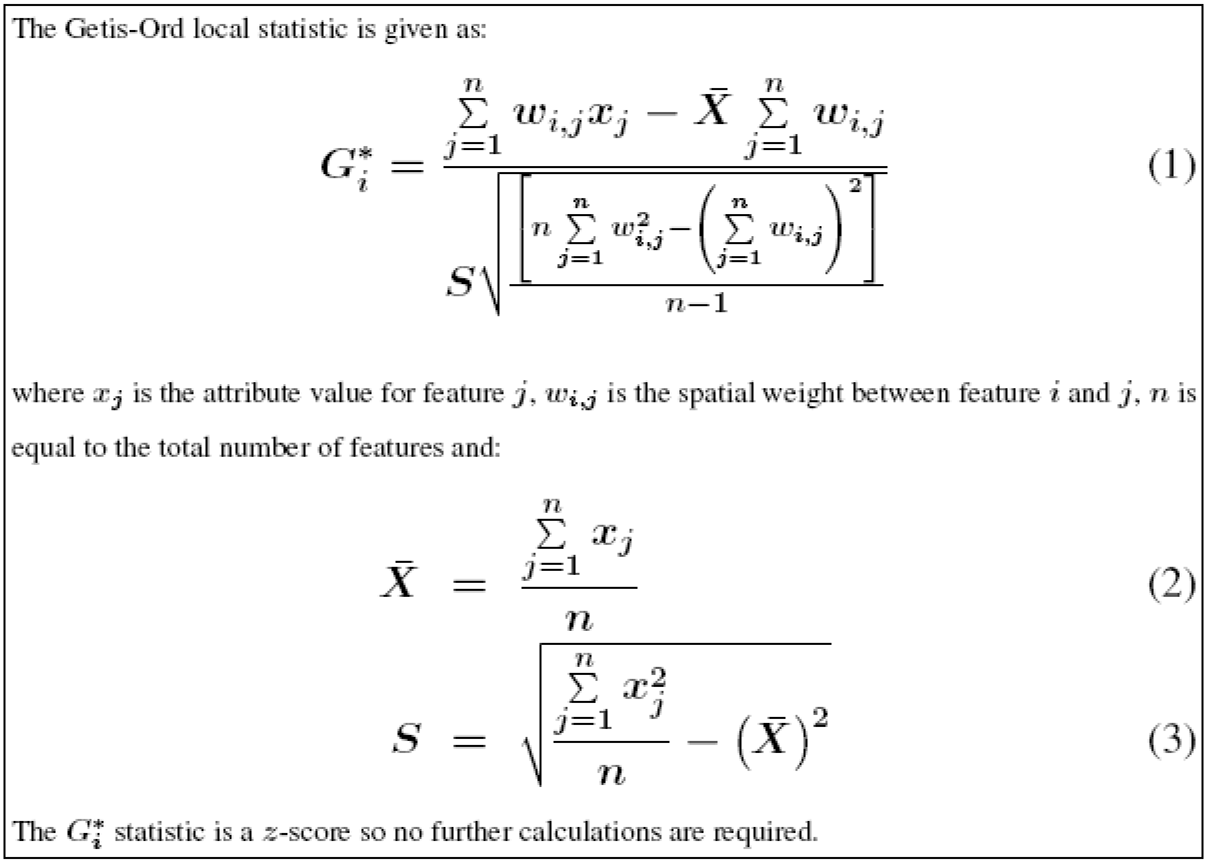
Getis-Ord Gi* statistic

The Hot Spot Analysis tool calculates the Getis-Ord Gi* statistic (pronounced G-i-star) for each feature in a dataset. The resultant z-scores and p-values tell 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 and be surrounded by other features with high values as well. The local sum for a feature and its neighbors is compared proportionally to the sum of all features; when the local sum is very different from the expected local sum, and 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:

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

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:

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

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

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

Interpretation

The Gi* 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). For more information about determining statistical significance, see What is a z-score? What is a p-value?

Output

This tool creates a new **Output Feature Class** with a z-score and p-value for each feature in the **Input Feature Class**. If there is a selection set applied to the Input Feature Class, only selected features will be analyzed, and only selected features will appear in the Output Feature Class. This tool also returns the z-score and p-value field names as derived output values for potential use in custom models and scripts.

When this tool runs in ArcMap, the **Output Feature Class** is automatically added to the table of contents with default rendering applied to the z-score field. The hot to cold rendering applied is defined by a layer file in `<ArcGIS>/ArcToolbox/Templates/Layers`. You can reapply the default rendering, if needed, by importing the template layer symbology.

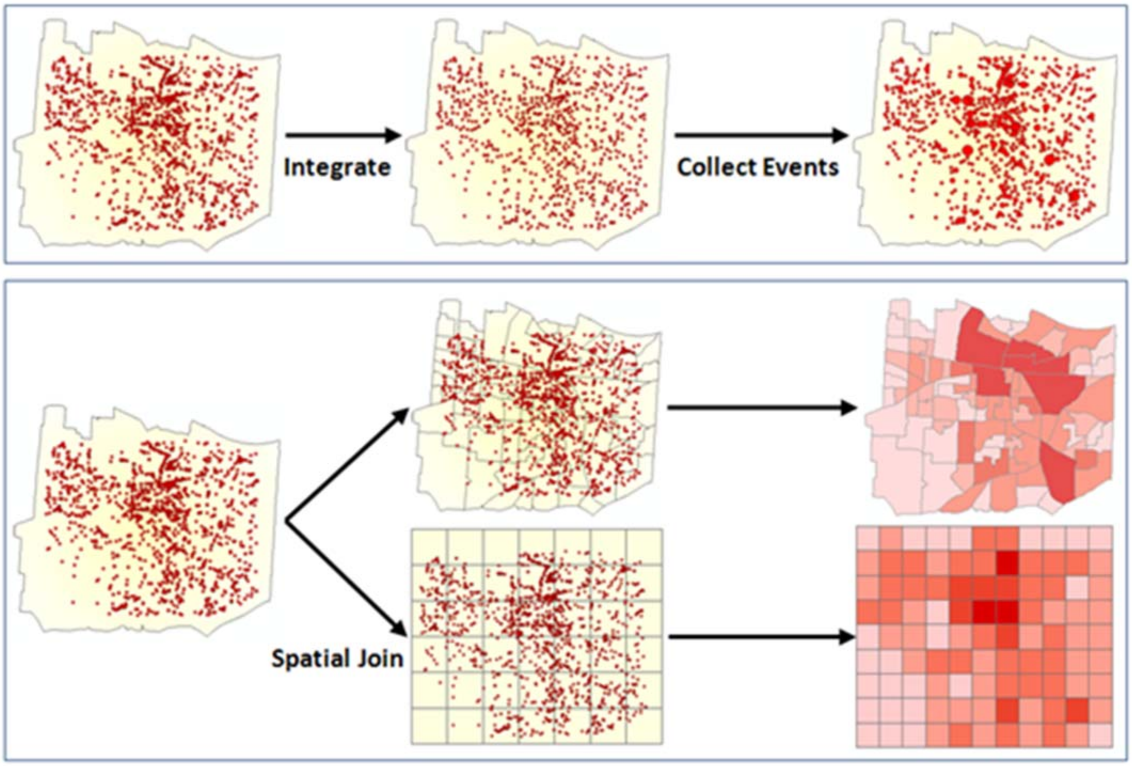
Hot spot analysis considerations

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

1. What is the Analysis Field (**Input 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 polygon features for your study area, you can use the Spatial Join tool to count the number of events in each polygon. The resultant field containing the number of events in each polygon becomes the **Input Field** for analysis.
    - Use the Create Fishnet tool to construct a polygon grid over your point features. Then use the Spatial Join tool to count the number of events falling within each grid polygon. Remove any grid polygons that fall outside 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 **Input Field** for analysis.

        Note:

        If you are concerned that your coincident points may be redundant records, the Find Identical tool can help you to locate and remove duplicates.



Strategies for aggregating incident data

2. Which **Conceptualization of Spatial Relationships** is appropriate? What **Distance Band or Threshold Distance** value is best?

   The recommended (and default) **Conceptualization of Spatial Relationships** for the Hot Spot Analysis (Getis-Ord Gi*) tool is **Fixed Distance Band**. Space-Time Window, Zone of Indifference, Contiguity, K Nearest Neighbor, and Delaunay Triangulation may also work well. For a discussion of best practices and strategies for determining an analysis distance value, see Selecting a Conceptualization of Spatial Relationships and Selecting a Fixed Distance. For more information about space-time hot spot analysis, see Space-Time Analysis.

3. What is the question?

   This may seem obvious, but how you construct the **Input Field** for analysis 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 (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 (for example, the number of crimes is a function of population; the number of foreclosures is a function of housing stock) and identify unexpected hot/cold spots.

Best practice guidelines

- Does the **Input Feature Class** contain at least 30 features? Results aren't reliable with less than 30 features.
- Is the **Conceptualization of Spatial Relationships** you selected appropriate? For this tool, the **Fixed Distance Band** method is recommended. For space-time hot spot analysis, see Selecting a Conceptualization of Spatial Relationships.
- Is the **Distance Band or Threshold Distance** appropriate? See Selecting a Fixed Distance.
  - All features should have at least one neighbor.
  - No feature should have all other features as neighbors.
  - Especially if the values for the **Input Field** are skewed, you want features to have about eight neighbors each.

Potential applications

Applications can be found in crime analysis, epidemiology, voting pattern analysis, economic geography, retail analysis, traffic incident analysis, and demographics. Some examples include the following:

- Where is the disease outbreak concentrated?
- Where are kitchen fires a larger than expected proportion of all residential fires?
- Where should the evacuation sites be located?
- Where/When do peak intensities occur?
- Which locations and at during what time periods should we allocate more of our resources?

Additional resources

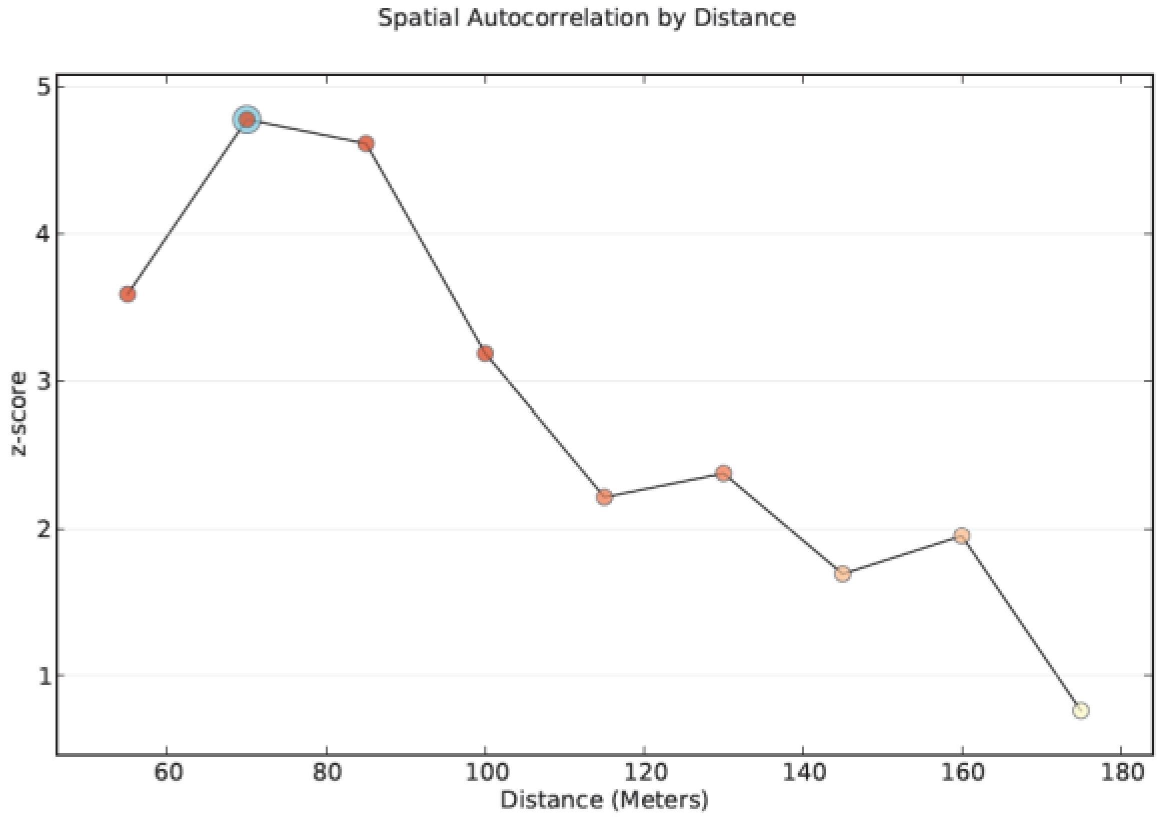
Mitchell, Andy. *The ESRI Guide to GIS Analysis,* Volume 2. ESRI Press, 2005.
Getis, A. and J.K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics" in *Geographical Analysis* 24(3).
Ord, J.K. and A. Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application" in *Geographical Analysis* 27(4).

How Incremental Spatial Autocorrelation works

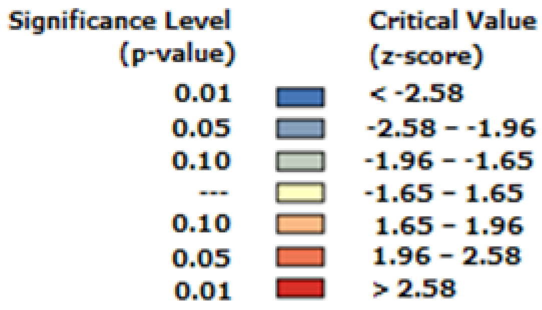
Desktop » Geoprocessing » Tool reference » Spatial Statistics toolbox » Analyzing Patterns toolset

With much of the spatial data analysis you do, the scale of your analysis will be important. The default **Conceptualization of Spatial Relationships** for the Hot Spot Analysis tool, for example, is`FIXED_DISTANCE_BAND` and requires you to specify a distance value. For many density tools you will be asked to provide a **Radius**. The distance you select should relate to the scale of the question you are trying to answer or to the scale of remediation you are considering. Suppose, for example, you want to understand childhood obesity. What is your scale of analysis? Is it at the individual household or neighborhood level? If so, the distance you use to define your scale of analysis will be small, encompassing the homes within a block or two of each other. Alternatively, what will be the scale of remediation? Perhaps your question involves where to increase after-school fitness programs as a way to potentially reduce childhood obesity. In that case, your distance will likely be reflective of school zones. Sometimes it’s fairly easy to determine an appropriate scale of analysis; if you are analyzing commuting patterns and know that the average journey to work is 12 miles, for example, then 12 miles would be an appropriate distance to use for your analysis. Other times it is more difficult to justify any particular analysis distance. This is when the Incremental Spatial Autocorrelation tool is most helpful.

Whenever you see spatial clustering in the landscape, you are seeing evidence of underlying spatial processes at work. Knowing something about the spatial scale at which those underlying processes operate can help you select an appropriate analysis distance. The Incremental Spatial Autocorrelation tool runs the Spatial Autocorrelation (Global Moran’s I) tool for a series of increasing distances, measuring the intensity of spatial clustering for each distance. The intensity of clustering is determined by the z-score returned. Typically, as the distance increases, so does the z-score, indicating intensification of clustering. At some particular distance, however, the z-score generally peaks. Sometimes you will see multiple peaks.



Peaks reflect distances where the spatial processes promoting clustering are most pronounced. The color of each point on the graph corresponds to the statistical significance of the z-score values.



Significance Level (p-value)		Critical Value (z-score)
0.01	🟦	< -2.58
0.05	🟦	-2.58 - -1.96
0.10	🟩	-1.96 - -1.65
---	🟨	-1.65 - 1.65
0.10	🟧	1.65 - 1.96
0.05	🟧	1.96 - 2.58
0.01	🟥	> 2.58

One strategy for identifying an appropriate scale of analysis is to select the distance associated with the statistically significant peak that best reflects the scale of your question. Often this is the first statistically significant peak.

How do I select the Beginning Distance and Distance Increment values?

All distance measurements are based on feature centroids and the default **Beginning Distance** is the smallest distance that will ensure every feature has at least one neighboring feature. This is generally a good choice, unless your dataset includes spatial outliers. Determine whether or not you have spatial outliers, then select all but the outlier features and run Incremental Spatial Autocorrelation on just the selected features. If you find a peak distance for the selection set, use that distance to create a spatial weights matrix file based on all of your features (even the outliers). When you run theGenerate_Spatial_Weights_Matrix tool to create the spatial weights matrix file, set the **Number of Neighbors** parameter to some value so that all features will have at least that many neighboring features.

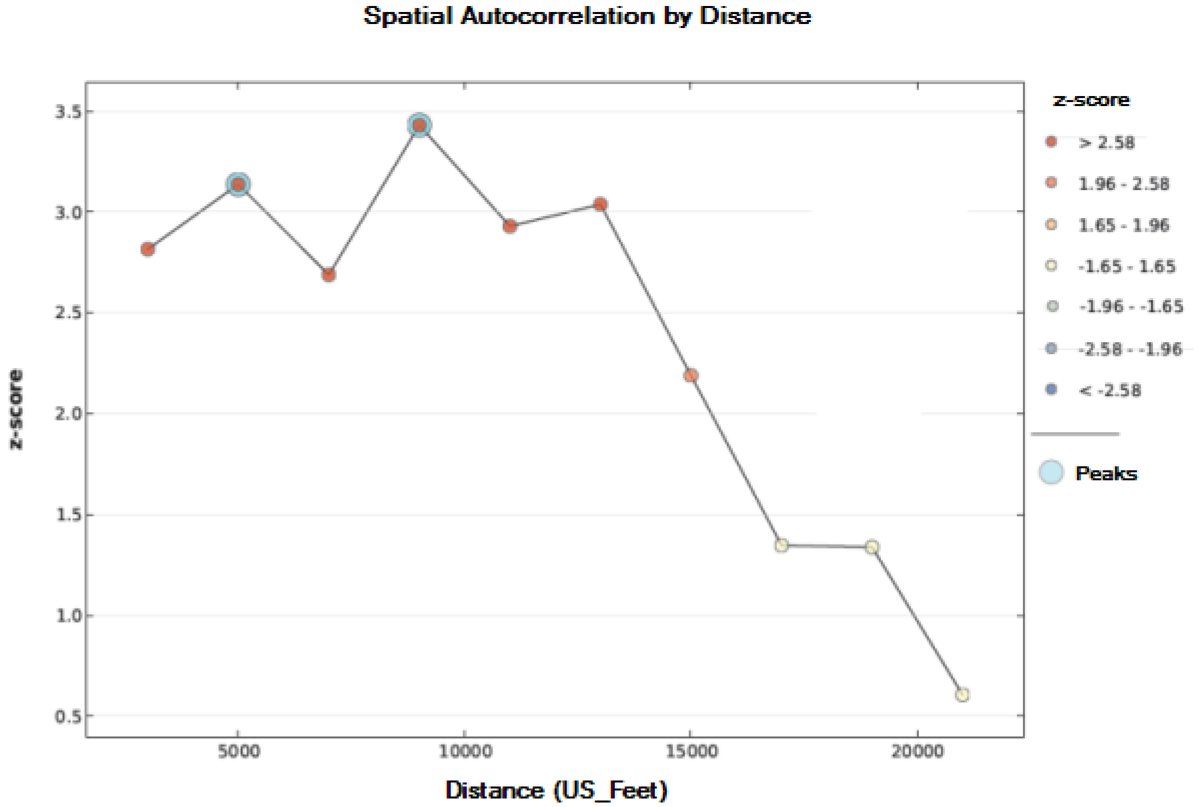
The default **Increment Distance** is the average distance to each feature's nearest neighboring feature. If you've determined an appropriate starting distance using the strategies above and still don't see a peak distance, you may want to experiment with smaller or larger increment distances.

What if the graph never peaks?

In some cases, you will use the Incremental Spatial Autocorrelation tool and get a graph with a z-score that just continues to rise with increasing distances; there is no peak. This most often happens in cases where data has been aggregated and the scale of the processes impacting your **Input Field** variable are smaller than the aggregation scheme. You can try making your Distance Increment smaller to see if this captures more subtle peaks. Sometimes, however, you won't get a peak because there are multiple spatial processes, each operating at a different distance, in your study area. This is often the case with large point datasets that are noisy (no clear spatial pattern to the point data values you're analyzing). In this case, you will need to justify your scale of analysis using some other criteria.

Interpreting results

When you run the Incremental Spatial Autocorrelation tool in the foreground, the z-score results for each distance are written to the ***Progress*** window. This output is also available from the Results window. If you right-click on the Messages entry in the Results window and select **View**, the tool results are displayed in a **Message** dialog box. When you specify a path for the optional **Output Table** parameter, a table is created that includes fields for **Distance**, **MoransI**, **ExpectedI**, **Variance**, **z_score**, and **p_value**. By examining the z-score values in the ***Progress*** window, **Message** dialog box, or **Output Table**, you can determine if there are any peak distances. More typically, however, you would identify peak distances by looking at the graphic in the optional **Output Report** file. The report has three pages. An example of the first page of the report is shown below. Notice that this graph has three peak z-scores associated with distances of 5000, 9000, and 13000 feet. A halo will be drawn to highlight both the first peak distance and the maximum peak distance, but all peaks represent distances where the spatial processes promoting clustering are most pronounced. You can select the peak that best reflects the scale of your analytical question. In some cases, there will only be one halo because the first and the maximum peaks are found at the same distance. If none of the z-score peaks are statistically significant, then none of the peaks will have the light blue halo. Notice that the color of the plotted z-score corresponds to the legend showing the critical values for statistical significance.



On page two of the report, the distances and z-score values are presented in table format. The last page of the report documents the parameter settings used when the tool was run. To get a report file, provide a path for the **Output Report** parameter.