify hot spots or cold spots
- Identify and locate spatial outliers
- Use bivariate and differential Local Moran's *I* to identify if spatiotemporal autocorrelation exists and if changes cluster over time

- Apply these tools using ArcGIS
- Interpret the results from both the statistical significance and spatial analysis standpoints

4.1 Spatial Autocorrelation

Definition

Spatial autocorrelation is the degree of spatial dependency, association or correlation between the value of an observation of a spatial entity and the values of neighboring observations of the same variable. The terms "spatial association" and "spatial dependence" are often used to reflect spatial autocorrelation as well.

Why Use

Spatial autocorrelation is examined to determine if relationships exist among the attribute values of nearby locations and if these values form patterns in space.

Interpretation

According to the first law of geography, objects in a neighborhood tend to have more similarities and interactions than those lying further away. This is what we call "spatial dependency." To measure spatial dependency, we use spatial autocorrelation metrics. Put simply, spatial autocorrelation measures how much the value of a variable in a specific location is related to the values of the same variable at neighboring locations.

The spatial autocorrelation concept is similar to that of the statistical correlation used for nonspatial variables. Still, there is a major difference. While statistical correlation refers to two distinct variables with no reference to location, spatial autocorrelation refers to the value of a single variable at a specific location in relation to the values of the same variable at neighboring locations.

In statistical correlation, if two variables tend to change in similar ways (e.g., higher income correlated to higher educational attainment), we have positive correlation. Likewise, if similar values of a variable (either high or low) in a spatial distribution tend to collocate, we also have positive spatial autocorrelation. Positive spatial autocorrelation is the state where "data from locations near one another in space are more likely to be similar than data from locations remote from one another"(O'Sullivan & Unwin 2010 p. 34). In other words, autocorrelation (or self-correlation) exists when an attribute variable of a spatial dataset, correlates with itself at specific distances, called lags. This means that location affects the values of the variable in such a way that promotes values

clustering in specific areas. A typical example of positive spatial autocorrelation is the income distribution within a city. Households with higher incomes generally tend to cluster in specific regions of the city, while households with lower incomes tend to cluster in other regions.

With negative spatial autocorrelation, on the other hand, neighboring spatial entities tend to have different values. This is similar to negative correlation, where high values of one variable indicate low values of the other. When there is no spatial autocorrelation, there is a random distribution of the values in relation to their locations, with no apparent association among them.

Discussion and Practical Guidelines

Spatial autocorrelation analysis is extremely important in geographical studies. If spatial autocorrelation did not exist, geographical analysis would be of little interest (O'Sullivan & Unwin 2010 p. 34). Think about it. We perform geographical analysis because we assume that location matters. If it did not, geography would be irrelevant. In most cases, phenomena do not vary randomly across space. For example, population concentrates on cities, income concentrates on cities, temperature displays small fluctuations inside a small area, and rain is uniform for a relatively small area. A student searching for a seat in an auditorium is most likely to sit next to a friend (who is already sitting). If you visit a restaurant, you will sit at an empty table. All these facts reveal nonrandom patterns. This is why geography is worth studying. If spatial arrangements were random, the global population could be located in every single location of the world with the same probability. If this were the case, more people would be living in the Antarctic or at heights above 5,000 m. If the temperature were random, you might be able to experience 30°C weather while standing outside the front door of your house and 1°C weather by jumping into the backyard. If rain were random, you might get wet while sunbathing at a beach in Santorini (Greece) during summer while the fellow right next to you lies on the sunbed sweating underneath the sun. In an auditorium, you will hardly see a student sit in a position that is already occupied. In a restaurant, it is rare (though not unusual) to sit at a table with people entirely unknown to you.

The aforementioned examples show that location matters and that a certain state influences what follows in a nonrandom way. They also remind us of the first- and second-order effects (see Section 3.2.1). By studying location, we reveal trends and patterns regarding the phenomenon at hand – for example, the spatial distribution of household income, the pattern of a disease outbreak, the relationship between residential location and mental well-being (Liu et al. 2019), or the linkage between built environment and physical health (Wang et al. 2019). Spatial autocorrelation is quite common in geographical analysis. This does not necessarily mean that it will occur across the entire study area (known as global autocorrelation). Spatial autocorrelation is sometimes

evident only in subregions of the study area (known as local autocorrelation). In any case, spatial autocorrelation reveals the nonrandom distribution of the phenomenon being studied.

The nonrandom geographical distribution of the values of the variables under study has significant effects on the accuracy of classical statistics. In conventional statistics, the observed samples are assumed to be independent. In the presence of spatial autocorrelation, this assumption is violated. Observations are now spatially clustered or dispersed. This typically means that classical statistical tools are no longer valid. For example, linear regression would lead to biased estimates or exaggerated precision (Gangodagamage et al. 2008 p. 34). Bias leads to overestimate or underestimate of a population parameter, and exaggerated precision leads to a higher likelihood of getting statistically significant results when, in reality, we should have gotten less (de Smith 2018 p. 122). In addition, spatial autocorrelation infers redundancy in the dataset. Each newly selected sample is expected to provide less new information, affecting the calculation of confidence intervals (O'Sullivan & Unwin 2010). That is why we should use spatial statistics when analyzing spatial data and perform a spatial autocorrelation analysis before conducting any conventional statistical analysis.

Several diagnostic measures can be used to identify spatial autocorrelation. Those that estimate spatial autocorrelation by a single value for the entire study area are named **global spatial autocorrelation measures**. The most commonly used are

- Moran's *I* index
- General G-Statistic
- Geary's C index

As mentioned, it is unlikely that any spatial process will be homogeneous in the entire area due to the nonuniformity and noncontinuity of space. The magnitude of spatial autocorrelation may vary from space to space due to spatial heterogeneity. To estimate spatial autocorrelation at the local level, we use **local measures** of spatial autocorrelation, like

- Local Moran's *I* index
- Getis-Ord G_i and G_i^* statistics

With such metrics, we describe spatial heterogeneity (in the distribution of the values of a variable) as they identify hot or cold spots, clusters and outliers.

Spatial autocorrelation (either positive or negative) is a key concept in geographic analysis. A test of global or local spatial autocorrelation should be conducted prior to any other advanced statistical analysis when dealing with spatial data. Note that correlation does not necessarily imply causation; it implies only association. Relationships of cause and effect should always be established

only after thorough analysis in order to avoid erroneous linkages (see Section 2.3.4 for more discussion on this).

4.2 Global Spatial Autocorrelation

4.2.1 Moran's *I* Index and Scatter Plot

Definition

Moran's *I* index computes global spatial autocorrelation by taking into account feature locations and attributes values (of a single attribute) simultaneously (Moran 1950). It is calculated by the following formula (4.1):

$$I = \frac{n}{\sum_i^n \sum_j^n w_{ij}} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i^n (x_i - \bar{x})^2} \tag{4.1}$$

where

n is the number of the spatial features
x_i is the attribute value of feature i, (remember that a variable is also called attribute in the spatial analysis context)
x_j is the attribute value of feature j
$\bar{x}$ is the mean of this attribute
$w_{i,j}$ is the spatial weight between feature i and j
$\sum_i^n \sum_j^n w_{ij}$ is the aggregation of all spatial weights

The tool calculates the mean $\bar{x}$, the deviation from the mean $(x_i - \bar{x})$ and the data variance $\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ (denominator). Deviations from all neighboring features are multiplied to create cross-products (the covariance term). Then, the covariance term is multiplied by the spatial weight. All other parameters are used to normalize the value of the index. For example, the aggregation of spatial weights is used to normalize for the number of adjacencies. By the same means, the variance is used to ensure that the value index will not be large just because of a large variability in x values (O'Sullivan & Unwin 2010 p. 206).

Why Use

Global Moran's *I* is used as a metric for global spatial autocorrelation. It is mostly used for aerial data, along with ratio or interval data.

Interpretation and Moran's *I* Scatter Plot

Moran's *I* index is an inferential statistic. It is interpreted based on the expected value calculated (Eq. 4.2) under the null hypothesis of no spatial autocorrelation (complete spatial randomness) and is statistically evaluated using a *p*-value and a *z*-score (just as any common inferential statistic – see Section 2.5). The expected value for a random pattern is (4.2):

$$E(I) = \frac{-1}{n-1} \tag{4.2}$$

where n denotes the number of spatial entities.

The expected value is the value that would have resulted if the specific dataset were the result of complete spatial randomness. The more spatial objects there are, the more the expected value tends to zero. The observed index value is the Moran's *I* index value calculated for the specific dataset through Equation (4.1). Positive Moran's *I* index values (observed) significantly larger than the expected value *E*(*I*) indicate clustering and positive spatial autocorrelation (i.e., nearby locations have similar values). Negative Moran's *I* index values (observed) significantly smaller than the expected value *E*(*I*) indicate negative spatial autocorrelation, meaning that the neighboring locations have dissimilar values. Values close to the expected value indicate no autocorrelation.

The difference between the observed and expected values has to be evaluated based on a *z*-score and a *p*-value. Through these metrics, we assess if this difference is statistically significant.

- If the *p*-value is large (usually $p > 0.05$), the results are not statistically significant, and we cannot reject the null hypothesis. The interpretation in statistical jargon is that *we cannot reject the null hypothesis that the spatial distribution of the values is the result of complete spatial randomness due to a lack of sufficient evidence*.
- A small *p*-value (usually $p < 0.05$) indicates that we can reject the null hypothesis of complete spatial randomness and accept that spatial autocorrelation exists:
  - In such a case, when the *z*-value is positive, there is positive spatial autocorrelation and a clustering of high or low values. Nearby locations will have similar values on the same side of the mean (O'Sullivan & Unwin 2010 p. 206).
  - If the *z*-value is negative, there is negative spatial autocorrelation and a dispersed pattern of values. Nearby locations will have dissimilar attribute values on the opposite sides of the mean (i.e., a feature with a high value repels other features with low values).

Let us consider an example. Imagine a spatial arrangement of 25 spatial objects (e.g., postcodes) with attribute values of either 1 (white) or 0 (black; see Figure 4.1).

- In a **perfectly dispersed** pattern, squares are located so that each one has neighbors of the opposite value. Spatial autocorrelation exists, as the squares have a competitive spatial relationship. If one square is black, the neighbor is white. Moran's *I* gets a negative value, smaller than the

expected value, and there is negative spatial autocorrelation (see far left in Figure 4.1).

- If the squares are **grouped** (as in the far right in Figure 4.1), then clustering occurs. There is again spatial autocorrelation, but it is positive. The spatial relationship is that similar values tend to cluster. Moran's *I* gets a positive value, significantly larger than the expected one.
- When the values are **randomly scattered**, then Moran's *I* value is close to the expected values, and there is zero spatial autocorrelation (see central section in Figure 4.1).
- In intermediate states, there is no perfectly uniform status. However, an indication of clustering, dispersion or randomness can be assessed depending on the index value. Positive higher index values show a tendency toward clustering. Lower negative index values show a tendency toward dispersion. Note that we might obtain a low positive Moran's *I* value while also observing local clusters (see the following discussion). In all cases, we should evaluate the difference between the observed and expected values through *p*-values and *z*-scores, as mentioned before.

Moran's *I* scatter plot is used to visualize the spatial autocorrelation statistic (see Figure 4.2). It allows for a visual inspection of the spatial associations in the neighborhood of each observation (data point). In other words, it provides a representation used to assess how similar an attribute value at a location is to its neighboring ones. Data points are points that have as coordinates the values of the variable *X* for the *x*-axis and the spatial lag of the variable *X* (*Lag-X*) for the *y*-axis. *Lag-X* is the weighted average values of *X* in a specified neighborhood. Both *X* and *Lag-X* variables are used in a standardized form. As such, the weighted average is plotted at the center of the graph on the coordinates (0, 0). Distances in the plot are expressed as number of standard deviations from

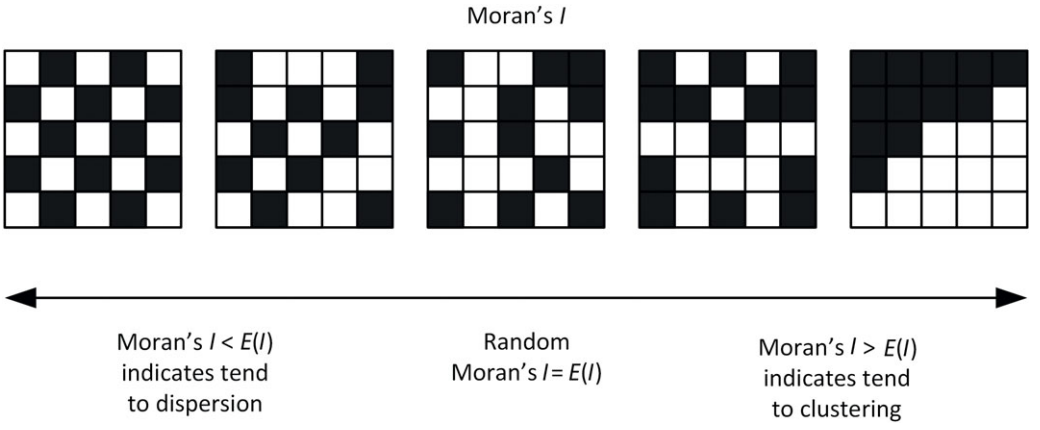


Figure 4.1 Global Moran's *I* and spatial autocorrelation.

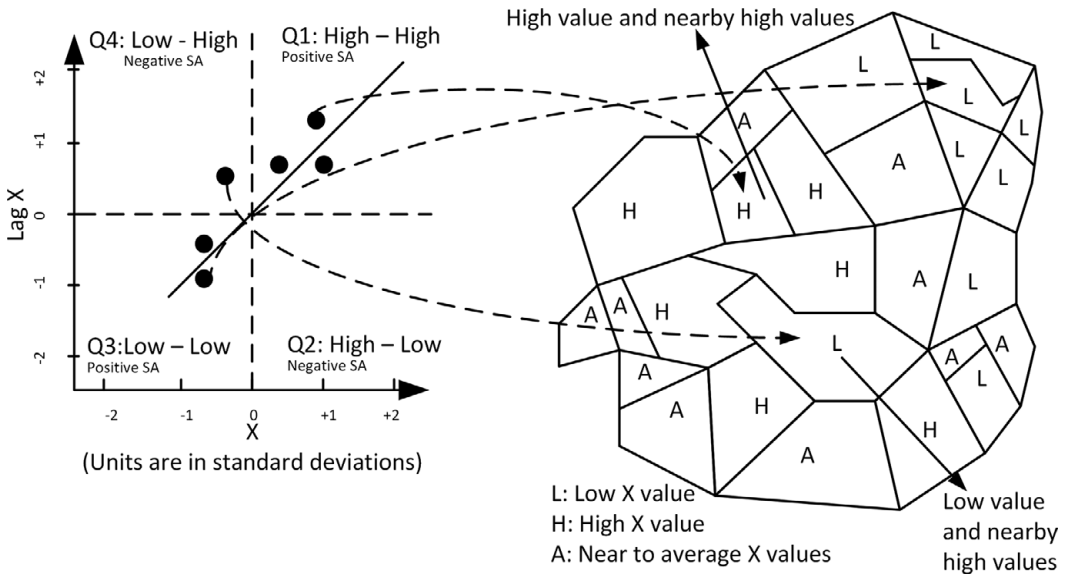


Figure 4.2 Moran's *I* scatter plot. The four quadrants divide the space into four types of spatial autocorrelation. Each dot in the scatter plot stands for one polygon in the map (in this graph, fewer dots are depicted for clarity's sake). Polygons with high values surrounded by polygons with high values are placed in the upper-right quadrant (Q1). Not all surrounding polygons need to have high values, but the more polygons with similar values there are, the stronger their associations.

the origin (0, 0). The produced scatter plot identifies which type of spatial autocorrelation exists according to the place where a dot lies (the dot standing for a spatial entity, such as a polygon). In the case of a polygon layer, a dot in the upper-right corner (Q1) indicates a polygon that has high *X* and high *Lag-X* (also called "High-High"). In other words, this polygon has a high value of *X* and is surrounded by other polygons that also have high values of *X*. That is why the *Lag-X* (average values of these neighboring polygons) is also high. In this case, there is positive spatial autocorrelation. If a dot lies in the lower-left corner (Q3), the polygon has a low value of *X* and is surrounded by polygons with low values of X (i.e., Low-Low). We thus again have positive spatial autocorrelation. A dot in the upper-left corner (Q4) indicates a polygon with low *X* surrounded by polygons with high *X* (i.e., Low-High). This is negative spatial autocorrelation and a strong indication of outlier presence. Finally, a dot in the lower-right corner (Q2) indicates a polygon with high *X* surrounded by polygons with low *X* (i.e., High-Low). There is negative spatial autocorrelation and an indication of outlier presence again.

Dots can also be compared to a superimposed regression line. The slope of the regression line over the data points equals the Moran's *I* index value when calculated using binary and row standardized weights. The closer a dot is to the line, the closer the polygon is to the general spatial autocorrelation trend. The

further away a dot lies from the line, the greater the deviation of this spatial unit from the general trend. Points that deviate greatly from the regression line can be regarded as outliers. Potential outliers with respect to the regression line may function as leverage points that distort the Moran's *I* value. Such observations should be examined further, as in any case of an outlier's presence.

Moran's *I* index values can be bounded to the range −1.0 to +1.0 when the weights are row standardized (see Section 1.9). For most real-world problems, it is hard to find perfectly dispersed (−1) or clustered (+1) patterns. An index score higher than 0.3 is an indication of relatively strong positive autocorrelation, while a score lower than −0.3 is an indication of relatively strong negative autocorrelation (O'Sullivan & Unwin 2010 p. 206).

If we do not row standardize the weights, Moran's *I* Index might have values beyond the [−1,1] boundaries. This typically indicates problems with the tool's parameter settings. Examples of such problems are as follows:

- Values of the attribute in question are skewed. Check the histogram to see if this is the case.
- Some features do not have neighbors or have relatively few. The conceptualization of the spatial relationships or distance band should be checked to fix this problem. In skewed distributions, each object must have at least eight neighbors. This is not always sufficient (in skewed distributions), however, and has to be determined by the user.
- Selecting inverse distance often performs well but may produce very small values, something that should be avoided.
- Row standardization is not applied but should be (i.e,. row standardization is suggested most of the times when data refer to polygons).

Discussion and Practical Guidelines

Moran's *I* index is a global statistic and assesses the overall pattern of a spatial dataset. By contrast, local spatial autocorrelation metrics focus on each spatial object separately within a predefined neighborhood. Statistically significant results for the local Moran's *I* (e.g., detecting clustering) do not imply statistically significant results for the Global Moran's *I*. Although clusters may exist and be evident at the local level, these clusters may remain unnoticed when we examine the pattern at the global level. Global statistics are more effective when there is a consistent trend across the study area. If global statistics fail to reveal a pattern in the spatial distribution, this does not mean that local statistics will perform in similar ways. On the contrary, we should use them to find localized trends and patterns hidden at the global level.

It should also be mentioned that we use neighborhoods for Global Moran's *I* calculation, but this does not make the statistic local. The term "global" implies that a single value for the index is produced for the entire pattern. The term "local" means that a value is produced for each spatial object separately. We can thus map local Moran's *I*, as each spatial object has a local Moran's *I* value,

but we cannot map Global Moran's *I*. To map Global Moran's *I*, we use the Moran's *I* scatter plot.

The neighborhood is also defined in Global Moran's *I*, for two main reasons:

(a) To reduce computational cost: The more objects there are on the data, the more time is required to compute the results.
(b) As we use a distance function (e.g., distance decay), the further away the objects lie, the less impact they have on each other. As a result, there is no need to calculate the index when weights are practically close to zero. Selecting the right cutoff distance is not trivial. It is suggested to start from a distance so that each object has at least one neighbor. In the case of skewed data, each object has to have at least eight neighbors. Alternatively incremental spatial autocorrelation is a useful method of determining the appropriate cut of distance, as explained in Section 4.3.

The following are some practical guidelines:

- Results are reliable if we have at least 30 spatial objects.
- Row standardization should be applied if necessary. Row standardization is common when we have polygons.
- When the *p*-value is low (usually $p < 0.05$), we can reject the null hypothesis of zero spatial autocorrelation:
  (a) If the *z*-score is positive, there is positive spatial autocorrelation (clustering tendency).
  (b) If the *z*-score is negative, there is negative spatial autocorrelation (dispersion tendency).

Finally, potential case studies for which Moran's *I* index could be used include

- Examining if income per capita is clustered (socioeconomic analysis)
- Analyzing consumption behavior (geomarketing analysis)
- Analyzing house values (economic analysis)

4.2.2 Geary's C Index

Definition

Geary's C index is a statistical index used to compute global spatial autocorrelation (Geary 1954) and is calculated by the following formula (4.3):

$$C = \frac{(n-1)}{2\sum_i^n \sum_j^n w_{ij}} \frac{\sum_i^n \sum_j^n w_{ij}(x_i - x_j)^2}{\sum_i^n (x_i - \bar{x})^2} \tag{4.3}$$

where

n is the total number of spatial objects
x_i is the attribute value of feature i, x_j is the attribute value of feature j
$\bar{x}$ is the mean of this attribute

w_{ij} is the spatial weight between feature i and j

$\sum_i^n \sum_j^n w_{ij}$ is the aggregation of all spatial weights

Why Use

To trace the presence of global spatial autocorrelation in a spatial dataset.

Interpretation

Geary's C index varies between 0 and 2. A value of 1 typically indicates no spatial autocorrelation. Values significantly smaller than 1 indicate positive spatial autocorrelation, while values significantly larger than 1 indicate negative spatial autocorrelation (O'Sullivan & Unwin 2010 p. 211).

Discussion and Practical Guidelines

Moran's *I* offers a global indication of spatial autocorrelation, while Geary's C is more sensitive to differences in small neighborhoods (Zhou et al. 2008 p. 69). As such, when we search for global spatial autocorrelation, Moran's *I* is usually preferred over Geary's C.

4.2.3 General G-Statistic

Definition

General G-Statistic is a statistical index used to compute global spatial autocorrelation. The General G-Statistic detects clusters of low values (cold spots) or high values (hot spots) and is an index of spatial association (Getis & Ord 1992, O'Sullivan & Unwin 2010 p. 223).

$$G(d) = \frac{\sum_i^n \sum_j^n w_{ij}(d) x_i x_j}{\sum_i^n \sum_j^n x_i x_j}, \forall j \neq i \tag{4.4}$$

where

n is the total number of observations (spatial objects)
x_i is the attribute value of feature i
x_j is the attribute value of feature j
d is the distance that all pairs (x_i, x_j) lie within
w_{ij} is the spatial weight between feature i and j
$\forall j \neq i$ indicates that features i,j cannot be the same

Why Use

This index is used to distinguish if the positive spatial autocorrelation detected is due to clustering of high values or due to clustering of low values. When clusters of low values coexist with clusters of high values in the same study area, they tend to counterbalance each other. Moran's *I* is more suitable to trace this association on the global scale.

Interpretation

The General G-Statistic is inferential, and its results are interpreted based on a rejection (or not) of the null hypothesis that there is complete spatial randomness and, thus, clusters do not exist. A *z*-score and a *p*-value are calculated along with the expected index value. The expected value is the value that would result if the spatial distribution of the values (of the variable being studied) were the outcome of complete spatial randomness. The difference between the observed and expected values is evaluated based on a *z*-score and a *p*-value that test if this difference is statistically significant.

- When the *p*-value is small (usually $p < 0.05$), the null hypothesis is rejected, and there is statistically significant evidence for clustering:
  (a) If the *z*-value is positive, the observed General G-Statistic value is larger than expected, indicating a concentration of high values (hot spots).
  (b) If the *z*-value is negative, the observed General G-Statistic value is smaller than expected, indicating that low values (cold spots) are clustered in parts of the study area. In both cases, this is an indication of positive autocorrelation.

In cases where the weights are binary or less than 1, the index value is bounded between 0 and 1. This happens because the denominator includes all (x_i, x_j) pairs, regardless of their vicinity. The numerator will be always less than or equal to the denominator. The final outcome regarding spatial association should be concluded only after an examination of the *p*-value and the *z*-score.

Discussion and Practical Guidelines

The General G-Statistic measures the overall (hence "general") clustering of all pairs (x_i, x_j) within a distance d of each other (Getis & Ord 1992). Moran's I index cannot distinguish between high- and low-value clustering, but the General G-Statistic can. On the other hand, the General G-Statistic is appropriate only when there is positive spatial autocorrelation, as it detects the presence of clusters. If the General G-Statistic does not produce statistically significant results, we cannot reject the null hypothesis of complete spatial randomness. Negative spatial autocorrelation might exist through a competitive process (i.e., where high values and low values for the same variable are nearby). As a result, the General G-Statistic is more an index of spatial association or an index of positive spatial autocorrelation, rather than a pure spatial autocorrelation index.

Here are some practical guidelines to follow:

- General G-Statistic works only with positive values.
- A binary weights matrix is more appropriate for this statistic. It is thus recommended to use fixed distance band, polygon contiguity, *k*-nearest neighbors or Delaunay triangulation that produce binary weighting schemes. For example, if we set a fixed distance of 2 km, each object

inside this distance will be a neighbor and have a weight of 1. All other objects lying further away than 2 km will not be neighbors and have zero weight. In this case, row standardization is not necessary, as the weights already lie in a 0-to-1 range.

- When we use binary weighting and fixed distance, the size of the polygons might matter. For example, if large polygons tend to have lower values for the attribute being studied (e.g., population density) than the small polygons, we might obtain higher observed values of the General G-Statistic because more smaller polygons are creating pairs at the same distance set. We would thus obtain higher *z*-values and stronger clustering results than what the real situation would justify.
- It is more common to use the local version of the General G-Statistic index, as it provides the exact locations of the clusters. There are two versions of the Local G-Statistic: the G_i and the G_i^*.

Finally, potential case studies include

- Analyzing PM2.5 in an urban environment (environmental analysis)
- Analyzing educational patterns (demographic analysis)
- Analyzing house rents (economic analysis)

4.3 Incremental Spatial Autocorrelation

Definition

Incremental spatial autocorrelation is a method based on Global Moran's *I* index to test for the presence of spatial autocorrelation at a range of band distances (ESRI 2015).

Why Use

It is used to approximate the appropriate scale of analysis (appropriate analysis distance). Instead of arbitrary selecting distance bands, this method identifies an appropriate fixed distance band for which spatial autocorrelation is more pronounced. In other words, it allows us to identify the farthest distance at which an object still has a significant impact on another one. After the appropriate scale of analysis is established, local spatial autocorrelation indices and other spatial statistics can be calculated more accurately.

Interpretation

Incremental spatial autocorrelation is used to calculate Global Moran's *I* for a series of incremental distances (see Figure 4.3). For each distance increment, the method produces Global Moran's *I*, Expected *I*, variance, a *z*-score and a *p*-value. With this method, we can plot a graph of *z*-scores over an increasing distance.

z-score peaks reflect distances at which a clustering process seems to be occurring. The higher the *z*-score, the stronger the clustering process at that distance. By locating the peak in the graph, the *z*-score and the corresponding distance, we can better define the distance band, which can be used in many spatial statistics such as hot spot analysis (see Section 4.4.2). The distance in the first peak of the *z*-score graph is often selected as the appropriate scale for further analysis (but this is not always the case; see the following discussion).

Discussion and Practical Guidelines

Selecting the appropriate scale of analysis for the problem at hand is one of the most challenging tasks in spatial analysis. The scale of analysis defines the size and shape of the neighborhoods for which spatial statistics are calculated and is closely related to the problem in question (see Section 1.3). Researchers and analysts lacking an in-depth understanding of spatial statistics often tend to apply spatial statistics tools based on the predefined default values (i.e., distance band) of the software being used. The uncritical selection of spatial parameters (e.g., the distance of analysis) leads to incorrect estimates and conclusions. The difficulty is not running a software tool but setting it up properly and interpreting the results based on statistical theory. The graph produced by incremental spatial autocorrelation allows for a statistically sound estimation of the scale of the analysis, which is superior to an arbitrary or intuitive selection.

It is quite common that more than one peak may occur. Different distance bands (peaks) might reveal underlying processes on different scales of analysis. Hypothetically, unemployment clustering statistically significant at 100 m and 1,000 m peaks reflect patterns of clustering at both the census block level and the postcode level. If we are interested only in the census block level, we could apply the 100 m distance in our analysis. Greater distances reflect broader, regional trends (e.g., east to west), while smaller distances reflect local trends

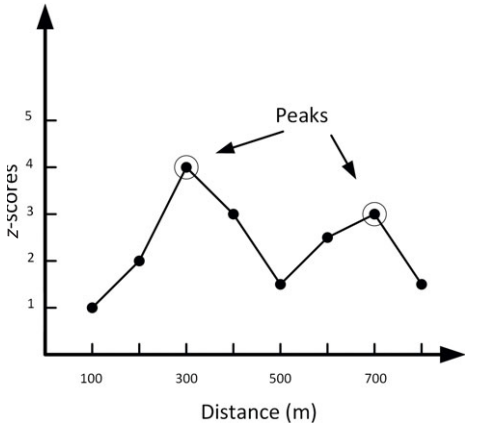


Figure 4.3 *z*-scores over incremental distance.

(e.g., between neighborhoods). For example, the scale of analysis is usually small when the analysis is of children going to school (usually close to their homes), while the scale is larger when analyzing commuting patterns.

There are some practical guidelines regarding this method. An initial distance (the value from which distances will start incrementing) should be set. The initial default distance value should ensure that each object has at least one neighbor. In this case, if locational outliers (objects lying far away from others) exist, the initial distance calculated may be large enough. A large initial distance value results in a graph with no peaks, simply because a peak lies before the beginning distance. In addition, an increment distance should also be set. We can use either a distance increment that fits the needs of the study (i.e., 100 m for a local study) or try the average distance to each feature's nearest neighbor (usually the software's default value). Locational outliers may distort this value, leading to very large increments that might not be representative. For this reason, before running incremental spatial autocorrelation, we should check if locational outliers exist.

To define the appropriate scale of analysis, the following procedure can be applied:

Step 1: Check for locational outliers (see Section 3.1.6 and Exercise 3.4). If no locational outliers exist, perform incremental spatial autocorrelation as described earlier. If locational outliers exist, go to step 2.
Step 2: Select all features except outliers and perform *Incremental Spatial Autocorrelation* (only for the selected features).
Step 3: Locate a peak and keep the relevant distance
Step 4: Create a *Spatial Weights Matrix* for the entire dataset (including the locational outliers) with the distance defined in step 3 (see Section 1.9). Set the *Number of Neighbors* to a value so that each object has at least this number of neighbors.
Step 5: Run spatial statistic (e.g., Local Moran's *I* index) using the *Spatial Weights Matrix* created in step 4.

When locational outliers are removed, the *z*-scores graph might change significantly, yielding a completely different scale of analysis. Through this procedure, *Spatial Weights* are calculated based on the distance threshold that results when outliers are removed. For objects that have no neighbors at this distance, the *Number of Neighbors* parameter will be used instead. This practically means that outliers will be treated differently but will be included in the study so that they do not negatively impact the rest of the objects. As such, local spatial autocorrelation indices will be calculated based on the final weights matrix for all objects in the dataset.

Other practical guidelines include the following (ESRI 2015):

- There might be more than one peak. Each one reflects spatial autocorrelation at different distances. We will typically select the one with the highest *z*-score, which is usually the first. We may also select a peak

(distance) that better reflects the regional or local perspective for the problem at hand.
- Select as the beginning distance the one that will ensure that each object has at least one neighbor.
- Look for locational outliers before setting the initial distance. A locational outlier inflates distance metrics, with negative impacts on the graph. Remove outliers, and run the incremental spatial autocorrelation tool.
- If there are no peaks, we can use smaller or larger distance increments. If we still cannot locate any peaks, we should avoid the incremental spatial autocorrelation method. We should rely on other criteria or our common sense based on previous knowledge to define the appropriate distance band. We can also use optimized hot spot analysis, which enables distance band definition even if no peaks exist (see Section 4.4.3).
- The final distance selected should provide an adequate number of neighbors for each feature and an appropriate scale of analysis.

4.4 Local Spatial Autocorrelation

Global indices of spatial autocorrelation identify whether there is clustering in a variable's values, but they do not indicate where clusters are located. To determine the location and magnitude of spatial autocorrelation, we have to use local indices instead. Local Moran's *I* and local Getis-Ord G_i^* are the most widely used local indices of spatial autocorrelation.

4.4.1 Local Moran's *I* (Cluster and Outlier Analysis)

Definition

Local Moran's *I* is an inferential spatial statistic used to calculate local spatial autocorrelation. For *n* spatial objects in a neighborhood (Anselin 1995), the local Moran's *I* of the *i* object is given as (4.5):

$$I_i = \frac{x_i - \bar{X}}{m_2} \sum_j w_{ij}\left(x_j - \bar{X}\right) \tag{4.5}$$

$$m_2 = \frac{\sum_i \left(x_j - \bar{X}\right)^2}{n} \tag{4.6}$$

where

n is the total number of observations (spatial objects).
x_i is the attribute value of feature *i*.
x_j is the attribute value of feature *j*.
$\bar{X}$ is the mean of this attribute.

$w_{i,j}$ is the spatial weight between feature *i* and *j*.
m_2 is a constant for all locations. It is a consistent but not unbiased estimate of the variance. (Anselin 1995 p. 99).

Tip: Keep in mind that m_2 is actually a scalar that does not affect the significance of the metric as it is the same for all locations (Anselin 2018). In some cases, the formula of this scalar may slightly change. For example, ArcGIS and GeoDa uses $n - 1$ in the denominator instead of n (de Smith et al. 2018).

Why Use

Local Moran's *I* can be used for an attribute to (a) identify if a clustering of high or low values exists and (b) to trace spatial outliers (Grekousis & Gialis 2018).

Interpretation

Local Moran's *I* index is interpreted based on the expected value, a pseudo *p*-value,and a *z*-score under the null hypothesis of no spatial autocorrelation (complete spatial randomness). The expected value for a random pattern is (Anselin 1995 p. 99):

$$E(I_i) = \frac{-1 \sum_j w_{ij}}{n-1} \tag{4.7}$$

where

n denotes the number of spatial entities
w_{ij} is the spatial weight between feature *i* and *j*

The expected value is the value that would have resulted if the specific attribute's values geographical distribution were the result of complete spatial randomness. The observed index value is the local Moran's *I* index value given by Equation (4.5).

Positive local Moran's *I* index values (observed) significantly larger than the expected value indicate potential clustering and positive spatial autocorrelation. Negative local Moran's *I* index values (observed) significantly smaller than the expected value indicate the potential presence of spatial outliers and negative spatial autocorrelation. Values close to the expected value indicate no autocorrelation. To finalize a conclusion regarding spatial autocorrelation's presence, we should evaluate the previously mentioned difference (expected vs. observed) based on the *z*-score and the *p*-value. Using these two metrics, we assess if this difference is statistically significant.

- If the *p*-value is large (usually $p > 0.05$), the results are not statistically significant (even if the difference is large), and we cannot reject the null hypothesis (see Table 4.1). The interpretation is that, *due to a lack of sufficient evidence, we cannot reject the null hypothesis that the spatial distribution of the values is the result of complete spatial randomness.*

Table 4.1 Interpretation of Moran's *I* *p*-values and *z*-scores. *z*-score values are indicative and can be differentiated based on data.

p-value	*z*-score	Interpret
$>\alpha$ (e.g., 0.05)		Can not reject the null hypothesis of complete spatial randomness.
$<\alpha$ (e.g., 0.05)	$z < 0$	Negative spatial autocorrelation. Low negative values (e.g., $z < -2.60$) is an indication of spatial outlier presence.
	$z > 0$	Positive spatial autocorrelation: clustered pattern. Large positive values (e.g., $z > 2.60$) indicate intense clustering of either low (cold spots) or high (hot spots) values.

- A small *p*-value (usually $p < 0.05$) indicates that we can reject the null hypothesis of complete spatial randomness and accept that spatial autocorrelation exists. In this case:
  (a) If the *z*-value is positive, we have positive spatial autocorrelation and clustering.
  b) iI the *z*-value is negative, we have negative spatial autocorrelation, and spatial outliers may exist especially for low *z*-score values.

A high positive *z*-score (e.g., greater than 2.60) for a spatial entity means that the neighboring spatial entities have similar values. If the values are high, then High-High clusters are formed, meaning that spatial entities with high values (for a specific variable) are surrounded by spatial entities of high values (of the same variable). If the values are low, then "Low-Low clusters" are formed, meaning that spatial entities with low values are surrounded by spatial entities of low values. Note that the spatial clusters formed in High-High or Low-Low arrangements depict only the core of a real cluster. This happens because the statistical value for each location is calculated based on the neighboring values. Thus, the locations (e.g., polygons) at the periphery of a cluster might not be assigned to a High-High or Low-Low cluster.

A low negative *z*-score, (e.g., less than -2.60) for a spatial entity indicates dissimilar nearby attribute values and potential spatial outliers. If the spatial entity has a low attribute value, then it is surrounded by features with high values, creating a Low-High arrangement. If a spatial entity has a high attribute value, then it is surrounded by features with low values, creating a High-Low arrangement.

Discussion and Practical Guidelines

Even with complete spatial randomness, clustering or outliers might exist due to randomness. To overcome this problem, we use a Monte Carlo random permutation procedure. Permutations are used to estimate the likelihood of generating, via complete spatial randomness, a spatial arrangement of values similar to the observed one. Using Monte Carlo, we generate multiple random patterns and then compare the results to those of the local Moran's *I* of the

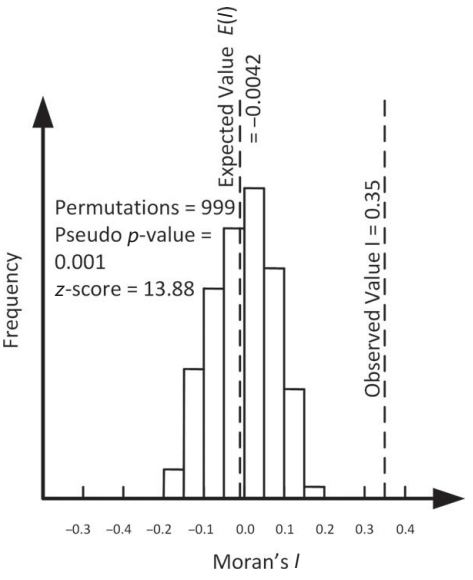


Figure 4.4 Permutation reference distribution for 999 random permutations under complete spatial randomness. The results of this example suggest that the observed value $I = 0.35$ is highly significant and not a result of spatial randomness, as it lies far away from the rest of the values (with a z-score of 13.88) and the expected (theoretical) Moran's I value $E(I) = -0.0042$. The pseudo p-value is 0.001, indicating that none of the 999 random patterns' local Moran's I values surpassed the observed value.

original dataset. By inspecting the reference distribution, we assess how unusual the observed value would be in relation to this randomized benchmark (see Figure 4.4).

For each permutation (i.e., of a total of 999), the values (of the attribute variable) are randomly rearranged around each feature, and the local Moran's I index is calculated. A reference distribution of the local Moran's I index values is then created (see Figure 4.4). The reference distribution should be centered at around zero, as it is supposed to be the result of complete spatial randomness with no spatial autocorrelation. The range of local Moran's I index values, which vary due to randomness, is depicted in the x-axis. If the local Moran's I observed value lies far away from the reference distribution and in relation to the z-score obtained (which quantifies the distance from the mean), we can reject the null hypothesis of complete spatial randomness and accept that the spatial autocorrelation observed is statistically significant.

A pseudo p-value ($p = \frac{R+1}{M+1}$) is calculated as the proportion of how many times (R) the computed local Moran's I values generated by the permutations are equal to or larger than the observed local Moran's I to the number of permutations (M; Anselin 2018). Typically, for 999 permutations, the pseudo p-value is set to 0.001; for 99 permutations, it is set to 0.01. A pseudo p-value cannot

be interpreted as a typical *p*-value, as it is the summary of the reference distribution (Anselin 2018). A pseudo *p*-value of 0.01 (i.e., $p = \frac{0+1}{99+1} = 0.01$), for example, means that none (zero) of the 99 random patterns yielded a local Moran's *I* value equal to or more extreme than the observed data. In other words, no pattern exhibited clustering (or dispersion) equal to or larger than the observed one.

Practical guidelines include the following (ESRI 2018a):

- Results are reliable for at least 30 spatial objects.
- We cannot perform this test for points events (e.g., points as crime incidents, without any attribute fields attached). Nevertheless, we can aggregate data into polygons and then continue with the analysis in the usual fashion (see discussion in Section 4.4.3).
- Each feature has to have at least one neighbor.
- No feature has to have all features as neighbors.
- When values are skewed, each feature should have around eight neighbors or more.
- The conceptualization of spatial relationships, distance bands and distance functions used should be done carefully.
- The false discovery rate (see Section 4.6) can be used to account for multiple comparison problems and spatial dependence.

Potential Case Studies Include

- Analyzing unemployment distribution
- Analyzing income inequalities
- Analyzing house values

4.4.2 Optimized Outlier Analysis

Definition

Optimized outlier analysis is a procedure used to optimally select the parameters of the local Moran's *I* index (ESRI 2018a). Similar to local Moran's *I*, it locates clusters of either high or low values and traces spatial outliers.

Why Use

Optimized outlier analysis is used to overcome the difficulties of setting the parameters of the local Moran's *I* index. Optimized outlier analysis performs an automated preliminary analysis of the data to ensure optimal results. The method is used to

(a) Identify how many locational outliers exist (if any).
(b) Estimate the distance band at which the spatial autocorrelation is more pronounced (scale of analysis) through incremental spatial autocorrelation.
(c) Adjust for spatial dependence and multiple testing through the false discovery rate correction method (see Section 4.6).

(d) Handle point events with no variables attached. These events are automatically aggregated into weighted features within some regions (i.e., grid; see discussion in Section 4.4.3). The weighted variable is then analyzed in the usual fashion.

Interpretation

Optimized outlier analysis applies the local Moran's *I* index, and the results can be interpreted accordingly (see Section 4.4.1).

Discussion and Practical Guidelines

Optimized outlier analysis applies incremental spatial autocorrelation to define the scale of analysis (see Section 4.3). The distance band selected is the one in which a peak occurs in the related graph. If multiple peaks are found, the distance related to the first peak is usually selected. If no peaks occur, the optimized outlier analysis applies a different procedure. The spatial distribution of the features is analyzed by calculating the average distance so that each feature has *K* neighbors. *K* is defined as 5% of the total number of (*n*) features ($K = 0.05 \times n$). *K* is adjusted so that it ranges between 3 and 30. If the average distance that ensures *K* neighbors for each feature is larger than one standard distance, the distance band is set to one standard distance (ESRI 2018a). If it is not, the *K* neighbor average distance reflects the appropriate scale of analysis. Finally, optimized outlier analysis is effective even for data samples. It is also effective in case of oversampling, as the associated tools have more data with which to compute accurate results.

4.4.3 Getis-Ord G_i and G_i^* (Hot Spot Analysis)

Definition

Getis-Ord G_i index and G_i^* index (pronounced G-i-star) comprise a family of statistics that identify statistically significant clusters of high values (hot spots) and clusters of low values (cold spots) and are used as measures of spatial association (Getis & Ord 1992, Ord & Getic 1995, O'Sullivan & Unwin 2010 p. 219). The process is also named hot spot analysis. The G_i index is given as (4.8):

$$G_i(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j}, j \neq i \tag{4.8}$$

where

d is the estimated range of observed spatial autocorrelation
$\sum_j w_{ij}(d)$ is the sum of weights for $j \neq i$ within distance d
n is the total number of observations
x_j is the attribute value of feature j

Getis-Ord G_i^* index is given as (4.9):

$$G_i^*(d) = \frac{\sum_j w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j} \tag{4.9}$$

Note that in contrast to G_i, in G_i^* the restriction $\forall j \neq i$ is lifted. In other words, the index takes into account the attribute value x_i in location i.

Why Use

Hot spot analysis identifies if low or high values of a variable are spatially clustered and create cold spots or hot spots respectively.

Interpretation

For each polygon, a *z*-score value is calculated along with a *p*-value to assess the statistical significance of the results. The null hypothesis is that *There is complete spatial randomness of the values associated with the features*. Having a high positive *z*-score and a small *p*-value is an indication of spatial clustering of high values (i.e., a hot spot), whereas having a low negative *z*-score with a small *p*-value reveals the presence of cold spots (spatial clustering of low values). In both cases, there is positive spatial autocorrelation. The higher the *z*-score (either positive or negative), the more intense the clustering at hand. *z*-scores values near to zero typically indicate no spatial clustering. When *p*-values are larger than 0.05 (or larger than another established significance level), the null hypothesis cannot be rejected, and the results are not statistically significant. Nonsignificant results mean that there is no indication of clustering, as the process at hand might be random. The results can be rendered in a map with three confidence level classes (99%, 95% or 90%) for hot spot polygons, three classes (99%, 95% or 90%) for cold-spot polygons and another class for rendering polygons with nonsignificant results.

Discussion and Practical Guidelines

G_i^* is used more widely than G_i. The use of G_i^* is also linked to hot spot analysis. Hot spot analysis is mainly used to identify if clusters of values of a specific variable are formed in space (Grekousis 2018). It is most commonly used with polygon features. For point features, though, it is more helpful to study the intensity of the objects rather than a specific attribute. In this respect, we use hot spot analysis to identify if hot spots or cold spots of events' intensity exist. Such analysis should begin by aggregating points into some regions (e.g., postcodes, census tracts). This can be easily done by overlaying the relevant (administrative) polygon layers and applying typical GIS techniques (such as spatial join to count how many points lie within each polygon). Alternatively, we can create a grid (in the absence of a polygon layer, or to avoid at some extent the modifiable areal unit problem – see Chapter 1) by using a fishnet tool and then perform spatial join. The grid size should be set in such a way that

most of the grids have more than one incident. The new attribute field created, containing the number of points per polygon/grid, can be used as the attribute field to be analyzed, similarly to any other polygon layer.

Some practical guidelines for the G_i^* (ESRI 2018a):

- Results are reliable if we have at least 30 objects.
- Fixed distance is recommended for the G_i^* index. The appropriate distance should be determined by using incremental spatial autocorrelation or optimized hot spot analysis (see Section 4.4.4). In case of locational outliers, a fixed distance band can be combined with a minimum number of neighbors per spatial feature. In this case, when the fixed distance band leaves some polygons with no neighbors, the minimum number of neighbors ensures that all polygons will have at least a specific number of neighbors.
- The false discovery rate (see Section 4.6) can also be used here, as in the case of Local Moran's *I*.
- This index cannot be applied for point objects (e.g., crime incidents), without any attribute fields attached. However, we can aggregate data by using spatial join to polygons and then continue with the analysis.
- The spatial relationships, distance bands and distance functions used should be conceptualized carefully.

Potential case studies include

- Human geography/demographics: Are there any areas where the unemployment rate form spatial clusters?
- Economic geography: How is income spatially distributed? Are there cold or hot spots?
- Health analysis: Are there any unusual patterns to heart attacks?
- Voting pattern analysis: Do people in favor of a specific party cluster together?

4.4.4 Optimized Hot Spot Analysis

Definition

Optimized hot spot analysis is a procedure used to optimally select the parameters of the Getis-Ord G_i^* index (ESRI 2018a). Similar to Getis-Ord G_i^*, it locates spatial clusters of low values (cold spots) and spatial clusters of high values (hot spots).

Why Use

Optimized hot spot analysis is used to overcome the difficulties of setting the parameters of the Getis-Ord G_i^* index. Optimized hot spot analysis performs hot spot analysis using the Getis-Ord G_i^* index in an automated way to ensure optimal results. The method is used to

(a) Identify how many locational outliers exist (if any).
(b) Estimate the distance band at which the spatial autocorrelation is more pronounced (scale of analysis).
(c) Adjust for spatial dependence and multiple testing through the false discovery rate correction method (see Section 4.6).
(d) Handle point events with no variables attached. These events are automatically aggregated into weighted features within some regions (i.e., grid; see discussion in Section 4.4.3). The weighted variable is then analyzed in the usual fashion.

Interpretation

Optimized hot spot analysis applies the Getis-Ord G_i^* index and results can be interpreted as the Getis-Ord G_i^* results are interpreted (see Section 4.4.3).

Discussion and Practical Guidelines

Optimized hot spot analysis applies incremental spatial autocorrelation to define the scale of analysis (see Section 4.3). This is done in the same way as that described for optimized outlier analysis (see Section 4.4.2).

As mentioned, for point features with no other attributes attached, optimized hot spot analysis aggregates the points into zones and identifies potential event concentration (clustering) or dispersion across space. It can thus be regarded as an alternative approach to point pattern analysis that indicates if spatial autocorrelation among point events is evident (see Section 3.2.1). Finally, the scale of analysis resulting from the optimized hot spot analysis can be applied in kernel density estimation as an alternative way to select the bandwidth h (see Section 3.2.4).

4.5 Space–Time Correlation Analysis

4.5.1 Bivariate Moran's *I* for Space–Time Correlation

Definition

Bivariate Moran's *I* measures the degree to which a variable in a specific location is correlated with the spatial lag (average value at nearby locations) of a different variable (Anselin 2018). A special case of Bivariate Moran's *I* (and a more useful one) occurs when a single variable (instead of two different variables) is used for two different time stamps. It measures the degree of the spatiotemporal correlation of a single variable.

Why Use

Bivariate Moran's *I* index is used to assess how the linear association (positive or negative) of two distinct variables varies in space. The extension of Bivariate

Moran's *I* for space–time correlation is used to trace if spatiotemporal autocorrelation exists for the same variable.

Interpretation

In the case of row standardized weights, Bivariate Moran's *I* for space–time correlation would lie between −1 and 1, with values close to zero indicating no correlation, values close to 1 strong positive correlation and values close to −1 strongly negative correlation. The significance of the statistic is determined through a permutations approach.

Discussion and Practical Guidelines

Before discussing the bivariate extension to time (single variable), let us briefly describe how the standard Bivariate Moran's *I* index works through the Bivariate Moran scatter plot. As explained before, Bivariate Local Moran's *I* analyzes two distinct variables. For instance, suppose we study the potential association between income at a specific location and land prices in surrounding areas. A Bivariate Moran's *I* scatter plot would relate the income values for each location (Income, horizontal axis) to the average land value at nearby locations (W_land, vertical axis; see Figure 4.5A). However, bivariate spatial correlation does not take into account the inherent correlation between the two variables (i.e., income and land price at the same location; Anselin 2018). Leaving unaccounted this correlation makes Bivariate Moran's *I* hard to interpret; this often leads to incorrect conclusions, as the statistic may overestimate the spatial effect of the correlation, which might be merely the result of the same location correlation (Anselin 2018). Thus, Bivariate Moran's *I* is more useful when time is included.

More analytically, a particular case of Bivariate Moran's *I* spatial correlation occurs when the correlation is calculated for variable *X* in a location with *Lag-X* within a time interval. *Lag-X* is the average value of *X* in a nearby location but in a previous time stamp (see Figure 4.5B). This is the Bivariate Moran's *I* for space–time correlation. Conceptually, this approach explains how neighboring values in a previous period affect the present value (Anselin 2018). To put it slightly differently, this approach would explain how the value at a location in a subsequent time is affected by the average values at nearby locations in a previous time. It can be regarded as the inward diffusion from the neighbors at a specific point in time to the core in the future. Switching the selection of variable settings in the scatter plot axes produces a scatter plot with $X(t-1)$ in the *x*-axis and the lag$X(t)$ in the *y*-axis. This approach studies how a location in a previous time affects the values of nearby locations in the future. It can be seen as an outward diffusion originating from the core at a specific time to the neighbors in the future (Anselin 2005). Although this approach is formally correct, the results might be misleading (Anselin 2018), mainly because the notion of spatial autocorrelation refers to how neighbors affect the value of a central location and not the contrary (Anselin 2018). These approaches are

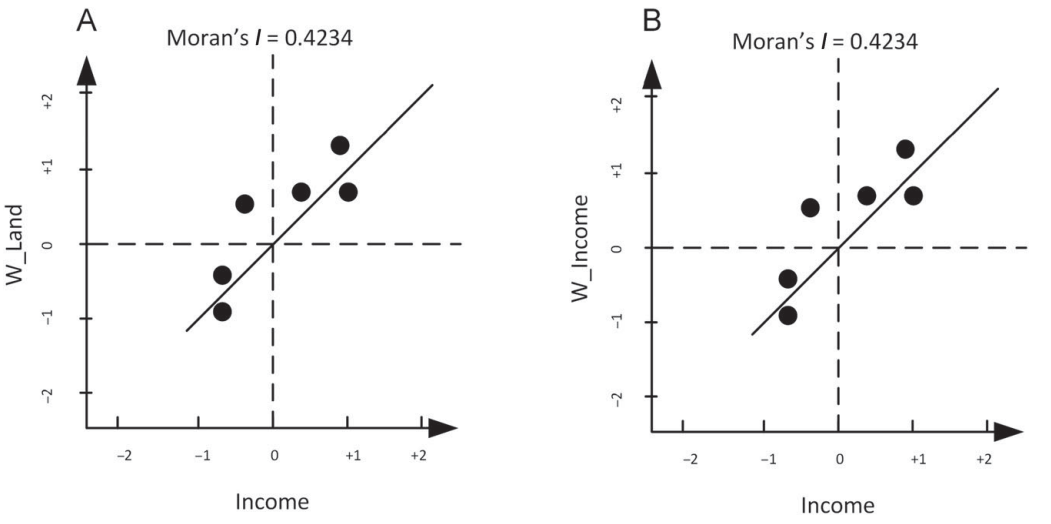


Figure 4.5 (A) Bivariate Moran's *I* for income and land price. The graph indicates that income (*x*-variable) is correlated with the weighted average value of land (*y*-W_Land) within its neighborhood. All variables are expressed in standardized forms, with zero mean and a variance of one. Spatial weights are also row standardized (Anselin 2018). (B) Spatiotemporal Bivariate Moran's *I*. Moran's *I* calculates the correlation of variable *X* in a location with *Lag-X* with the previous time stamp. The proper interpretation is that a value at a location for income (*x*-variable) is correlated with the weighted income value of its neighbors in a previous time (Anselin 2018).

slightly different, and the best one to use depends on the problem at hand and the underlying process being studied.

4.5.2 Differential Moran's *I*

Definition

For two time stamps, **Differential Moran's *I*** tests whether a variable's change at a specific location is related to the change of the same variable in neighboring locations.

Why Use

Differential Global or Local Moran's *I* is used to identify if changes over time are spatially clustered.

Interpretation

As with the interpretation of Moran's *I*, if a high change in a variable's value between two time stamps for a specific location is accompanied by a high change of the same variable in the surrounding area, there is positive spatial autocorrelation of the High-High type (i.e., hot spots) (see Figure 4.6). In other words, the change of a variable's value in a specific location follows a

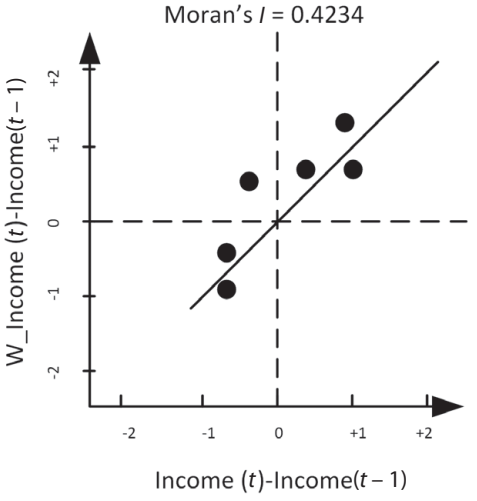


Figure 4.6 Differential Moran's I. The *x*-variable is the difference in the variable between two time stamps. The *y*-variable is the spatial lag of this difference (weighted difference calculated as the average difference for the nearby locations). A high value of the statistic indicates that changes in the variable cluster over time (and space).

trend similar to the change of the same variable in the neighboring area. If low changes in a variable's value between two time stamps are surrounded by low changes, then a Low-Low type of cluster is formed (i.e., cold spots). If a low change in a variable's value between two time stamps for a specific location is accompanied by a high change of the same variable in the surrounding area, there is negative spatial autocorrelation of the Low-High type. A High-Low type of negative spatial autocorrelation emerges when high values are surrounded by low values. In other words, the change in a location does not follow a trend similar to that followed by the neighboring area. Spatiotemporal outliers can also be traced in the case of High-Low or Low-High formations.

Discussion and Practical Guidelines

Differential Moran'sI is more descriptive in a spatiotemporal context than is mapping the local Moran's *I* index of a variable for each time stamp.

4.5.3 Emerging Hot Spot Analysis

Definition

Emerging hot spot analysis identifies spatiotemporal clusters in a spatial dataset (point counts or attribute values) using the Getis-Ord G_i^* statistic (hot spot analysis; ESRI 2017).

Why Use

A cluster may exist throughout the entire period, diminish after a specific time stamp, emerge after some other time stamp or disappear at some other point in time. Emerging hot spot analysis is used to trace such types of different hot spots or cold spots through time.

Interpretation

Emerging hot spot analysis groups locations according to the definitions as presented in Table 4.2 (ESRI 2017).

Discussion and Practical Guidelines

Emerging hot spot analysis is very useful for locating fluctuations in the density, distribution and total count of events in temporal point data. For example, crime events might concentrate on specific locations during the day and on other locations during the night. Through appropriate analysis, measures and related policies may be implemented to better handle the problems being studied.

4.6 Multiple Comparisons Problem and Spatial Dependence

Multiple Comparisons Problem

The multiple comparisons problem, also known as the multiple testing problem, is the problem of getting false significant results in multiple hypothesis testing. Local spatial statistics rely on tests conducted for every single spatial feature in the dataset. As multiple inferences (tests) are drawn for the same set of spatial features, there is a probability that some results will be declared statistically significant by chance, something that should be controlled (Mitchell 2005, Caldas de Castro & Singer 2006, ESRI 2018c). The multiple comparisons problem is a Type I error (see Section 2.5.5). In this type of error, we reject the null hypothesis when, in fact, it is true. In the context of geography, the more spatial objects, the more likely it is that some will be misclassified as statistically significant when a hypothesis is tested.

For example, if we run a test to detect spatial outliers and a spatial feature gets a *p*-value of 0.04, this spatial feature would be a spatial outlier based on statistically significant results at the 95% confidence level. However, there would be a 5% chance that this feature is not an outlier. When we run multiple statistical tests for a few spatial features, the multiple comparisons problem is not very severe. For many objects, however, the multiple comparisons problem is significant. Detecting spatial outliers in 10,000 spatial features infers 10,000 hypothesis tests and 10,000 *p*-values (one test and one *p*-value per object). For a 95% confidence level, the likelihood for each object to pass the test is 5%. In other words, 500 objects might be found to be significant by chance, largely altering the conclusions to be drawn.

Table 4.2 Emerging hot spot analysis groups.

Pattern	Description when statistically significant
No pattern	There is no indication for any hot or cold spots through the entire study period
New hot spot	This location is a new hot spot at the last time stamp. Before that, there was no pattern in this location.
Consecutive hot spot	This location is a hot spot for the final time steps.
Intensifying hot spot	This location is a hot spot for 90% of the time intervals including the last step. In addition, in each time step, the intensity of clustering is increasing.
Persistent hot spot	This location is a hot spot for at least 90% of the time intervals but with no notable fluctuations in the intensity of clustering.
Diminishing hot spot	A location that is a hot spot for at least 90% of the time intervals including the last one. The intensity of clustering is decreasing overall.
Sporadic hot spot	A location that is a hot spot for less than 90% of the time intervals. For none of the time intervals has this location been a statistically significant cold spot.
Oscillating hot spot	A hot spot for the final time step that has also been cold spot in some other time intervals.
Historical hot spot	This location is not a hot spot for the most recent time intervals. Still, it has been traced as statistically significant hot spot for at least 90% of the past time intervals.
New cold spot	This location is a new cold spot at the last time stamp. Before that, there was no pattern in this location.
Consecutive cold spot	This location is a cold spot at the final time steps.
Intensifying cold spot	This location is a cold spot for 90% of the time intervals including the last step. In addition, in each time step, the intensity of clustering is increasing.
Persistent cold spot	This location is a cold spot for at least 90% of the time intervals but with no notable fluctuations in the intensity of clustering.
Diminishing cold spot	A location that is a cold spot for at least 90% of the time intervals including the last one. The intensity of clustering is decreasing overall.
Sporadic cold spot	A location that is a cold spot for less than 90% of the time intervals. For none of the time intervals has this location been a statistically significant hot spot.
Oscillating hot spot	A cold spot for the final time step that has also been a hot spot in some other time intervals
Historical hot spot	This location is not a cold spot for the most recent time intervals. Still, it has been traced as statistically significant cold spot for at least 90% of the past time intervals.

Spatial Dependency

According to the first law of geography, spatial entities that are closer tend to be more similar than those lying further away. This is what we call spatial dependency (see Section 1.3). In local spatial statistics, spatial dependency is

highly likely to seem more evident than it really is. The reason is that local spatial statistics are calculated using the neighboring values of each spatial feature (using a spatial weights matrix). However, features that are near each other are likely to share common neighbors as well, leading to an overestimation of spatial dependence and an artificial inflation of statistical significance.

Dealing with Multiple Comparisons Problem and Spatial Dependence

Two approaches can be used to handle the multiple comparisons problem and spatial dependence:

Bonferroni correction (Bonferroni 1936): This correction divides the alpha significance level by the number of tests (in spatial analysis this equals the number of features). For example, for ten tests and alpha = 0.05, tests with *p*-values smaller than 0.05/10 = 0.005 are statistically significant. In other words, the *p*-value at which a result is declared statistically significant is stricter.

False discovery rate (FDR) correction: FDR correction has been particularly influential in statistics and has been applied to other research areas as well – e.g., genetics, biochemistry (Benjamini & Hochberg 1995, Benjamini 2010). False discovery rate correction is used to account for both spatial dependency and the multiple comparisons problem. It lowers the *p*-value at which a statistic is regarded as significant. FDR correction estimates the number of objects misclassified (false positive error, rejects the null hypothesis) for a given confidence level and then adjusts the critical *p*-value. Statistically significant *p*-values (less than alpha) are ranked from smallest (strongest) to largest (weakest). FDR calculates the expected error in rejecting the null hypothesis (false positive) and, based on this estimate, the weakest objects are eliminated. Within the spatial statistics context, applying FDR correction reduces the number of features with statistically significant *p*-values.

Many statisticians recommend ignoring both the multiple comparisons problem and the spatial dependence problem. For a small number of spatial objects (say, fewer than 100), few objects are likely to be misclassified, so correction may not be necessary. As the number of objects increases, correction should be considered. As software tools for applying corrections are readily available, it is more rational to utilize them and then compare the results with the non-corrected outputs. For example, applying FDR correction in hot spot analysis will probably reduce the features assigned to clusters relative to hot spot analysis without correction. It is advised to test which features are not included, along with their attributes and their neighbors. Finally, we should keep in mind that, even with corrections, we might still experience false results. The question of whether the FDR or Bonferroni correction should be applied in the calculation of spatial statistics depends on the problem, the knowledge of the study area, the intuition of the researcher and the results produced with and without corrections.

4.7 Chapter Concluding Remarks

- Spatial autocorrelation is the degree of spatial dependency, association or correlation between the value of an observation of a spatial entity and the values of neighboring observations of the same variable.
- A major difference with statistical correlation is that, while statistical correlation refers to two distinct variables with no reference to location (or for the same location), spatial autocorrelation refers to a value of a single variable at a specific location in relation to the values of the same variable to its neighboring locations.
- *Lag-X* is the weighted average values of *X* in a specified neighborhood.
- There are four types of arrangement in a Moran's *I* scatter plot: High-High and Low-Low, expressing positive spatial autocorrelation, and High-Low and Low-High, indicating negative spatial autocorrelation.
- Obtaining statistically significant results for the Local Moran's *I* (e.g., detecting clustering) does not mean that we will obtain statistically significant results for the Global Moran's *I* as well.
- When calculating spatial autocorrelation, "global" implies that a single value for the index is produced for the entire pattern, while "local" means that a value is produced for each spatial object separately.
- While Moran's *I* index cannot distinguish between high- or low-value clustering, the General G-Statistic can. On the other hand, the General G-Statistic is appropriate only when there is positive spatial autocorrelation, as it detects the presence of hot or cold spots.
- General G-Statistic is more an index of spatial association, or an index of positive spatial autocorrelation, than a pure spatial autocorrelation index.
- In incremental spatial autocorrelation, by locating the peak in the graph, the *z*-score and the corresponding distance, we can better define the distance band to be used in many spatial statistics such as hot spot analysis.
- More than one peak may occur. This is not wrong. Different distance bands (peaks) might reveal underlying processes at different scales of analysis.
- Smaller distances are often more suitable for geographical analysis at the local scale.
- Before running incremental spatial autocorrelation, we should check if locational outliers exist and remove them if necessary.
- Local Moran's *I* is used to identify if clusters or outliers exist in the spatial dataset. That is why this method is also called "cluster and outlier analysis."
- When calculating local spatial autocorrelation, permutations are used to estimate the likelihood of generating, through complete spatial randomness, a spatial arrangement of values similar to the observed one.

- Hot spot analysis cannot be used to locate outliers.
- Hot spot analysis cannot be directly applied to point objects (e.g., crime incidents) without any attribute fields attached. However, we can aggregate data by using spatial join to polygons and then continue the analysis in the usual fashion.
- Optimized hot spot analysis is a procedure used to optimally select the parameters of the Getis-Ord G_i^* index.
- It is easier to use optimized hot spot analysis instead of the Getis-Ord G_i^* index, as long as we comprehend the outputs.

Questions and Answers

The answers given here are brief. For more thorough answers, refer back to the relevant sections of this chapter.

Q1. What is spatial autocorrelation? What types of spatial autocorrelation exist? Which are the most commonly used metrics?

A1. Spatial autocorrelation is the degree of spatial dependency, association or correlation between the value of an observation of a spatial entity and the values of neighboring observations of the same variable. There are two types of spatial autocorrelation, namely global and local. Global spatial autocorrelation measures autocorrelation by a single value for the entire study area. To estimate spatial autocorrelation at the local level, we use local measures of spatial autocorrelation. The most common global spatial autocorrelation measures are the Moran's *I* index and the General G-Statistic. Local measures of spatial autocorrelation are the Local Moran's *I* index, the Getis-Ord G_i and the Getis-Ord G_i^* statistic.

Q2. Why is spatial autocorrelation important to geographical analysis and spatial statistics?

A2. Spatial autocorrelation analysis is extremely important in geographical studies. If spatial autocorrelation did not exist, geographical analysis would be of little interest. In conventional statistics, the observed samples are assumed to be independent. In the presence of spatial autocorrelation, this assumption is violated. Observations are now spatially clustered or dispersed. This typically means that classical statistical tools are no longer valid. For example, linear regression would lead to biased estimates or exaggerated precision. As such, spatial statistics should be used instead.

Q3. What is incremental spatial autocorrelation, and why is it used?

A3. Incremental spatial is a method based on Global Moran's *I* index to test for the presence of spatial autocorrelation at a range of band distances. It is used to approximate the appropriate scale of analysis. Instead of

arbitrarily selecting distance bands, this method identifies an appropriate fixed distance band, for which spatial autocorrelation is more pronounced. In other words, it allows us to identify the farthest distance at which an object still has a significant impact on another one. After the appropriate scale of analysis is established, local spatial autocorrelation indices and other spatial statistics can be calculated more accurately.

Q4. Why is a Moran's *I* scatter plot used?

A4. It is used to visualize the spatial autocorrelation statistic. It allows for a visual inspection of the spatial associations in the neighborhood of each observation (data point). In other words, it provides a representation used to assess how similar an attribute value at a location is to its neighboring ones. The slope of the regression line over the data points equals the Moran's *I* index value when calculated using binary and row standardized weights. The produced scatter plot identifies which type of spatial autocorrelation exists according to the place where a dot lies (the dot standing for a spatial entity, such as a polygon).

Q5. How can we identify a spatial outlier with a Moran's scatter plot? What does a Low-High arrangement mean?

A5. We can identify a spatial outlier by inspecting a Moran's *I* scatter plot in locations where High-Low or Low-High concentrations exist. A Low-High arrangement means that a spatial object depicted as a dot in the scatter plot has a low value (for the variable studied) and is surrounded by spatial objects with high values. It is probably a spatial outlier, but further analysis should be carried out to confirm it.

Q6. Why is it necessary to set up the right scale of analysis?

A6. The scale of analysis defines the size and shape of the neighborhoods for which spatial statistics are calculated. The scale of analysis is closely related to the problem in question. Hypothetically, unemployment clustering statistically significant at 100 m and 1,000 m peaks reflects patterns of clustering at both the census-block level and the postcode level. If we are interested only in the census-block level, we could apply the 100 m distance in our analysis. Greater distances reflect broader, regional trends (e.g., east to west), while smaller distances reflect local trends (e.g., between neighborhoods). If we use a large scale of analysis, when we are looking at a local level, we might generalize and lose hidden spatial heterogeneity.

Q7. What is cluster and outlier analysis? How can we interpret the results of the index used?

A7. It is an analysis applying local Moran's *I* to (a) identify if a clustering of high or low values exists and (b) to trace spatial outliers. If the *p*-value is large (usually $p > 0.05$), the results are not statistically significant. A small *p*-value (usually <0.05) indicates that we can reject the null hypothesis of complete spatial randomness and accept that spatial autocorrelation exists. In this case, when *z*-value is positive, we have positive spatial autocorrelation and clustering. If *z*-value is negative, we have negative

spatial autocorrelation and a dispersed pattern of values. A high negative value is an indication of a spatial outlier.

Q8. What is a High-High or Low-Low cluster in a Local Moran's *I*?

A8. A high positive *z*-score (e.g., greater than 2.60) for a spatial entity means that the neighboring spatial entities have similar values. If the values are high, then High-High clusters are formed, meaning that spatial entities with high values (for a specific variable) are surrounded by spatial entities of high values (of the same variable). If the values are low, then Low-Low clusters are formed, meaning that spatial entities with low values are surrounded by spatial entities of low values.

Q9. What is hot spot analysis, and what are a cold spot and a hot spot? Can this analysis be used with point data?

A10. Hot spot analysis identifies if low or high values of a variable are spatially clustered and create cold spots or hot spots respectively. In case of point features, it might be more interesting to study the intensity of the objects rather than a specific attribute. In this respect, we use hot spot analysis to identify if hot spots or cold spots of events' intensity exist. Such analysis should begin by aggregating points into some regions (e.g., postcodes, census tracts).

Q10. What are the main benefits of using optimized hot spot analysis?

A10. Optimized hot spot analysis performs hot spot analysis using the Getis-Ord G_i^* index in an automated way to ensure optimal results. The method is used to

(a) Identify how many locational outliers exist (if any).
(b) Estimate the distance band at which the spatial autocorrelation is more pronounced (scale of analysis).
(c) Adjust for spatial dependence and multiple testing through the false discovery rate correction method.
(d) Handle point events with no variables attached. These events are automatically aggregated into weighted features within some regions (i.e., grid). The weighted variable is then analyzed in the usual fashion.

LAB 4
SPATIAL AUTOCORRELATION

Overall Progress

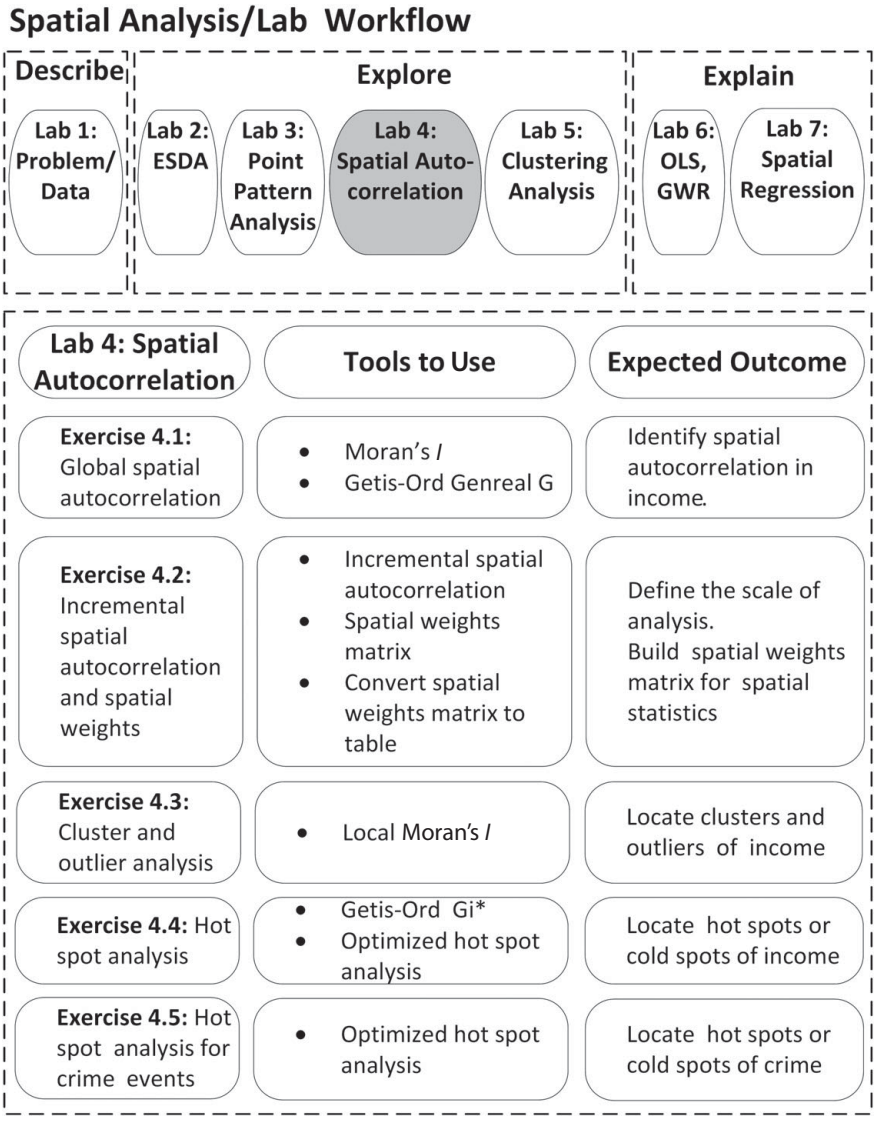


Figure 4.7 Lab 4 workflow and overall progress.

Scope of the Analysis

This lab deals with

- **Objective 1:** Locating high income areas (see Table 1.2)
- **Objective 2:** Locating low crime areas

We further analyze the income distribution in the city to identify whether spatial clustering and spatial autocorrelation exist and also to locate income hot spots through spatial statistics (see Figure 4.7). Moreover, we will study the spatial autocorrelation patterns of crime by locating cold and hot spots.

Section A ArcGIS

Exercise 4.1 Global Spatial Autocorrelation

In this exercise, we calculate the global spatial autocorrelation of income using the Moran's *I* index and the Getis-Ord General G-Statistic.

ArcGIS Tools to be used: Spatial Autocorrelation (Moran's I), High/Low Clustering (Getis-Ord General G))

ACTION: Calculate Global Moran's I

Navigate to the location you have stored the book dataset and click on Lab4_SpatialAutocorrelation.mxd

Main Menu > File > Save As > My_Lab4_SpatialAutocorrelation.mxd

In I:\BookLabs\Lab4\Output

ArcToolBox > Spatial Statistics Tools > Analyzing Patterns > Spatial Autocorrelation (Moran's I)

Input Feature Class = City (see Figure 4.8)

Input Field = Income

Generate Report = Check the box

Conceptualization of Spatial Relationships = INVERSE_DISTANCE (See Chapter 1 for theory.)

Distance: EUCLIDEAN_DISTANCE

Standardization = ROW (See Chapter 1 for theory. We should use ROW when we have polygons and data aggregated at this level.)

Distance Band or Threshold Distance = Leave blank. This is a cutoff distance for Inverse Distance and Fixed Distance conceptualization methods. Features outside the specified cutoff value for a target feature are ignored. The tool uses by default the distance ensuring that each single feature has at least one

Exercise 4.1 (*cont.*)

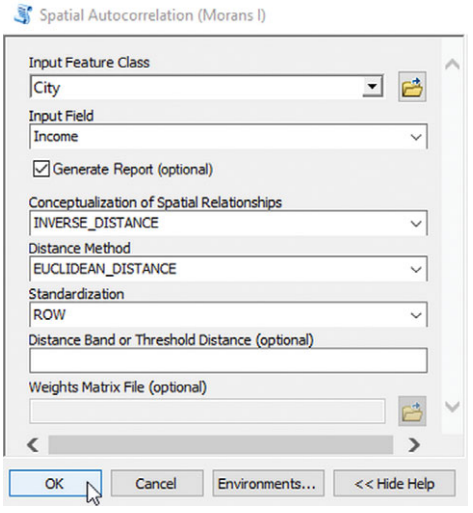


Figure 4.8 Global Moran's *I* tool.

neighbor. This distance is not necessarily the appropriate one. We can use this value to begin with and progressively increase it to test how values of the index will vary. See Chapter 1 for theory.

OK

Main Menu > Geoprocessing > Results > Current Session > Spatial Autocorrelation > DC on MoransI_Result.html (see Figure 4.9)

Interpreting results: The Moran's *I* Index is 0.61 (see Figure 4.9). Given the *z*-score of 10.86 and the *p*-value of 0.000000, there is a less than 1% likelihood (significance level α) that this clustered pattern is the result of random chance. In other words, we have a 99% probability (confidence level) that the distribution of income forms a clustered pattern. A slightly different way to interpret the results is as follows: The spatial arrangement of the income values has a tendency to cluster, and there is a likelihood of less than 1% that this pattern is the result of random chance. The distance threshold defined by the tool is set to 1050.79 m, so every postcode has at least one neighbor. Global Moran's provides a first indication of income clustering. However, we cannot locate where the clustering occurs just by using this index. Moreover, although we calculated spatial autocorrelation, we have not yet defined the appropriate scale of analysis.

Exercise 4.1 (*cont.*)

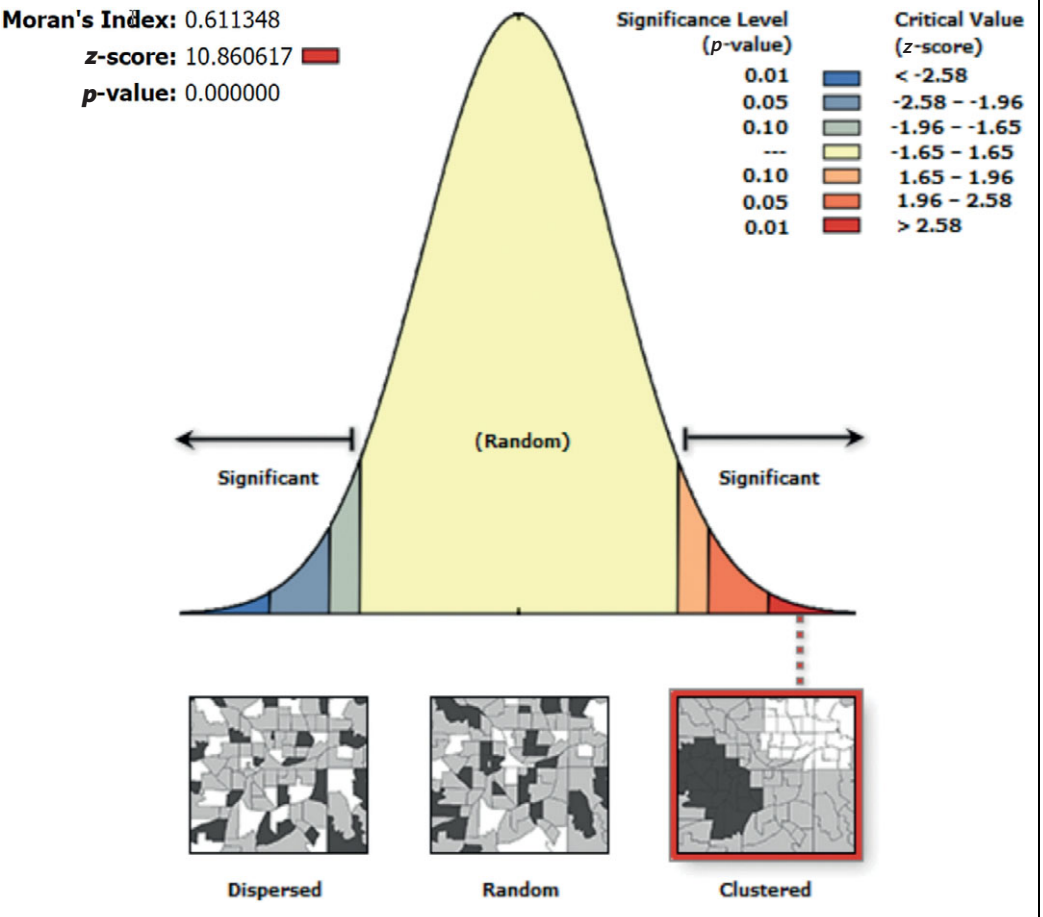


Figure 4.9 Global Moran's *I* report.

ACTION: Calculate High/Low Clustering (Getis-Ord General G)

ArcToolBox > Spatial Statistics Tools > Analyzing Patterns > High/Low Clustering (Getis-Ord General G)

Input Feature Class = City (see Figure 4.10)

Input Field = Income

Generate Report = Check the box

Conceptualization of Spatial Relationships = INVERSE_DISTANCE (See Chapter 1 for theory)

Exercise 4.1 (*cont.*)

`Distance: EUCLIDEAN_DISTANCE`

`Standardization = ROW`

`Distance Band or Threshold Distance = Leave blank`

`OK`

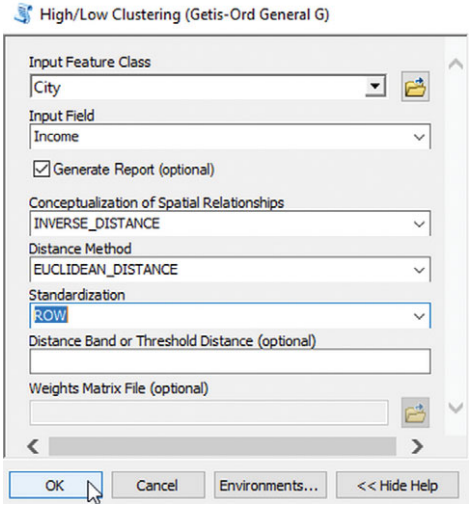


Figure 4.10 Getis-Ord General G tool.

`Main Menu > Geoprocessing > Results > Current Session > High/Low Clustering (Getis-Ord General G)> DC on GeneralG_Results.html (see Figure 4.11)`

`Main Menu > File > Save`

Interpreting results: The Getis-Ord General G Index is 0.01 (see Figure 4.11). Given the *z*-score of 7.052 (positive value) and the *p*-value of 0.000020, there is a less than 1% likelihood (significance level α) that this clustered pattern of high values is the result of random chance. In other words, we have a 99% probability (confidence level) that the distribution of income forms a clustered pattern of high values. As does the Global Moran's I, the Getis-Ord General G provides a first indication of income clustering.

Exercise 4.1 (*cont.*)

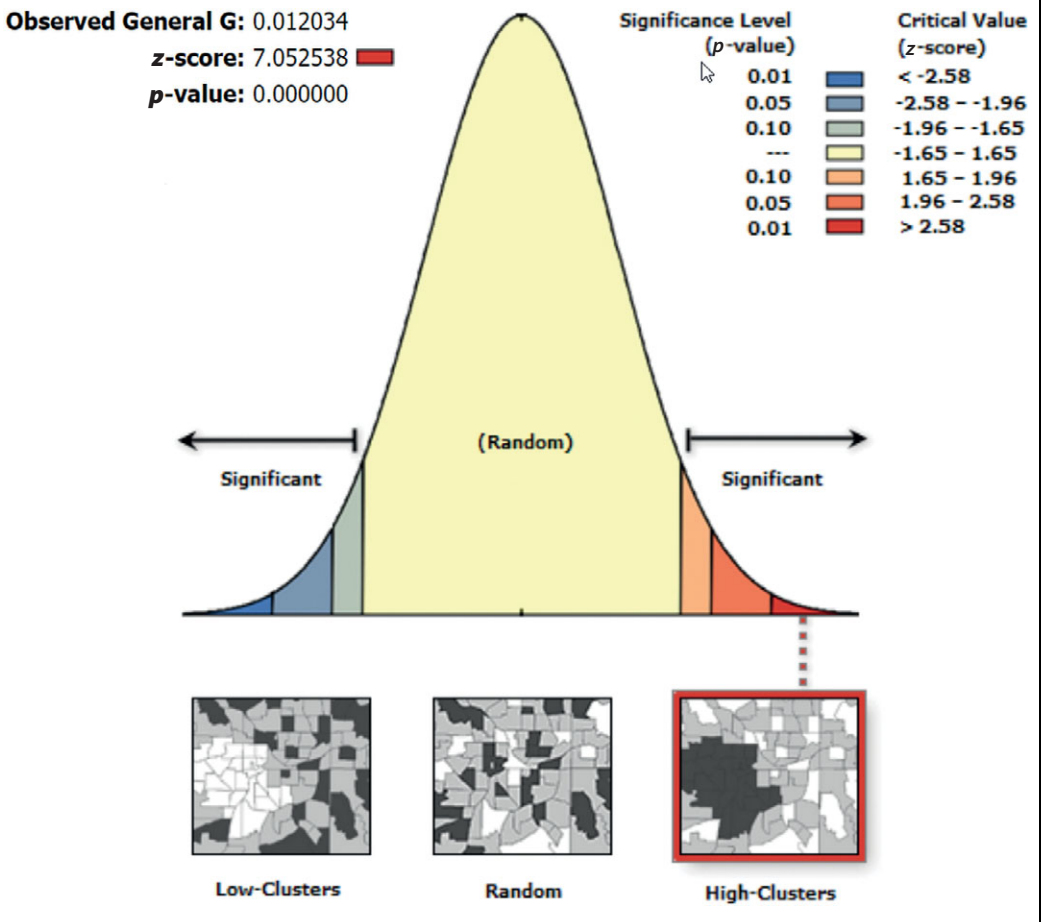


Figure 4.11 Getis-Ord General G report.

However, we cannot locate where the clustering occurs just by using this index. Moreover, although we calculated spatial autocorrelation, we have not yet defined the appropriate scale of analysis.

Exercise 4.2 Incremental Spatial Autocorrelation and Spatial Weights Matrix

In this exercise, we calculate the incremental spatial autocorrelation of income to define an appropriate scale of analysis. Based on this scale, the spatial weights matrix (see Section 1.9) is calculated, which is needed for local spatial statistics.

Exercise 4.2 (*cont.*)

ArcGIS Tools to be used: Incremental Spatial Autocorrelation, Generate Spatial Weights Matrix, Convert Spatial Weights Matrix to Table

ACTION: Incremental Spatial Autocorrelation

Navigate to the location you have stored the book dataset and click

My_Lab4_SpatialAutocorrelation.mxd

ArcToolBox > Spatial Statistics Tools > Analyzing Patterns > Incremental Spatial Autocorrelation

Input Features = City (see Figure 4.12)

Input Field = Income

Number of distance bands = 10

Beginning Distance = Leave blank

Distance Increment = Leave blank

Distance = EUCLIDEAN

Row Standardization = Check the box

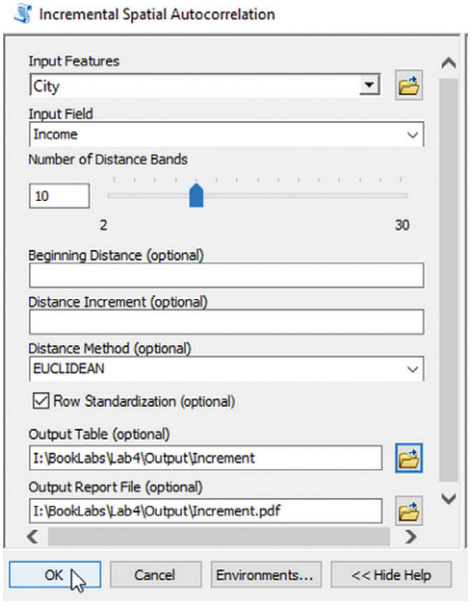


Figure 4.12 Incremental spatial autocorrelation dialog box.

Exercise 4.2 (*cont.*)

```
Output Table = I:\BookLabs\Lab4\Output\Increment

Output Report File = I:\BookLabs\Lab4\Output\Increment.pdf

OK

Main Menu > Geoprocessing > Results > Current Session >
Incremental Spatial Autocorrelation > DC on Output Report File:
Increment.pdf

Global Moran's I Summary by Distance
```

Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	*z*-score	*p*-value
1051.00	0.554941	-0.011236	0.003096	10.174893	0.000000
1157.38	0.515557	-0.011236	0.002437	10.670662	0.000000
1263.76	0.471874	-0.011236	0.002063	10.635359	0.000000
1370.14	0.453057	-0.011236	0.001729	11.166550	0.000000
1476.52	0.415797	-0.011236	0.001389	11.457432	0.000000
1582.90	0.369182	-0.011236	0.001170	11.120921	0.000000
1689.27	0.330625	-0.011236	0.001012	10.745499	0.000000
1795.65	0.296338	-0.011236	0.000852	10.537571	0.000000
1902.03	0.272382	-0.011236	0.000747	10.378124	0.000000
2008.41	0.241786	-0.011236	0.000641	9.991983	0.000000

```
First Peak (Distance, Value): 1157.38, 10.670662

Max Peak (Distance, Value): 1476.52, 11.457432

Distance measured in Meters
```

Interpreting results: Prior to incremental spatial autocorrelation, we should trace if locational outliers exist. We conducted this analysis in Exercise 3.4 and concluded that no locational outliers existed (the theoretical discussion in Section 4.3 explains how to handle locational outliers).

We observe that there are two peaks (see Figure 4.13): one at 1,157 m and one at 1,476 m. The Moran's *I* values for these distances are 0.51 and 0.41, respectively (with high *z*-scores), indicating intense clustering. Both distances reveal a form of clustering. As mentioned in the theoretical section, there is not a single correct distance at which to perform our analysis, as the

Exercise 4.2 (*cont.*)

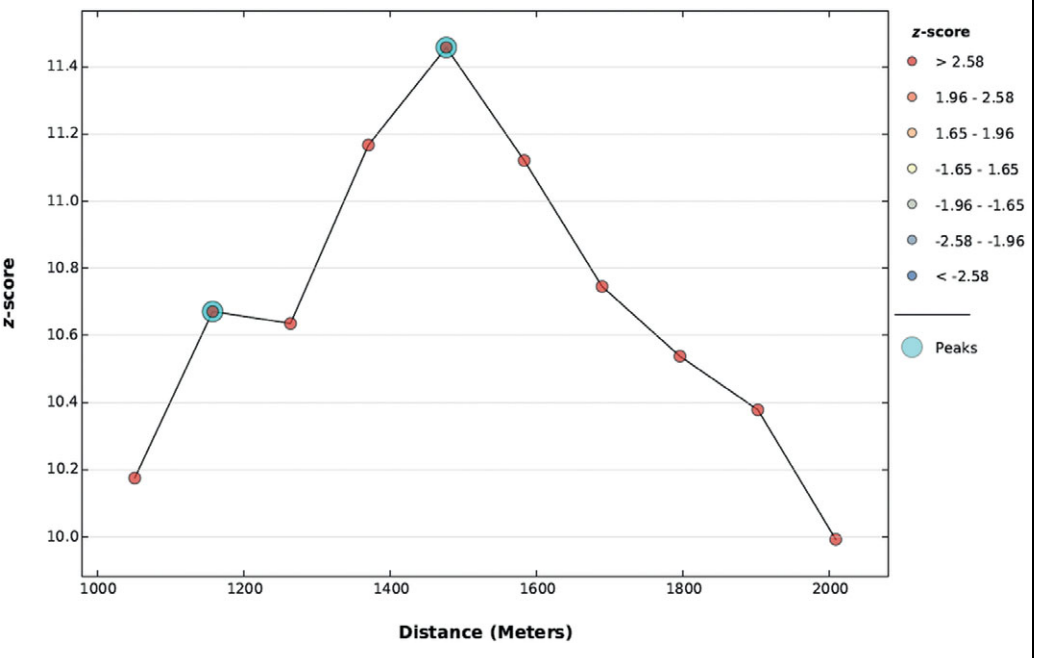


Figure 4.13 Incremental spatial autocorrelation graph. z-scores are plotted over incremental distances. Peaks are highlighted with a larger circle.

scale largely depends on our problem. It is quite common to select the first peak. As such, the scale of analysis of income can be set to 1,150 m (rounded 1157.38). The overall conclusion is that there is spatial autocorrelation of income and an underlying clustering process.

ACTION: Generate Spatial Weights Matrix

ArcToolBox > Spatial Statistics Tools > Modeling Spatial Relationships > Generate Spatial Weights Matrix

Input Feature Class = City (Navigate to I:\BookLabs\Data\City.shp) (see Figure 4.14)

Unique ID Field = PostCode

Output Spatial Weights Matrix File =

I:\BookLabs\Lab4\Output\CityWeights.swm

Conceptualization of Spatial Relationships = FIXED_DISTANCE

Distance Method = EUCLIDEAN

Exercise 4.2 (*cont.*)

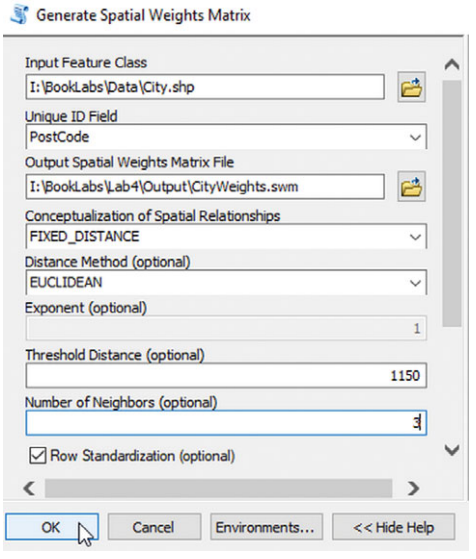


Figure 4.14 Generate spatial weights matrix dialog box.

`Exponent = 1`

`Threshold Distance = 1150 (as produced from incremental spatial autocorrelation in exercise 4.1)`

`Number of Neighbors = 3`

`Row Standardization = Check`

`OK`

ArcGIS tip: The `Number of neighbors` parameter (see Figure 4.14) is available only from the Generate Spatial Weights Matrix tool. The *k*-nearest neighbors option is also used in exploratory regression (analyzed in Chapter 6) to assess regression residuals. It takes a default value of 8.

Interpreting results: We set `FIXED_DISTANCE` as the function with which to conceptualize space, as this method is more appropriate for hot spot analysis (see Figure 4.14). The `Threshold Distance` is set to 1150 as the appropriate scale of analysis and is the result of the first part of this exercise (scale of analysis through incremental spatial autocorrelation). Objects lying further away than this distance will not be included in the calculation of the weights function. As we set a cutoff value for the `FIXED_DISTANCE`, some features may have no neighbors at this distance. To calculate the weights

Exercise 4.2 (*cont.*)

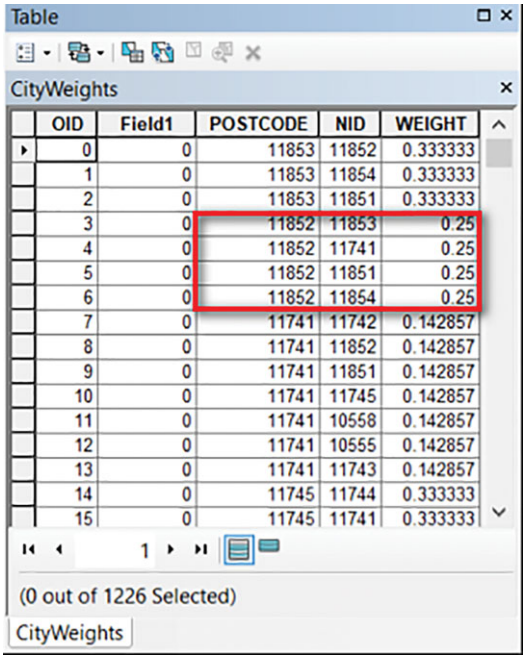
matrix, however, we should have at least some minimum number of neighbors for all features. To ensure that each object has at least a minimum number of neighbors, we use the parameter `Number of Neighbors`. By this combination (threshold and number of neighbors), features with no neighbors (or fewer than three) inside the threshold value will finally be attached to their neighbors. In other words, the threshold value is temporarily extended to ensure that each feature will have at least a minimum number of neighbors defined. In general, a spatial weights matrix is automatically generated when we apply spatial statistics by defining a conceptualization method. Nevertheless, to have more control over the weights, it is recommended to create a user-defined spatial weights matrix that can be applied thereafter. If uniformity or the isotropic environment is violated in a part of our case study, we might need to change the weights. For instance, two objects might have large weights, indicating high interaction and small distance. Due to a natural barrier (e.g., river, lake, island polygons), these objects might be close, but their interaction might be low. In such a case, we could edit the weight matrix accordingly.

`ACTION: Convert Spatial Weights Matrix to Table`

A typical spatial weights matrix in ArcGIS has three columns (see Figure 4.15): ID (unique ID of the spatial object), NID (the ID of the neighboring object with which there is a relationship) and Weight (the value of the weight that quantifies the spatial relationship). Nonexistent spatial relationships (weight = 0) are not included in the matrix to keep the table short. The output file is in a unreadable format. To read and edit the weights for each set of spatial objects, we must convert the .swm file into a table using the Convert Spatial Weights to Table tool.

```
ArcToolBox > Spatial Statistics Tools > Utilities > Convert
Spatial Weights Matrix to Table

Input Spatial Weights Matrix File =

I:\BookLabs\Lab4\Output\CityWeights.swm

Output Table = I:\BookLabs\Lab4\Output\CityWeights

OK

TOC > List By Source > RC CityWeitghs > Open

Close table

Main Menu > File > Save
```

Exercise 4.2 (*cont.*)



OID	Field1	POSTCODE	NID	WEIGHT
0	0	11853	11852	0.333333
1	0	11853	11854	0.333333
2	0	11853	11851	0.333333
3	0	11852	11853	0.25
4	0	11852	11741	0.25
5	0	11852	11851	0.25
6	0	11852	11854	0.25
7	0	11741	11742	0.142857
8	0	11741	11852	0.142857
9	0	11741	11851	0.142857
10	0	11741	11745	0.142857
11	0	11741	10558	0.142857
12	0	11741	10555	0.142857
13	0	11741	11743	0.142857
14	0	11745	11744	0.333333
15	0	11745	11741	0.333333

Figure 4.15 Spatial weights table.

Interpreting results: NID is the ID of the neighboring object, and WEIGHT is the calculated weight (see Figure 4.15). For example, postcode 11852 has four neighbors (11853,11741,11851,11854) within a fixed distance of 1,150 m (the distance between the polygon centroids); that is why the weight is 0.25 on each.

Exercise 4.3 Cluster and Outlier Analysis (Anselin Local Moran's *I*)

In this exercise, we calculate the local spatial autocorrelation of income using Local Moran's *I* to identify if clusters and outliers exist.

ArcGIS Tools to be used: `Cluster and Outlier Analysis`

`ACTION: Cluster and Outlier Analysis`

`Navigate to the location you have stored the book dataset and click`

Exercise 4.3 (*cont.*)

My_Lab4_ SpatialAutocorrelation.mxd

ArcToolBox > Spatial Statistics Tools > Mapping Clusters > Cluster and Outlier Analysis

Input Feature Class = City (see Figure 4.16)

Input Field = Income

Output Feature Class = I:\BookLabs\Lab4\Output\LocalMoranI.shp

Conceptualization of Spatial Relationships =

GET_SPATIAL_WEIGHTS_FROM_FILE

Weights Matrix File = I:\BookLabs\Lab4\Output\CityWeights.swm

Apply False Discovery Rate (FDR) Correction = Check

OK
TOC > RC LocalMoranI > Open Attribute Table

Close Table

Main Menu > File > Save

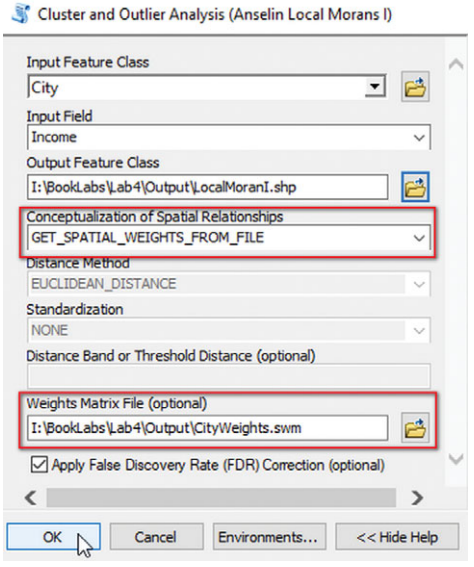


Figure 4.16 Local Moran's *I* dialog box.

Exercise 4.3 (*cont.*)

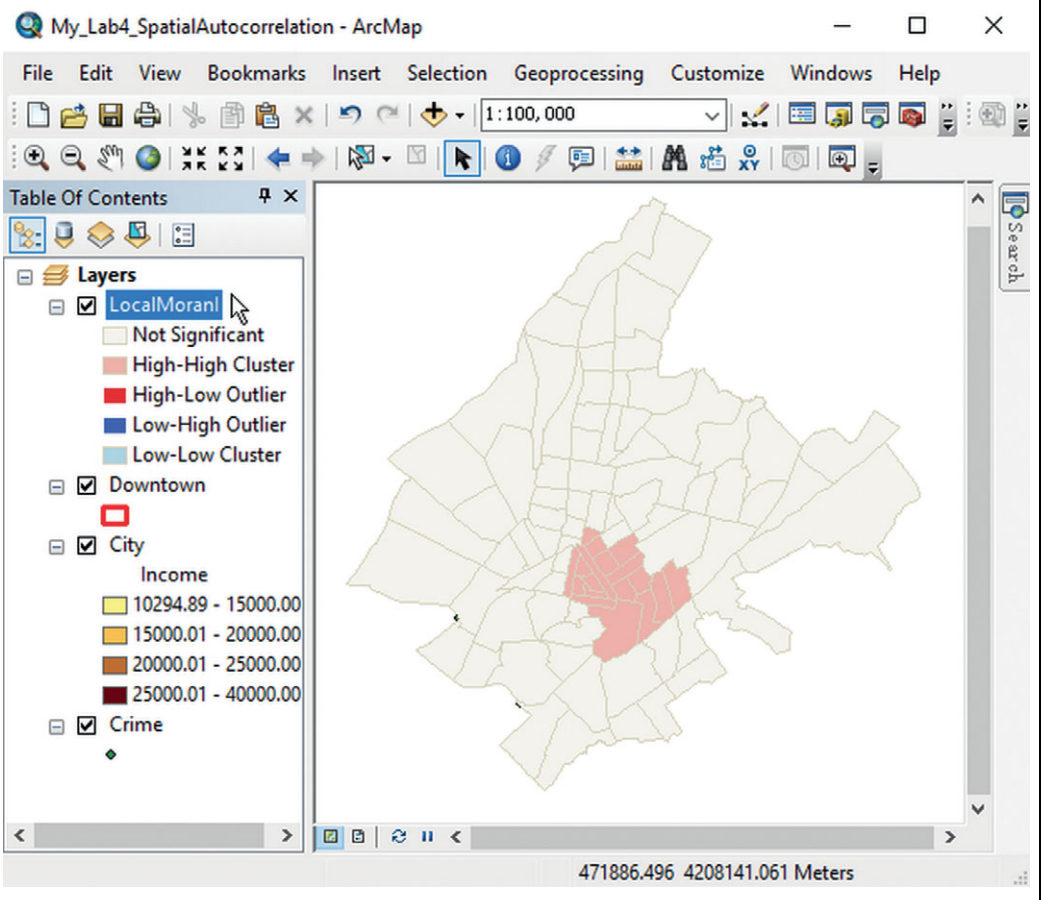


Figure 4.17 Local Moran's *I* output map.

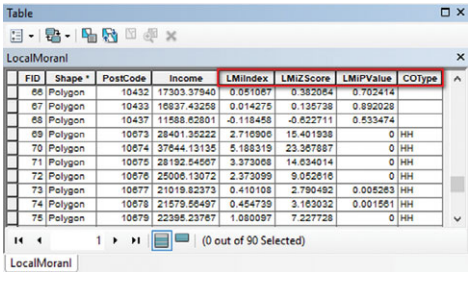


Figure 4.18 Local Moran's *I* table with local Moran's *I* index value, *z*-score, *p*-value and type of cluster assigned to each object.

Exercise 4.3 (*cont.*)

Interpreting results: A new layer is added to the table of contents (see Figure 4.17). The COType field in the Output Feature Class indicates if a postcode is an outlier or if it belongs to a cluster (see Figure 4.18). If the postcode has high income and is surrounded by postcodes with low incomes, it is marked as HL. If the postcode has low income and is surrounded by postcodes with high income, it is marked as LH. If postcodes are clustered, the COType field is HH for a statistically significant cluster of high-income values and LL for a statistically significant cluster of low-income values. The attribute table also shows the local Moran's *I* value, the *z*-score and the *p*-value. In this example and with FDR applied, income is positively spatially autocorrelated, and a statistically significant clustering of High-High values is observed in the center of the city at the 99% confidence level. No outliers or clusters of low values are detected elsewhere. In other words, people with high incomes tend to live in the red areas located in and around the downtown area.

Exercise 4.4 Hot Spot Analysis (Getis-Ord Gi*I) and Optimized Hot Spot Analysis

In this exercise, we calculate local spatial autocorrelation to identify income hot spots and cold spots using the local Getis-Ord G_i^* index.

ArcGIS Tools to be used: Hot Spot Analysis, Optimized Hot Spot Analysis

ACTION: Hot Spot Analysis

Navigate to the location you have stored the book dataset and click

My_Lab4_ SpatialAutocorrelation.mxd

ArcToolBox > Spatial Statistics Tools > Mapping Clusters > Hot Spot Analysis

Input Feature Class = City (see Figure 4.19)

Input Field = Income

Output Feature Class = I:\BookLabs\Lab4\Output\HotSpotIncome.shp

Conceptualization of Spatial Relationships =

Exercise 4.4 (*cont.*)

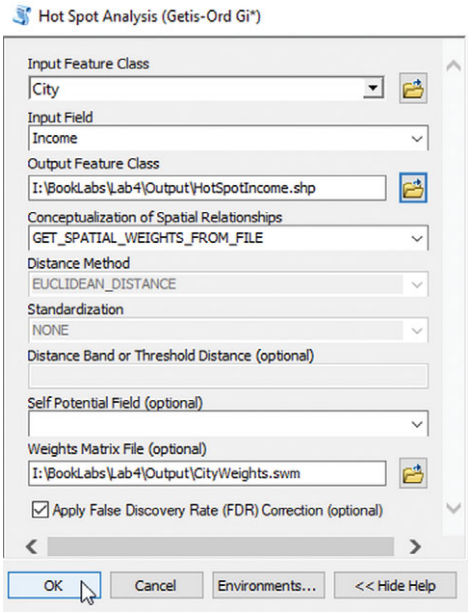


Figure 4.19 Hot spot analysis dialog box.

GET_SPATIAL_WEIGHTS_FROM_FILE

Self Potential Filed = Leave blanc

Weights Matrix File = I:\BookLabs\Lab4\Output\CityWeights.swm

(We can also directly use the FIXED_DISTANCE_BAND in the conceptualization method and add the distance threshold. Still, this option does not allow for specifying a minimum number of nearest neighbors. It is advised to use a spatial weights matrix from file – see Exercise 4.2).

Apply False Discovery Rate (FDR) Correction = Do not check (Check what happens when FDR is checked and refer back to theory)

OK

Interpreting results: A new layer is added to the table of contents (see Figure 4.20). The Gi_Bin field in the Output Feature Class indicates whether there is a hot spot or a cold spot and the related confidence level. In our example, we locate a statistically significant hot spot in and around the city center and a statistically significant cold spot in the western part of the city.

Exercise 4.4 (*cont.*)

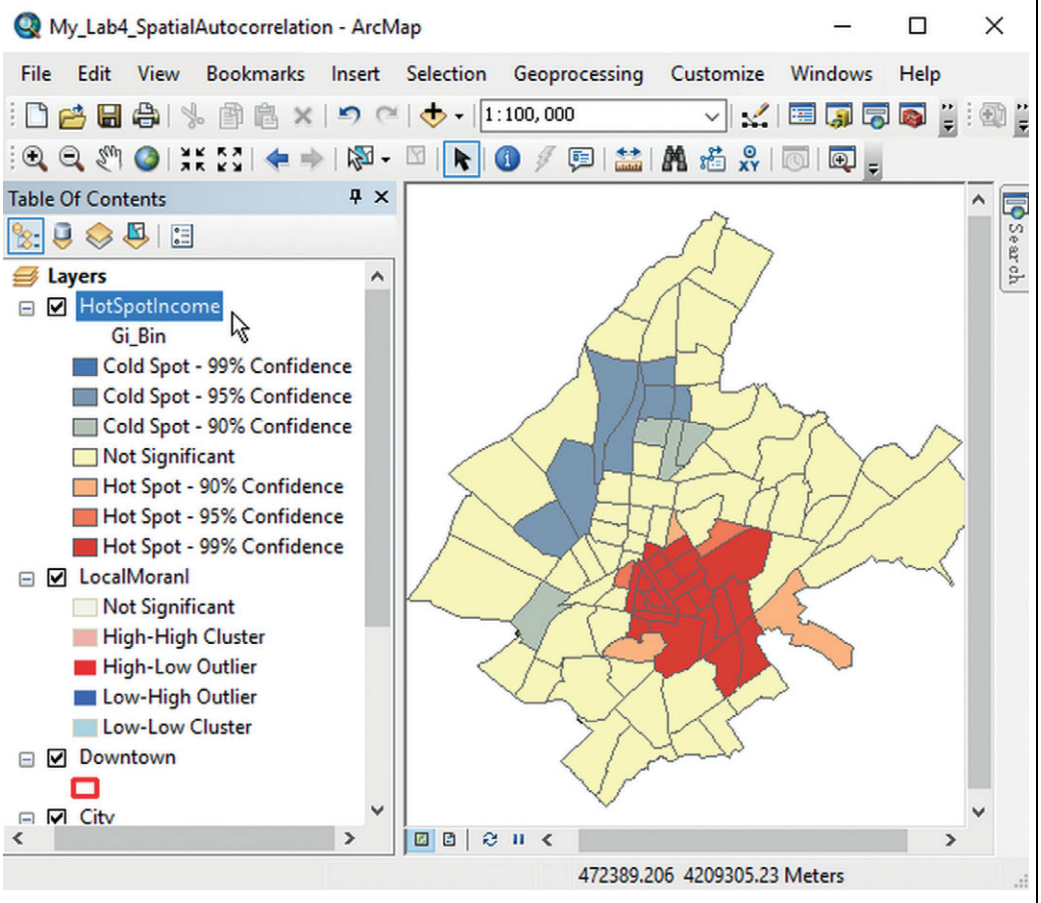


Figure 4.20 Hot spot analysis output map indicating cold (blue) and hot (red) spots.

Nonsignificant results mean that there is no indication of income clustering in these postcodes. An income cold spot means that polygons with low income values are surrounded by polygons with low income values. An income hot spot means that polygons with high income values are surrounded by polygons with high values. We notice that the hot spot analysis identifies a cluster of low values in addition to those identified in the cluster and outlier analysis in Exercise 4.3. This means that it is better to run both statistics and evaluate the results comparatively. As we are looking for high-income areas, the red areas might be more appropriate as locations for the coffee shop.

ACTION: Optimized Hot Spot Analysis

`ArcToolBox > Spatial Statistics Tools > Mapping Clusters > Optimized Hot Spot Analysis`

Exercise 4.4 (*cont.*)

Input Features = City (see Figure 4.21)

Output Features =

I:\BookLabs\Lab4\Output\C3_IncomeOpitmizedHotSot.shp

Analysis Field = Income

OK

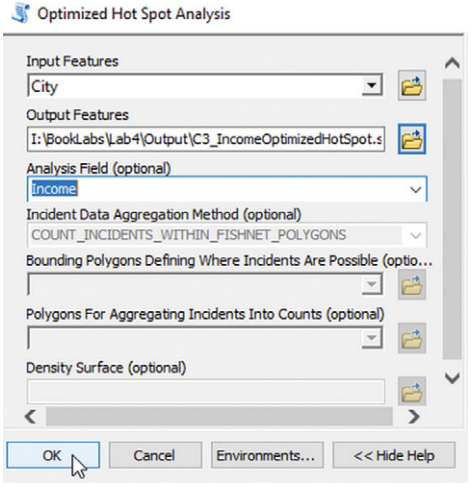


Figure 4.21 Optimized hot spot analysis dialog box.

```
**********Initial Data Assessment**********
Making sure there are enough weighted features for analysis...
    - There are 90 valid input features.

Evaluating the Analysis Field values...
    - INCOME Properties:

        Min:       10294.8949
        Max:       37644.1314
        Mean:      16316.7536
        Std. Dev.:  4947.8961

Looking for locational outliers...
    - There were no outlier locations found.
```

Exercise 4.4 (*cont.*)

**********Scale of Analysis**********
Looking for an optimal scale of analysis by assessing the intensity of clustering at increasing distances....

- The optimal fixed distance band is based on peak clustering found at 1157.3791 Meters

**********Hot Spot Analysis**********

Finding statistically significant clusters of high and low INCOME values....

- There are 38 output features statistically significant based on an FDR correction for multiple testing and spatial dependence.

**********Output**********

Creating output feature class:

I:\BookLabs\Lab4\Output\C3_IncomeOptimizedHotSpot.shp

- Red output features represent hot spots where high INCOME values cluster.

- Blue output features represent cold spots where low INCOME values cluster.
The above results can be also found through:

Main Menu > Geoprocessing > Results > Current Session > Optimized Hot Spot Analysis > Messages

Close Results

Main Menu > File > Save

Exercise 4.4 (*cont.*)

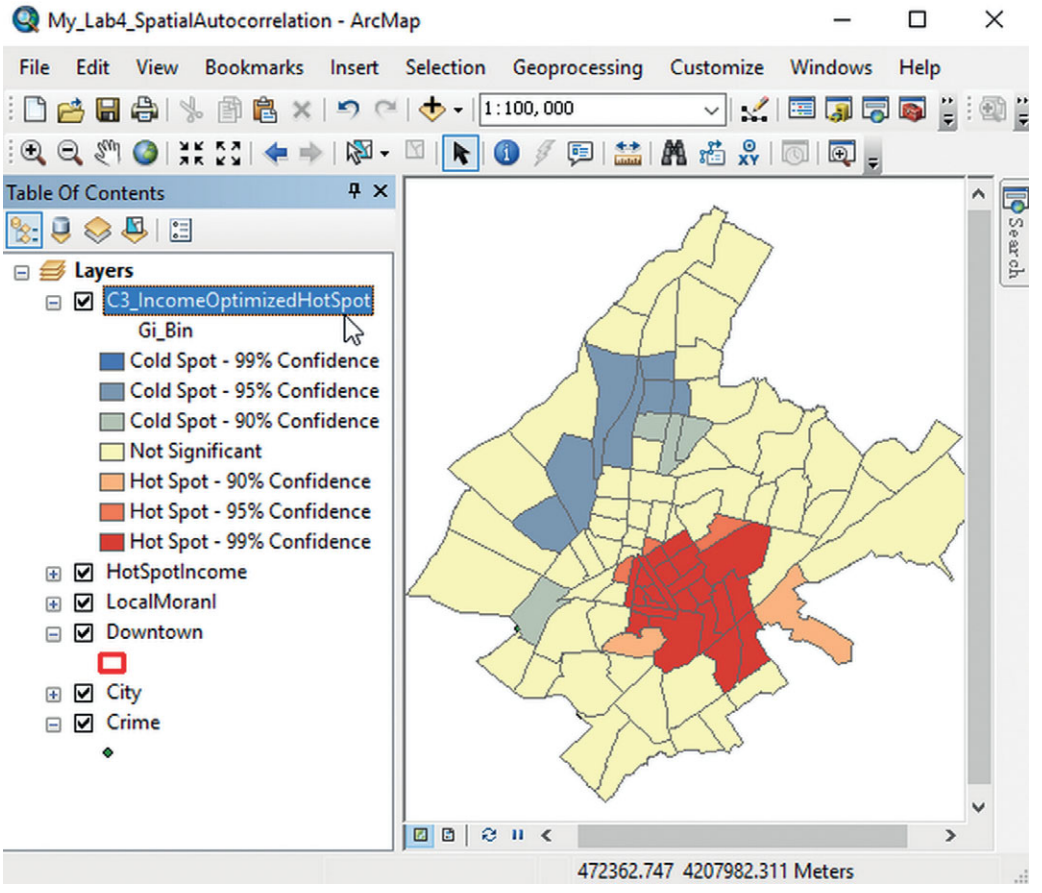


Figure 4.22 Optimized hot spot analysis output map.

Interpreting results: Optimized hot spot analysis runs the entire procedure in an automated way, so it saves significant analysis time. The results are similar to those of the hot spot analysis (because the hot spot analysis used a similar procedure but in a non-automated way; see Figure 4.22). The tool reports the scale of analysis (1,157 m) as well as the presence (or not) of locational outliers. Hot spots of income are potential areas for the location of the new coffee shop (see Box 4.1).

Exercise 4.4 (*cont.*)

Box 4.1 Analysis Criterion C3 to Be Used in Synthesis Lab 5.4: The location of the coffee shop should lie within income hot spots areas. [C3_IncomeHotSpot.lyr]

TOC > RC C3_IncomeOptimizedHotSpot > Save As Layer File > C3_IncomeHotSpot.lyr > Save

Exercise 4.5 Optimized Hot Spot Analysis for Crime Events

In this exercise, we perform optimized hot spot analysis of 539 crime events for a period of two years to identify hot spots and cold spots of crime incidents (see Figure 4.24).

ArcGIS Tools to be used: Optimized Hot Spot Analysis, Hot spot analysis

ACTION: Hot Spot Analysis

Navigate to the location you have stored the book dataset and click

My_Lab4_ SpatialAutocorrelation.mxd

TOC > List By Drawing Order > Drag Crime.shp on the top of all layers.

ArcToolBox > Spatial Statistics Tools > Mapping Clusters > Optimized Hot Spot Analysis

Input Feature Class = Crime (see Figure 4.23)

Output Feature Class =

I:\BookLabs\Lab4\Output\C4_CrimeOptimizedHotSpot.shp

Leave other blank / default

OK

Exercise 4.5 (*cont.*)

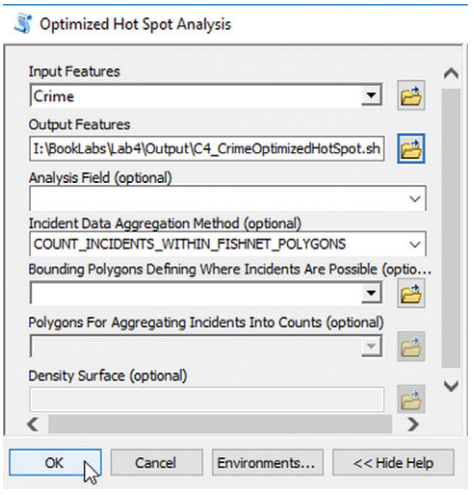


Figure 4.23 Optimized hot spot analysis dialog box.

```
**********Initial Data Assessment**********

Making sure there are enough incidents for analysis....

    - There are 768 valid input features.

Looking for locational outliers....

    - There were 5 outlier locations; these will not be used to
compute the polygon cell size.

**********Incident Aggregation**********

Creating fishnet polygon mesh to use for aggregating in-
cidents....

    - Using a polygon cell size of 233.0000 Meters

Counting the number of incidents in each polygon cell....

    - Analysis is performed on all polygon cells containing at
least one incident.
```

Exercise 4.5 (*cont.*)

```
Evaluating incident counts and number of polygons....

    - The aggregation process resulted in 457 weighted polygons.

    - Incident Count Properties:

        Min:       1.0000

        Max:       5.0000

        Mean:      1.6805

        Std. Dev.: 0.8868

**********Scale of Analysis**********

Looking for an optimal scale of analysis by assessing the
intensity of clustering at increasing distances....

   - The optimal fixed distance band is based on peak clustering
found at 933.0000 Meters

**********Hot Spot Analysis**********

Finding statistically significant clusters of high and low
incident counts....

     - There are 82 output features statistically significant
based on an FDR correction for multiple testing and spatial
dependence.

**********Output**********

Creating output feature class:

I:\BookLabs\Lab4\Output\C4_CrimeOptimizedHotSpot.shp

   - Red output features represent hot spots where high incident
counts cluster.

      - Blue output features represent cold spots where low
incident counts cluster.
```

Exercise 4.5 (*cont.*)

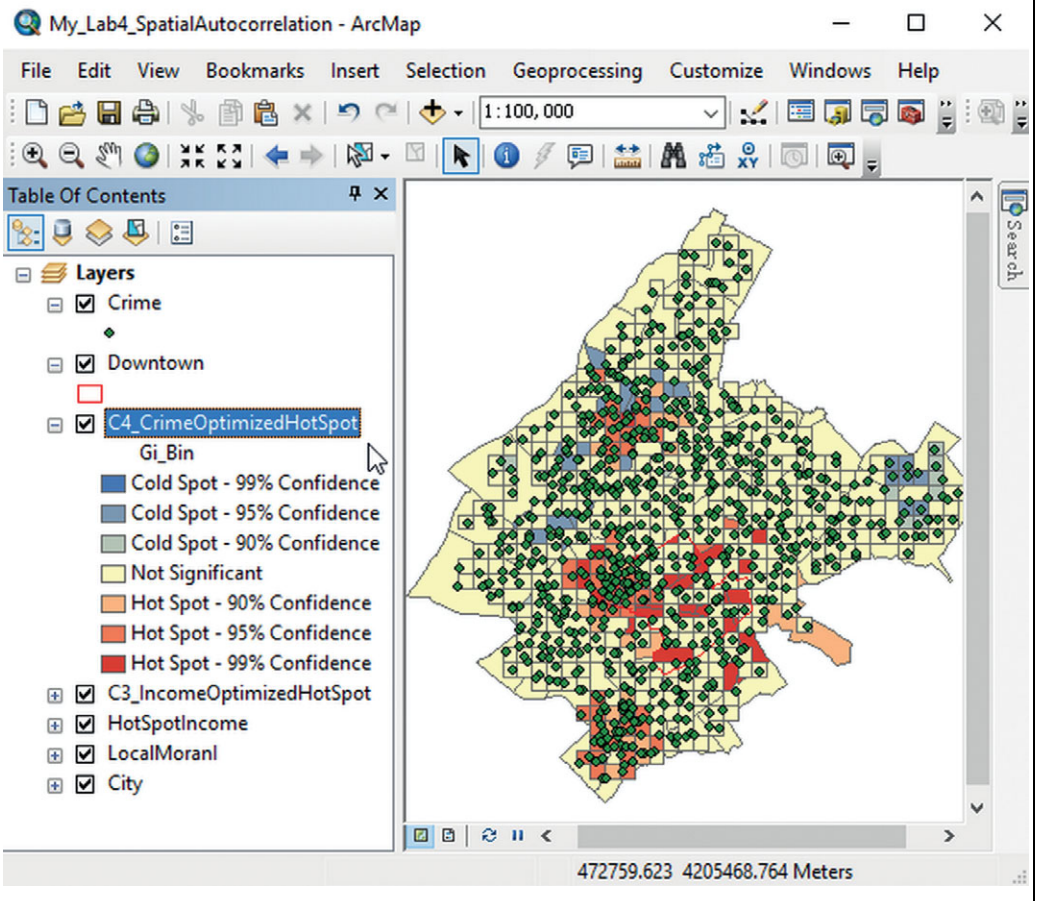


Figure 4.24 Optimized hot spot analysis dialog box.

The above results can be found through:

Main Menu > Geoprocessing > Results > Current Session > Optimized Hot Spot Analysis > Messages

Close Results
Main Menu > File > Save

Exercise 4.5 (*cont.*)

Interpreting results: Optimized hot spot analysis for point entities aggregates events into polygon cells (see Figure 4.24). To calculate the polygon cell size, five locational outliers are removed. Finally, crime events are aggregated on polygon cells with a size of 233 m. The optimal fixed distance is identified at 700 m, and FDR correction is applied. Crime events are scattered all over the study area, but one cold spot and three hot spots of crime are identified, as shown on the map (see Figure 4.24). The distance (700 m) at which autocorrelation is more pronounced reveals that the hot spots are quite large (relative to the case study area's size) and that crime is a significant problem in three regions of the city. These hot spots also reflect the center of the real clusters of crime, and, in this sense, crime might be evident in polygons adjacent to the hot spot polygons as well. Crime hot spots should be excluded as candidates for the new coffee shop's location (see Box 4.2).

Box 4.2 Analysis Criterion C4 to Be Used in Synthesis Lab 5.4: The location of the coffee shop should not lie within crime hot spots areas. [C4_CrimeHotSpot.lyr]

TOC > RC C4_CrimeOptimizedHotSpot > Save As Layer File > C4_CrimeHotSpot.lyr > Save

Section B GeoDa

Exercises 4.1 and 4.2 Global Spatial Autocorrelation and Spatial Weights Matrix

In this exercise, we calculate the global spatial autocorrelation of income using the Moran's *I* index and the Getis-Ord General G-Statistic. Before doing so, we should create the spatial weights matrix. Unlike the exercises in Section A, Exercises 4.1 and 4.2 are presented in reverse order because GeoDa requires that the spatial weights be created by the user, while ArcGIS allows for automatic calculation when the spatial autocorrelation tools are executed.

Exercises 4.1 and 4.2 (*cont.*)

GeoDa Tools to be used: Weights Manager , Univariate Moran's I

ACTION: Calculate Global Moran's I

Navigate to the location you have stored the book dataset and click the Lab4_SpatialAutocorrelation_GeoDa.gda

Main Menu > Tools > Weights Manager > Create >

Select ID Variable = PostCode (see Figure 4.25)

TAB = Distance Weight

TAB = Distance band > Specify bandwidth > Leave default value:1050.6848

Check the "Use inverse distance". Set Power to 1.

Create

File name = CityGeoDa (see Figure 4.26)

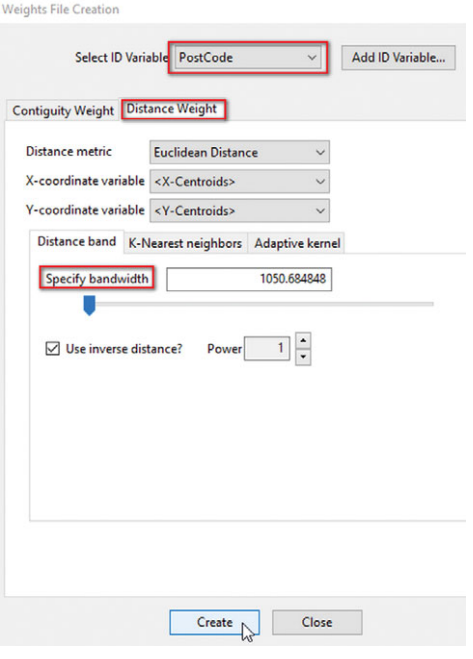


Figure 4.25 Calculating spatial weights dialog box.

Exercises 4.1 and 4.2 (*cont.*)

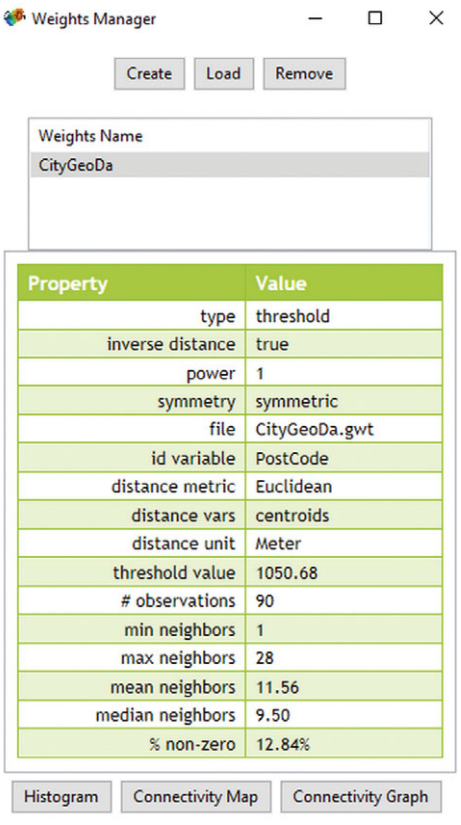


Figure 4.26 Weights manager showing the CityGeoDa spatial weights file.

Save as type = gwt (inside folder Lab4/GeoDa)

The weights manager dialog box is updated

Save > Close > Close Weights Manager window

ACTION: Calculate Global Moran's I

Main Menu > Space > Univariate Moran's I >

First Variable (X) = Income (see Figure 4.27)

Weights = GityGeoDa

OK

Exercises 4.1 and 4.2 (*cont.*)

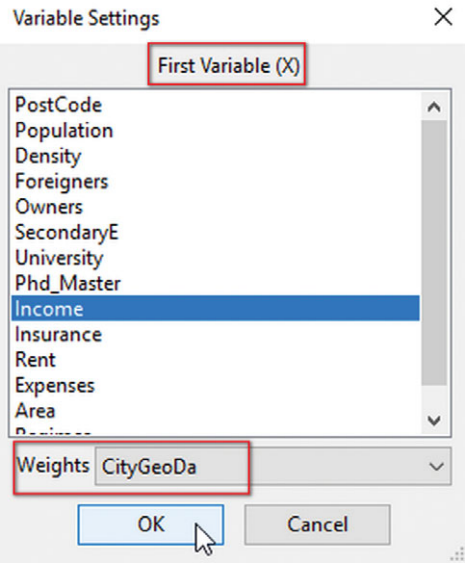


Figure 4.27 Setting the variable and weights file for Moran's *I*.

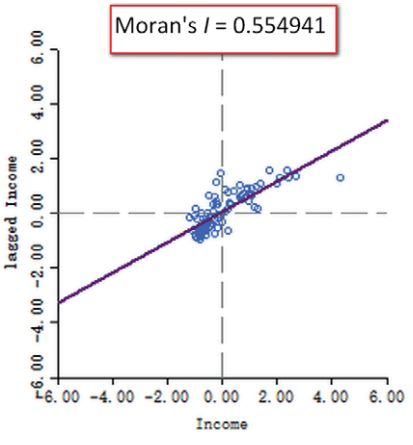


Figure 4.28 Moran's *I* scatter plot with a Moran's *I* index of 0.555. Graph reveals high positive autocorrelation.

Exercises 4.1 and 4.2 (*cont.*)

`You can save the graph as image file if you wish.`

Permutations are used to estimate how likely it is that a spatial arrangement of values similar to that we observe would be produced through complete spatial randomness. We use Monte Carlo and the following procedure.

`RC on the scatter plot > Randomization > 999 permutations (see Figure 4.29)`

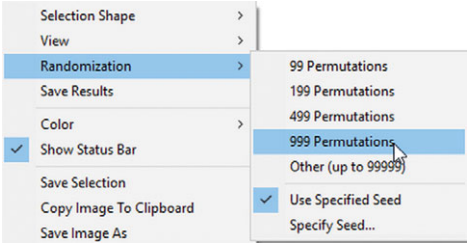


Figure 4.29 Mont Carlo simulation with 999 permutations.

Interpreting results: The results of this example suggest that the observed value $I = 0.5549$ is highly significant and not a result of spatial randomness, as it lies far away from the rest of the values (see Figure 4.30) (z-score = 9.8410; the green vertical line at the right depicts the Moran's I statistic value). The rest of the values depict the distribution that results from complete spatial randomness. The number of permutations, the setting of the pseudo p-value significance level, the expected (theoretical) Moran's I value $E(I)$, the observed Moran's I (I), the standard deviation, the z-value and the mean of the reference distribution are also presented. With 999 permutations, the pseudo p-value is set to 0.001, indicating that none of the 999 random patterns' local Moran's I values surpassed the observed value. As such, the spatial arrangement of the income values has a tendency to cluster. The distance threshold defined by the tool is set to 1,050.79 m so that every postcode has at least one neighbor. This is a first indication of income clustering. However, we cannot locate where the clustering occurs just by using this index. Moreover, although we calculated spatial autocorrelation, we have not yet defined the appropriate scale of analysis. GeoDa does not offer an automated tool for incremental spatial autocorrelation as ArcGIS does; this analysis is therefore not carried out here. As the scale of analysis, we will use the outcome of the incremental spatial autocorrelation as presented in Exercise 4.2 in Section A, which is 1150 m.

Exercises 4.1 and 4.2 (*cont.*)

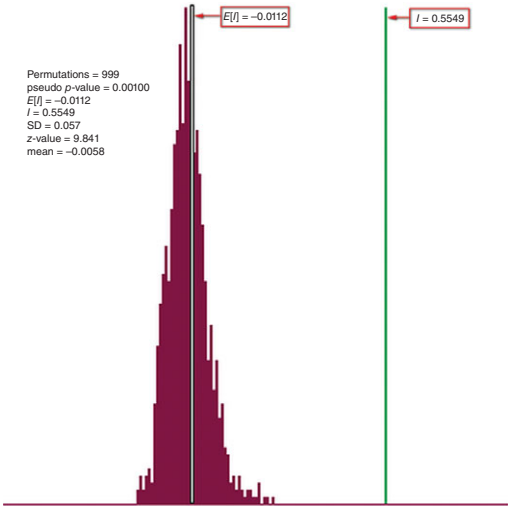


Figure 4.30 Mont Carlo reference distribution for 999 permutations.

Tip: Using the random seed for permutations might cause the results to differ slightly among different computers or sometimes even when using the same machine.

Exercise 4.3 Cluster and Outlier Analysis (Anselin Local Moran's *I*)

In this exercise, we calculate the local spatial autocorrelation of income using Local Moran's *I* to identify clusters and outliers. As mentioned in the previous exercise, the scale of analysis is 1,150m. Before we calculate the Local Moran's *I*, we should recalculate the spatial weights to reflect the adopted scale of analysis.

GeoDa Tools to be used: `Weights Manager`, `Univariate Moran's I`, `Moran's scatter plot`

Exercise 4.3 (*cont.*)

ACTION: Weights Manager

Navigate to the location you have stored the book dataset and click Lab4_SpatialAutocorrelation_GeoDa.gda

Main Menu > Tools > Weights Manager > Create

Select ID Variable = PostCode

TAB = Distance Weight

TAB = Distance band > Specify bandwidth = 1150

Check "Use inverse distance". Set Power to 1.

Create

File name = CityGeoDa1150

Save as type = gwt (inside folder Lab4/GeoDa)

Save > OK > Close > Close Weights Manager window

ACTION: Cluster and Outlier Analysis

Main Menu > Space > Univariate Local Moran's I > Income > Weights = CityGeoDa1150 > OK

Check: Significance Map

Check: Cluster Map

Check: Moran Scatter Plot

OK

Save Project (see Figure 4.31)

Interpreting results: If a postcode has high income and is surrounded by postcodes with low income, it is marked as High-Low. If the postcode has low income and is surrounded by postcodes with high income, it is marked as Low-High. Where postcodes are clustered, they are labeled as High-High for a statistically significant cluster of high-income values and Low-Low for a statistically significant cluster of low-income values.

In this example and without FDR correction, income is positively spatially autocorrelated, and a statistically significant clustering of high values is observed in the center of the city at the 99% confidence level. In other words, people with high incomes tend to live in the red areas located in and around

Exercise 4.3 (*cont.*)

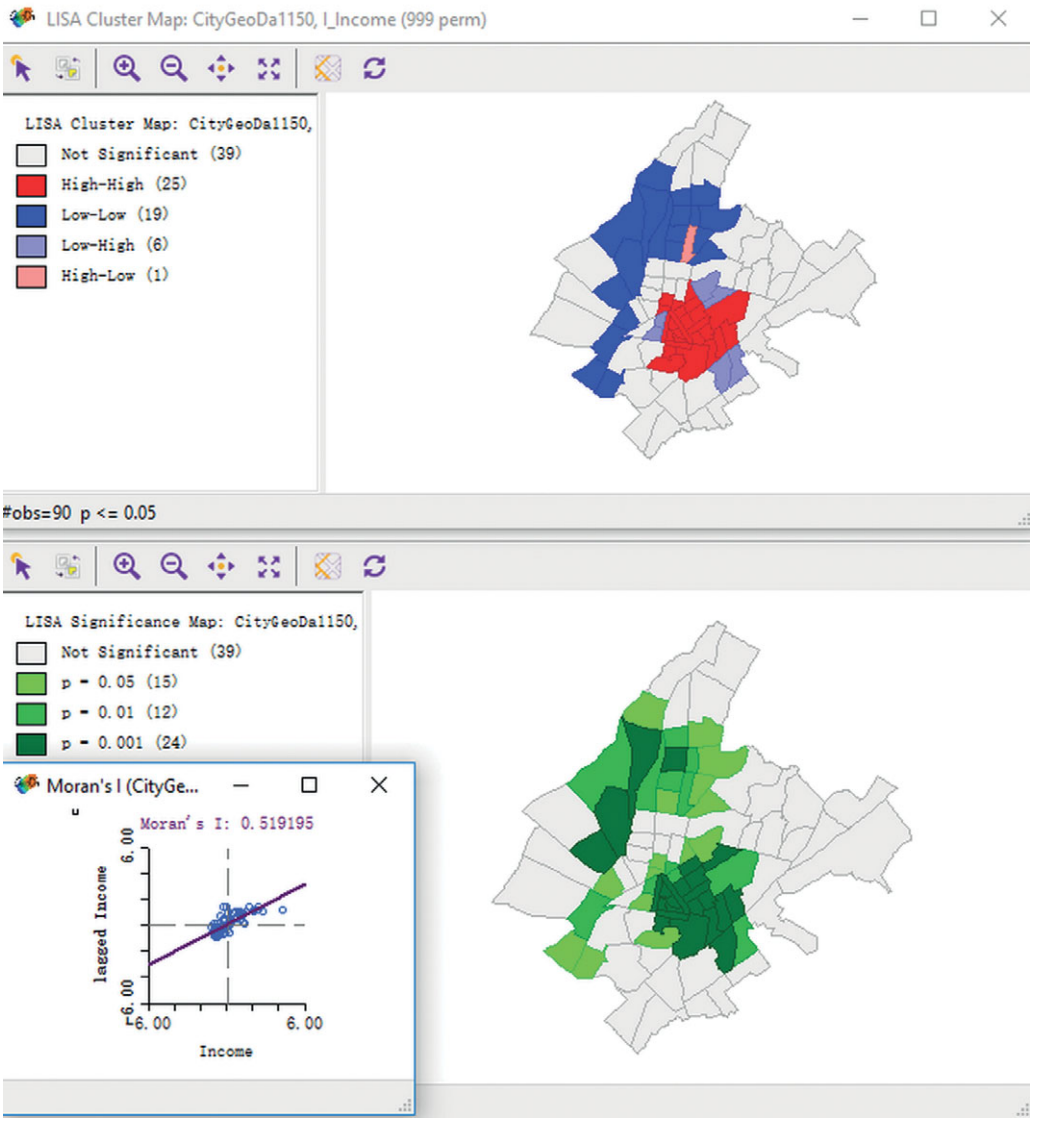


Figure 4.31 Clusters and significance map for three levels of significance (0.05, 0.01, 0.001). Unlike with the ArcGIS output in Figure 4.17, we have not applied FDR correction, and more postcodes are thus statistically significant. FDR can be applied manually in GeoDa.

the downtown area. A cluster of low values is detected in the western parts of the city (this cluster is not identified with ArcGIS due to FDR corrections). One outlier of Low-High values and five outliers of High-Low values are also located in the study area.

Exercise 4.4 Hot Spot Analysis (Getis-Ord Gi*I)

In this exercise, we calculate the local spatial autocorrelation of income using the local Getis-Ord G_i^* index to identify hot spots and cold spots (optimized hot spot analysis is not carried out as GeoDa does not offer such a tool).

GeoDa Tools to be used: `Local G*`

`ACTION: Local G*`

`Navigate to the location you have stored the book dataset and click Lab4_SpatialAutocorrelation_GeoDa.gda`

`Main Menu > Space > Local G* > Income > Weights = CityGeoDa1150 > OK`

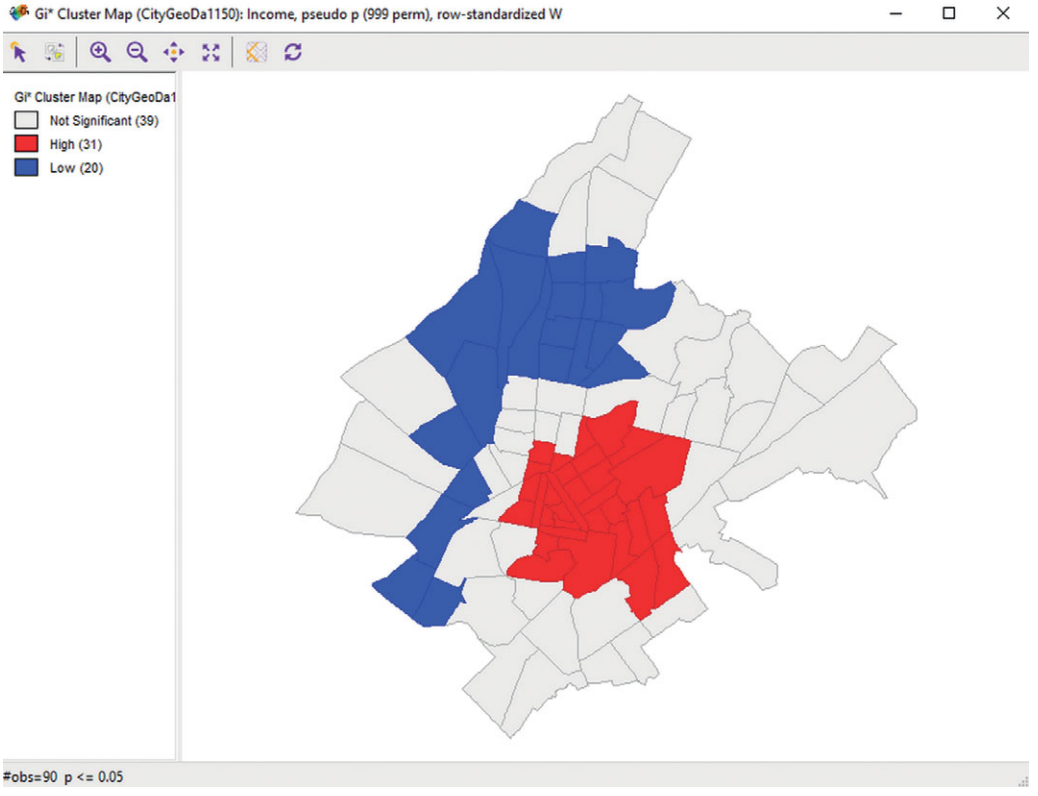


Figure 4.32 Hot spot analysis output map indicating cold (blue) and hot (red) spots significant at the $p <= 0.05$ level. There are minor differences from the results shown in Figure 4.20 using ArcGIS due to the slightly different weights matrix used. However, the main conclusions regarding the presence of hot and cold spots remain unchanged.

Exercise 4.4 (*cont.*)

```
Check: Cluster Map

Check: using row-standardized weights

Save project
```

Interpreting results: We locate a statistically significant hot spot in and around the city center and a statistically significant cold spot in the western part of the city (see Figure 4.32). Nonsignificant results mean that there is no indication of income clustering for these postcodes. A cold spot of income means that polygons with low values of income are surrounded by polygons with low values of income. A hot spot of income means that polygons with high values of income are surrounded by polygons with high values. As we are looking for areas with high income, the red areas might be more appropriate as locations for the coffee shop.

Remark: Optimized hot spot analysis is not offered in GeoDa, and Exercise 4.5 is presented only in Section A.