Lab Assignment 5 – Areal Analysis

Due Date: 10/16/2020

Overview

This assignment is designed to help you practice various analytical methods and tools regarding areal data (i.e. polygon feature type). The topics covered in this lab assignment include: geometric measurements, Join Count Analysis, and other spatial autocorrelation statistics such as Moran's I and G-Statistic (both global and local measures). Some of these methods can help determine if there is any clustering in the study region while others can help identify the specific clustering areas (i.e. hot spots or cold spots).

You will practice using these tools in ArcGIS Pro to familiarize yourself with spatial autocorrelation. In addition, to better understand how these methods actually work, you will need to understand the mathematics behind. The best way of learning is to manually do some calculations. Then, you can compare the conclusion based on your own calculations to the result while using ArcGIS tools. This will help you thoroughly understand those various methods and tools related to spatial autocorrelation.

Part I – Preparing Data

In real world, when you obtain GIS data from various sources, you often need to clean up and process the data before you actually use them for analysis. In this assignment, you are provided with some raw data and need to prepare them.

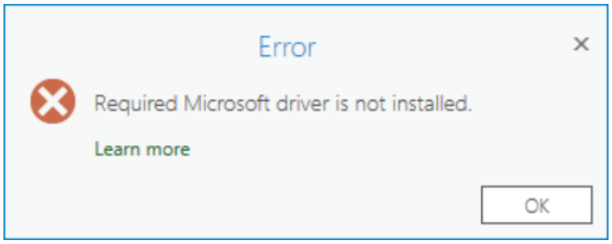
Launch ArcGIS Pro.

Add the Maryland county data.

Open the attribute table and study the information contained. There are fields such as county names, county numbers, etc. However, there is no social-economic information such as population and income data which is often very useful.

Fortunately, these information can be found from a different source, for example, this Wikipedia page - Maryland counties ranked by per capita income. You can easily copy and paste this table into Excel and then create your own table.

Now, try to add your Excel table in ArcGIS Pro. It is very likely you will get an error message (see below).



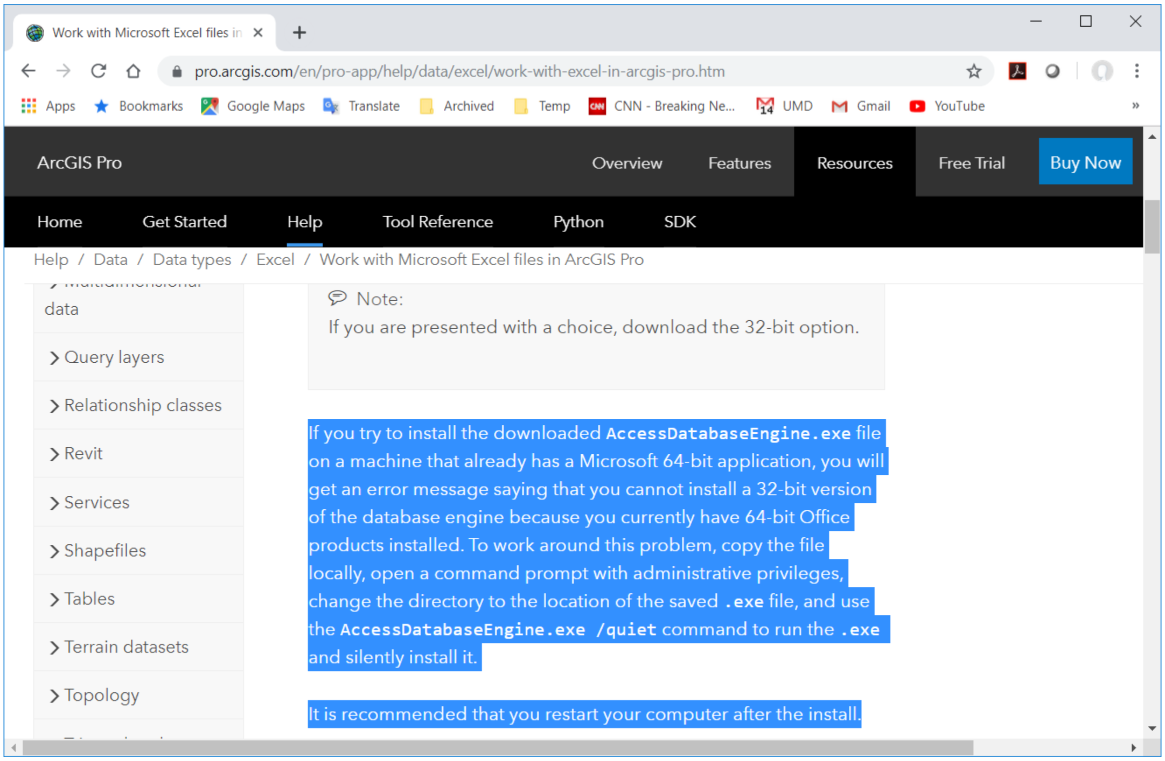
Error

Required Microsoft driver is not installed.

To make it work, you will need to install the Microsoft Access Database Engine driver. This driver allows you to directly open Microsoft Office Excel tables in ArcGIS Pro and work with them in the same way as other tabular data sources.

The tricky part is that, depending on the setting up of your computer, you may be automatically prompted to download and install the 32-bit version of the driver although your computer is 64-bit. In this case, you will **have to** install the driver of 32-bit version. (I had encountered the same issue. I tried to install the 64-bit version myself but it didn't work properly in the end. However, the 32-bit version works perfectly fine.)

To install this 32-bit driver, you will need to use MS-DOS (i.e. Windows Command). Make sure you follow the exact instructions highlighted in the screenshot below.


Work with Microsoft Excel files in
pro.arcgis.com/en/pro-app/help/data/excel/work-with-excel-in-arcgis-pro.htm
Apps Bookmarks Google Maps Translate Archived Temp CNN - Breaking Ne... UMD Gmail YouTube
ArcGIS Pro
Overview Features Resources Free Trial Buy Now
Home Get Started Help Tool Reference Python SDK
Help / Data / Data types / Excel / Work with Microsoft Excel files in ArcGIS Pro
data
Query layers
Relationship classes
Revit
Services
Shapefiles
Tables
Terrain datasets
Topology

Note:
If you are presented with a choice, download the 32-bit option.

If you try to install the downloaded `AccessDatabaseEngine.exe` file on a machine that already has a Microsoft 64-bit application, you will get an error message saying that you cannot install a 32-bit version of the database engine because you currently have 64-bit Office products installed. To work around this problem, copy the file locally, open a command prompt with administrative privileges, change the directory to the location of the saved `.exe` file, and use the `AccessDatabaseEngine.exe /quiet` command to run the `.exe` and silently install it.

It is recommended that you restart your computer after the install.

This will require you to have some basic knowledge of using Windows Command. If you are not sure how, you may refer to the tutorials below:

- https://www.bleepingcomputer.com/tutorials/windows-command-prompt-introduction/
- https://www.computerhope.com/issues/chusedos.htm

If you really have difficulty to make it work, you can use ArcMap to complete this part of the lab. I will do a demo next week to make sure you successfully install it on your computer.

Once you are able to add the Excel table in ArcGIS Pro, you can open it and study the information inside. You will notice that it contains exactly the kind of social-economic information we are interested in.

Note:

- The data is from the 2010 United States Census Data and the 2010-2014 American Community Survey 5-Year Estimates.

Now, the issue is that you have two data in different formats: a GIS data layer and an Excel file. It will be desriable to add the information from the Excel table to the attribute table of the GIS data layer.

Your first task is to join the Excel table to the Maryand county data.

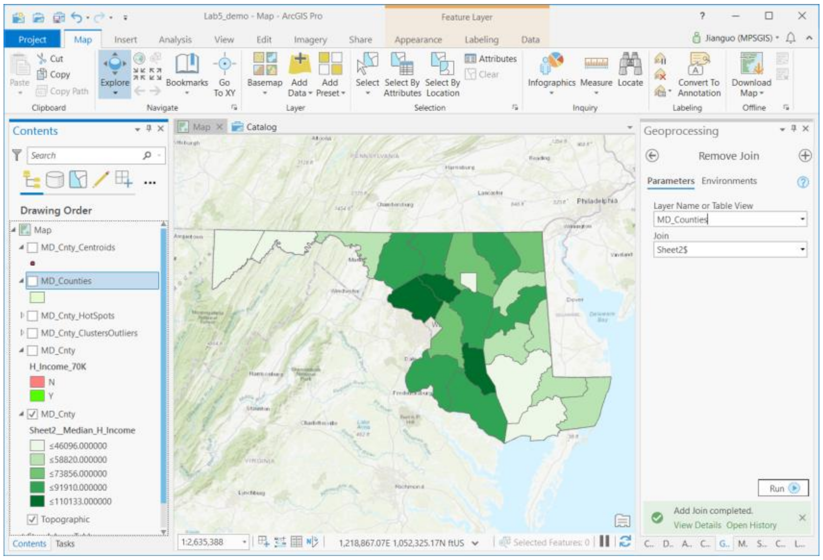
When you do the tabular join, it is important to pay attention to the details. For example, if you use the county name as the common filed (i.e. key) for joining, you need to make sure the county names are spelled exactly the same in both tables. Please also note that the texts are case sensitive.

To save you time, I am providing some hints. Otherwise, it may take you a few trials to realize the mistakes.

- Some counties are spelled in all upper case in the attribute table, for example, Howard county is entered as "HOWARD". This does not match with the county name "Howard" in the Excel table.
- Also, there is one county – Baltimore City is actually entered as "Baltimore city". This will create trouble when joining the tables if you don't notice the minor difference.

After you join the two tables, you will be able to display the income information on the map.

Create a Choropleth map with graduate colors based on the value – Median Household Income. For classification, you can use five classes. The map should be similar to the one below.



Map
- MD_Cnty_Centroids
- MD_Counties
- MD_Cnty_HotSpots
- MD_Cnty_ClustersOutliers
- MD_Cnty
  - H_Income_70K
  - 🟥 N
  - 🟩 Y
- ☑ MD_Cnty
  - Sheet2__Median_H_Income
  - ≤46096.000000
  - ≤58820.000000
  - ≤73856.000000
  - ≤91910.000000
  - ≤110133.000000
- ☑ Topographic

Layer Name or Table View: MD_Counties

Join: Sheet2$

Add Join completed.
View Details Open History

1:2,635,388 1,218,867.07E 1,052,325.17N ftUS Selected Features: 0

Make a screen shot of your map.

[1] Include the screen shot in the report.

From the map you just created, you can get a general idea how the income values are distributed by counties. However, the map can be misleading because it may look very differently if you use different number of classes or use different classification methods. To really find out if there is any clustering of income values or not, you will need to use more robust methods or tools, for example, Moran's I. You will get to them later.

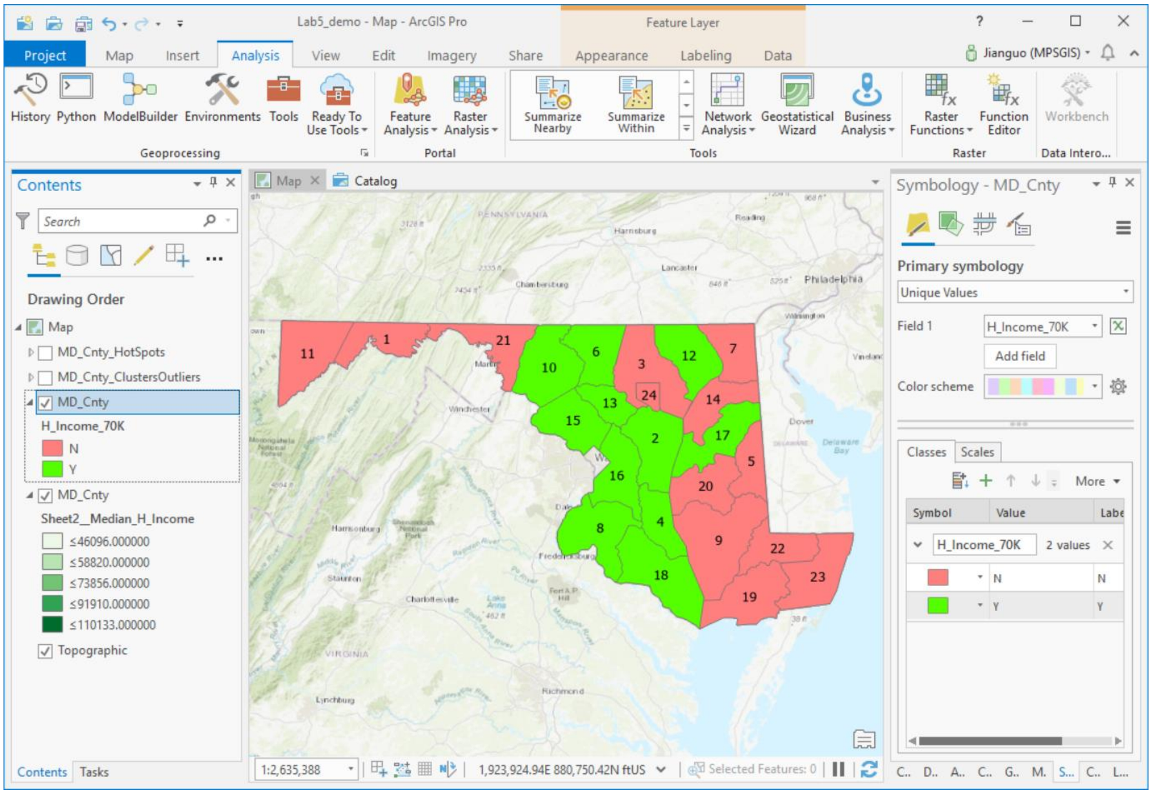
Now, go back to the joined table and continue. Your next task is to create a new field to help you identify those counties have a median household income of at least $70,000 and also those with lower income. So, basically you are going to reclassify or categorize the income values into two groups. That means the values of this new field are binary.

Because the values are binary, you should set the data type as "string" when creating the new field. For example, "Yes" and "No", or "+" and "-".

After you have created the new field, you need populate the values based on the income.

Then, you will create a thematic map using Unique Values based on this new field. In addition, you should label the counties using county numbers which already exist in the attribute table.

You will create a map similar to the one below.



Lab5_demo - Map - ArcGIS Pro

Contents — Drawing Order
- Map
  - ☐ MD_Cnty_HotSpots
  - ☐ MD_Cnty_ClustersOutliers
  - ☑ MD_Cnty
    - H_Income_70K
    - 🟥 N
    - 🟩 Y
  - ☑ MD_Cnty
    - Sheet2_Median_H_Income
    - ≤46096.000000
    - ≤58820.000000
    - ≤73856.000000
    - ≤91910.000000
    - ≤110133.000000
  - ☑ Topographic

Symbology - MD_Cnty
Primary symbology: Unique Values
Field 1: H_Income_70K

Symbol	Value	Label
	H_Income_70K	2 values
🟥	N	N
🟩	Y	Y

1:2,635,388 1,923,924.94E 880,750.42N ftUS Selected Features: 0

Make a screen shot of your map.

[2] Include the screen shot in the report.

Again, from the map you just created, you can get a general idea how the income values are distributed by counties. However, the map still can be misleading because it may look very differently if you choose a different cut-off value instead of $70,000. So, simple visualization methods cannot be used as a reliable way of discovering spatial patterns such as clustering. We will have to refer to spatial autocorrelation methods.

Part II – Geometric Measurements

1. Area

When using geodatabase, the area and perimeter of polygons automatically are calculated and displayed in the attribute table as "Shape_Area" and "Shape_Length".

You can verify this by checking the attributes of the Maryland county data.

However, if you are using shapefiles, you will need to calculate them by yourself.

Add the data – "UMD_Buildings".

Open the attribute table. You will see that the basic geometric measures (e.g. area and perimeter) of those polygons are missing.

Your first task is to calculate the area and then add it to the attribute table. You need to decide which data type is the most appropriate.

Rank the buildings based on their areas. Select and zoom in to the building with the largest area.

Label this building with its name.

Make a screen shot at this point.

[3] Include the screen shot in the report.

2. Perimeter

Similarly, you will need to calculate the perimeter and then add it in the attribute table.

Rank the buildings based on their perimeter. Select and zoom in to the building with the largest perimeter. Check the corresponding area to see how it ranks compared with the previous answer.

Label this building with its name.

Make a screen shot at this point.

[4] Include the screen shot in the report.

Note:

- You will notice that the building with the largest area does not necessarily have the largest perimeter. This leads to the introduction of another geometric measurement - Compactness in the following exercise, an indicator of the shape of the area.

3. Compactness

Compactness measures how different the shape of an area or polygon is from a circle that has the same perimeter as the polygon.

The compactness index can be calculated as below:

$$C = \sqrt{4\pi \frac{A}{P^2}}$$

Where:

C - Compactness ratio
A - The area of the polygon.
P - Perimeter of the polygon.

Compactness is a number between 0 and 1. If the index is 0, the polygon is actually a line. If the index is 1, then the polygon is a circle. Therefore, the closer a polygon look like a circle, the closer its compactness index closer to 1.

Meanwhile, those long and narrow polygons should have much lower compactness index. You can verify this.

Now, create a new field and name it as “Compactness”. You need to decide the appropriate data type.

In the Field Calculator, you need to type the syntax based on the equation.

After you finish calculating the compactness index, you will rank the buildings based on this index in descending order.

Now, zoom in to the building that has the largest compactness index. You will verify that the shape of this particular building is very close to the shape of a circle.

Label this building and make a screen shot at this point.

[5] Include the screen shot in the report.

Now, zoom in to the building that has the smallest compactness index. You will see that the shape of this particular building is narrow and long. This means that this building has relatively larger perimeter and at same time, smaller area.

Label this building and make a screen shot at this point.

[6] Include the screen shot in the report.

4. Centroid

Centroid is commonly used to represent the central point of an area. It is also often used to calculate the distance between different areas (polygons).

Now you will need to create a new data layer showing the centroids of all those buildings on UMD campus. There are many ways or tools to complete this task.

You can use the **Feature to Point** tool to create the centroids.

Note:

- Make sure you don’t check the small box in front of “Inside (optional)”. If the Inside option on the dialog box is checked, the points created are not necessarily centroids even though most of them match with centroids. It has been tested and verified during the lab session.

Once you have created the Centroids data layer, find LeFrak Hall and zoom in.

Make a screen shot at this point.

[7] Include the screen shot in the report.

There are some buildings with the centroids located outside of their boundaries. For example, find “A.V. Williams Building” (FID = 36) and zoom in. You will see the centroid of this polygon is actually located outside.

Make a screen shot at this point.

[8] Include the screen shot in the report.

Note:

- Because the location of centroid is dependent on the shape of the polygon, for polygon feature type, we are not interested in their locations or distance. Instead, we focus on the attribute values and also the neighborhood between polygons. This will lead to the exercises in next section.

Part III – Spatial Autocorrelation

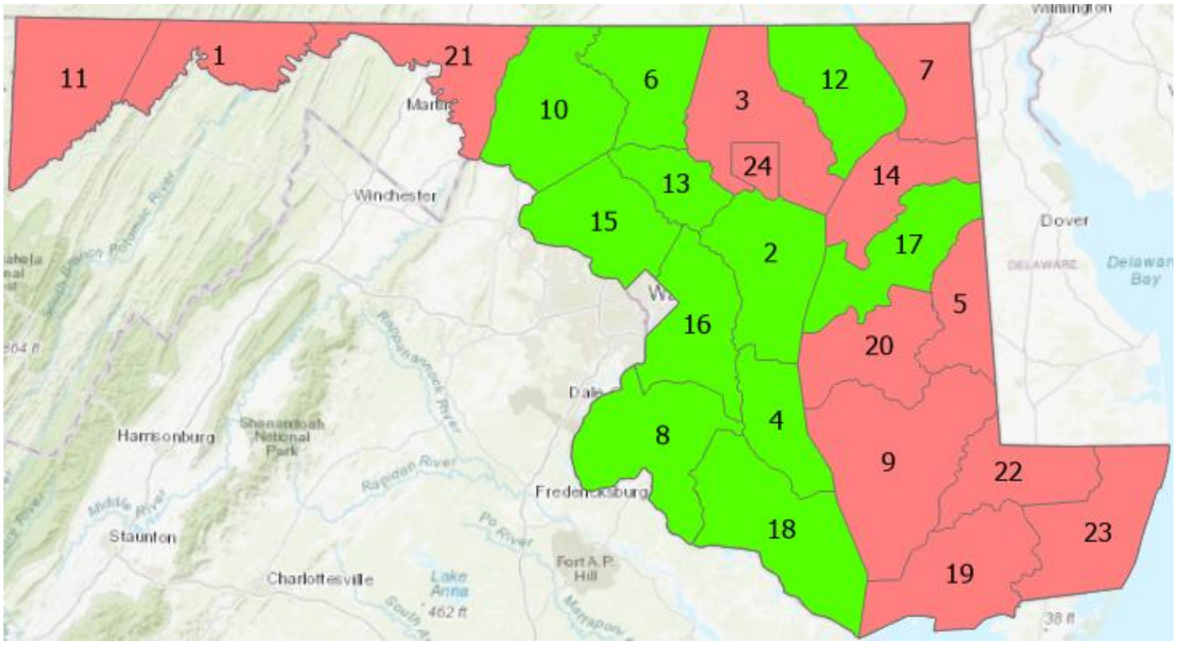
Spatial Autocorrelation is used to indicate whether the distribution of values of polygons is dependent on the spatial distribution of the features. In another word, it investigates whether particular values are likely to occur in one location (clustering), or are equally likely to occur at any location (random).

Depending on the data types of the attribute values, there are different measurements. If the data type is nominal or categorical, Join Count Analysis can be used. And if the values are ratio or interval data type, there are a few options to choose from. Of course, you can always reclassify the ratio values to create categorical values.

1. Join Count Analysis

Join Count Analysis is used for contiguous areas that have category attributes (i.e. nominal data). The number of joins of each type is compared with the expected number based on the probability of two adjacent areas having the same value by chance. If the counted number of joins for areas having the same value is greater than the expected number, then that value is clustered.

In Part I, you had created the map showing counties based on their median household income values. The values of the polygons are binary (Y or N). If the value of a county is "Y", the median household income of this county is at least $70,000. And if the value of a county is "N", then the median household income of that county is less than $70,000.

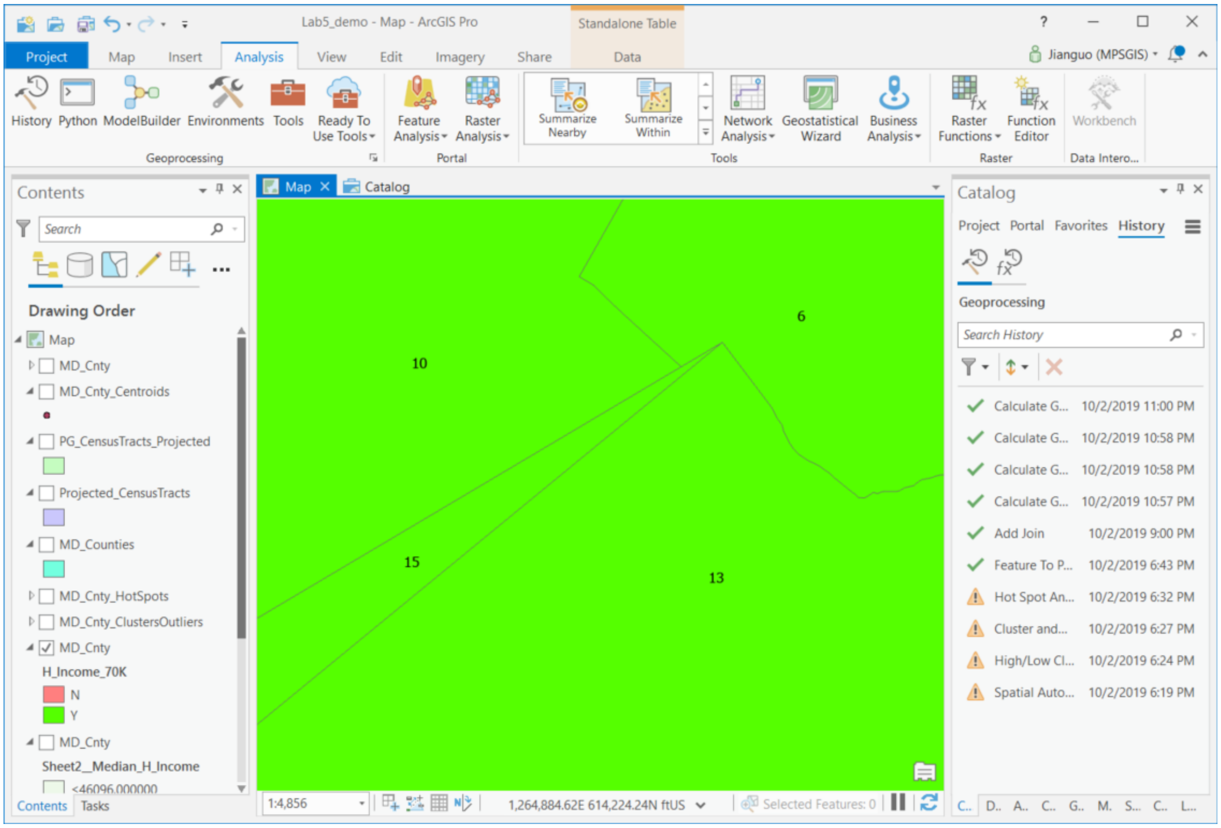


The joint count statistic compares the observed number of YN (corresponding to BW in the formula below) joins with the number of YN joins that would be expected if no spatial autocorrelation were present.

So, your first step is to find out the number of YY, YN, ad NN joins. These will be the observed counts.

Please note what you see on the map is not necessarily what you get.

If you zoom in the area as shown below, you will notice that actually County #6 and #15 actually share a border. Therefore, in this case, this pair is actually a join (YY).



Calculate G... 10/2/2019 11:00 PM
Calculate G... 10/2/2019 10:58 PM
Calculate G... 10/2/2019 10:58 PM
Calculate G... 10/2/2019 10:57 PM
Add Join 10/2/2019 9:00 PM
Feature To P... 10/2/2019 6:43 PM
Hot Spot An... 10/2/2019 6:32 PM
Cluster and... 10/2/2019 6:27 PM
High/Low Cl... 10/2/2019 6:24 PM
Spatial Auto... 10/2/2019 6:19 PM

It may take some time to count those joins. You will really need to pay attention to avoid miscount.

Then, you will calculate the expected count for YN (corresponding to BW) as below.

$$E(BW)=\frac{2JBW}{N(N-1)}$$

Where J is the total number of joins, B(Y) is the number of polygons with a value of B(Y), W(N) is the number of polygons with a value of W(N), and N is the total number of polygons.

Then, you will calculate the Z-score to test the significance of the result.

$$Z = \frac{(Observed\,BW) - E(BW)}{\sqrt{Var(BW)}}$$

Now, fill this table.

Count of YY joins	
Count of YN joins	
Count of NN joins	
Total count of all joins	
Expected count of YN joins	
Z-score	

[9] Include the filled table in the report.

Based on the calculated Z score, what is your conclusion on the clustering analysis?

[10] Include the answer in the report.

Now, you can look back the maps you created in Part I to see if the Join Count Analysis verifies your impression or not.

There are inherent flaws when using Join Count Analysis because it doesn't take into account the magnitude of the values in each area. It is based simply on whether the area falls into one category or the other. For example: In a two-candidate election, an area with 49.9% of voting "Yes" to Candidate A has the same category value as another area with only 2% of voting "Yes" to the same candidate. Therefore, when the feature values are reclassified into two categories, there will be information loss which is unavoidable.

Other factors may also influence the Join Count statistic results, for example, when there are less than 30 features in the study area. In this exercise, actually that's exactly what has happened as there are only 24 polygons (counties) in the data.

2. Global Moran's I with Manual Calculations

When the attribute values are categorical or nominal, we used Join Count Analysis. However, when the attribute values are interval or ratio data type, we will use different methods and tools.

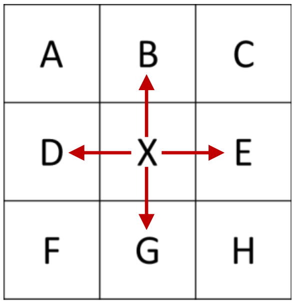
Spatial autocorrelation statistics can measure and test how clustered/dispersed the points (or polygons) are with respect to their attributes. Global measures are used to describe the level of spatial autocorrelation for the entire study region.

The Moran's I statistic compares the values for each feature in a neighboring features to the mean value for the dataset.

$$I = \frac{n\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_i \sum_j w_{ij})\sum_i (x_i - \bar{x})^2}$$

Where:
n - the total number of points/polygons.
w_{ij} - the weight for the pair of polygon *i* and *j*’s attributes,
x_i, x_j - the value of polygon *i* and *j*,
$\bar{x}$ – the mean value

To define the neighborhood, we can use Rook’s case here because it is the most common one:



<u>Note:</u>

- For Exercise 9 you will do the calculations for the other two cases: Queen and Bishop.

Assume that there are *n* polygons (areal units) in the study area. Then there are *n* x *n* pairs of relationships to be captured.

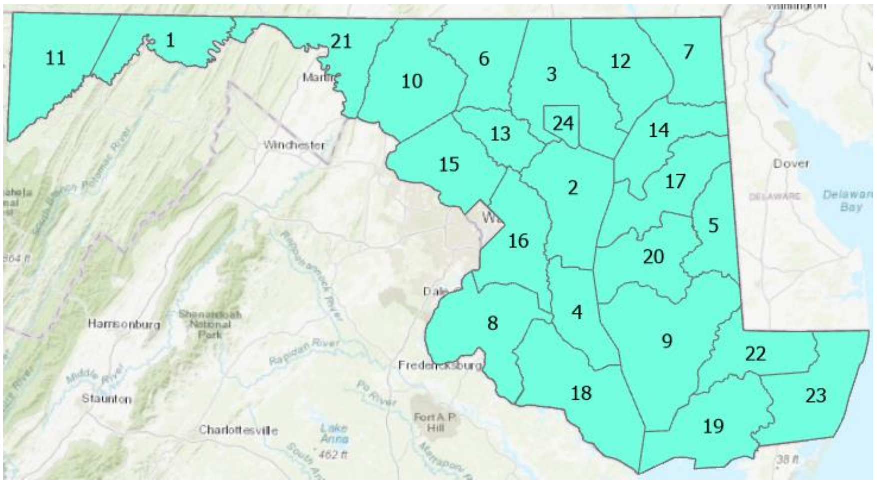
Create a Spatial Weight Matrix using the Rook's case. The values of the matrix are binary. A value of “1” is assigned to the cell that indicates adjacency relationship for a pair of polygons. Otherwise, a value of “0” is assigned to the cell.

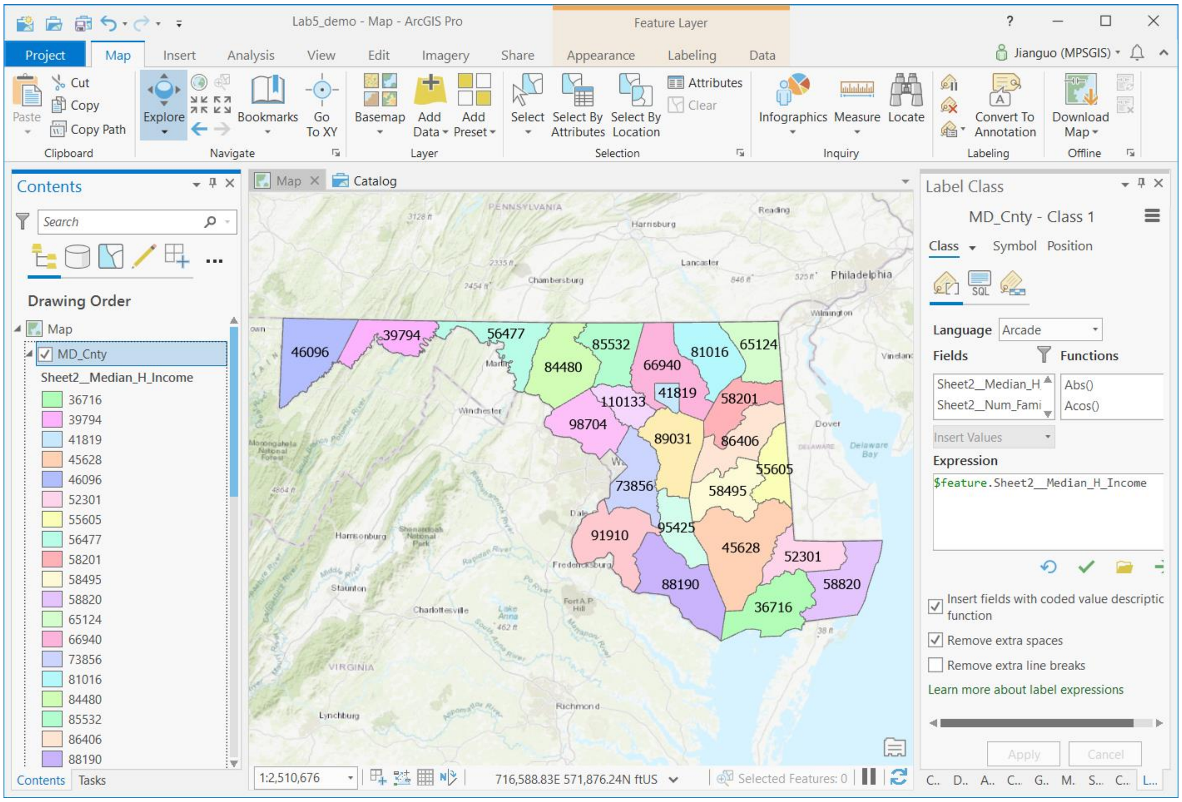
These values are the same as the weights for the pairs of polygons in the study area.

Calculate Moran's I for the set of polygons below using the example in the lecture as a guide.

Once you calculated the Moran's *I* value:
If I > 0, then the values of features are clustered.
If I = 0, then the values of features are random.
If I < 0, then the values of features are dispersed.

We will use the same data to calculate Moran’s I value. The values (median household income) used for calculation are ratio data type now, instead of binary (Y/N) values in Join Count Analysis. You can display the income values as labels.





Map
☑ MD_Cnty
Sheet2_Median_H_Income

🟩	36716
🟪	39794
🟦	41819
🟧	45628
🟦	46096
🟥	52301
🟨	55605
🟩	56477
🟥	58201
🟨	58495
🟪	58820
🟩	65124
🟥	66940
🟦	73856
🟦	81016
🟩	84480
🟩	85532
🟧	86406
🟪	88190

MD_Cnty - Class 1

Language: Arcade

Fields: Sheet2_Median_H, Sheet2_Num_Fami

Functions: Abs(), Acos()

Expression:

```
$feature.Sheet2_Median_H_Income
```

☑ Insert fields with coded value description function

☑ Remove extra spaces

☐ Remove extra line breaks

1:2,510,676 716,588.83E 571,876.24N ftUS

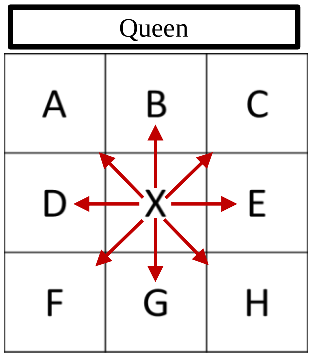
Your task is to calculate the Moran's *I* value to investigate to see whether there is any clustering or not based on the spatial distribution of the median household income values by county. You may want to use Excel to organize and track all those steps.

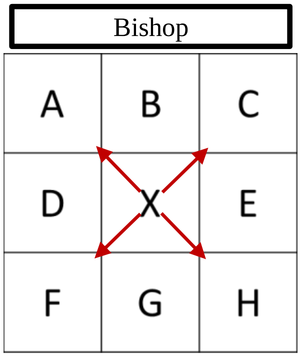
Questions:

1) What is the Moran's *I* value you have calculated?
2) What is your conclusion on the clustering analysis?

[11] Include the answers in the report.

There are different definitions about the neighborhood/adjacency. The other two are Bishop's case and Queen's Case.





Now, let's assume that you use Queen's case to define the adjacency/neighborhood. Then, based on the equation, you will calculate the *I* value.

Questions: (<span>**This part is optional.)**</span>

1) What is the *I* value you calculated in Queen's case?
2) What is your conclusion on the clustering analysis?

The conclusion might be completely different from that of Rook's case.)

3. Global Moran's I with ArcGIS Tool

Now that you have done the manual calculations of Global Moran's I, you can use ArcGIS to do the same clustering analysis and then compare with your calculated result. It will be great to see if they match or not. Time to test ArcGIS software.

The tool to be used: ***Spatial Autocorrelation (Morans I)***.

You will use the same data – Maryland County.

When running this tool, make sure you choose settings carefully.

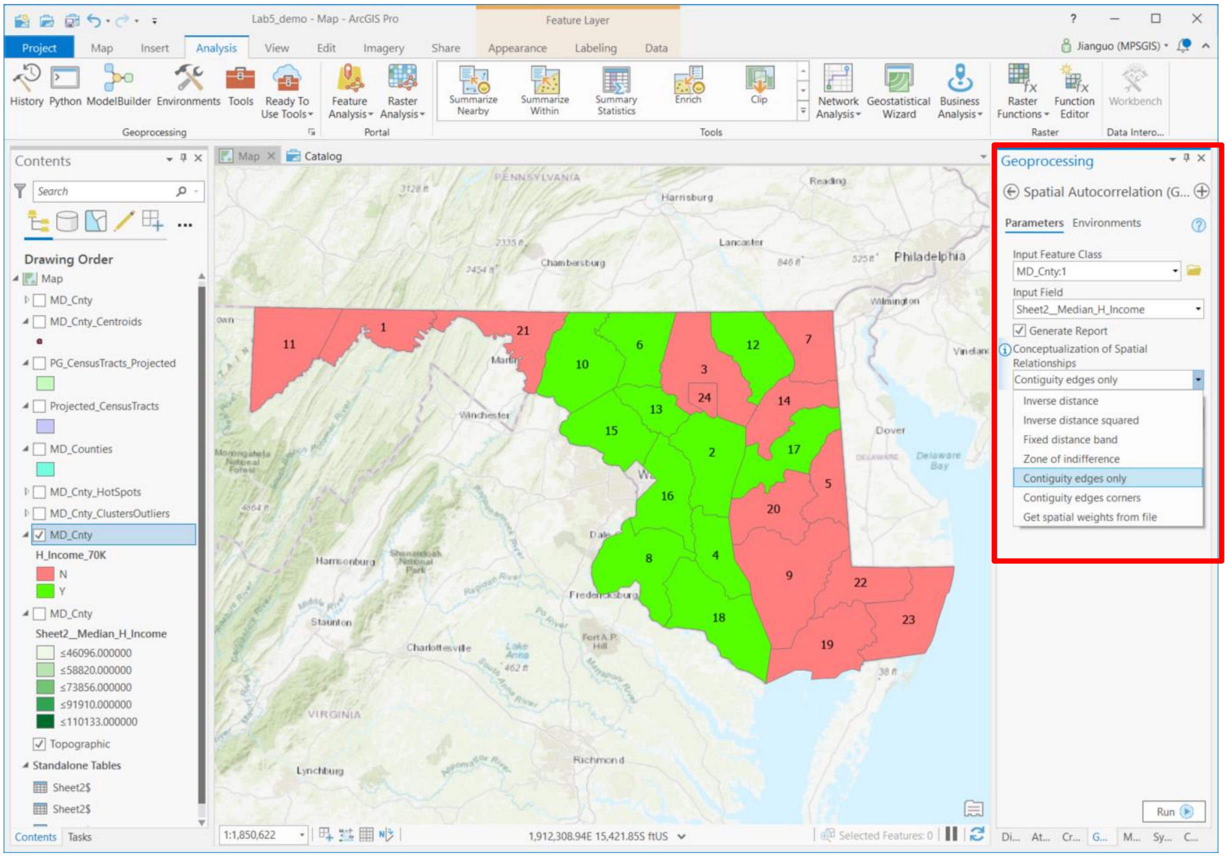
Note:

- An important difference between spatial and traditional (aspatial or nonspatial) statistics is that spatial statistics integrate space and spatial relationships directly into their mathematics. Consequently, many of the tools in the Spatial Statistics toolbox require you to select a value for the Conceptualization of Spatial Relationships parameter prior to analysis.

In this case, you will need to choose "Continuity edges only" from the drop-down list for Conceptualization of Spatial Relationships. (see the screenshot below) This indicates that the neighborhood relationship is defined using Rook's case. For Contiguity edges only, polygons that share an edge (that have coincident boundaries) are included in computations for the target polygon. Polygons that do not share an edge are excluded from the target feature computations.

For Standardization, chose "None".

Also, make sure you check the small box in front of "Generate Report".



Geoprocessing

Spatial Autocorrelation (G...

Parameters Environments

Input Feature Class: MD_Cnty:1

Input Field: Sheet2__Median_H_Income

☑ Generate Report

Conceptualization of Spatial Relationships: Contiguity edges only

After you run the tool, check out the report in a web browser.

Are the income values clustered based on Moran's I? The diagram in the report should clearly show the conclusion.

There are also some numbers in the report. The Moran's I calculates an index value in addition to both a Z score and p-value evaluating the significance of that index. The z-score and p-value indicate whether or not you can reject the null hypothesis.

Make a screen shot of the report to show the diagram.

[12] Include it in the report.

Does this result from ArcGIS match with the result from your own calculation? **They should match exactly!**

By now, you should have a very good understanding about the mathematic behind Moran's I whenever you use this tool with ArcGIS in the future. You will also be able to accurately interpret those values in the report.

4. General G Statistic

General G Statistic also applies to contiguous areas with ratio data values. Unlike Moran's I, this statistic measures how concentrated the high or low values are for a given study area.

In ArcGIS, the tool to be used: ***High/Low Clustering (Getis-Ord General G)***

Use this tool to test the clustering with the same data. It measures the degree of clustering for either high values or low values using the Getis-Ord General G statistic.

You will need to choose "Continuity edges only" from the drop-down list for Conceptualization of Spatial Relationships.

For Standardization, chose "None".

After you run the tool, check out the report in a web browser.

Are the income values clustered based on General G statistic? The diagram in the report should clearly show the conclusion.

Make a screen shot of the analysis report to show the diagram.

[13] Include it in the report.

The null hypothesis for the General G statistic states "there is no spatial clustering of the values". The z-score and 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.

Does this result verify the result from Moran's I? It should.

Note:

- The input data must be projected before carrying out the spatial autocorrelation analysis.
- These global measures can help you find out if there is clustering in the study area but do not show where exactly the clusters (hot spots or cold spots) are. The local version of these statistics can address this issue. It leads to the next section.

5. Local Moran's I

Local Moran's I identifies statistically significant hot spots, cold spots, and spatial outliers.

In ArcGIS, the tool to be used: ***Cluster and Outliers (Anselin Local Morans I)***

In the output table, this tool calculates a Local Moran's I value, a Z score, a p-value, and a code representing the cluster type for each feature.

You will need to choose "Continuity edges only" from the drop-down list for Conceptualization of Spatial Relationships.

For Standardization, chose "None".

Use this tool and calculate the result.

Make a screen shot of the result.

[14] Include it in the report.

How does this map compare to the Choropleth map in Part I?

Now, open the attribute table of the output layer. Scroll to the last few columns (fields) of the table and you should see four new fields: "LMiIndex", "LMiZScore", "LMiPValue", and "COType".

The Z score and p-value represent the statistical significance of the computed index value. 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. A high positive z-score for a feature indicates that the surrounding features have similar values (either high values or low values).

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), and outlier in which a low value is surrounded primarily by high values (LH).

Note:

- You can conduct different queries based on the "COType" field. Then, you can compare the Z-scores, I values and P values for those different groups – "HH", "HL", and "LL".
- This will help you understand how those values are interpreted.

6. Hot Spot Analysis (Getis-Ord Gi*)

Similar to Local Moran's I, this tool identifies statistically significant spatial clusters of high values (hot spots) and low values (cold spots). It creates a new Output Feature Class with a z-score and p-value for each feature in the Input Feature Class.

This tool works by looking at each feature within the context of neighboring features. 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.

In ArcGIS, the tool to be used: ***Hot Spot Analysis (Getis-Ord Gi*)***

The z-scores and p-values are measures of statistical significance which tell you whether or not to reject the null hypothesis, feature by feature. In effect, they indicate whether the observed spatial clustering of high or low values is more pronounced than one would expect in a random distribution of those same values.

A high z-score and small p-value for a feature indicates a spatial clustering of high values. A low negative z-score and 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.

You will need to choose “Continuity edges only” from the drop-down list for Conceptualization of Spatial Relationships.

For Standardization, chose “None”.

Use this tool to test the clustering with the same data. Make a screen shot of the analysis result.

[15] Include it in the report.

Open the attribute table of the result layer. Scroll to the last few columns (fields) of the table and you should see two new fields: “GiZScore” and “GiPValue”.

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).

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