Lab Assignment 2 – Point Pattern Analysis

Due Date – 09/23/2020

Overview

This assignment is designed to help you practice various analytical methods regarding point data. The focus of the point pattern analysis is to investigate the distribution of points and find out if there is a significant occurrence of clustering. The analysis is primarily based on either frequency (count), distance (location) or attribute values of the points. Different scenarios require different methods and tools. This assignment will provide you an opportunity to practice how to specifically use ArcGIS conduct point pattern analysis. The topics covered include: geometric measurements, clustering analysis, kernel density analysis, and nearest neighbor analysis. You will use some of the skills you practiced in Lab 1.

Objectives

The objectives of this lab assignment are to help you:

- Become familiar with some of the common point pattern analysis techniques.
- Develop a better understanding about the theory (statistics) behind those techniques.
- Know not only how to use those various techniques, but, more importantly, know the differences between them.
- Understand the possibilities and limitations of those analysis techniques.

Part I – Geometric Measurements of Points

1. Distance

Scenario: There are many schools in Prince George's County. Unfortunately, there are also many car accidents happening daily. It is important to get a better understanding about the spatial relationship between these two sets of point features. This information can help find ways of reducing potential harm to the students.

Before you use the data, make sure you study and understand your data first, for example, the spatial reference, extent, attributes, etc.

Tasks: In this exercise, you are required to perform an analysis to find out all car accidents that have occurred **within 1 mile** to each public school and calculate the corresponding distances. The distances should be Euclidean distances.

Launch ArcGIS Pro.

You will use the data from "MD_data" folder: "PG_CarAccidents" and "PG_PublicSchools".

Now, you are required to find out the **nearest car accident to each school** and the accident should occur within 1 mile to that school.

You will use the tool **Near** and calculate the distance from each feature in the Input Features to the nearest features in the Near Features, within the Search Radius.

Note:

- There is used to be a tool called **Point Distance** which can help determine the distances between point features in the Input Features to all points in the Near Features. The result table will have a total number (N1xN2) of distances calculated. N1 and N2 are the number of points in each layer respectively. If a Search Radius is specified, then only the distance for those Near Features that are within the Search Radius of the Input Features will be calculated.
- This tool can be found in ArcToolbox when using ArcMap. However, in the latest ArcGIS Pro, this is a deprecated tool. This functionality has been replaced by Near and Generate Near Table tools that now calculate distances between point, polyline and polygon features.

You need to specify the Input Features and Near Features based on the requirement as well as the Search Radius.

Make sure to check the small box in front of "Location (optional)". By doing this, it will allow the *x,y* coordinates of the nearest feature to be added to the Input Features, as well as NEAR_FID and NEAR_DIST. So, you will be able to know exactly which one is the nearest feature and also the distance.

After you finish, open the attribute table of the Input Feature and scroll the table until the newly added fields (i.e. NEAR_FID, NEAR_DIST, NEAR_X, and NEAR_Y) are visible.

[1] Make a screen shot at this point and include it in the report.

In the exercise above, the Input Features and Near Features were different. However, actually they can be the same as well.

The next task is to find out the nearest distances for each hospital in PG County.

Note:

- Before you use this data, you should check the spatial reference. You will notice that this data has different coordinate system (GCS) from the data you user earlier.
- It is recommended that the data should be projected. You should project this data so that it matches with the spatial reference of the other data.

Use "PG_Hospitals" to define both the Input Features and the Near Features.

[2] Make a screen shot of the result, i.e. the attribute table, and include it in the report.

2. Central Tendency

In this exercise, you are going to practice how to describe point distribution using various centers.

2.1 Mean Center

This measurement identifies the geographic center (or the center of concentration) for a set of features. The average x-coordinate and average y-coordinate for all points in the study area will be calculated.

Now, use the data layer – "MD_Cities".

This tool - **Mean Center** can be found through search.

Define the Input and Output Features, and leave other options as default.

After you did this analysis, change the symbology of the mean center point and make it easily distinguished from other points.

[3] Make a screen shot of the result and include it in the report.

2.2 Weighted Mean Center

The Weighted Mean Center is the geographic center of a set of points as adjusted for the influence of a value associated with each point. A weight indicates how important a geospatial object is. The centers are biased toward locations with high weights.

Now use the same data to calculate the Mean Center again. For the Weight Field, choose "Pop1990".

After you did this analysis, change the symbology of this population-weighted mean center point and make it easily distinguished from other points.

[4] Make a screen shot of the result and include it in the report.

Now, let's investigate to see if the mean center varies depending on factors such as race.

Do the Mean Center analysis. You can set the "Weight Field" as "White".

Change the symbology of this weighted mean center point and make it easily distinguished from other points.

Now, do the Mean Center analysis again. This time, you can set the "Weight Field" as "Asian_PI".

Change the symbology of this weighted mean center point and make it easily distinguished from other points.

You should be able to see that the weighted mean center of White population is very different from that of Asian population.

[5] Make a screen shot of the two mean centers and include them in the report.

Now think about how the calculations have been done behind ArcGIS software. You can refer to the equations in the lecture slides which you should have a better understanding now.

2.3 Median Center

This measurement is also called the center of minimum distance. The median center is a location that has the shortest total distance to all points.

This tool - **Median Center** can be found through search.

Use the same data as in the previous session and find the Median Center of the MD cities.

After you did this analysis, change the symbology of the median center point and make it easily distinguished from other points and also the Mean Center.

[6] Make a screen shot of the result and include it in the report.

2.4 Central Feature

This measurement identifies the most centrally located feature in a point, line, or polygon feature class. It is the point at which the total distance to all other points is the shortest.

This tool - **Central Feature** can be found through search.

You are going to use the same data as above. Define the Input and Output Features, and leave other options as default.

After you did this analysis, change the symbology of the Central Feature point and make it easily distinguished from other points. Also label the cities.

[7] Make a screen shot of the result and include it in the report.

Note:

- The Central Feature is a real point (location of an event).
- The Mean Center or Median Center exists mathematically and is not a location of any real object.

3. Spatial Dispersion

This measurement identifies the degree to which features are concentrated or dispersed around the geometric mean center.

Now, create a new map.

Add the dataset from "DC_data" folder.

Find the tool - **Standard Distance**.

Define the Input as "DC_Crimes_2017".

Now, set the "Circle Size" as "1 Standard Deviation". Leave other options as default.
Click Ok. You will get the result which is a circle polygon.

Note:

- The "Standard Deviation" is actually the average standard distance that is calculated based on the equation below. It represents the average distance to the mean center.

$$S_D = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{X})^2 + \sum_{i=1}^{N}(y_i - \overline{Y})^2}{N}}$$

Repeat these steps and create two more circles by setting the "Circle Size" as "2 Standard Deviation" and "3 Standard Deviation" respectively.

Rearrange the order of these three new polygons and change the colors, etc, in order to make them distinguishable. Be creative!

[8] Make a screen shot of the results and include them in the report.

Now, remember in the lecture we talked about that at least 60% of the points will fall within 1-standard distance from the mean center, 95% within 2-standard distance, and 99% within 3-standard distance. Let's do a test and prove it!

To do this, you must be able to count how many points fall within each polygon. There are different solutions. Based on Lab 1, you should know which to use.

Questions:

1. How many points (i.e. crimes) fall within 1 standard distance from the mean center? What is the percentage of total points?
2. How many points (i.e. crimes) fall within 2 standard distances from the mean center? What is the percentage of total points?
3. How many points (i.e. crimes) fall within 3 standard distances from the mean center? What is the percentage of total points?

You may want to make a table to organize/present the results.

[9] Include the answers in the report.

Standard Distance measures the dispersion of the incidents around the mean center, but it does not capture any directional bias or the shape of the distribution. This leads to Standard Deviational Ellipse.

4. Direction (Standard Deviational Ellipse)

This measurement tells us whether a distribution of features exhibits a directional trend, i.e. whether features are farther from a specified point in one direction than in another direction.

Continue to use ArcGIS Pro.

You are going to use the same data as above.

Find the tool - **Directional Distribution** (Standard Deviational Ellipse).

Define the Input and Output Features.

For Ellipse Size, use “1 Standard Deviation”. Leave other options as default.

[10] Make a screen shot of the result and include it in the report.

Now, open the attribute table of the Output Feature. You will find the standard deviations along the x and y axes.

Based on the numbers in the table, you will calculate the flatness of this ellipse. Refer to the equation in the lecture slides.

[11] Include the answer (the flatness of the ellipse) in the report.

Part II –Nearest Neighbor Analysis

This analysis calculates a nearest neighbor index based on the average distance from each feature to its nearest neighboring feature. It will enable you to investigate the spatial pattern. If the distance is less than the distance for a hypothetical random distribution, the distribution of the features being analyzed is considered clustered. If the distance is greater than a hypothetical random distribution, the features are considered dispersed.

Continue to use ArcGIS Pro.

Find the tool - Average Nearest Neighbor.

For the Input Point Feature, set it as “DC_Crimes_2017”.

Make sure to check the small box in front of “Generate Report”. Leave the other option as default.

Make a screen shot of the NNA result (i.e. the graphic output). You can find the report by placing the History tab on the menu bar. This will list all the recent tasks performed. Find the one you just did. Click it and then a pop-u window will show up. In this window, you can find a URL behind Report File. Click on the link will lead you to a graph report.

The graph report will clearly show the conclusion whether those points are clustered or not. Pay attention to those statistical parameters. Make sure you understand them and know how to interpret them.

Make a screen shot of the graph report.

[12] Include the screen shot in your report.

Notes:

- This NNA analysis is based on distance and applies only to point feature types.
- In Lab 5, when we work with areal (polygon) data, we will revisit clustering analysis. In that case, it will be based on neighborhood or attribute values, for example, the number of crimes in each census tract. We will use different tools to investigate whether there is clustering based on such values.

To investigate whether points are clustered or not based on their values, we need to use different tools and methods. This leads to next exercise.

Part III – Clustering Analysis (Spatial Autocorrelation)

Use the data from the folder – “NYS_data”.

First, let’s ask a simple question: Are those farms are clustered or not based on distance?

To answer this question, you can use the same tool – NNA as in the previous session.

Make a screen shot of the graph report.

[13] Include the screen shot in your report.

Now, let’s redesign the question: Are those farms clustered or not based on the number of cows in each farm? In another word, do those large farms tend to cluster with large farms? And do smaller farms tend to cluster with smaller farms?

To answer this question, we need to use a different tool – **Spatial Autocorrelation (Moran's I)**. This tools measures spatial autocorrelation based on feature locations and attribute values using the Global Moran's I statistic.

For the Input Field, you should choose the attribute – "MILKCOWS".

Make sure you check the small box in front of **Generate Report**.

Make a screen shot of the graph report.

[14] Include the screen shot in your report.

We can also use another tool to answer the same question – **High/Low Clustering (Getis-Ord General G)**.

This tools measures the degree of clustering for either high values or low values using the Getis-Ord General G statistic.

For the Input Field, you should choose the attribute – "MILKCOWS".

Make sure you check the small box in front of **Generate Report**.

Make a screen shot of the graph report.

[15] Include the screen shot in your report.

Part IV – Kernel Density Analysis

While the previous analysis can help find out whether there is clustering or not, but it does not tell where the clusters are located if there are any. The Kernel Density Analysis calculates a magnitude per unit area from point or polyline features using a kernel function to fit a smoothly tapered surface to each point or polyline. The density is estimated by counting the number of events in a kernel, centered at the location where the estimate is to be made. It is particularly and also commonly used to identify hot spots.

Find the tool - Kernel Density.

Use the MD data. The Input Point Feature should be "MD_Cities". Set the "Population Field" as "None" which is the default. Also leave other options as default.

Make a screen shot of the result.

[16] Include the screen shot in your report.

In this case, what do the cell values of the density surface mean? In another words, what are those numbers referring to?

Now, do the same analysis. However, this time you should use the field "MEDIANRENT" to set the "Population Field".

Make a screen shot of the result.

[17] Include the screen shot in your report.

In this case, what do the cell values of the density surface mean? In another words, what are those numbers referring to?

Note:

- We set the radius as default. However, if you set the radius to be very large or very small, the smoothness of the surface will vary.
- Again, if you are interested or have time, you should do some tests by changing the radius. It is optional.

Part V – Population Center of Maryland (1900-2008)

During the lecture, I showed you an example of calculating the mean centers of US population from 1790 to 2000. You can check out an animation video here - https://archive.org/details/SVS-3163

There are different ways to calculate the Population Mean Center. In the original publication, the following formulae were used:

$$\bar{\phi} = \frac{\sum w_i \phi_i}{\sum w_i} \qquad \bar{\lambda} = \frac{\sum w_i \lambda_i Cos(\phi_i)}{\sum w_i Cos(\phi_i)}$$

Where *Φ, λ, w* are the latitude, longitude, and population.

Now, you can do a similar study to find out the change of the Maryland population centers for the past 110 years (1900 to 2008) to see if you can discover something interesting.

You could have used the existing tool in ArcGIS to calculate the weighted mean centers. But, to really understand how it works, it will be helpful if you do the calculation "manually" using these formulae.

The points can be represented by the centroids of those 24 counties in Maryland.

For practicing purpose, I intentionally provided incomplete data. So, you will need to process, clean up, and create the dataset for your analysis.

The population can be found on the QuickFacts web site by the US Census Bureau: https://www.census.gov/quickfacts/fact/table/US/AGE295219 . The population data is not continuous as the historical population data was only available by every decade. And the population from 2000 and 2008 are estimates. To save your time, I have already downloaded the population data for Maryland Counties. The data are in two separate files: one is an Excel file and the other a .TXT file. You will extract the information from these two sources and then compile your table. Then, you can use the techniques that you practiced from the Lab 1 to eventually create your GIS data.

You can start the process by deriving the centroids of all counties with the data – "MD_Cnty_Bnd" (Maryland County Boundary). You can either use Field Calculator to do the calculations. Or, you can export the table and do the calculations in Excel and then create an event layer and eventually export it into a GIS layer.

Make a screen shot of the map which shows the population centers of Maryland during the past 110 years.

You should be able to identify the specific county which is the most recent population center in State of Maryland.

Hints:

- You should get the coordinates of the centroids in decimal degrees. Before you plug these latitude and longitude coordinates into the formulae, make sure they are converted into radians first.
- I would recommend to use Excel for the calculations.
- When using Excel to do the calculation, I would recommend to use a function - "sumproduct()" in case you have never used it before. This will save you a lot of time.

Include the screen shot and the answer in your report.

----- THE END -----