



Working with spatial statistics

Estimated time: 25 minutes

Copyright © ESRI

All rights reserved.

Course version . VERSION RELEASE DATE NOT SET.

Printed in the United States of America.

The information contained in this document is the exclusive property of ESRI. This work is protected under United States copyright law and other international copyright treaties and conventions. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information storage or retrieval system, except as expressly permitted in writing by ESRI. All requests should be sent to Attention: Contracts and Legal Services Manager, ESRI, 380 New York Street, Redlands, CA 92373-8100 USA.

EXPORT NOTICE: Use of these Materials is subject to U.S. export control laws and regulations including the U.S. Department of Commerce Export Administration Regulations (EAR). Diversion of these Materials contrary to U.S. law is prohibited.

The information contained in this document is subject to change without notice.

U. S. GOVERNMENT RESTRICTED/LIMITED RIGHTS

Any software, documentation, and/or data delivered hereunder is subject to the terms of the License Agreement. The commercial license rights in the License Agreement strictly govern Licensee's use, reproduction, or disclosure of the software, data, and documentation. In no event shall the U.S. Government acquire greater than RESTRICTED/LIMITED RIGHTS. At a minimum, use, duplication, or disclosure by the U.S. Government is subject to restrictions as set forth in FAR §52.227-14 Alternates I, II, and III (DEC 2007); FAR §52.227-19(b) (DEC 2007) and/or FAR §12.211/12.212 (Commercial Technical Data/Computer Software); and DFARS §252.227-7015 (NOV 1995) (Technical Data) and/or DFARS §227.7202 (Computer Software), as applicable. Contractor/Manufacturer is ESRI, 380 New York Street, Redlands, CA 92373-8100, USA.

@esri.com, 3D Analyst, ACORN, Address Coder, ADF, AML, ArcAtlas, ArcCAD, ArcCatalog, ArcCOGO, ArcData, ArcDoc, ArcEdit, ArcEditor, ArcEurope, ArcExplorer, ArcExpress, ArcGIS, ArcGlobe, ArcGrid, ArcIMS, ARC/INFO, ArcInfo, ArcInfo Librarian, ArcLessons, ArcLocation, ArcLogistics, ArcMap, ArcNetwork, *ArcNews*, ArcObjects, ArcOpen, ArcPad, ArcPlot, ArcPress, ArcReader, ArcScan, ArcScene, ArcSchool, ArcScripts, ArcSDE, ArcSdl, ArcSketch, ArcStorm, ArcSurvey, ArcTIN, ArcToolbox, ArcTools, ArcUSA, *ArcUser*, ArcView, ArcVoyager, ArcWatch, ArcWeb, ArcWorld, ArcXML, Atlas GIS, AtlasWare, Avenue, BAO, Business Analyst, Business Analyst Online, BusinessMAP, CommunityInfo, Database Integrator, DBI Kit, EDN, ESRI, ESRI—Team GIS, ESRI—The GIS Company, ESRI—The GIS People, ESRI—The GIS Software Leader, FormEdit, GeoCollector, Geographic Design System, Geography Matters, Geography Network, GIS by ESRI, GIS Day, GIS for Everyone, GISData Server, JTX, MapIt, Maplex, MapObjects, MapStudio, ModelBuilder, MOLE, MPS—Atlas, PLTS, Rent-a-Tech, SDE, SML, Sourcebook·America, Spatial Database Engine, StreetMap, Tapestry, the ARC/INFO logo, the ArcGIS logo, the ArcGIS Explorer logo, the ArcPad logo, the ESRI globe logo, the ESRI Press logo, the GIS Day logo, the MapIt logo, The Geographic Advantage, The Geographic Approach, The World's Leading Desktop GIS, *Water Writes*, www.esri.com, www.geographynetwork.com, www.gis.com, www.gisday.com, and Your Personal Geographic Information System are trademarks, registered trademarks, or service marks of ESRI in the United States, the European Community, or certain other jurisdictions.

Other companies and products mentioned herein may be trademarks or registered trademarks of their respective trademark owners.

Spatial Pattern Analysis

Estimated time: 25 minutes

Analyzing the Spatial Patterns of Dengue Fever

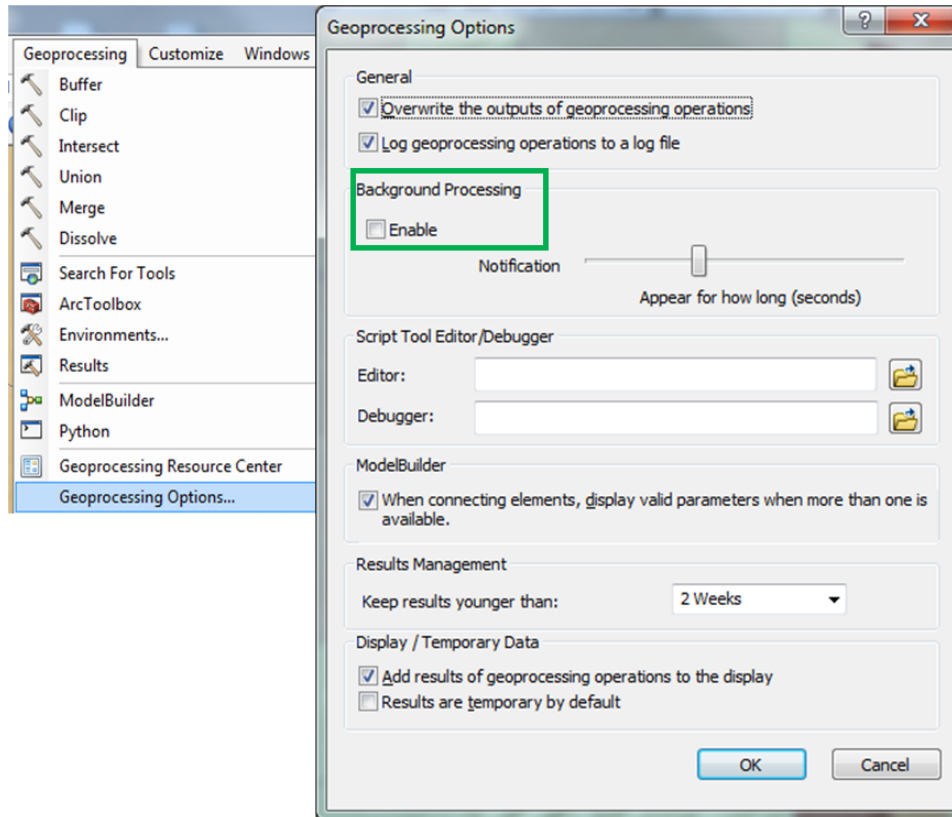
In this tutorial we are going to use some of the spatial statistics tools to better understand the pattern of Dengue Fever for Pennathur, a village in Southern India. This village is one of 44 villages that are part of a Dengue Fever study. Dengue Fever is a painful, potentially fatal illness that is spread by a tiny mosquito, and unfortunately it is quite common in Southeast Asia and Central America. It is estimated that as many as 100 million people contract this disease each year...and the CDC is still years away from finding a vaccine. So the project you are looking at is an attempt to better understand the disease in order to identify strategies that might reduce the disease until a vaccine can be found.

Step 1: Analysis

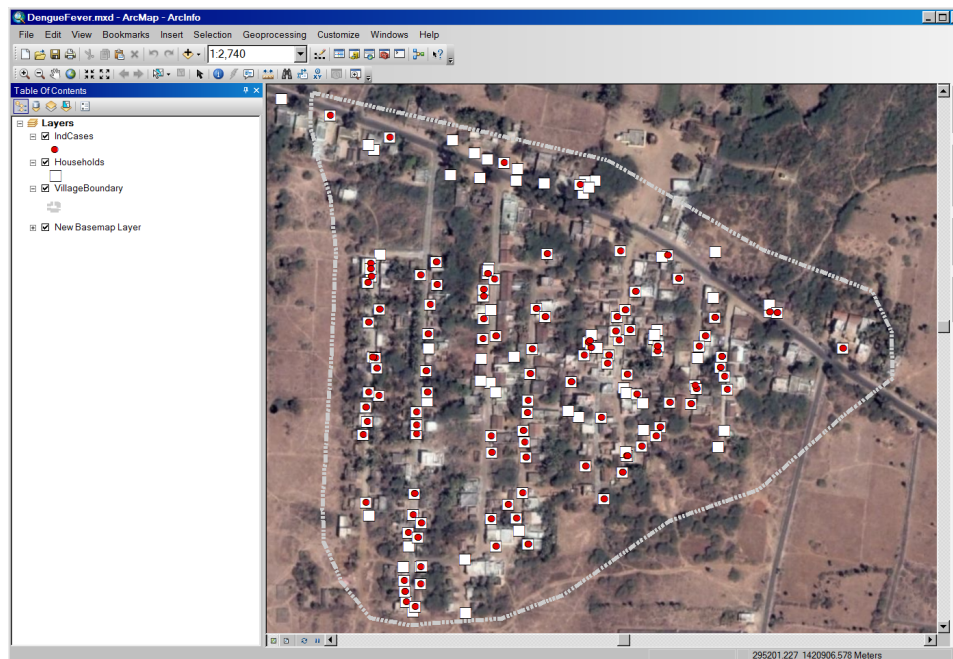
The steps in this tutorial document assume the tutorial data is stored at C:\SpatialStats. If a different location is used, substitute "C:\SpatialStats" with the alternate location when entering data and environment paths.

☐ Open ...\PatternAnalysis\DengueFever.mxd

- Make sure that Background Geoprocessing is NOT enabled
(Geoprocessing>Geoprocessing Options)



We are going to begin by looking at the spatial pattern of Dengue in this village. The white squares are homes, and the bright red dots are individual dengue fever cases over a 35 day time period.



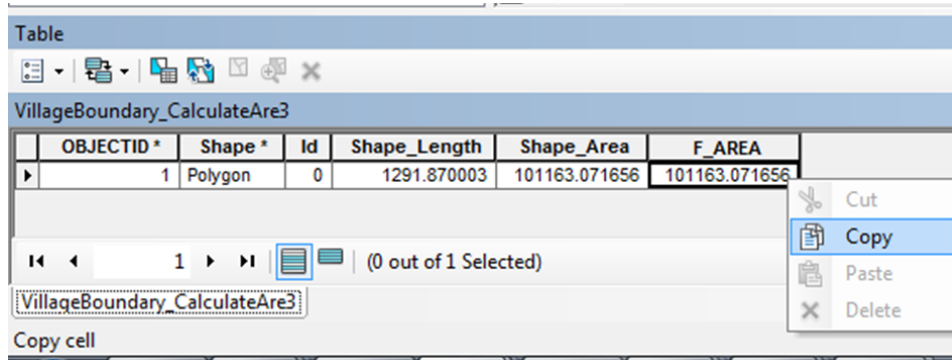
For our first analysis, we'll use the Average Nearest Neighbor tool to see if the cases of dengue fever cluster in the village.

- ☐ Open the Average Nearest Neighbor tool, in the Spatial Statistics Toolbox in the Analyzing Patterns toolset.

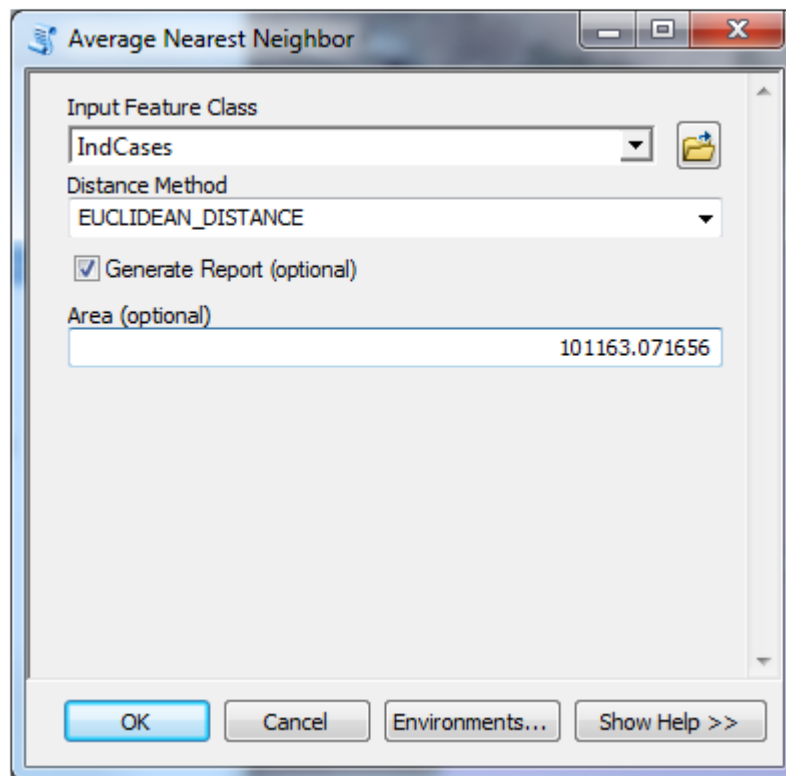
Notice that the tool asks for an optional AREA value. The average nearest neighbor tool calculates the distance between each feature and its nearest neighbor, then computes the average for all nearest neighbor distances. It then compares the computed average distance to a theoretical one that would be obtained if the points were randomly distributed inside a circle with the same AREA value. We'll use the area for the village boundary polygon. To find out what that AREA is, we can run the Calculate Area tool:

- ☐ Run the Calculate Area tool, in the Spatial Statistics Toolbox, in the Utilities Toolset
 - Input Feature Class: Village Boundary
 - Output Feature Class: accept the default or choose an output location

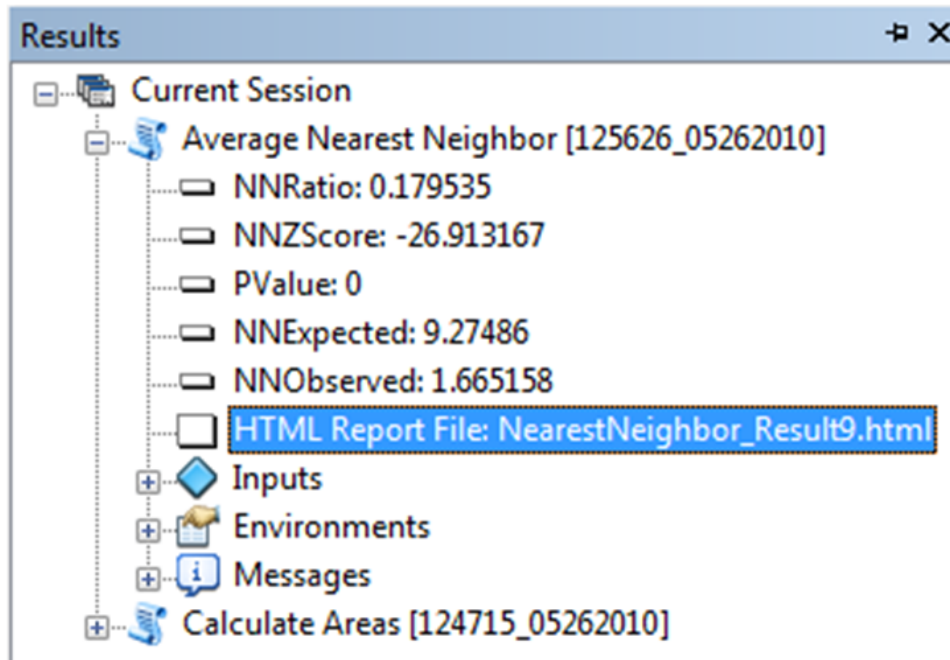
- ☐ Open the attribute table for the feature class created and copy the F_Area value into Area parameter for the Average Nearest Neighbor tool



- ☐ Turn off the output from the Calculate Area tool (uncheck the layer or remove the layer from the TOC)
- ☐ Run Average Nearest Neighbor tool with the following options:
 - Input Feature Class: IndCases
 - DistanceMethod: EUCLIDEAN_DISTANCE
 - Generate Report checked ON
 - Area: 101163.071656



- ☐ Open the results window and double click the HTML report.



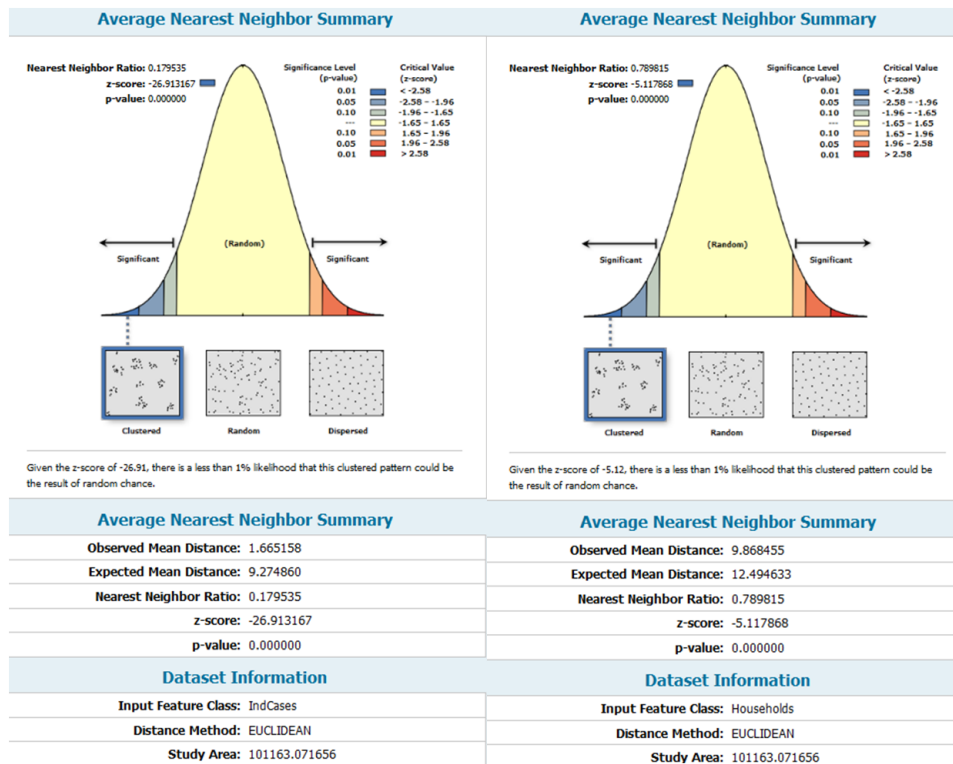
Notice that the report from the Average Nearest Neighbor tool shows us that the Dengue cases are, in fact clustered.

We'll talk more about the numeric output from the Average Nearest Neighbor tool in a minute. What I'd like you to notice right now, however, is that the dengue cases are reported by household. The Average Nearest Neighbor tool indicated that cases were clustered, but if the households are clustered, and if we are collecting data by household, do you think that could impact our results? Absolutely!

To make sure that the clustering is valid, let's run the average nearest neighbor tool on the households and compare the result to the clustering for individual cases.

- ☐ Open the Ave Nearest Neighbor tool
 - Input Feature Class: Households
 - DistanceMethod: EUCLIDEAN_DISTANCE
 - Generate Report checked ON
 - Area: Same as before (This should still be on your clipboard to paste again, but if not the value is: 101163.071656)
- ☐ Open the results window and double click the HTML report

- Compare the HTML report for Households with the HTML report for the Individual Cases



Notice that the report for both Households and Individual Cases indicates statistically significant clustering. The Z scores, however, are quite different. For cases the Z score is -26.9 and for households it is 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 on the tail of the normal curve the Z score falls; 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, in the location with the largest area under the curve and so the probability that the spatial pattern is random would be large. With very small Z scores there is a very small probability (less than 1% likelihood) that the spatial pattern is random. For this tool, when the Z score is in the left tail, the spatial pattern is more clustered than we would find with a random pattern. If the Z score is positive, 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 our case, we are comparing the spatial pattern of homes to the spatial pattern of Dengue fever cases and we find that the clustering is much more intense for the Dengue fever cases than it is for the homes. We can conclude, then, that the spatial pattern of dengue fever cases is clustered, and that the observed clustering is more pronounced than we 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 on 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? 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, in the Mapping Clusters toolset

- 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.

- Conceptualization of Spatial Relationships: 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 ensure each feature has at least one neighbor. We can use the Calculate Distance Band from Neighbor tool to find this distance.

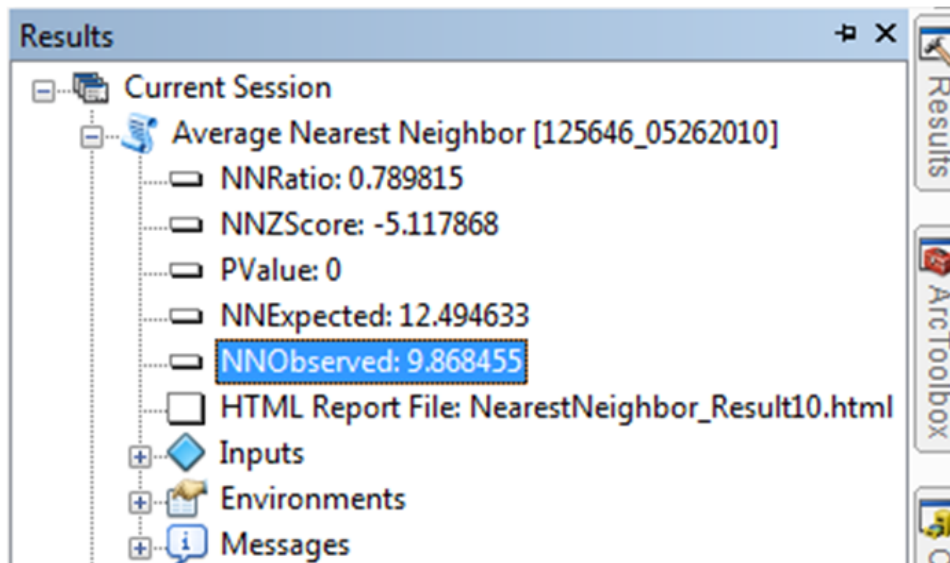
☐ Open the Calculate Distance Band from Neighbor tool, in the Spatial Statistics toolbox, in the Utilities toolset

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

- ☐ Run the tool

The maximum distance will ensure 1 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).

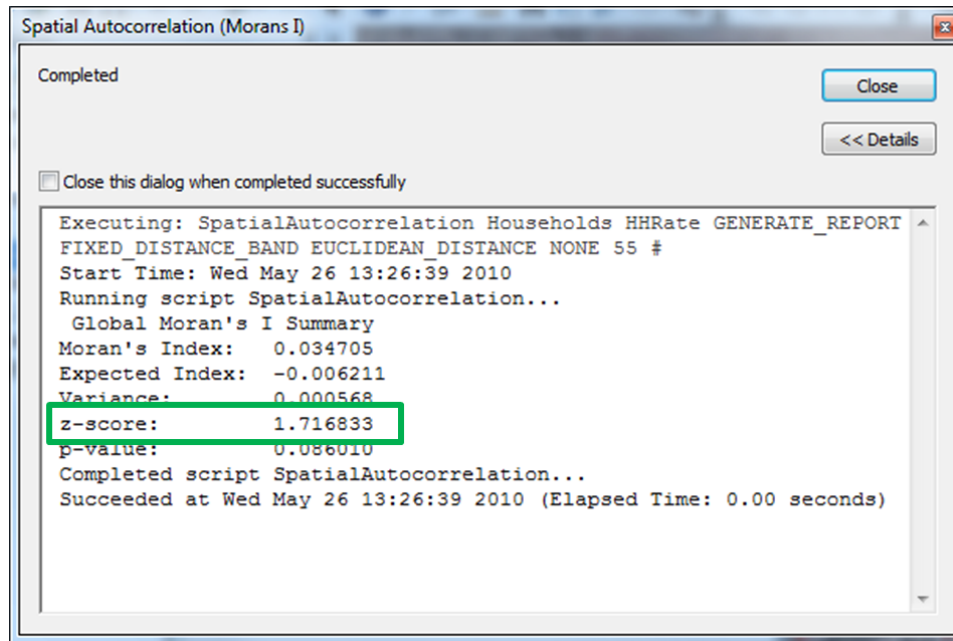


Notice that the Average Nearest Neighbor for households in the village is about 10 meters (9.868), so each time we run the tool, we will increase the distance by 10 meters.

- ☐ Open the Spatial Autocorrelation tool, in the Spatial Statistics toolbox in the Analyzing Patterns toolset
 - Input Feature Class: Households
 - Input Field: HHRate
 - Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND
 - Distance Band: 55 (remember, we used the Calculate Distance Band from Neighbor Tool to figure this starting distance out)

- ☐ Run the tool

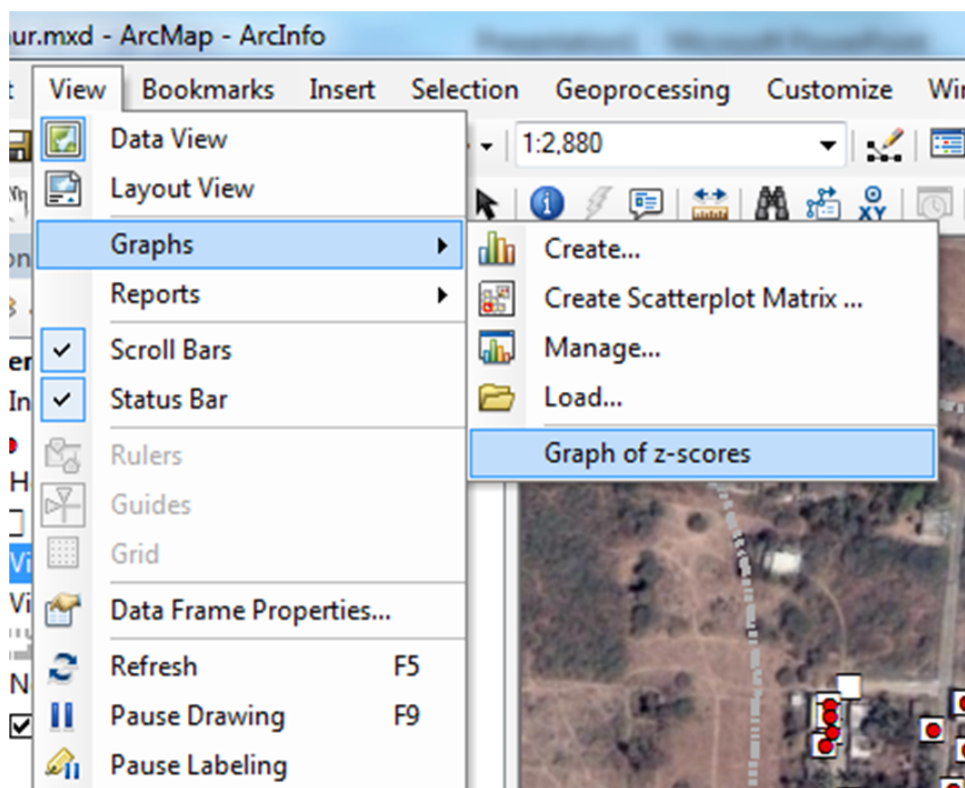
- ☐ Note the Z score of 1.7



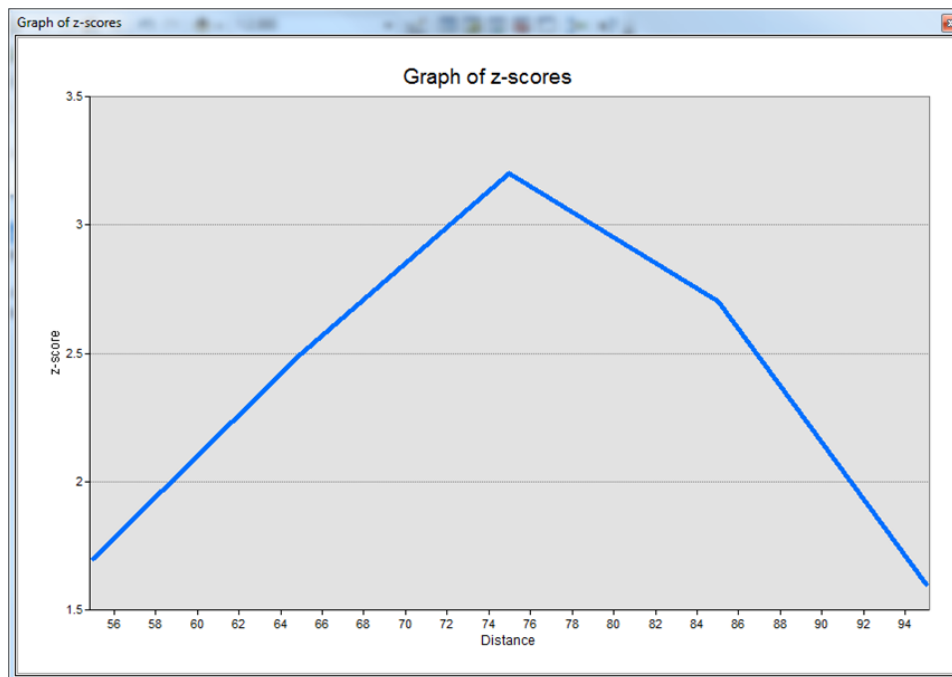
- ☐ Repeat: Open the Spatial Autocorrelation tool, in the Spatial Statistics toolbox in the Analyzing Patterns toolset
 - Input Feature Class: Households
 - Input Field: HHRate
 - Conceptualization of Spatial Relationships: FIXED_DISTANCE_BAND
- ☐ Distance Band: Now use a distance band of 65 (Remember, we decided to use increments of 10 because it is the observed Average Nearest Neighbor Distance)
- ☐ Run the tool, then note the Z score of 2.5

We would continue doing this, increasing the distance by 10 meters, until we found a peak in the Z score value. A graph has already been created that plots the z-scores at every distance up to 95 meters.

- Open the graph called Graph of z-scores (View>Graphs>Graph of z-scores)



This graph shows where I've plotted the Z score associated with each distance up to 95 meters, and we can see that there is a very distinct peak at 75 meters, so we'll use this distance for our hot spot analysis.

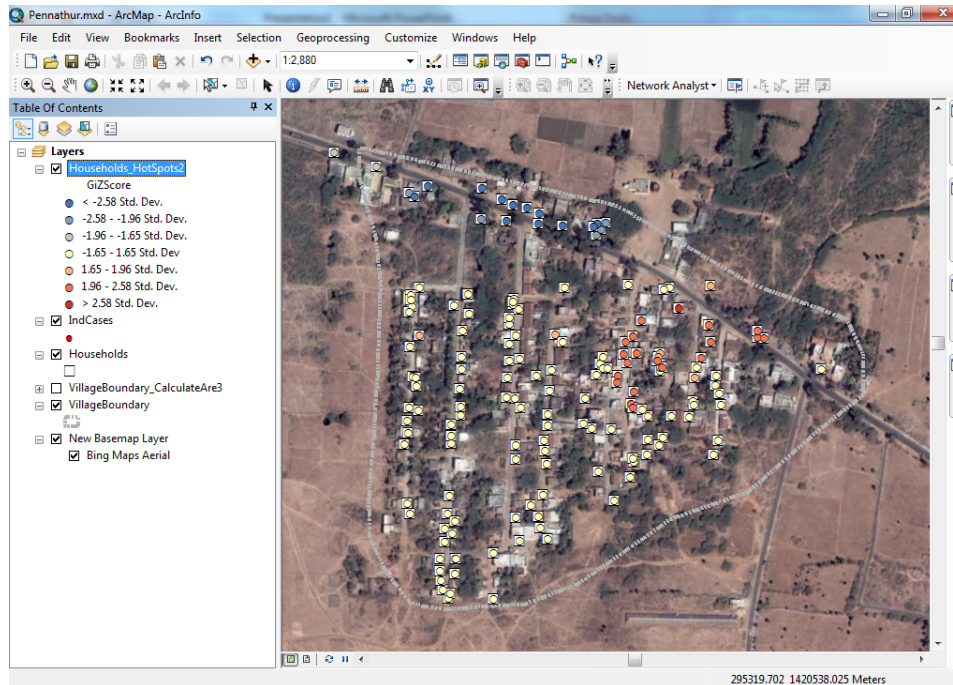


☐ Close the graph and open the hot spot analysis tool from before (should be minimized)

- 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

☐ Run the tool

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.