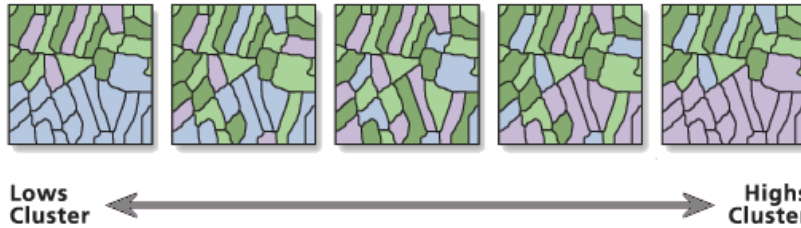


Getis-Ord General G (Global Statistic)

Measures the degree of clustering for either high values or low values



How High/Low Clustering: Getis-Ord General G works

The **High/Low Clustering (Getis-Ord General G) tool (ArcToolbox → Spatial Statistics → Analyzing Patterns)** measures how concentrated the high or low values are for a given study area. As an example, you can use this statistic to monitor the number of emergency room visits & look for unexpected spikes which might indicate an outbreak of a local or regional health problem.

Calculations

The measures how concentrated the high or low values are for a given study area. As an example, you can use this statistic to monitor the number of emergency room visits & look for unexpected spikes which might indicate an outbreak of a local or regional health problem.

The General G statistic of overall spatial association is given as:

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{i,j} x_i x_j}{\sum_{i=1}^n \sum_{j=1}^n x_i x_j}, \quad \forall j \neq i \quad (1)$$

where x_i and x_j are attribute values for features i and j , and $w_{i,j}$ is the spatial weight between feature i and j .

The z_G -score for the statistic is computed as:

$$z_G = \frac{G - E[G]}{\sqrt{V[G]}} \quad (2)$$

where:

$$E[G] = \frac{\sum_{i=1}^n \sum_{j=1}^n w_{i,j}}{n(n-1)}, \quad \forall j \neq i \quad (3)$$

$$V[G] = E[G^2] - E[G]^2 \quad (4)$$

Additional calculations are as follows:

$$E[G^2] = \frac{A + B}{C} \quad (5)$$

$$A = D_0 \left(\sum_{i=1}^n x_i^2 \right)^2 + D_1 \sum_{i=1}^n x_i^4 + \quad (6)$$

$$D_2 \left(\sum_{i=1}^n x_i \right)^2 \sum_{i=1}^n x_i^2$$

$$B = D_3 \sum_{i=1}^n x_i \sum_{i=1}^n x_i^3 + D_4 \left(\sum_{i=1}^n x_i \right)^4 \quad (7)$$

$$C = \left[\left(\sum_{i=1}^n x_i \right)^2 - \sum_{i=1}^n x_i^2 \right]^2 \times \quad (8)$$

$$n(n-1)(n-2)(n-3)$$

$$D_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3W^2 \quad (9)$$

$$D_1 = -[(n^2 - n)S_1 - 2nS_2 + 6W^2] \quad (10)$$

$$D_2 = -[2nS_1 - (n+3)S_2 + 6W^2] \quad (11)$$

$$D_3 = 4(n-1)S_1 - 2(n+1)S_2 + \quad (12)$$

$$8W^2$$

$$D_4 = S_1 - S_2 + W^2 \quad (13)$$

$$W = \left(\sum_{i=1}^n \sum_{j=1, j \neq i}^n w_{i,j} \right) \quad (14)$$

$$S_1 = (1/2) \sum_{i=1}^n \sum_{j=1, j \neq i}^n (w_{i,j} + w_{j,i})^2 \quad (15)$$

$$S_2 = \sum_{i=1}^n \left(\sum_{j=1, j \neq i}^n w_{i,j} + \sum_{j=1}^n w_{j,i} \right)^2 \quad (16)$$

The p-value is a numerical approximation of the area under the curve for a known distribution, limited by the test statistic.

Interpretation

The High/Low Clustering tool is an inferential statistic, which means that the results of the analysis are interpreted within the context of a null hypothesis. The null hypothesis for the General G statistic states "there is no spatial clustering of the values". When the absolute value of the Z score is large & the p-value is very small, the null hypothesis can be rejected (see What is a Z score? What is a p-value?). If the null hypothesis is rejected, then the sign of the Z score becomes important. If the Z score value is positive, it means that high values cluster together in the study area. If the Z Score value is negative, it means that low values cluster together.

Output

This tool calculates the High/Low General G value (observed & expected) & the associated Z score & p-value for a given input feature class. These values are written to the message section of the command line window, & the observed General G, Z score, & p-value are passed as derived output.

Potential applications

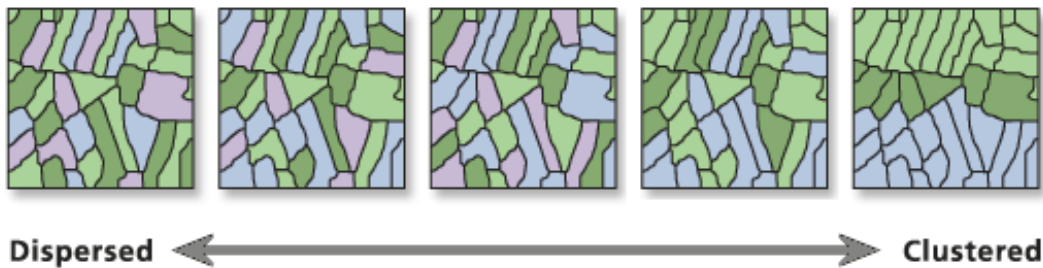
- Comparing the spatial pattern of different types of retail within a city to see which types cluster with competition to take advantage of comparison shopping (automobile dealerships, for example) & which types repel competition (fitness centers/gyms, for example).
- Summarizing the level at which spatial phenomena cluster in order to examine changes at different times or in different locations. For example, it is known that cities & their populations cluster. Using High/Low Clustering analysis, you can compare the level of clustering for western versus eastern cities (urban morphology) or the level of population clustering within a single city over time (analysis of urban growth & density).

Usage Tips

- This tool honors the environment output coordinate system. Feature **geometry** is projected to the output coordinate system prior to analysis, so the units associated with values entered for the Distance Band/Threshold Distance parameter should match those specified in the output coordinate system. All mathematical computations are based on the output coordinate system **spatial reference**.
- Calculations based on either Euclidean or Manhattan distance require projected data to accurately measure distances.
- If you will be running several analyses on a single dataset (e.g., analyzing several different fields) or if you have a dataset with more than 3000 features, it is recommended that you construct the spatial weights matrix file prior to analysis.
- The General G tool calculates the value of the General G index, associated Z score & p-value for a given input feature class. These values are written to the Command Line message window & passed as derived output.
- The Z score & p-value are measures of statistical significance which tell you whether or not to reject the null hypothesis. For this tool, the null hypothesis states that the values associated with features are randomly distributed.
- The Z score is based on the Randomization Null Hypothesis computation. For more information on Z scores, see What is a Z score? What is a p-value?.
- The higher (or lower) the **Z score**, the stronger the intensity of the clustering. A Z score near zero indicates no apparent clustering within the study area. A positive Z score indicates clustering of high values. A negative Z scores indicates clustering of low values.
- For line & polygon features, true geometric feature centroids are used in computations.
- The input field you select should only contain positive numeric values. The General G statistic was designed to work with non-negative values only.
- The input field should contain a variety of non-negative values. The math for this statistic requires some variation in the variable being analyzed; it cannot solve if all input values are 1, for example. If you have incident data, & want to analyze incident intensity, consider aggregating your incident data or using Integrate with the Collect Events tool prior to analysis.
- Whenever using shapefiles keep in mind that they cannot store null values. Tools or other procedures that create shapefiles from non-shapefile inputs may store or interpret null values as zero. This can lead to unexpected results.
- The Conceptualization of Spatial Relationships used for analysis should be based on your understanding of spatial interaction among the features being analyzed.
- For the Fixed Distance option, the distance band used for analysis should be based on your understanding of spatial interaction among the features being analyzed. Alternatively, features may be evaluated for a range of distance values or at the specific distance where spatial autocorrelation is maximized.
- For Inverse Distance conceptualization options: when zero is entered for the "Distance Band or Threshold Distance" parameter all features are considered neighbors of all other features; when this parameter is left blank, a default threshold distance will be applied.
- When the spatial conceptualization is an Inverse Distance method (Inverse Distance, Inverse Distance Squared, or Zone of Indifference) any two points that are coincident will be given a weight of one to avoid zero division. This assures features are not excluded from analysis.
- With inverse distance conceptualizations, weights for distances less than 1 become unstable. The weighting for features separated by less than 1 unit of distance (common with Geographic Coordinate System projections), are given a weight of 1.
- Analysis on features with a Geographic Coordinate System projection is not recommended with any of the inverse distance based spatial conceptualization methods.
- The "Display Output Graphically" parameter will only work on the Windows operating system. When set to true it will display the results of the tool graphically.
- When output is shown graphically, a separate graphics dialog box will be displayed. If you use the tool in a script, set the Display_Output_Graphically parameter to "false", otherwise your script will not complete until you click "Close" on the popup graphic.
- Current map layers may be used to define the input feature class. When using layers, only the currently selected features are included in the analysis.

Moran's I (Global Statistic)

Measures spatial autocorrelation based on feature locations & attribute values.



How Spatial Autocorrelation: Moran's I works

The **Spatial Autocorrelation (Morans I) tool** ([ArcToolbox](#) → [Spatial Statistics](#) → [Analyzing Patterns](#)) measures spatial autocorrelation (feature similarity) based on both feature locations & feature values simultaneously. Given a set of features & an associated attribute, it evaluates whether the pattern expressed is clustered, dispersed, or random. The tool calculates the Moran's I Index value & both a Z score & p-value evaluating the significance of that index. In general, a Moran's Index value near +1.0 indicates clustering while an index value near -1.0 indicates dispersion. However, without looking at statistical significance you have no basis for knowing if the observed pattern is just one of many, many possible versions of random.

In the case of the Spatial Autocorrelation tool, the null hypothesis states that "there is no spatial clustering of the values associated with the geographic features in the study area". When the p-value is small & the absolute value of the Z score is large enough that it falls outside of the desired confidence level, the null hypothesis can be rejected. If the index value is greater than 0, the set of features exhibits a clustered pattern. If the value is less than 0, the set of features exhibits a dispersed pattern.

Calculations

The Moran's I statistic for spatial autocorrelation is given as:

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{i,j} z_i z_j}{S_0 \sum_{i=1}^n z_i^2} \quad (1)$$

where z_i is the deviation of an attribute for feature i from its mean ($x_i - \bar{X}$), $w_{i,j}$ is the spatial weight between feature i and j , n is equal to the total number of features, and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{i,j} \quad (2)$$

The z_I -score for the statistic is computed as:

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}} \quad (3)$$

where:

$$E[I] = -1/(n - 1) \quad (4)$$

$$V[I] = E[I^2] - E[I]^2 \quad (5)$$

Additional calculations are as follows:

$$E[I^2] = \frac{A - B}{C} \quad (6)$$

$$A = n [(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2] \quad (7)$$

$$B = D [(n^2 - n)S_1 - 2nS_2 + 6S_0^2] \quad (8)$$

$$C = (n - 1)(n - 2)(n - 3)S_0^2 \quad (9)$$

$$D = \frac{\sum_{i=1}^n z_i^4}{\left(\sum_{i=1}^n z_i^2\right)^2} \quad (10)$$

$$S_1 = (1/2) \sum_{i=1}^n \sum_{j=1}^n (w_{i,j} + w_{j,i})^2 \quad (11)$$

$$S_2 = \sum_{i=1}^n \left(\sum_{j=1}^n w_{i,j} + \sum_{j=1}^n w_{j,i} \right)^2 \quad (12)$$

The p-value is a numerical approximation of the area under the curve for a known distribution, limited by the test statistic.

Potential applications

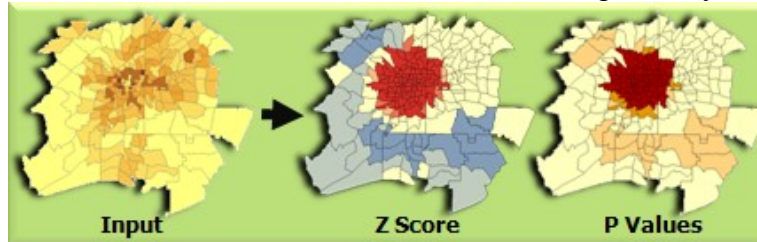
- Determine the feasibility of using a particular statistical method (for example, linear regression analysis & many other statistical techniques require independent observations).
- Perform spatial filtering prior to regression analysis to avoid violating assumptions of data independence.
- Help identify an appropriate neighborhood distance for a variety of spatial analysis methods.
- Determine the feasibility of using a particular statistical method. For example, linear regression analysis & many other statistical techniques require independent observations. Also useful for testing regression residuals.
- Help identify an appropriate neighborhood distance for a variety of spatial analysis methods. For example, find the distance where spatial autocorrelation is strongest.
- Measure broad trends in ethnic or racial segregation over time—is segregation increasing or decreasing.
- Summarize the diffusion of an idea, disease or trend over space & time—is the idea, disease, or trend remaining isolated & concentrated or spreading & becoming more diffuse.

Usage Tips

- This tool honors the environment [output coordinate system](#) even though no feature class output is created. Feature [geometry](#) is projected to the output coordinate system prior to analysis, so values entered for the Distance Band/Threshold Distance parameter should match those specified in the output coordinate system. All mathematical computations are based on the output coordinate system [spatial reference](#).
- Calculations based on either Euclidean or Manhattan distance require [projected data](#) to accurately measure distances.
- If you will be running several analyses on a single dataset (e.g., analyzing several different fields) or if you have a dataset with more than 3000 features, it is recommended that you construct the [spatial weights matrix file](#) prior to analysis.
- The Moran's I value & associated Z score & p-value are written to the Command window & passed as derived output.
- Given a set of features & an associated attribute, Global Moran's I evaluates whether the pattern expressed is clustered, dispersed, or random. When the Z score or p-value indicates statistical significance, a positive Moran's I index value indicates tendency toward clustering while a negative Moran's I index value indicates tendency toward dispersion.
- The Global Moran's I tool calculates a [Z score](#) & p-value to indicate whether or not you can reject the null hypothesis. In this case, the null hypothesis states that feature values are randomly distributed across the study area. In this tool, the Z score is based on the Randomization Null Hypothesis computation. For line & polygon features, true feature geometric centroids are used in computations.
- The input field should contain a variety of non-negative values. The math for this statistic requires some variation in the variable being analyzed; it cannot solve if all input values are 1, for example. If you have incident data, & want to analyze incident intensity, consider aggregating your incident data or using [Integrate](#) with the [Collect Events](#) tool prior to analysis.
- Whenever using shapefiles keep in mind that they cannot store null values. Tools or other procedures that create shapefiles from non-shapefile inputs may store or interpret null values as zero. This can lead to unexpected results.
- The Conceptualization of Spatial Relationships used for analysis should be based on your understanding of spatial interaction among the features being analyzed.
- For Inverse Distance conceptualization options: when zero is entered for the "Distance Band or Threshold Distance" parameter all features are considered neighbors of all other features; when this parameter is left blank, a default threshold distance will be applied.
- When the spatial conceptualization is an Inverse Distance method (Inverse Distance, Inverse Distance Squared, or Zone of Indifference) any two points that are coincident will be given a weight of one to avoid zero division. This assures features are not excluded from analysis.
- With inverse distance conceptualizations, weights for distances less than 1 become unstable. The weighting for features separated by less than 1 unit of distance (common with [Geographic Coordinate System](#) projections), are given a weight of 1.
- Analysis on features with a Geographic Coordinate System projection is not recommended with any of the inverse distance based spatial conceptualization methods.
- The "Display Output Graphically" parameter will only work on the Windows operating system. When set to true it will display the results of the tool graphically.
- When output is shown graphically, a separate graphics dialog box will be displayed. If you use the tool in a script, set the `Display_Output_Graphically` parameter to "false", otherwise your script will not complete until you click "Close" on the popup graphic.
- In ArcGIS version 9.2, the "Global" standardization option was removed. Global standardization returns the same results as no standardization. Models built with previous versions of ArcGIS that use the Global standardization option may need to be rebuilt.
- See the [Modeling Spatial Relationships](#) help page for further explanation of this tool's parameters.
- Current map layers may be used to define the input feature class. When using layers, only the currently selected features are included in the analysis.

Getis-Ord G_i^* (Local Statistic)

Calculates the Getis-Ord G_i^* statistic for hot spot analysis.



How Hot Spot Analysis: Getis-Ord G_i^* (Spatial Statistics) works

The **Hot Spot Analysis (Getis-Ord G_i^*)** tool ([ArcToolbox](#) → [Spatial Statistics](#) → [Mapping Clusters](#)) calculates the Getis-Ord G_i^* statistic for each feature in a dataset. The resultant Z score tells you where features with either high or low values cluster spatially. This tool works by looking at each feature within the context of neighboring features. A feature with a high value is interesting, but may not be a statistically significant hot spot. To be a statistically significant hot spot, a feature will have a high value & be surrounded by other features with high values as well. The local sum for a feature & its neighbors is compared proportionally to the sum of all features; when the local sum is much different than the expected local sum, & that difference is too large to be the result of random chance, a statistically significant Z score results.

Calculations

The Getis-Ord local statistic is given as:

$$G_i^* = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{i,j}}{S \sqrt{\frac{n \sum_{j=1}^n w_{i,j}^2 - \left(\sum_{j=1}^n w_{i,j} \right)^2}{n-1}}} \quad (1)$$

where x_j is the attribute value for feature j , $w_{i,j}$ is the spatial weight between feature i and j , n is equal to the total number of features and:

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \quad (2)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2} \quad (3)$$

The G_i^* statistic is a z-score so no further calculations are required.

The p-values are numerical approximations of the area under the curve for a known distribution, limited by the test statistic.

Interpretation

The G_i^* statistic returned for each feature in the dataset is a Z score. For statistically significant positive Z scores, the larger the Z score is, the more intense the clustering of high values (hot spot). For statistically significant negative Z scores, the smaller the Z score is, the more intense the clustering of low values (cold spot).

Hot Spot Analysis

There are 3 things to consider when undertaking any hot spot analysis:

1) What is the Analysis Field?

The hot spot analysis tool assesses whether high or low values (the number of crimes, accident severity, or dollars spent on sporting goods, for example) cluster spatially. The field containing those values is your Analysis Field. For point incident data, however, you may be more interested in assessing incident intensity than in analyzing the spatial clustering of any particular value associated with the incidents. In that case you will need to aggregate your incident data prior to analysis.

There are several ways to do this:

- If you have census blocks or other polygon features for your study area, consider doing a [Spatial Join](#) to count the number of events in each block. The resultant field containing the number of events in each polygon becomes the Analysis Field.
- Use the [Create Fishnet](#) tool to construct a polygon grid over your point features. Then do a [Spatial Join](#) to count the number of events falling within each grid polygon. Remove any grid polygons that fall outside of your study area. Also, in cases where many of the grid polygons within the study area contain zeros for the number of events, increase the polygon grid size if appropriate, or remove those zero count grid polygons prior to analysis.
- Alternatively, if you have a number of coincident points or points within a short distance of one another, you can use [Integrate](#) with the [Collect Events](#) tool to (1) snap features within a specified distance of each other together & then (2) create a new feature class containing a point at each unique location with an associated count attribute to indicate the number of events/snapped points. Use the resultant ICOUNT field as your Analysis Field.

2) Which Conceptualization of Spatial Relationships is appropriate? What Distance value is best?

The recommended (& default) Conceptualization of Spatial Relationships for the Hot Spot Analysis tool is Fixed Distance. Zone of Indifference, Contiguity, K Nearest Neighbor & Delaunay Triangulation may also work well. For a discussion of best practices & strategies for determining an analysis distance value, see [Selecting a Conceptualization of Spatial Relationships: Best Practices](#) & also [Selecting a Fixed Distance](#).

3) What is the question?

This may seem obvious, but how you construct the Analysis Field determines the types of questions you can ask. Are you most interested in determining where you have lots of incidents, or where high/low values for a particular attribute cluster spatially? If so, run Hot Spot Analysis on the raw values or raw incident counts. This type of analysis is particularly helpful for resource allocation types of problems. Alternatively (or in addition), you may be interested in locating areas with unexpectedly high values in relation to some other variable. If you are analyzing foreclosures, for example, you probably expect more foreclosures in locations with more homes (or said another way: at some level, you expect the number of foreclosures to be a function of the number of houses). If you divide the number of foreclosures by the number of homes, & then run the Hot Spot Analysis tool on this ratio, you are no longer asking "Where are there lots of foreclosures?"; instead you are asking "Where are there unexpectedly high numbers of foreclosures, given the number of homes?". By creating a rate or ratio prior to analysis, you can control for certain expected relationships (e.g., the number of crimes is a function of population; the number of foreclosures is a function of housing stock) & identify unexpected hot/cold spots.

Potential Applications

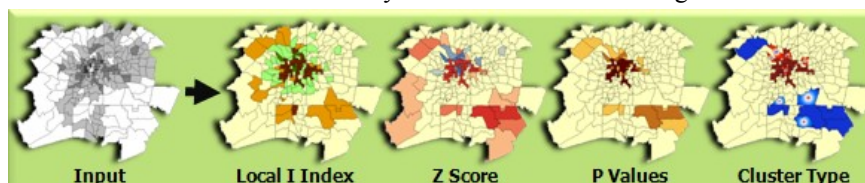
Applications can be found in crime analysis, epidemiology, voting pattern analysis, economic geography, retail analysis, traffic incident analysis, & demographics.

Usage Tips

- This tool honors the environment [output coordinate system](#). Feature [geometry](#) is projected to the output coordinate system prior to analysis, so values entered for the Distance Band/Threshold Distance parameter should match those specified in the output coordinate system. All mathematical computations are based on the output coordinate system [spatial reference](#).
- Calculations based on either Euclidean or Manhattan distance require [projected data](#) to accurately measure distances.
- If you will be running several analyses on a single dataset (e.g., analyzing several different fields) or if you have a dataset with more than 3000 features, it is recommended that you construct the [spatial weights matrix file](#) prior to analysis.
- Given a set of weighted features, the Getis-Ord G_i^* statistic identifies spatial clusters of high values (hot spots) & spatial clusters of low values (cold spots).
- This tool creates as derived output the Z score & p-value fieldnames.
- The output from the Hot Spot Analysis tool is a Z score & p-value for each feature. These values represent the statistical significance of the spatial clustering of values, given the conceptualization of spatial relationships & the scale of analysis (distance parameter).
- A high Z score & small p-value (probability) for a feature indicates a spatial clustering of high values. A low negative Z score & small p-value indicates a spatial clustering of low values. The higher (or lower) the Z score, the more intense the clustering. A Z score near zero indicates no apparent spatial clustering.
- The Z score is based on the Randomization Null Hypothesis computation.
- The input field should contain a variety of non-negative values. The math for this statistic requires some variation in the variable being analyzed; it cannot solve if all input values are 1, for example. If you have incident data, & want to analyze incident intensity, consider aggregating your incident data or using [Integrate](#) with the [Collect Events](#) tool prior to analysis.
- Whenever using shapefiles keep in mind that they cannot store null values. Tools or other procedures that create shapefiles from non-shapefile inputs may store or interpret null values as zero. This can lead to unexpected results.
- The Conceptualization of Spatial Relationships used for analysis should be based on your understanding of spatial interaction among the features being analyzed. For this tool the fixed distance or contiguity spatial conceptualization methods are generally more appropriate than the inverse distance conceptualization methods.
- For the Fixed Distance option, the distance band used for analysis should be based on your understanding of spatial interaction among the features being analyzed. Alternatively, features may be evaluated for a range of distance values or at the [specific distance](#) where spatial autocorrelation is maximized.
- Use a Conceptualization of Spatial Relationships &/or Distance Band value that will ensure every feature has at LEAST one neighbor. Especially if the input data is skewed (does not create a nice bell curve when you plot the values as a histogram), you want to make sure that the number of neighbors is neither too small (most features have only one or two neighbors) nor too large (several features include all other features as neighbors), because that would make resultant Z scores less reliable. The Z scores are reliable (even with skewed data) as long as each feature is associated with several neighbors (approximately 8, as a rule of thumb). This tool can be applied to skewed data because it is "asymptotically normal".
- For Inverse Distance conceptualization options: when 0 is entered for the "Distance Band or Threshold Distance" parameter all features are considered neighbors of all other features; when this parameter is left blank, a default threshold distance will be applied
- When the spatial conceptualization is an Inverse Distance method (Inverse Distance, Inverse Distance Squared, or Zone of Indifference) any two points that are coincident will be given a weight of one to avoid zero division. This assures features are not excluded from analysis.
- With inverse distance conceptualizations, weights for distances less than 1 become unstable. The weighting for features separated by less than 1 unit of distance (common with [Geographic Coordinate System](#) projections), are given a weight of 1.
- Analysis on features with a Geographic Coordinate System projection is not recommended with the inverse distance spatial conceptualization methods.
- This tool computes the G_i^* statistic where each feature is its own neighbor; however, if you specify a Self Potential field in which all values are zero, the tool performs the G_i statistic (local calculations for a feature exclude the feature's own value).
- In ArcGIS version 9.2, the "Global" standardization option was removed. Global standardization returns the same results as no standardization. Models built with previous versions of ArcGIS that use the Global standardization option may need to be rebuilt.
- See the [Modeling Spatial Relationships](#) help page for further explanation of this tool's parameters.
- Current map layers may be used to define the input feature class. When using layers, only the currently selected features are included in the analysis.
- When this tool runs in ArcMap, the output feature class is automatically added to the Table of Contents (TOC) with default rendering applied to the Z Score field. The hot to cold rendering applied is defined by a layer file in ArcToolbox/Templates/Layers. You can reapply the default rendering, if needed, by [importing](#) the template layer symbology.

Anselin's Local Moran's I (Local Statistic)

Given a set of weighted features, identifies where high or low values cluster spatially, & features with values that are very different from surrounding feature values.



How Cluster & Outlier Analysis: Anselin's Local Moran's I works

Given a set of weighted features, the **Cluster & Outlier Analysis (Anselin Local Morans I)** tool ([ArcToolbox](#) → [Spatial Statistics](#) → [Mapping Clusters](#)) identifies clusters of features with values similar in magnitude. The tool also identifies spatial outliers. To do this, the tool calculates a Local Moran's I value, a Z score, a p-value, & a code representing the cluster type for each feature. The Z score & p-value represent the statistical significance of the computed index value.

Calculations

The Local Moran's I statistic of spatial association is given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X}) \quad (1)$$

where x_i is an attribute for feature i , \bar{X} is the mean of the corresponding attribute, $w_{i,j}$ is the spatial weight between feature i and j , and:

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n w_{i,j}^2}{n-1} - \bar{X}^2 \quad (2)$$

with n equating to the total number of features.

The z_{I_i} -score for the statistics are computed as:

$$z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \quad (3)$$

where:

$$E[I_i] = - \frac{\sum_{j=1, j \neq i}^n w_{i,j}}{n-1} \quad (4)$$

$$V[I_i] = E[I_i^2] - E[I_i]^2 \quad (5)$$

Additional calculations are as follows:

$$E[I^2] = A - B \quad (6)$$

$$A = \frac{(n - b_{2i}) \sum_{j=1, j \neq i}^n w_{i,j}^2}{n-1} \quad (7)$$

$$B = \frac{(2b_{2i} - n) \sum_{k=1, k \neq i}^n \sum_{h=1, h \neq i}^n w_{i,k} w_{i,h}}{(n-1)(n-2)} \quad (8)$$

$$b_{2i} = \frac{\sum_{i=1, i \neq j}^n (x_i - \bar{X})^4}{\left(\sum_{i=1, i \neq j}^n (x_i - \bar{X})^2 \right)^2} \quad (9)$$

The p-values are numerical approximations of the area under the curve for a known distribution, limited by the test statistic.

Interpretation

A positive value for I indicates that the feature is surrounded by features with similar values. Such a feature is part of a cluster. A negative value for I indicates that the feature is surrounded by features with dissimilar values. Such a feature is an outlier. The Local Moran's index can only be interpreted within the context of the computed Z score or p-value

The COType field distinguishes between a statistically significant (0.05 level) cluster of high values (HH), cluster of low values (LL), outlier in which a high value is surround primarily by low values (HL), & outlier in which a low value is surrounded primarily by high values (LH).

Tool Info

- The Cluster/Outlier Analysis with Rendering tool combines the [Clusters & Outlier Analysis](#) & [Z Score Rendering](#) functions.
- Current map layers may be used to define the input feature class. When using layers, only the currently selected features are used in the Local Moran's I operation.
- This tool will only work on the windows operating system. If you are not using windows, please run the [Clusters & Outlier Analysis](#) tool by itself.

Potential applications

Applications can be found in economics, resource management, biogeography, political geography, & demographics.

Usage tips

- This tool honors the environment [output coordinate system](#). Feature [geometry](#) is projected to the output coordinate system prior to analysis, so values entered for the Distance Band/Threshold Distance parameter should match those specified in the output coordinate system. All mathematical computations are based on the output coordinate system [spatial reference](#).
- Calculations based on either Euclidean or Manhattan distance require [projected data](#) to accurately measure distances.
- If you will be running several analyses on a single dataset (e.g., analyzing several different fields) or if you have a dataset with more than 3000 features, it is recommended that you construct the [spatial weights matrix file](#) prior to analysis.
- The cluster & outlier analysis output is a Local Moran's I index value, Z score, P-value & cluster type code for each feature.
- This function creates as derived output the names of the index, z score, p-value, & cluster type result fields.
- The Z scores & p-values are measures of statistical significance which tell you whether or not to reject the null hypothesis, feature by feature. They, in effect, indicate whether the apparent similarity (or dissimilarity) in values for a feature & its neighbors is greater than one would expect in a random distribution.
- The Z Score is based on the Randomization Null Hypothesis computation.
- A high positive Z score for a feature indicates that the surrounding features have similar values (either high values or low value). The COType field indicates HH for a statistically significant (0.05 level) cluster of high values & LL for a statistically significant (0.05 level) cluster of low values.
- A low negative Z score for a feature indicates a statistically significant (0.05 level) spatial outlier. The COType field indicates if the feature has high value & is surrounded by features with low values (HL) or if the feature has a low value & is surrounded by features with high values (LH).
- For line & polygon features, true feature geometric centroids are used in computations.
- The input field should contain a variety of non-negative values. The math for this statistic requires some variation in the variable being analyzed; it cannot solve if all input values are 1, for example. If you have incident data, & want to analyze incident intensity, consider aggregating your incident data or using [Integrate](#) with the [Collect Events](#) tool prior to analysis.
- Whenever using shapefiles keep in mind that they cannot store null values. Tools or other procedures that create shapefiles from non-shapefile inputs may store or interpret null values as zero. This can lead to unexpected results.
- The Conceptualization of Spatial Relationships used for analysis should be based on your understanding of spatial interaction among the features being analyzed.
- For the Fixed Distance option, the distance band used for analysis should be based on your understanding of spatial interaction among the features being analyzed. Alternatively, features may be evaluated for a range of distance values or at the [specific distance](#) where spatial autocorrelation is maximized.
- For Inverse Distance conceptualization options: when zero is entered for the "Distance Band or Threshold Distance" parameter all features are considered neighbors of all other features; when this parameter is left blank, a default threshold distance will be applied.
- When the spatial conceptualization is an Inverse Distance method (Inverse Distance, Inverse Distance Squared, or Zone of Indifference) any two points that are coincident will be given a weight of one to avoid zero division. This assures features are not excluded from analysis.
- With inverse distance conceptualizations, weights for distances less than 1 become unstable. The weighting for features separated by less than 1 unit of distance (common with [Geographic Coordinate System](#) projections), are given a weight of 1.
- Analysis on features with a Geographic Coordinate System projection is not recommended with any of the inverse distance based spatial conceptualization methods.
- In ArcGIS version 9.2, the "Global" standardization option was removed. Global standardization returns the same results as no standardization. Models built with previous versions of ArcGIS that use the Global standardization option may need to be rebuilt.
- See the [Modeling Spatial Relationships](#) help page for further explanation of this tool's parameters.
- Current map layers may be used to define the input feature class. When using layers, only the currently selected features are included in the analysis.
- When this tool runs in ArcMap, the output feature class is automatically added to the Table of Contents (TOC) with default rendering applied to the Z Score field. The hot to cold rendering applied is defined by a layer file in **ArcToolbox/Templates/Layers**. You can reapply the default rendering, if needed, by [importing](#) the template layer symbology.

Additional Resources

The information in this document was copied directly from ArcGIS Desktop Help

➤ Global Statistics

• Getis-Ord General G

- Getis, Arthur, & J.K. Ord. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24, no. 3. 1992
- Mitchell, andy. *The ESRI Guide to GIS Analysis, Volume 2*. ESRI Press, 2005.

• Moran's I

- Goodchild, Michael F. *Spatial Autocorrelation*. *Catmog* 47, Geo Books. 1986.
- Griffith, Daniel. *Spatial Autocorrelation: A Primer*. Resource Publications in Geography, Association of American geographers. 1987.
- Mitchell, andy. *The ESRI Guide to GIS Analysis, Volume 2*. ESRI Press, 2005.

➤ Local Statistics

• Getis-Ord Gi*

- Mitchell, andy. *The ESRI Guide to GIS Analysis, Volume 2*. ESRI Press, 2005.
- [Hot Spot Analysis of 911 Emergency Call Data](#) (5 minute video; select Using Spatial Statistics Tools)
- Scott, L. & Warmerdam, N. [Extend Crime Analysis with ArcGIS Spatial Statistics Tools](#) in [ArcUser Online](#), April - June 2005.

• Anselin's Local Moran's I

- Mitchell, Andy. *The ESRI Guide to GIS Analysis, Volume 2*. ESRI Press, 2005.
- Anselin, Luc. "Local Indicators of Spatial Association – LISA," *Geographical Analysis*, 27(2): 93–115, 1995.