X

Working with spatial statistics

Exercise 1: Spatial pattern analysis
Estimated time: 25 minutes

Exercise 1: Spatial pattern analysis

Estimated time: 25 minutes

Analyzing the spatial patterns of dengue fever

In this tutorial, we are going to use some spatial statistics tools to better understand the pattern of dengue fever within Pennathur in Southern India, one of 44 villages that are part of a dengue fever study.

Dengue fever is a painful, potentially fatal illness spread by a tiny mosquito. Unfortunately, it's quite common in Southeast Asia and Central America. It's estimated that as many as 100 million people contract it annually. And the CDC (Centers for Disease Control and Prevention) is still years away from finding a vaccine.

The goal of this project is to better understand dengue fever in order to identify strategies that might reduce the disease until a vaccine can be found.

Step 1: Analysis

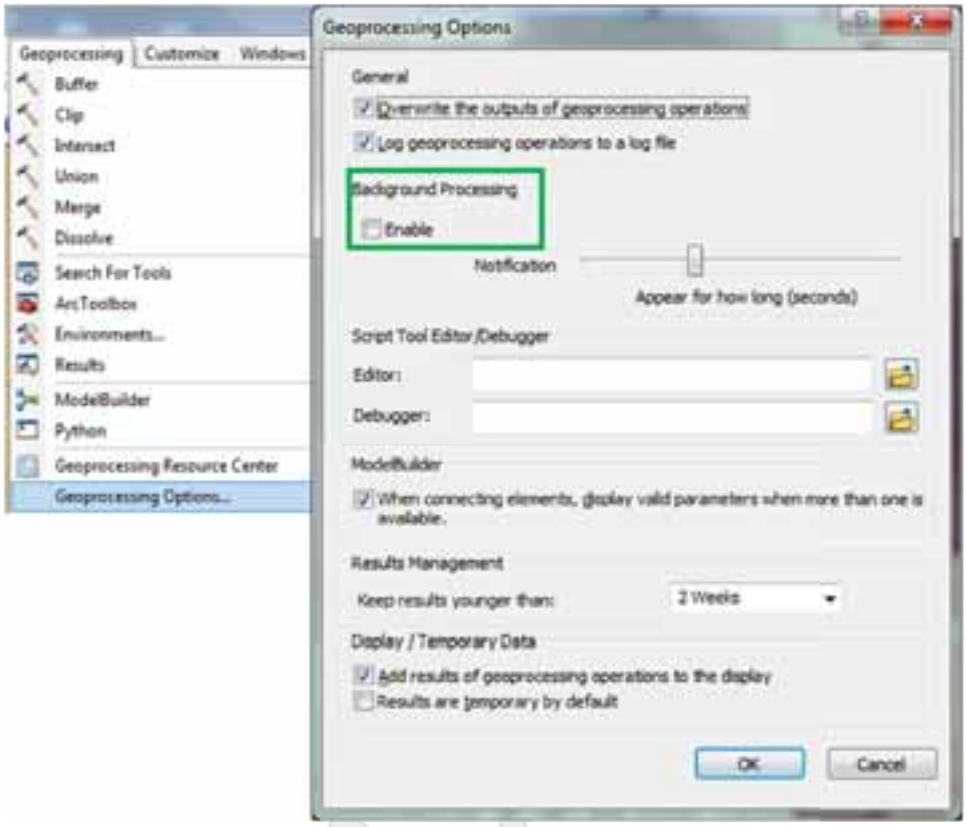
The data for this tutorial project is stored at C:\SpatialStats. If a different location for storing the project data is used, substitute the alternate location for "C:\SpatialStats" when entering data and environment paths below.

☐ Open ArcMap, then open **...\PatternAnalysis\DengueFever.mxd**.

Draft

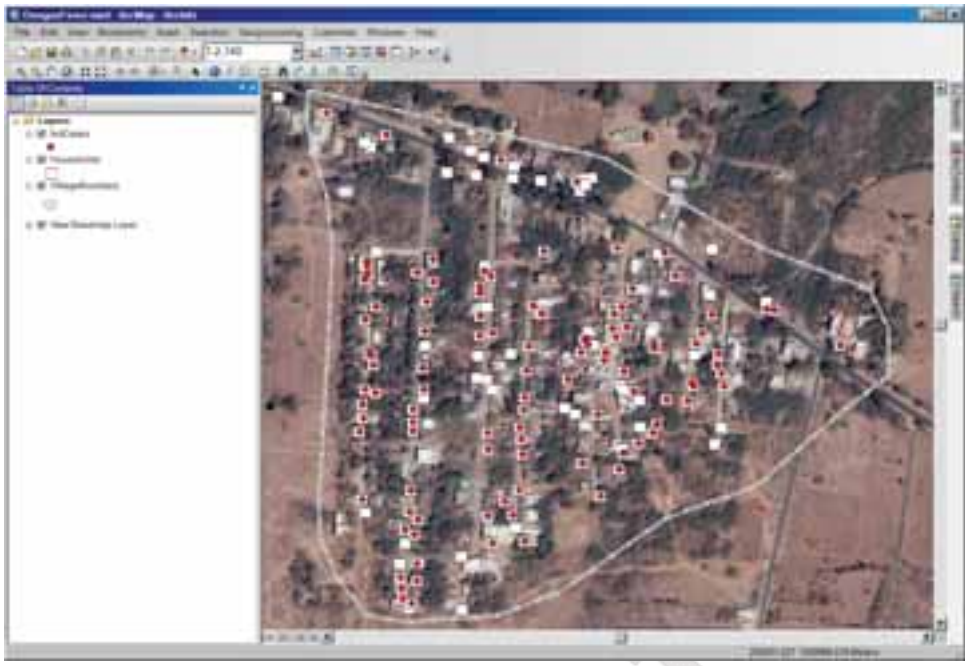
Copyright © Esri

☐ Make sure that Background Geoprocessing is NOT enabled within the Geoprocessing Options dialog box that is accessed from the Geoprocessing menu.



You'll begin by looking at the spatial pattern of dengue fever in Pennathur. The white squares in the graphic below are homes, and the bright red dots are individual dengue fever cases that occurred over a 35-day period.

Copyright © Esri



For your first analysis, you'll use the Average Nearest Neighbor tool to see whether the cases of dengue fever cluster in the village.

☐ From the ArcToolbox on the Standard toolbar, open the Spatial Statistics Toolbox, expand the Analyzing Patterns toolset, and then double-click the Average Nearest Neighbor tool.

The Average Nearest Neighbor tool calculates the distance between each feature and its nearest neighbor, then computes the average for all nearest neighbor distances. It then compares the computed average distance to a theoretical one that would be obtained if the points were randomly distributed inside a circle with the same AREA value. (Notice that the tool also has an optional AREA value.) For this project, you'll use the village boundary polygon for the area. To find out what that AREA is, you'll use the Calculate Areas tool.

☐ Within the Spatial Statistics Toolbox, expand the Utilities Toolset, then double-click Calculate Areas to run the tool with the following parameters:

- Input Feature Class: Village Boundary
- Output Feature Class: accept the default or choose an output location

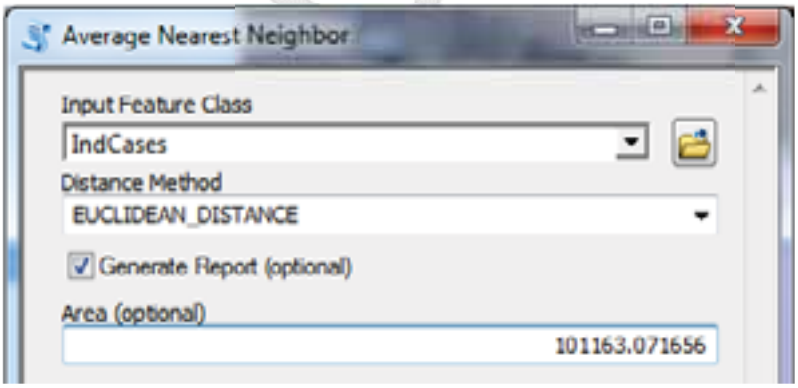
Copyright © Esri

☐ From the table of contents, open the attribute table for the feature class you've just created and copy the F_Area value into the Area field within the Average Nearest Neighbor tool. Then close the attribute table.

☐ Turn off the output from the Calculate Areas tool by unchecking or removing the layer from the table of contents.

☐ Run the Average Nearest Neighbor tool with the following parameters:

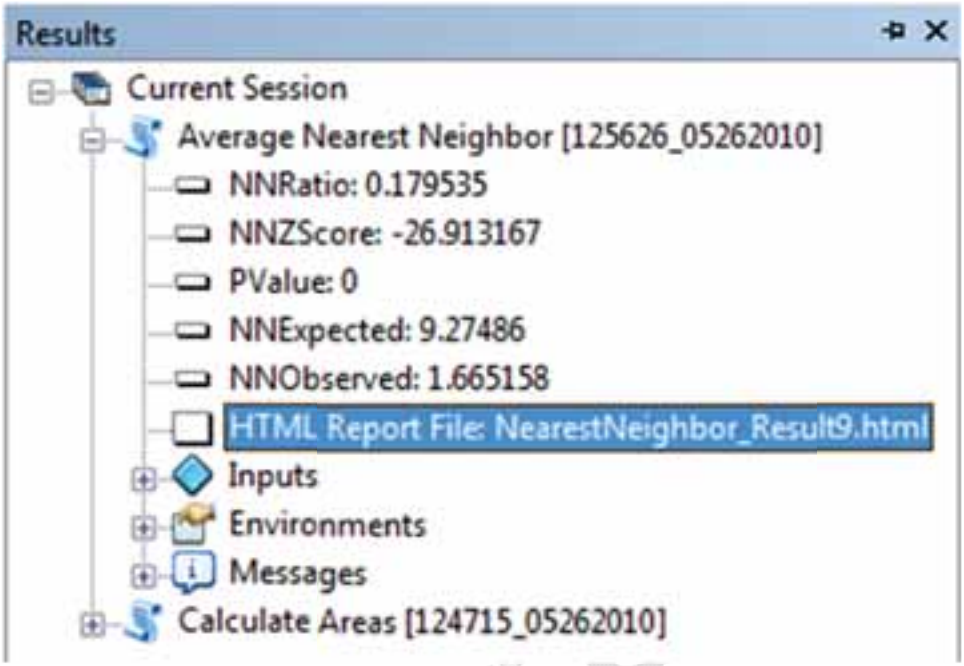
- Input Feature Class: IndCases
- DistanceMethod: EUCLIDEAN_DISTANCE
- Generate Report: check this ON
- Area: 101163.071656



Draft

Copyright © Esri

☐ From the ArcMap menu, choose Geoprocessing > Results to open the Results window. Double-click the HTML Nearest Neighbor report that you've just generated and view the results.



Notice that the report from the Average Nearest Neighbor tool shows that the dengue fever cases are, in fact, clustered.

The numeric output from the Average Nearest Neighbor tool will be addressed below. For now, however, notice that the dengue cases are reported by household. The Average Nearest Neighbor tool indicates that cases are clustered, but if the households themselves are clustered to begin with, and you're collecting data by household, do you think that could impact your results? Absolutely!

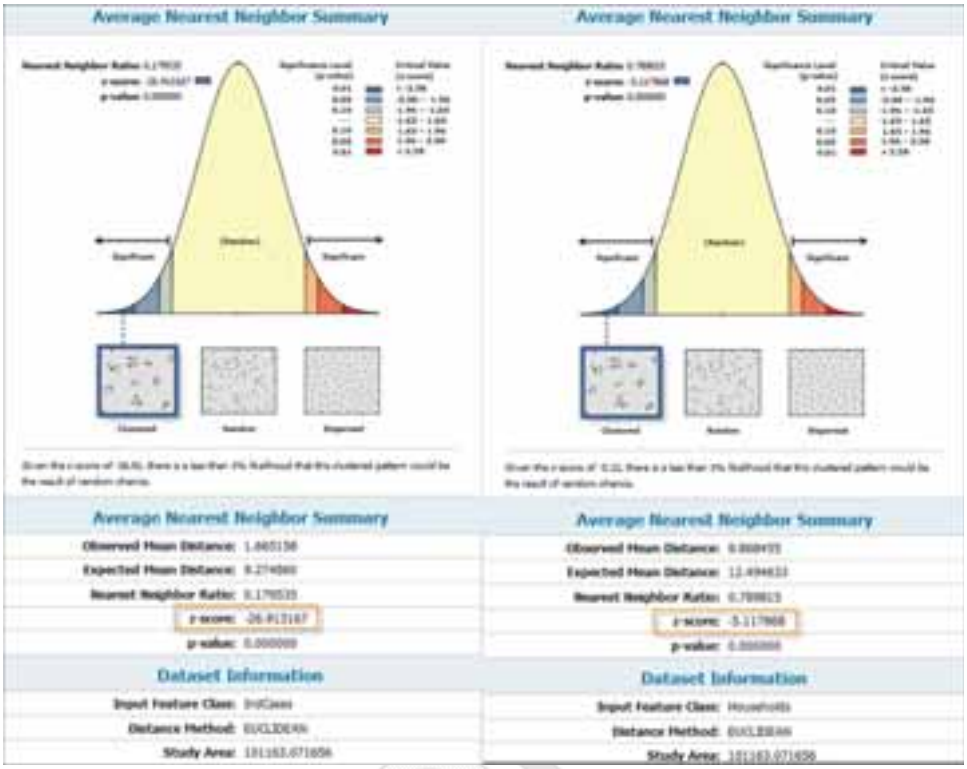
To make sure that this clustering is valid, you'll run the Average Nearest Neighbor tool on the households and compare those results to the clustering for individual cases.

☐ Run the Average Nearest Neighbor tool again with the following parameters:

- Input Feature Class: Households
- DistanceMethod: EUCLIDEAN_DISTANCE
- Generate Report: check this ON
- Area: same as before (This should still be on your clipboard to paste again, but if not, the value is: 101163.071656.)

Copyright © Esri

☐ Open the Results window and double-click the new HTML report. Then compare the HTML report for Households with the HTML report for Individual Cases.



Notice that the report for both Households and for Individual Cases indicates statistically significant clustering.

The z-scores, however, are quite different. For Individual Cases, the z-score is -26.9 and for Households, it's only -5.1. Z-scores are standard deviations and can be plotted on a normal curve. The smaller the number (-26 is smaller than -5), the farther down the z-score falls on the tail of the normal curve. The area under the curve gets very small in the tails, and this area represents the probability that the spatial pattern of your points is randomly distributed.

A z-score of 0 would fall right in the middle of the curve within the location with the largest area under the curve. There's a large probability with this result that the spatial pattern is random.

With very small z-scores, there is a very small probability (less than 1% likelihood) that the spatial pattern is random.

Copyright © Esri

For this tool, when the z-score is in the left tail, the spatial pattern is more clustered than you would find with a random pattern. If the z-score is positive and in the right tail, the spatial pattern is more dispersed than we would find in a random pattern.

It is very important to note that the z-score calculation is *strongly* influenced by the size of the study area. So the best way to use this tool is to compare different distributions within a fixed study area.

In this scenario, you are comparing the spatial pattern of homes to the spatial pattern of dengue fever cases, and you find that the clustering is much more intense for the dengue fever cases than it is for the homes. You can conclude, then, that the spatial pattern of dengue fever cases is clustered, and that the observed clustering is more pronounced than you would expect given the underlying clustering of the homes.

☐ Close the HTML reports.

While having a number that quantifies the clustered spatial pattern for a set of feature is useful, especially when we compare that number to a different set of features in the study area, often we are most interested in WHERE the clustering occurs.

To evaluate where spatial clusters, or hot spots, of dengue occur, we can use a different tool - the hot spot analysis tool. Unlike the Average Nearest Neighbor tool which works on incidents, the hot spot analysis tool evaluates the attributes of points. So we can't just use the individual cases, but instead we need to look at the number of cases within each home.

☐ Right-click Households, and open the attribute table to see the variables that have been calculated.

Each household includes information about the number of cases, the number of people, and a rate indicating the proportion of people who got Dengue in a particular household. So if there are 4 people living in a home, and 2 of them contracted Dengue, the rate would be 50% or 0.5. By the way, for this village as a whole, almost 30% of the population contracted Dengue fever in the 35 day time period (30%!!).

Now, if a disease is random, if getting dengue is purely a function of bad luck, we would expect the number of dengue cases to be a function of the number of people. In other words, we would expect the rates for each household to be around 30%…maybe a little higher for this home, and a little lower for its neighbor…but overall, the rate for all homes would be around 30%. If however, the disease is not random, if it hits harder in particular areas of the village, we can use this information to try to figure out what the causes might be. Does it strike more often in homes with small children or the elderly?

Draft

Copyright © Esri

Are there environmental factors like standing water, or different materials used for housing? The first step in trying to figure out these risk factors is to see if the spatial pattern of the disease is random or not.

☐ Close the Households attribute table.

☐ Open the Hot Spot Analysis tool (in the Spatial Statistics toolbox > Mapping Clusters toolset) and specify the following:

- Input Feature Class: Households
- Input Field: HHRate
- Output Feature Class: accept the default or choose a new location

The next few parameters for the Hot Spot Analysis tool are related to the way that we represent our scale of analysis. The hot spot analysis tool assesses each feature, each household in this case, within the context of its neighbors. The first parameter is the conceptualization of spatial relationships, and we are going to choose the Fixed Distance Band because it ensures that we have the same scale of analysis across the entire study area.

☐ For Conceptualization of Spatial Relationships, choose FIXED_DISTANCE_BAND.

The next parameter we want to set is the distance band, or the actual scale that we are going to use for the analysis. This is how we identify which households are considered neighbors in our analysis.

There are a number of strategies for picking a good distance band. We said the disease is spread by mosquitoes, so if we have information about the distance this mosquito can travel, this would be a good value to use for the distance parameter. Another strategy is to let the data help us understand the spatial scale of the processes we are analyzing.

The Spatial Autocorrelation tool will measure the degree of clustering for different distances. If we run that tool for a series of distances, and write down the Z score for each one, we can find the distance where the Z score peaks. The largest Z score indicates the scale where clustering is most intense, or in other words, the scale where the processes promoting clustering are most pronounced (in this case the spatial processes include the activity of mosquitoes and the mobility of infected people in the village).

☐ Minimize the Hot Spot Analysis tool.

The hot spot analysis tool works by assessing each feature, each household, within the context of neighbors, so we want to start looking for the peak using a distance that will

Draft

Copyright © Esri

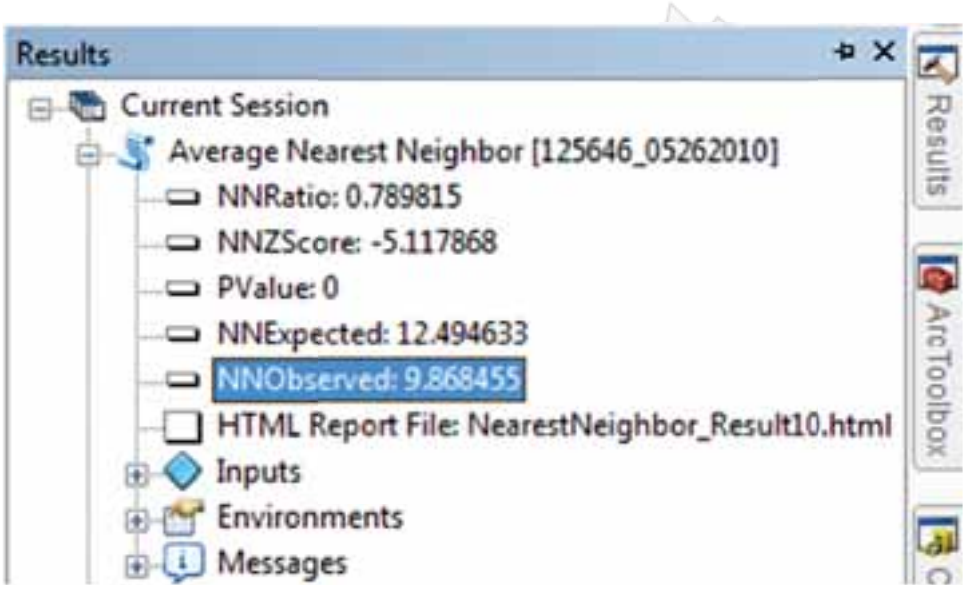
ensure each feature has at least one neighbor. We can use the Calculate Distance Band from Neighbor tool to find this distance.

☐ Run the Calculate Distance Band from Neighbor tool (in the Spatial Statistics toolbox > Utilities toolset) with the following parameters:

- Input Features: Households
- Neighbors: 1
- Distance Method: EUCLIDEAN_DISTANCE

The maximum distance will ensure one neighbor for every feature and we see that's 52.7, so we'll round up and start with a distance of 55.

☐ In the results window, point to the last run of Average Nearest Neighbor, and look at the Observed Nearest Neighbor Distance (NNObserved).



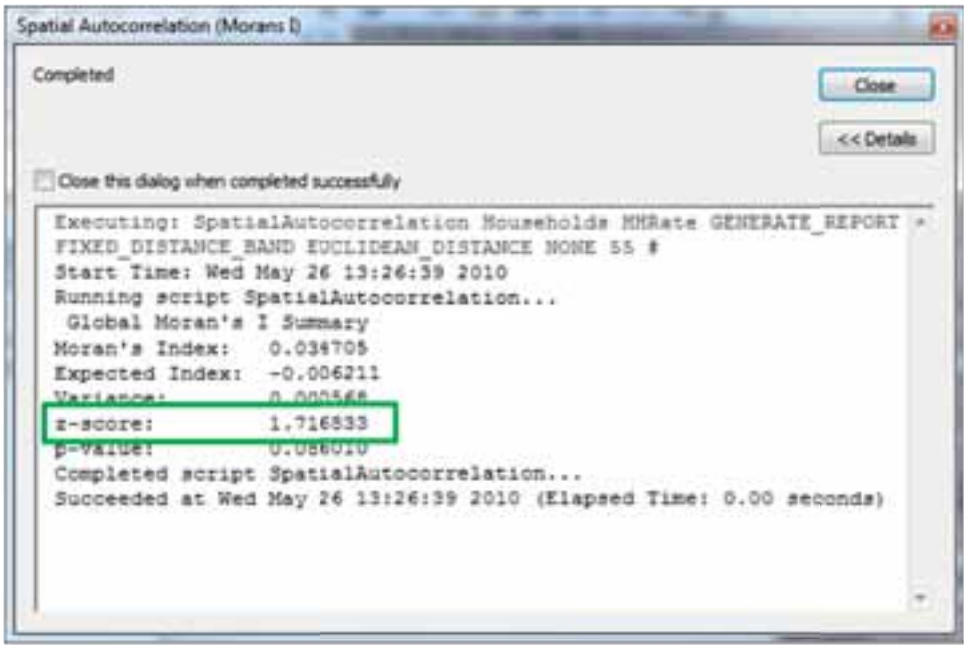
Notice that the Average Nearest Neighbor for households in the village is about 10 meters (9.868), so each time we run the tool, we will increase the distance by 10 meters.

Copyright © Esri

☐ Run the Spatial Autocorrelation tool (in the Spatial Statistics toolbox > Analyzing Patterns toolset):

- Input Feature Class: Households
- Input Field: HHRate
- Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND
- Distance Method: EUCLIDEAN_DISTANCE
- Distance Band: 55 (Remember, we used the Calculate Distance Band from Neighbor Tool to figure this starting distance out.)

☐ Note the Z score of 1.7.



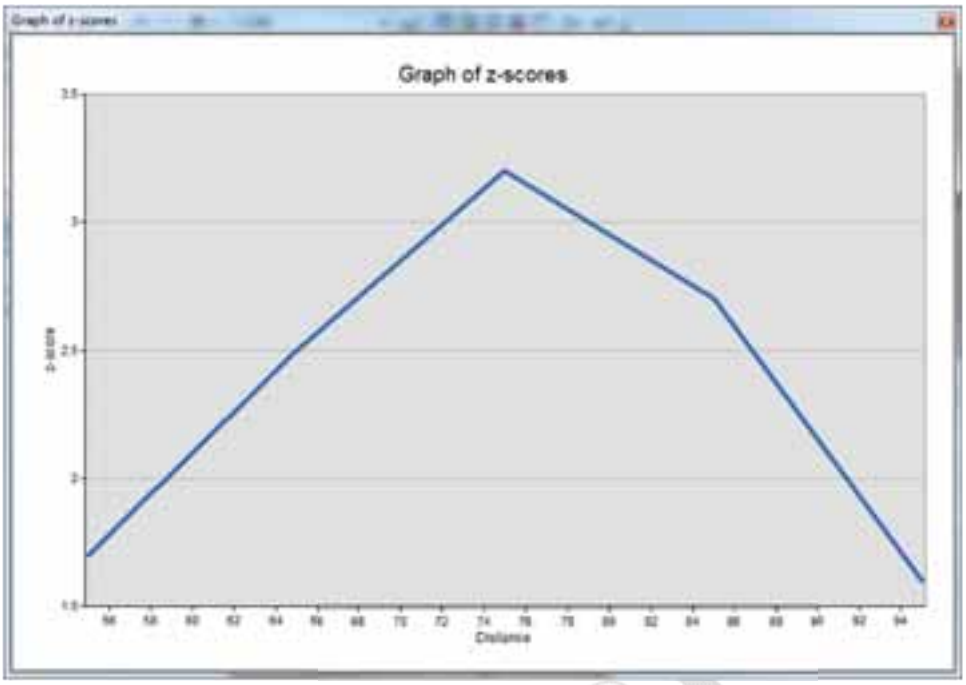
☐ Run the Spatial Autocorrelation tool again with the following parameters:

- Input Feature Class: Households
- Input Field: HHRate
- Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND
- Distance Band: 65 (Remember, we decided to use increments of 10 because it is the observed Average Nearest Neighbor Distance.)

☐ Note the Z score of 2.5.

We could continue to run the Spatial Autocorrelation (Morans I) tool over and over, for 65, 70, 75 meters, etc. We could then create a line graph relating distance to each z-score result. The graph might look like the following:

Copyright © Esri



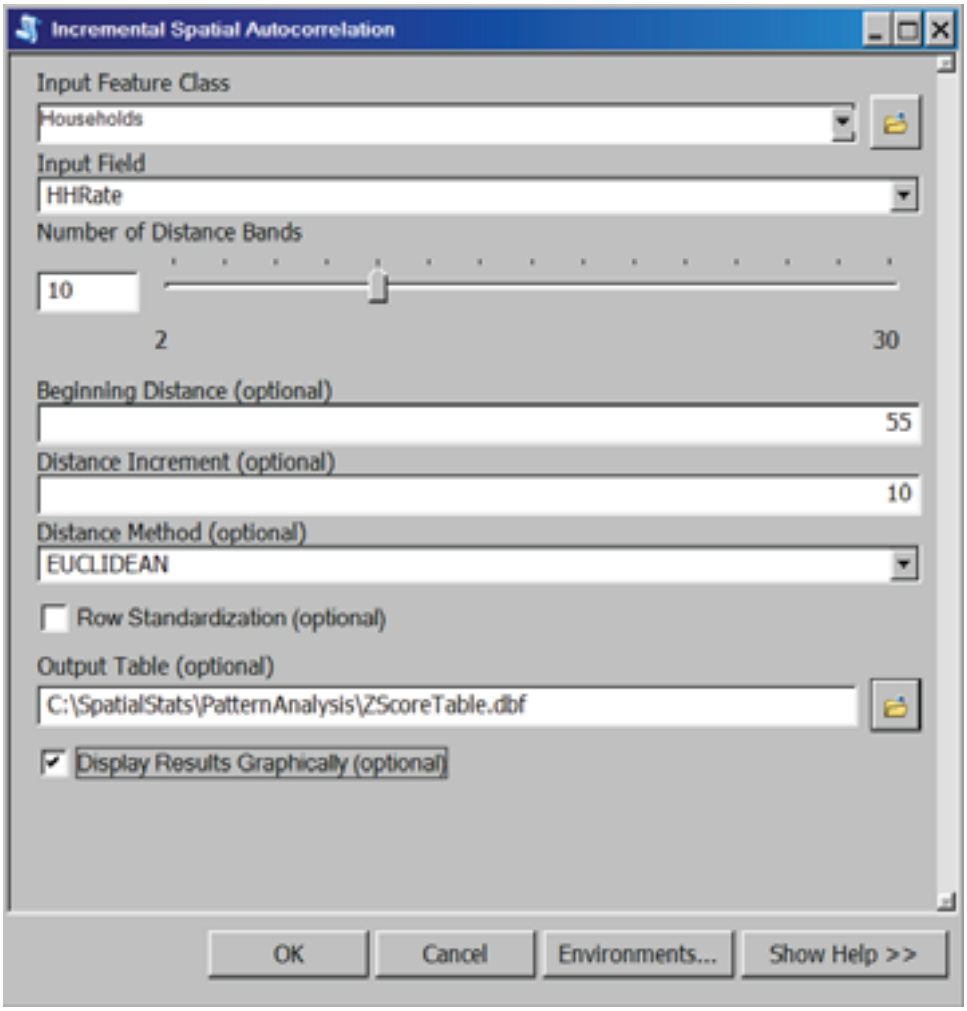
Because finding the peak z-score is such a common task, we've created a sample script tool to run Spatial Autocorrelation multiple times and create the graph for you.

☐ In the Catalog window, navigate to the Supplementary Spatial Statistics toolbox in the C:\SpatialStats\PatternAnalysis\SupplementarySpatialStatistics folder.

☐ Find the Incremental Spatial Autocorrelation tool in the Analyzing Patterns toolset.

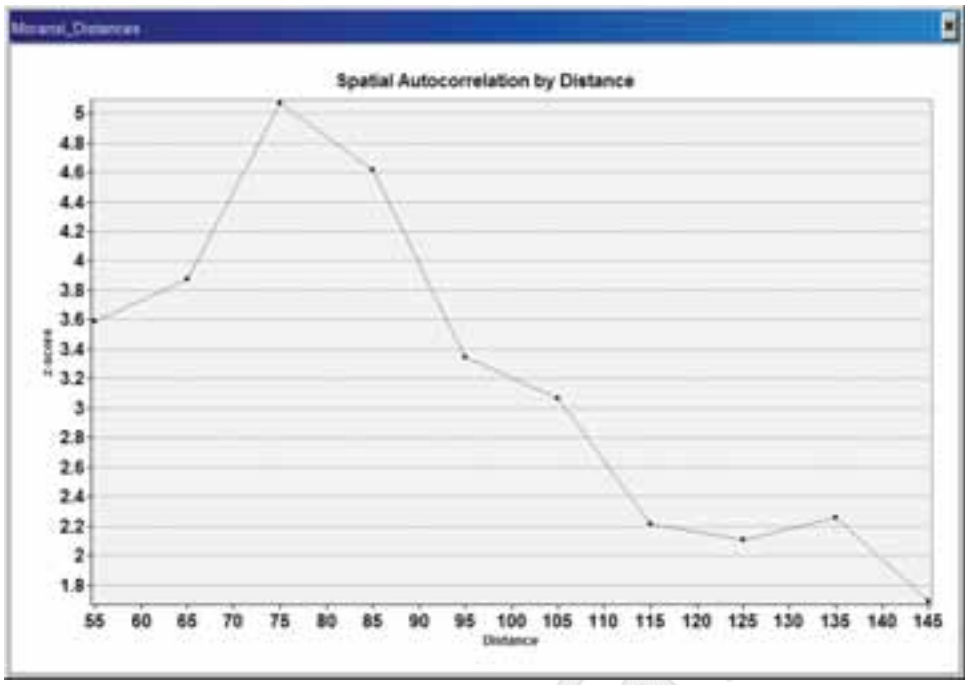
Copyright © Esri

☐ Open the Incremental Spatial Autocorrelation tool and fill it out as follows. You will start at 55 meters, increment by 10 meters, uncheck Row Standardization, save the results to a table, and create a graph:



Copyright © Esri

☐ Verify that your line graph matches the following:



Notice the peak at 75 meters. We'll use this distance for our hot spot analysis.

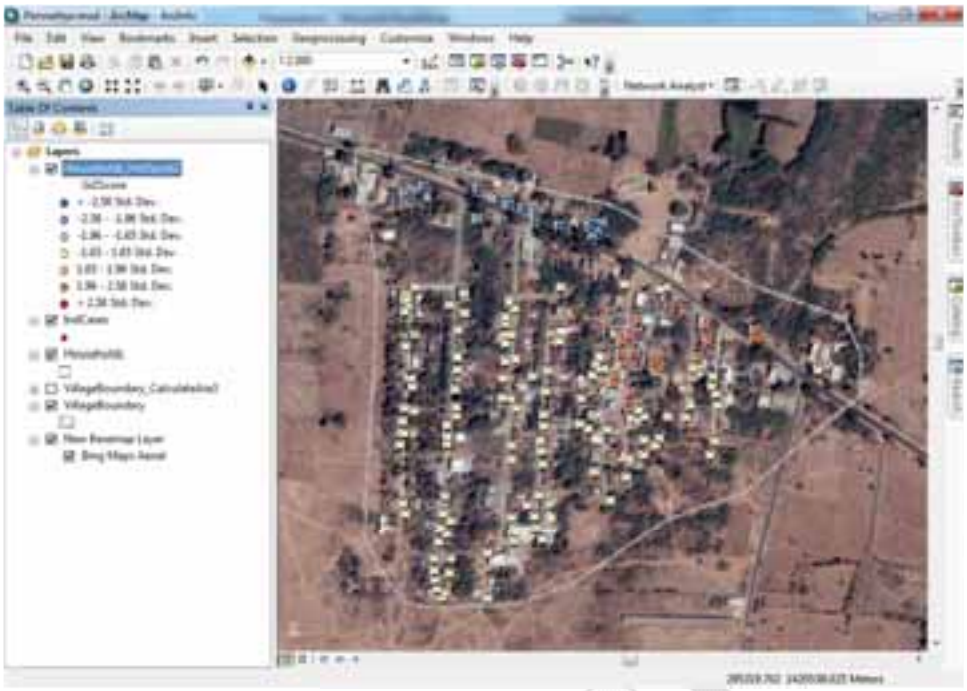
☐ Close the graph.

☐ Restore the Hot Spot Analysis tool that you minimized, and run the tool with the following parameters:

- Input Feature Class: Households
- Input Field: HHRate
- Output Feature Class: accept the default or choose a new location
- Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND
- Distance Band: 75 meters

Draft

Copyright © Esri



Notice that we DO see hot spots (red) and cold spots (blue).

What does this tell us? Well, the first thing it tells us is that getting dengue is not just about bad luck! And where we find the hot and cold spots is a first clue in trying to determine the risk factors. The next step would be analysis to try to figure out the factors that are contributing to higher rates of dengue in these hot spot areas.

Conclusion

Whenever one of these 44 villages has an outbreak of dengue fever, a team of epidemiologists rush in to collect data. There isn't a vaccine, but if we can learn more about the spatial pattern of this disease, and then use what we learn to identify risk factors, we will be in a better position to protect the people from this disease—perhaps we can experiment with different strategies (bed nets, vitamin supplements, pesticides) to see if any of these efforts break the pattern and improve outcomes.

Copyright © Esri