Geostatistical Analysis

15

Introduction

Geostatistics is a very useful approach that allows users to obtain meaningful information related to data in term of its distribution and patterns in GIS. This chapter includes some applications of Spatial Statistics from the GIS environment based on groundwater data. The intention is to focus on the application of GIS rather than emphasizing on complex mathematical and statistical theories. Nevertheless, some of the tools such as Measuring Geographic Distribution, Analysis Patterns, and Mapping Clusters of the Spatial Statistical analysis will be explained and applied using groundwater data.

Measuring Geographic Distribution Toolset

Using geographic distribution tools in ArcGIS aiming to perform statistical approaches to assist researchers in measuring the distribution of features. The tools allow users, for example, to calculate a value that represents a characteristic of the distribution. Such as the center of groundwater wells tapping an aquifer. By doing this, you can see how the wells are dispersed throughout the basin. There are three types of centers that can be calculated: Mean Center, Median Center, and Central feature.

1. **Mean Center**: is the average of the X-coordinate and Y-coordinates values of all features. The resulting X, Y coordinate pair is the mean center. For example in the Jarash area, there are several wells (Figure below) that spread through the area and in order to find the mean center, we calculate the averages of both X and Y coordinates (table below).

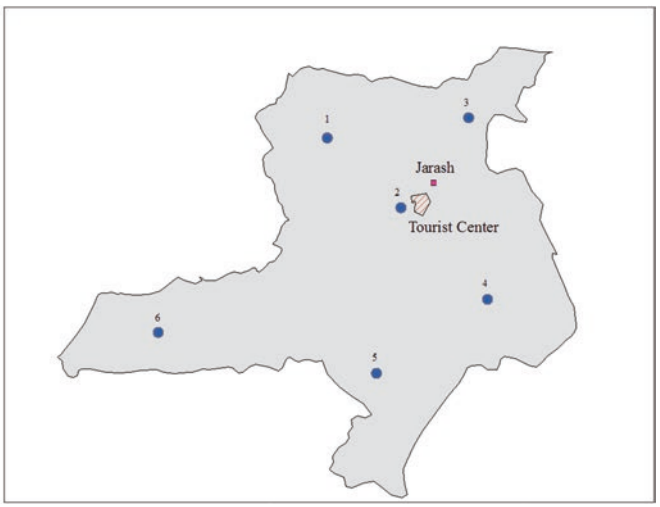
Electronic Supplementary Material: The online version of this chapter (https://doi.org/10.1007/978-3-319-61158-7_15) contains supplementary material, which is available to authorized users.

W. Bajjali, *ArcGIS for Environmental and Water Issues*, Springer Textbooks in Earth Sciences, Geography and Environment,
https://doi.org/10.1007/978-3-319-61158-7_15

ID	Depth	TDS	X-Coordinate	Y-Coordinate
1	376	1307	228970.98	1189834.73
2	200	4160	232948.09	1185966.58
3	123	3965	236816.24	1191033.32
4	90	1950	237796.90	1180899.85
5	150	2015	231640.55	1176868.26
6	60	1872	219545.77	1179210.94
Average			**231286.42**	**1183968.95**

2. **Central feature**: is the feature that associated with the smallest accumulated distance to all other features in a study area. For example there are 6 wells in the Jarash area (Figure below) and to calculate the central feature, the 6 wells (table below) will be organized into a table. The 6 wells are represented as records and columns and then the distance between the wells will be recorded. The sum of total distance of each well from the rest of the wells is then recorded, and the central feature will be the well that has the lowest total distance from all other wells. In the table below, you can see that well No 2 is selected as the central feature.

	Well 1	Well 2	Well 3	Well 4	Well 5	Well 6	Sum
Well 1	0.00	5.55	7.94	12.56	13.24	14.20	53.49
Well 2	5.55	0.00	6.37	7.01	9.19	15.00	**43.12**
Well 3	7.94	6.37	0.00	10.18	15.08	20.92	60.49
Well 4	12.56	7.01	10.18	0.00	7.36	18.33	55.44
Well 5	13.24	9.19	15.08	7.36	0.00	12.32	57.19
Well 6	14.20	15.00	20.92	18.33	12.32	0.00	80.77



3. **Median Center** Median Center is a slightly different way to calculate the middle (actually quite complicated to compute) and is a point in a pattern which minimizes the distance between itself and all other points. Median Center identifies the location that minimizes overall Euclidean distance to the features in a dataset.

Calculate Mean Center with and Without Weight

This section will explore the Mean Center with or without weight.

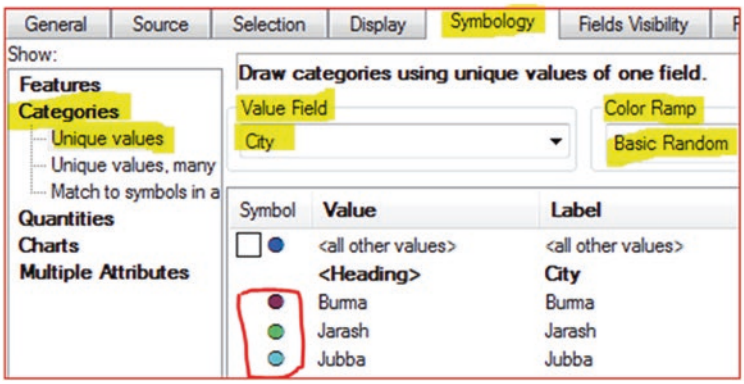
Mean Center

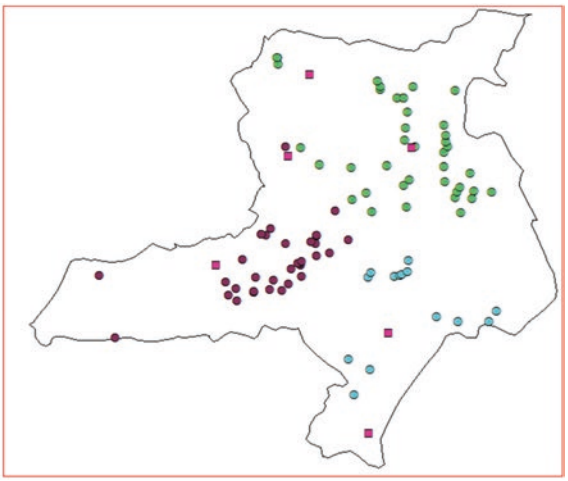
In order to find the center of a random distributed features over an area, you need to use the Mean Center tool resides in the spatial statistics tools. Calculating the center has many applications in applied sciences, especially in geoscience. The center is a feature in the middle of a given set of data and can service all other feature with a shortest time. For example, a set of groundwater wells is located in a particular study area, and finding the center of the wells will help building a water tower that will collect water from the surrounding wells faster and with less expense.

Scenario 1: In Jordan, in summer time, the demand for potable water increases. The Water Authority in the Jarash governorate has decided to build an extra water tower in the area to be used as a distribution center during the summer. The water tower should be supplied with water from a high quality groundwater source and it should be located in the center of the wells that belong to the major cities in the governorate.

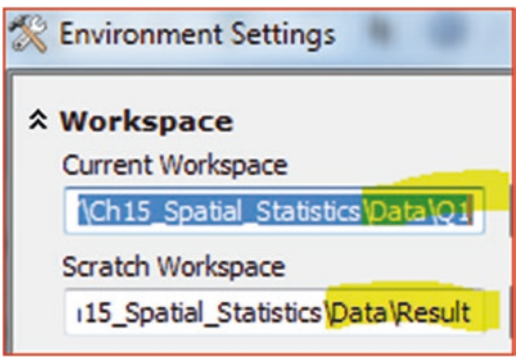
GIS Approach

1. Launch ArcMap and call the Layers Data Frame "**Mean Center**"
2. Integrate from \\Ch15\Data\Q1 folder **Governorate.shp**, **Town.shp**, and **Well.shp**
3. Click the symbol of Town layer in the TOC, select Square 2, color pink, size 9, OK
4. D-click Well layer in TOC/Symbology tab/choose categories and unique values/Value Field, enter "city"/click on "Add All Values"/uncheck <all other value>
5. Click on "Symbol" and choose properties for all symbols/choose circle 2 and change the size to 10/OK/choose Basic Random from the Color Ramp/OK
6. Click the symbol of Governorate/click Hollow and then OK



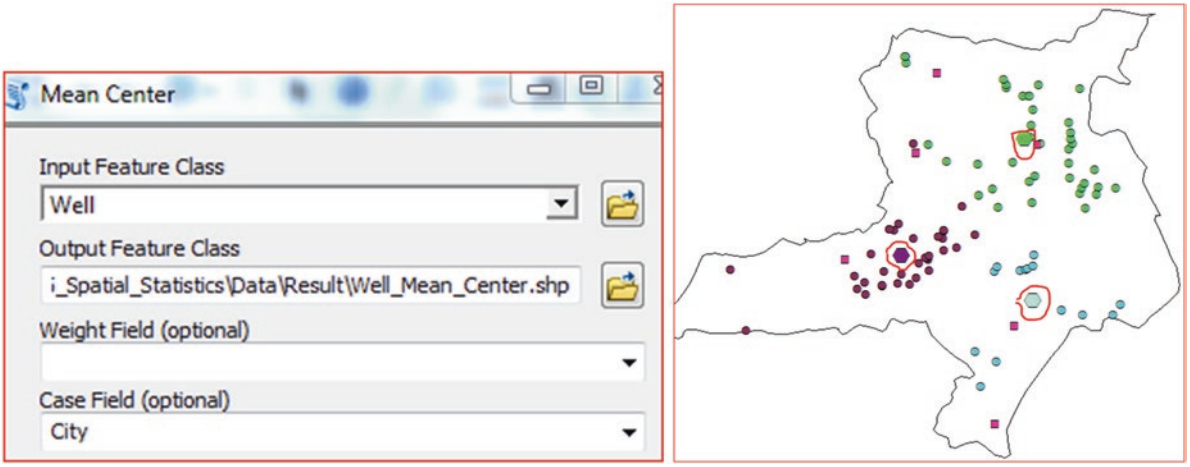


7. Launch ArcToolbox, click an empty place at the bottom of ArcToolbox
8. Click Environment/Workspace
9. Current Workspace \\Ch15\Data\Q1
10. Scratch Workspace \\Result
11. OK



12. ArcToolbox/Spatial Statistics Tools/Measuring Geographic Distributions
13. D-click Mean Center
14. Input Feature Class: Well
15. Output Feature Class: \\Result\Well_Mean_Center.shp

16. Case Field: City
17. OK
18. D_click on **Well_Mean_Center** layer/symbology tab/Categories/unique values/Choose "City" as the value field/click on add all values. Click on "symbol" and "properties for all symbols"/Change the symbols to hexagon 2, make the size as 20, then OK. Choose different colors to match the color of the wells in each town.

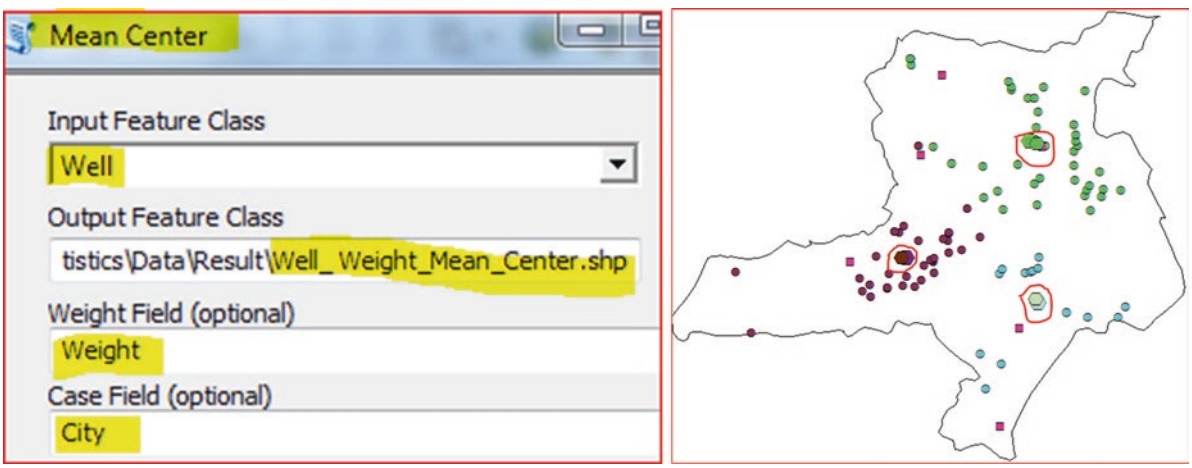


Mean Center with Weight

The Next Step is to run the **Mean Center** again on the well layer using the **Weight** field. The **Weight** field has values from 1 to 2. Value 2 indicates the most important criteria, and signifies wells that have TDS and NO_3^- less than 1000 and 45 mg/L respectively and depth of wells less than 300 m.

Run the **Mean Center** function again using the **Well** layer and the **Weight** field, call the output **Well_ Weight_Mean_ Center.shp**

1. D-click Mean Center
2. Input Feature Class Well
3. Output Feature Class \\Result\Well_Weight_Mean_Center.shp
4. Weighted Field: Weight
5. Case Field: City
6. OK
7. D-click on "Well_Weight_Mean_Center layer/Symbology/Categories, unique values" and again choose the city as the value field.
8. Click add all values and click on symbol and choose properties for all symbols.
9. Change the symbol to hexagon 2 size 18/choose the color ramp as before so the well weight mean center has the same color scheme as the cities.



Questions:

1. Are the **Well_ Weight_Mean_Center.shp** aligns with the **Well_Mean_Center**?
2. If they don't align how far are they from each other?
   Hint: Use the "measure" tool to measure the distance between the mean center and weighted mean center for each wells in the three cities.
3. Run the Central Feature tool and compare the result with the result of the mean center and comment on the result

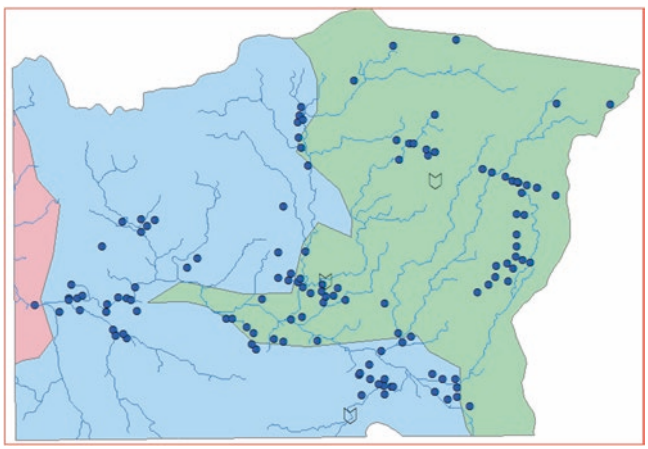
Standard Distance and Mean Center

The standard distance calculate the mean center of the displayed features and then draws a buffer around the mean center with a radius equal to the standard distance value. There are three values of the standard deviation. The first standard deviation will cover at least 68% of the sample features, the second standard deviation cover at least 95% of the sample features, while the third standard deviation covers almost 99% of all the samples.

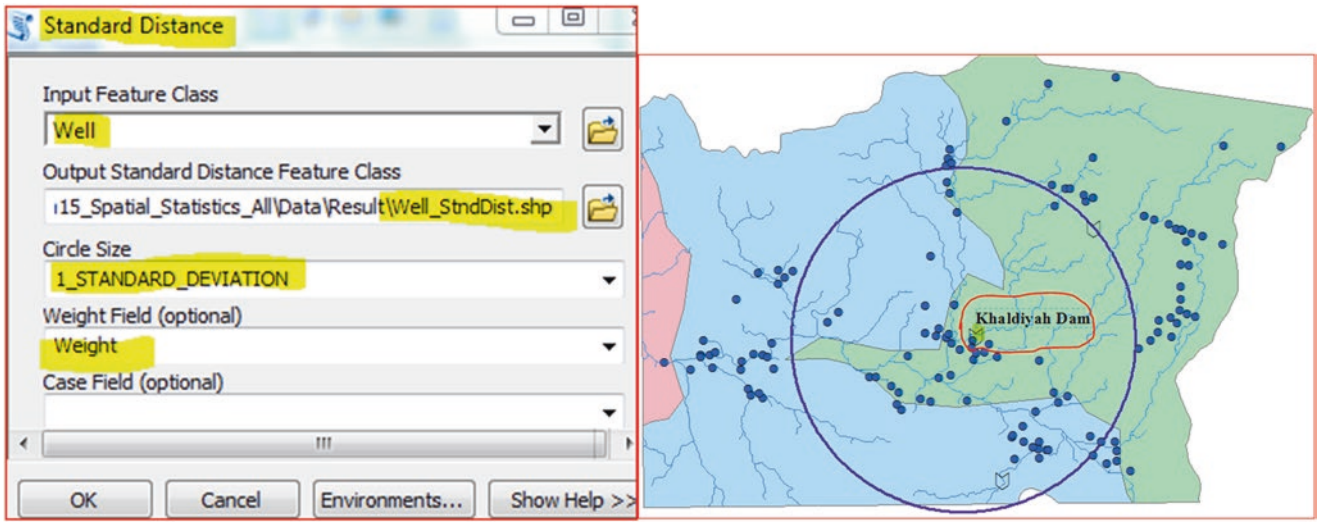
Scenario 2: You are a hydrogeologist and your goal is to replenish the groundwater and improve its water quality using an artificial recharge method. To perform this task, you decided to at least select one well that is located within 1 standard deviation from the mean center and in close proximity to the Khaldiyah dam. To execute the assignment, you decided to calculate the standard circle using a "**weight**" criteria. The weight is based on total depth of the wells and more emphasis is placed on the wells that are shallower than 100 m. A field called weight is added to the attribute table of the wells.

GIS Approach

1 Insert Data Frame and call it Standard Distance
2. Integrate **Dam.shp**, **Geology.shp**, **Stream.shp**, and **Well.shp** from \\Ch15\Data\Q2 folder
3. D-click the Geology layer in the TOC/Symbology/Categories/Unique values = Code/Add All Values/Uncheck <all other values/OK
4. Click the symbol of the Stream in TOC/select the River symbol in the Symbol Selector dialog box/then click OK
5. Click the symbol of the Dam in TOC/select the Dam Lock symbol in the Symbol Selector dialog box/then click OK (or search for "Dam Lock" in the Symbol Selector)
6. Click on "Symbol" of the Well in TOC/choose circle 2 and change the size to 10/choose blue color/then click OK



7. ArcToolbox/Spatial Statistics Tools/Measuring Geographic Distributions
8. D-click Standard Distance
9. Input Feature Class: Well
10. Output Standard Distance Feature Class: \\Result\Well_StndDist.shp
11. Circle Size: 1_Standard_Deviation
12. Weight Field: Weight
13. OK



Result: The **Standard distance** will calculate the mean center based on the weight, and then a buffer is created around the mean center using 68% of the wells in the study area.

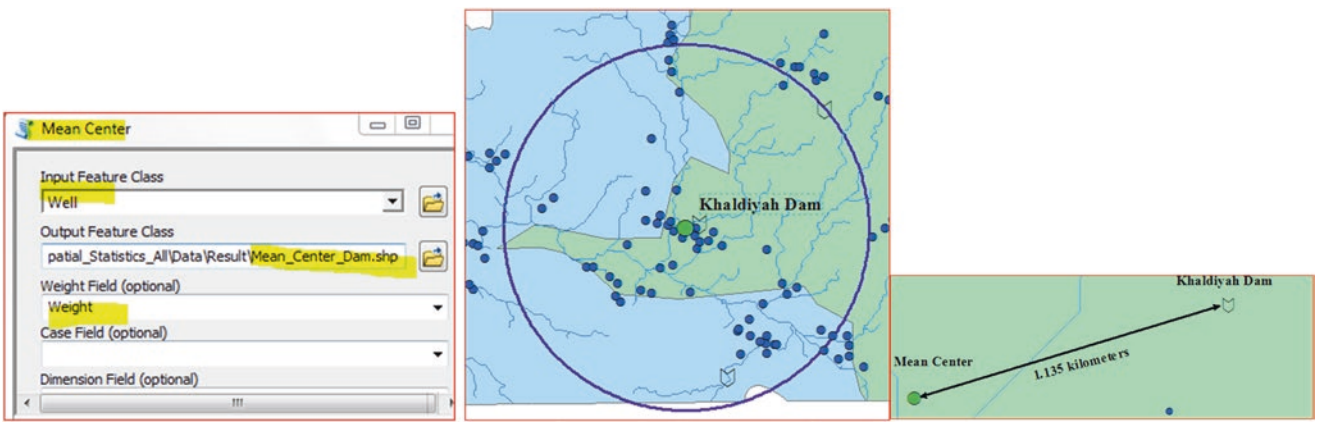
Note: In order to pursue with the analysis you have to answer the following questions:

1. How far the Khaldiyah dam from the center of the buffer
2. What is the closest well to the dam that can be used in the artificial recharge

Distance of the Khaldiyah Dam from the Center of the Buffer

To calculate the distance of the Khaldiyah Dam from the center of the buffer, you have to run the Mean Center tool as shown above and then use the measure tool to find the distance between the center of the buffer and the dam.

14. D-click Mean Center
15. Input Feature Class: Well
16. Weight Field: Weight
17. Output Feature Class: \\Result\Mean_Center_Dam.shp
18. OK



In order to find the well that is close to **Khaldyia dam**, you have to zoom in around the **Khaldyia dam** and use the measure tool to see, which well is the closest.

Result: It is clear that the Well No 109 has the shortest distance (419.88 m) to the **Khaldyia dam**.

Analyzing Pattern Toolset

Identifying Pattern Based on Location

Some statistical analyses aims to identify patterns, trends, and spatial relationship among features in any environment. Whether a certain set of data is more likely to show certain characteristics; Some Spatial Analyst tools are able to recognize the distribution patterns of geographic layers in a specific study area. In the geography Discipline there is a well documented practice that demonstrates how features are located near each other are more similar than features situated farther away from one another (Tobler's First Law of Geography). This idea is common sense, nevertheless, there is always exception to the rules. For example the weather in the Jordan Rift Valley, which is about 400 m below sea level is not similar to weather in the Ajloun High Lands, which is more than 1000 m above sea level and is only 20 km far away from each other. At the same time, the climate of the city of Aqaba (Jordan) is similar to the city of Jeddah (Saudi Arabia), even if the two cities are 970 km away from each other.

Average Nearest Neighbor tool

The "***Average Nearest Neighbor***" tool will detect if features are clustered or dispersed; this tool is used and tested with some degree of confidence level. The statistical approach behind this method is that the tool will measure the distance from each feature in the dataset to its single nearest feature neighbor and then calculating the average distance of all measurements. The tool then creates a hypothetical dataset with same number of features, but placed **randomly** within the **study area**. The tool then will be run again and measure the nearest distance to its nearest neighbor feature and calculate the average. The average distance of the random hypothetical data will be assessed with the real data. Two parameters will be generated: I and Z-score

A: a nearest neighbor index (I) is generated as follow

$$I = \frac{D_r}{D_h}$$

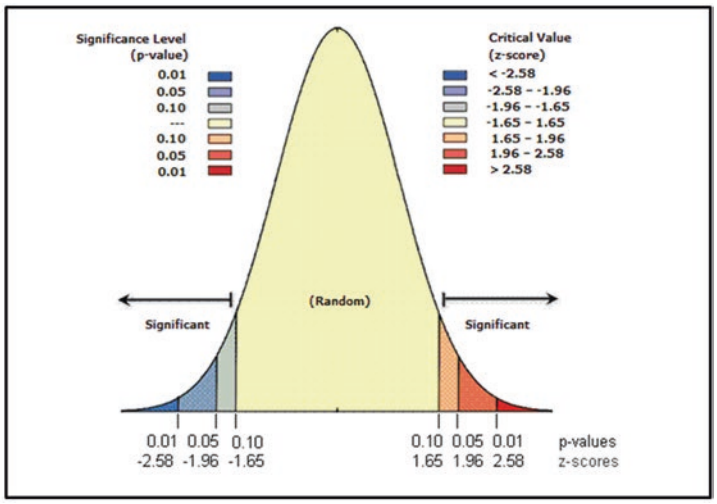
D_r is the calculated average distance of the real data
D_h is the average distance from the hypothetical data

If I < 1 the data show clustering
If I > 1 the data show dispersion
If I = 1 the data is randomly distributed

A pattern that falls at a point between dispersed and clustered is said to be random

B: A z-score will be calculated and is vital to making a decision to accept or reject the null hypothesis. The z-score is associated with the confidence level and is up to the researchers to adopt which confidence level they are willing to test with their hypothesis. Each confidence level is associated with z-score which is simply a standard deviation. For example, a 90% confidence level has a z-score between −1.65 and +1.65 and the 95% confidence level has a z-score between −1.96 and +1.96 (table below)

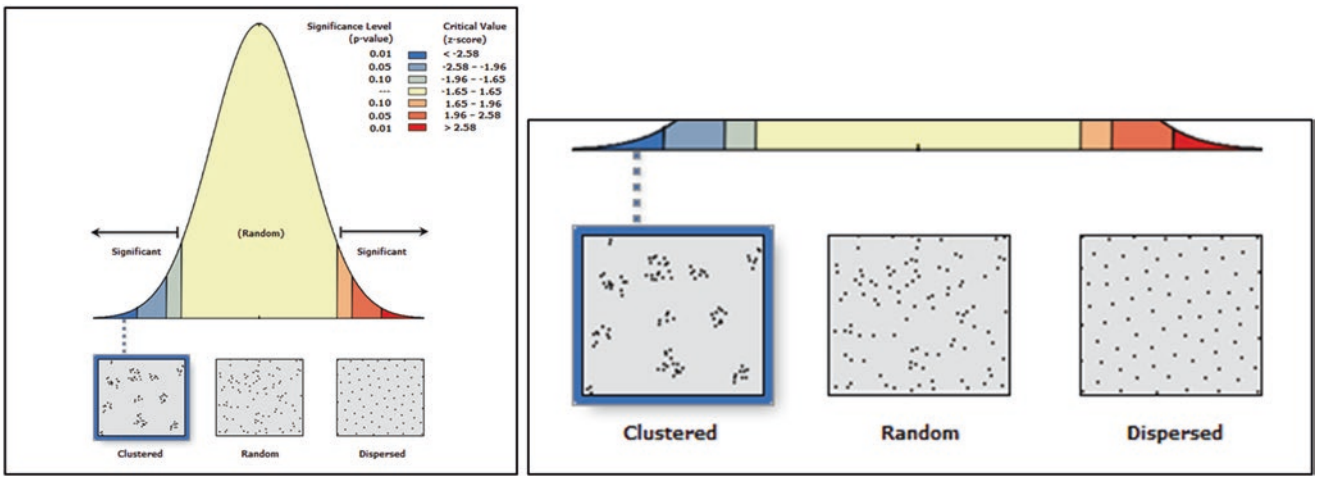
Z-score (Standard Deviations)	p-value (Probability)	Confidence level
-1.65 or +1.65	0.10	0.9
-1.96 or +1.96	0.05	0.95
-2.58 or +2.58	0.01	0.99



Null Hypothesis

In any statistical testing, you have to propose a **null hypothesis** and the null hypothesis states that features in the study area lacking any pattern. This means that the features are not clustering or dispersed, but randomly distributed.

Let us assume you are willing to test your hypothesis with 95% confidence level and you are assuming (Null Hypothesis) that the features are **randomly distributed**. After running the test the Z-score value that generated is between **−1.96 and +1.96** and your p-value is larger than **0.05**. Based on the result you have to **accept** the null hypothesis, which means that your features are randomly distributed. But if the Z-score fell outside that range for example −2.0 or +2.0 standard deviations, you have to **reject** the null hypothesis and your observed features are clustering or dispersed.

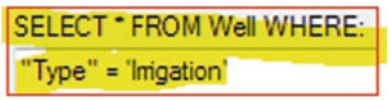


Scenario 3: You are a hydrogeologist and you have observed a heavy groundwater abstraction from the wells that are used for irrigation in Wala catchment area. This practice has dramatically lowered the water table in these wells, which affected the water balance in the whole basin. You have decided to examine if the wells' locations are one of the reasons that generate a dramatic dropdown. The nearness of the wells from each other could affect the zone of influence created by the wells' pumping. Your question is, are the wells used for agriculture randomly distributed or do they have a certain pattern (clustered or dispersed). To find out the answer you have to do the following:

GIS Approach

Propose a ***Null Hypothesis*** stating that wells are randomly distributed in the basin

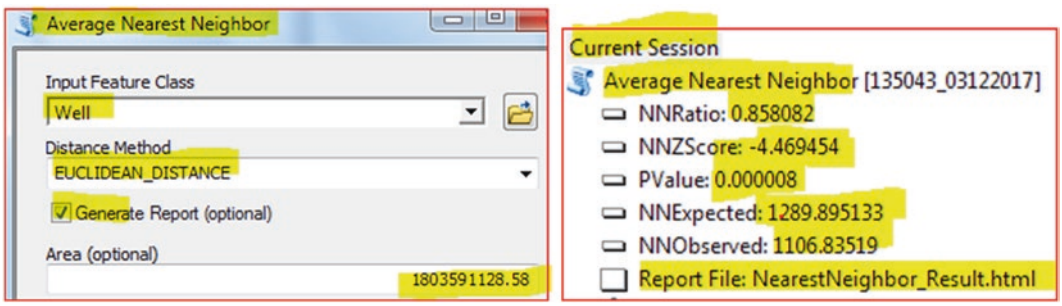
1. Insert Data Frame and call it "**Nearest Neighbor**"
2. Integrate the **Well.shp** and **WalaWatershed.shp** into from \\Ch15\Data\Q3 folder
3. R-click Well layer/Open Attribute Table (there are 333 wells)
4. R-click Well layer in the TOC/Properties tab/click Definition Query/click Query Builder
5. Write the SQL statement: "Type" = 'Irrigation'
6. OK/OK



Result: The well layer that used for irrigation is the only one is displayed and the rest of the wells are hidden now. You can verify that by opening the attribute table.

7. R-click Well layer/Open Attribute Table (there are 271 wells)
8. Open attribute table of WalaWatershed, the Shape_Area is **1,803,591,128.58 (unit is m^2)**

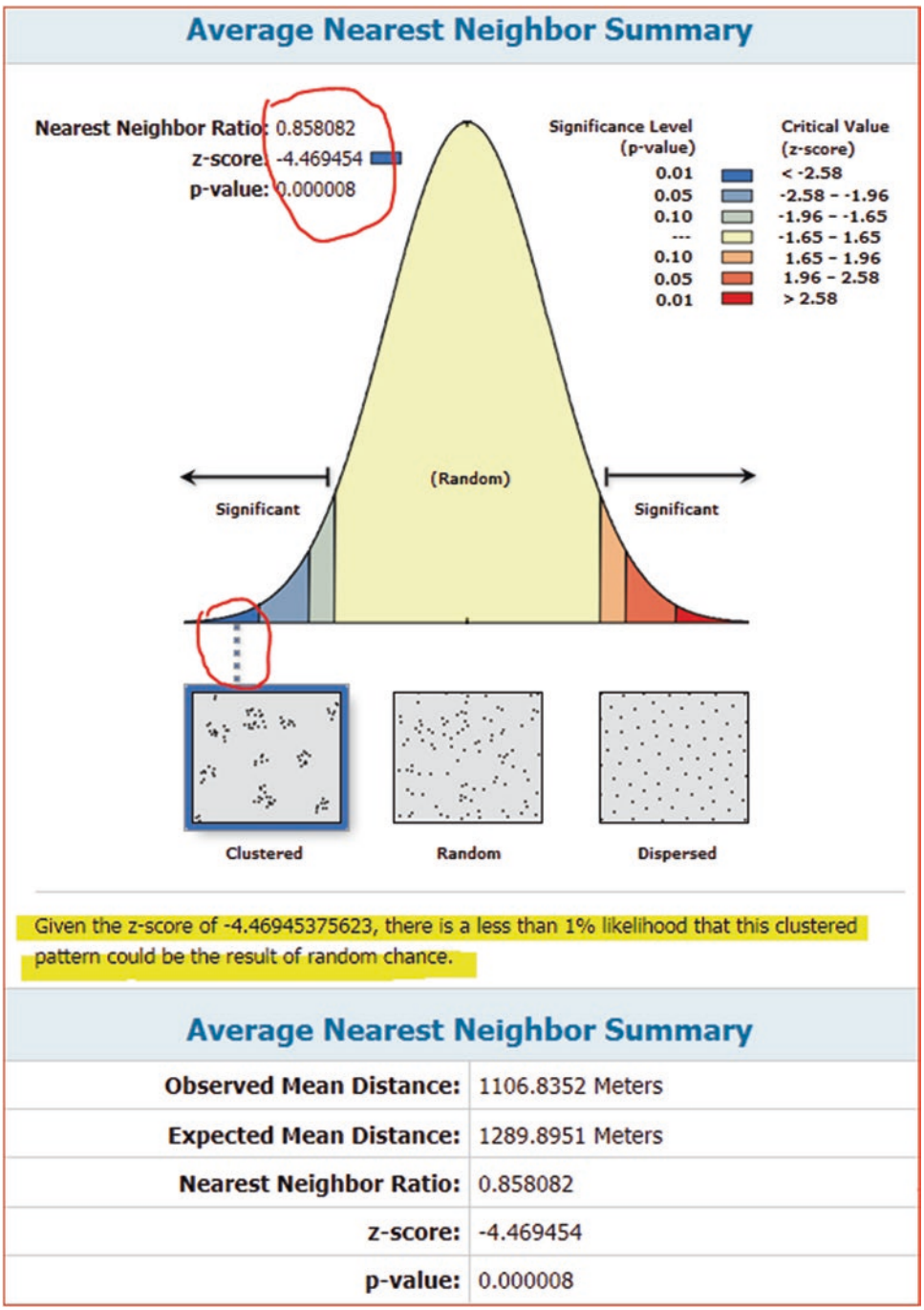
9. ArcToolbox/Spatial Statistics Tools/Analyzing Patterns
10. D-click "Average Nearest Neighbor"
11. Input Feature Class: Well
12. Distance Method: EUCLIDEAN_DISTANCE
13. AREA 1803591128.58
14. Check "Generate Report"
15. OK
16. Click Geoprocessing menu/and point to Results
17. Open the Current Session
18. Open Average Nearest Neighbor



19. Write down the following values:

Observed Mean Distance (NNObserved) = 1106.84
Expected Mean Distance (NNExpected) = 1289.89
Nearest Neighbor Ratio (NNRatio) = 0.858
Z-Score: (NNZScore) = −4.47

20. D-click Report File: Nearest Neighbor_Result.html
Result: The following graph displays



Interpretation: Given the z-score of −4.47, there is less than 1% possibility that the irrigation wells in the Wala catchment area could be the result of random chance. Because the "I" ratio is also 0.858, which is less than 1, we reject the null hypothesis and we consider the distribution of the irrigation wells clustered.

Identify Pattern Based on Values (Getis-Ord General G)

The location of features is not only determined by the clustering but also the values associated with the feature within a key distance of each other. The General G-statistics tool will be used to identify high or low values over the entire study area. The distance, which is based on either Euclidean or Manhattan will be calculated and it will reveal whether if it is significant or not. The tool allows user also to specify how spatial relationship among features are defined. For example, the "Fixed Distance Band" Each feature is analyzed within the context of neighboring features. Neighboring features inside the specified critical distance (Distance Band or Threshold Distance) receive a weight of one and exert influence on computations for the target feature. Neighboring features outside the critical distance receive a weight of zero and have no influence on a target feature's computations. The distance is an important part of the General G-statistics as it will show over which the tool will be ascertained to be significant. The ideal distance will be determined using the "Calculate Distance Band from Neighbor Count" tool.

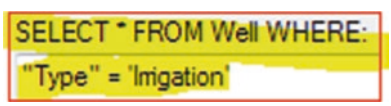
The "Calculate Distance Band from Neighbor Count" tool returns the minimum, the maximum, and the average distance to the specified Nth nearest neighbor (N is an input parameter) for a set of features, for example 5 wells.

The General G tool calculates the value of the General G index, Z score and p-value for a given input feature class. 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 the features are randomly distributed. The Z score value means the following:

a) A Z score near zero indicates no apparent clustering within the study area.
b) A positive Z score indicates clustering of high values.
c) A negative Z scores indicates clustering of low values.

Scenario 4: In the previous scenario we found out that the irrigation wells are clustering, this time you have to see if the "**Weight**" field has an influence on the clustering of the same wells in the watershed and at what distance the clustering taking place. The "**Weight**" field has values from 1 to 5, with 5 representing the wells that have the highest yields. In this exercise, you have to run the "Calculate Distance Band from Neighbor Count" tool to find the ideal distance to run the General G-statistics. This will show over the tool which will be ascertained to be significant. After identifying the average distance which will return the minimum, the maximum, and the average distance to the 5th nearest neighbor (N = 5 wells), you should use value higher and lower than the average return value and run all of them to decide which distance is ideal to use the General G-statistics.

1. Insert Data Frame and call it "**General G-Statistics**"
2. Integrate the **Well.shp** and **WalaWatershed.shp** from \\Ch15\Data\Q4 folder
3. R-click Well layer/Open Attribute Table (there are 333 wells)
4. R-click Well layer in the TOC/Properties tab/click Definition Query /click Query Builder
5. Write the SQL statement: "Type" = 'Irrigation'
6. OK/OK



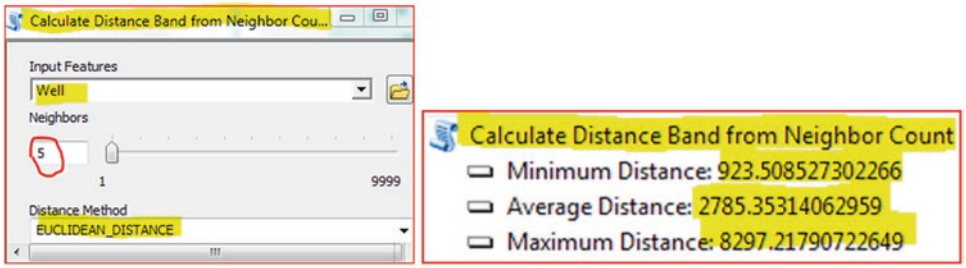
Result: The well layer that used for irrigation is the only one is displayed and the rest of the wells are hidden. You can verify that by opening the attribute table.

Null Hypothesis: Irrigation wells with high-ranking values represented by the "**Weight**" field are randomly distributed in the study area.

Find the Ideal Distance

Before running the General G-statistics, you have to run the "Calculate Distance Band from Neighbor Count" tool in order to find the best distance to use with the General G-statistics.

1. ArcToolbox/Spatial Statistics Tools/Utilities
2. D-click Calculate Distance Band from Neighbor Count
3. Input Features: Well
4. Neighbors: 5
5. Distance Method: EUCLIDEAN DISTANCE
6. OK

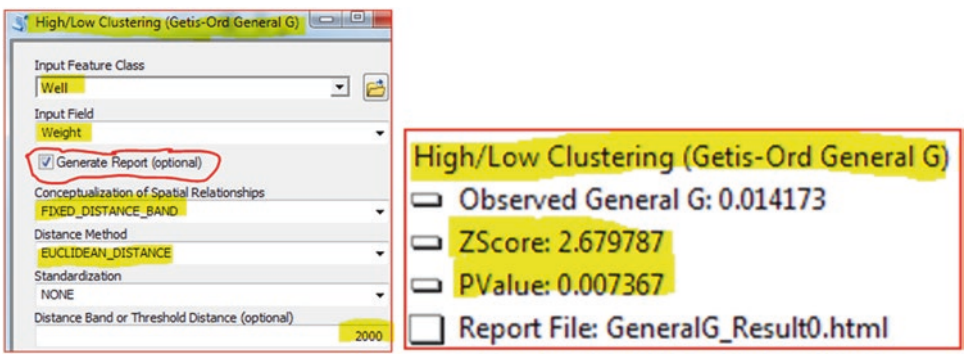


Result: If you open Geoprocessing menu/Result, you will notice the following:

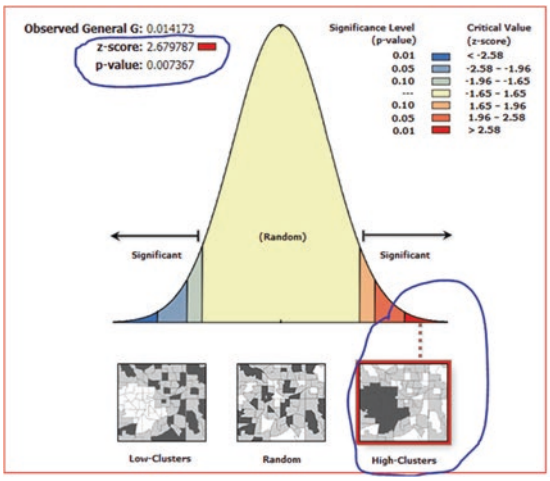
1. Minimum Distance: 923.50
2. Average Distance: 2785.35
3. Maximum Distance: 8297.21

The average distance is 2785 m with five neighbors, therefore a lower and higher number should be used to determine which value to use to run the General G statistical tool. The distance to try should be from 2000 to 3400 m at 400 m intervals. Run all these values using the "High/Low Clustering (Getis-Ord General G)" tool and fill the table at the end.

7. ArcToolbox/Spatial Statistics Tools/Analyzing Patterns
8. D-click "High/Low Clustering (Getis-Ord General G)"
9. Input Feature Class: Well
10. Input Field: Weight
11. Check the Generate Report
12. Conceptualization of Spatial Relationships Fixed_Distance_Band
13. Distance Method: Euclidian_Distance
14. Standardization: None
15. Distance band: 2000
16. OK



17. Click Geoprocessing menu/and point to Results
18. Open the Current Session
19. You see the result of the analysis
20. D-click on the Report File: GeneralG_Result



Distance (m)	Observed General G	Z-Score
2,000	2.679	0.007
2400	1.25	0.018
2800	1.32	0.024
3000	1.34	0.027

21. Repeat the previous step and replace the distance with 2400, 2800, 3000, and 3400

Interpretation: The best distance for clustering is the one that has the highest z-score.

Spatial Autocorrelation (Moran's I)

Moran's I index measure the spatial correlation using the feature location and an attribute value together to determine statistically if the data is clustered, dispersed or random. Using the spatial correlation helps define how the variables are arranged in a study area. The tool calculates the Moran's I Index value, Z score, and p-value. If the Moran's Index value is near +1.0 this indicates clustering while an index value near −1.0 indicates dispersion. The method has no output layer, but a report is established and demonstrate whether the well distribution in the watershed is clustered, dispersed, or random.

The tool will be run using the conceptualization of spatial relationship of Zone of indifference, Euclidian distance, and distance band of 500, 1000, 1500, 2000, 2500, and 4000. The concept of Zone of indifference that wells within the specified critical distance (Distance Band or Threshold Distance) of a target well receive a weight of one and influence computations for that well.

Scenario 5: Your supervisor asked you to look at the density of groundwater wells per block and would like to hear your judgment about at what distance these well densities cluster. This information is critical for management purposes as it helps to adjust the rate of pumping of the wells that are located close to each other in clustering pattern. Your duty is to do the following:

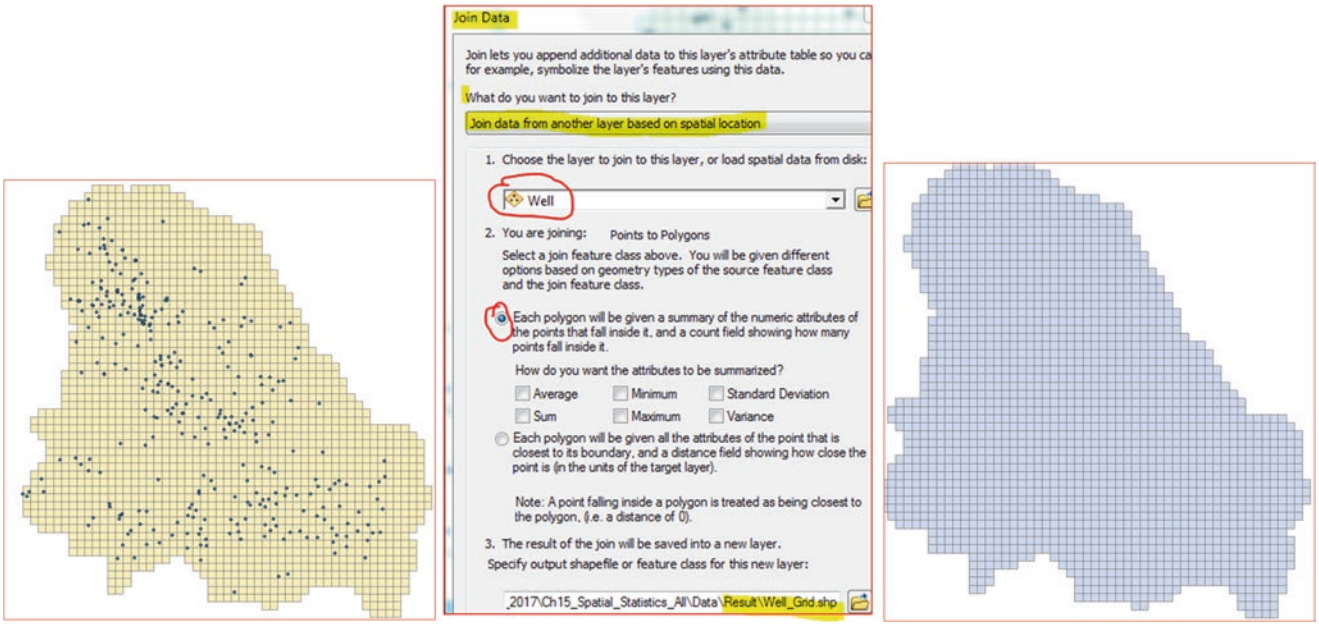
Prepare the Data for Analysis

1. Insert Data Frame and call it "**Moran's I Index**"
2. Integrate the **Grid_1000.shp**, **Well.shp**, and **WalaWatershed.shp** from \\Ch15\Data\Q5 folder

Spatial Join Between Grid_1000 and Well Layers

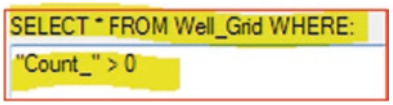
3. R-click Grid_1000/Join and Relate/click Join
4. Join Data from another layer based on spatial location
5. Choose the layer to join to this layer Well

6. Check “Each polygon will be given a summary of the numeric attributes of the points that fall inside it”.
7. Save it in \\Result as **Well_Grid.shp**
8. OK

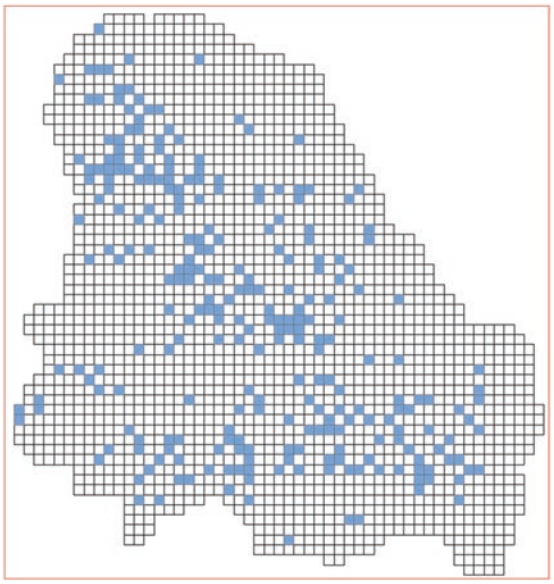


Result: The Well_Grid is created and it shows the density of the groundwater wells in the Wala watershed. If you open the attribute table, you will find a field called “**Count_**”. Some of the values are zero, which shows the cells that has no wells. These cells should be removed before running the statistics.

9. R-click Well_Grid/Properties/Definition Query/click Query Builder and type the following statement "Count_" > 0
10. OK/OK



Result: The Well_Grid displays showing only the cells in the grid that have wells inside them. Some cells contain 1 well and other more than one well. The maximum wells found in one cell is 7, and they are in the northern part of the watershed.



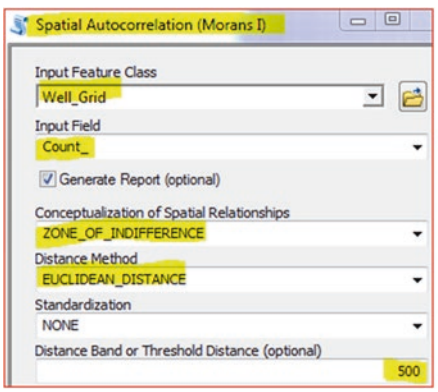
Well_Grid

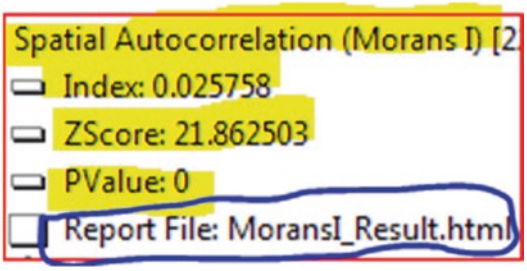
FID	Shape	Well_FID	PageNumber	Area	Length	Count_
33	Polygon	33	34	1000000	4000	1
92	Polygon	92	93	1000000	4000	1
93	Polygon	93	94	1000000	4000	2

Run Spatial Autocorrelation (Moran's I)

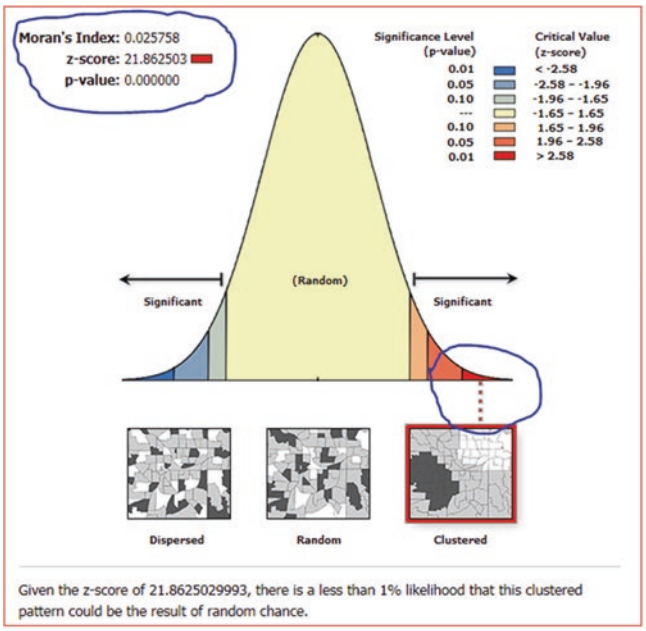
11. ArcToolbox/Spatial Statistics Tools/Analyzing Patterns
12. D-click "Spatial Autocorrelation (Morans I)"
13. Input Feature Class: Well_Grid
14. Input Field: Count_
15. Check the Generate Report
16. Conceptualization of spatial relationship: ZONE_OF_INDIFFERENCE
17. Distance Method: EUCLIDEAN DISTANCE
18. Distance Band: 500
19. OK
20. Click Geoprocessing menu/and point to Results
21. Open the Current Session
22. Open Spatial Autocorrelation (Moran I)

Result: The Spatial Autocorrelation tool returns the Moran's I Index, ZScore, and PValue, and the pattern. If you checked the Generate Report in the Spatial Autocorrelation (Moran I) tool, an HTML file with a graphical summary of results will be created.



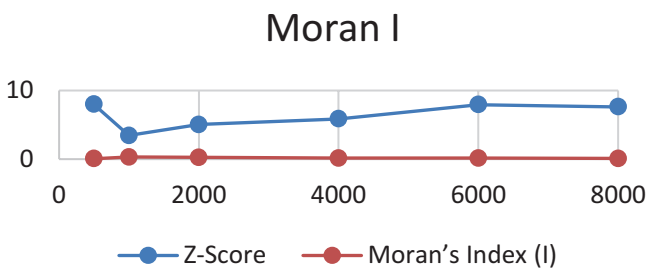


23. D-click on the HTML file, it will open the HTML file in the default Internet browser
24. Repeat the steps above and use 1000, 2000, 4000, 6000, and 8000 and fill the table below



Distance	Z-Score	Moran's Index (I)	Pattern
500	8	0.066	Clustered
1000	3.43	0.289	Clustered
2000	5.03	0.241	Clustered
4000	5.87	0.147	Clustered
6000	7.94	0.134	Clustered
8000	7.62	0.098	Clustered

Interpretation: The most significant clustering occur at a distance, where the Z-score is the highest and the Moran's Index (I) is the lowest.



Mapping Clusters

Cluster and Outlier Analysis (Anselin Local Moran I)

The cluster analysis will examine a dataset of features (such as wells) with a value associated with the features (such as depth or salinity). The output result of the analysis will be displayed as a feature class and the clustering will be highlighted. The new output feature class will have the following fields in the attribute table: Local Moran's I index (LMiIndex), z-score (LMiZScore), pseudo p-value (LMiPValue), cluster/outlier type (COType), in addition to other fields from the original layer. The z-scores and p-values are measures of statistical significance which tell users whether to accept or reject the null hypothesis. The interpretation of the result will be based on the following fields in the attribute table:

A high positive z-score in the attribute table indicates that the surrounding features have similar values (either deep wells or shallow wells).

The COType field will be HH for a statistically significant cluster of high values (deep wells) and LL for a statistically significant cluster of low values (shallow wells).

A low negative z-score (less than −1.4) for a well indicates a statistically significant spatial data outlier. The COType field indicate if the well has a deep well and is surrounded by wells with shallow depth (HL) or if the well has a shallow depth and is surrounded by wells with deep wells (LH).

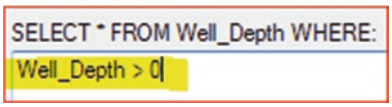
No permutations are used to determine how likely it would be to find the actual spatial distribution of the wells you are analyzing. For each permutation, the neighborhood values around each feature are randomly rearranged and the Local

Moran's I value calculated. The result is a reference distribution of values that is then compared to the actual observed Moran's I to determine the probability that the observed value could be found in the random distribution. The default is 499 permutations; however, the random sample distribution is improved with increasing permutations, which improves the precision of the pseudo p-value.

Scenario 6: In Amman-Zarqa basin, there are many groundwater wells drilled for agricultural development and they are tapping two aquifer systems: the carbonate and basalt aquifers. The wells that penetrate the basalt are in general deeper than the wells tapping the carbonate aquifer, especially in the eastern part of the basin. Your duty is to identify if there is a clustering based on the wells' depth.

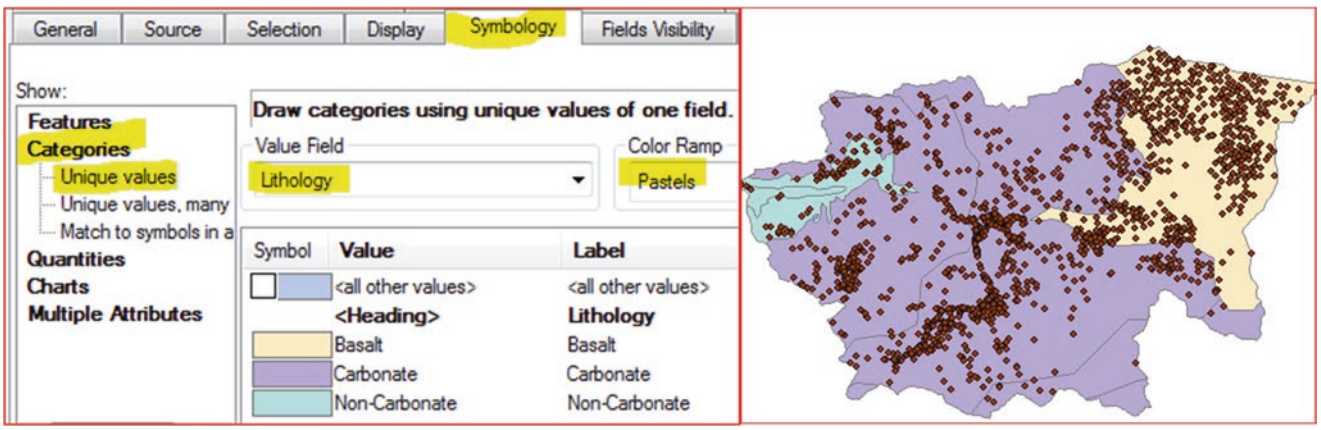
GIS Approach

1. Insert Data Frame and call it **Clustering**
2. Integrate the Well.shp, Geology.shp, and WWTP.shp from \\Ch15\Data\Q6 folder
3. R-click Well/Properties/Definition Query/click Query Builder
4. Type the following SQL statement "Well_Depth" > 0, then click OK/OK



Result: the well number decrease from 2039 to 1787.

5. D-click Geology layer/Symbology/Categories/Unique values, Value Field = Lithology/Add All Values/Color Ramp = Pastels/OK



Run Cluster and Outlier Analysis (Anselin Local Moran I)

This method allows you to use a distance of your choice in order to find a significant number of neighbors. A 1000 m and Euclidean distance will be used, wells outside the 1000 m for a target well are ignored in the analyses for that well.

6. ArcToolbox/Spatial Statistics Tools/Mapping Clusters
7. D-click "Cluster and Outlier Analysis (Anselin Local Moran I)"
8. Input Feature Class: Well
9. Input Field: Well_Depth
10. Output Feature Class: \\Result\MICluster1000.shp
11. Conceptualization of spatial relationship: FIXED_DISTANCE_BAND
12. Distance Method: EUCLIDEAN DISTANCE

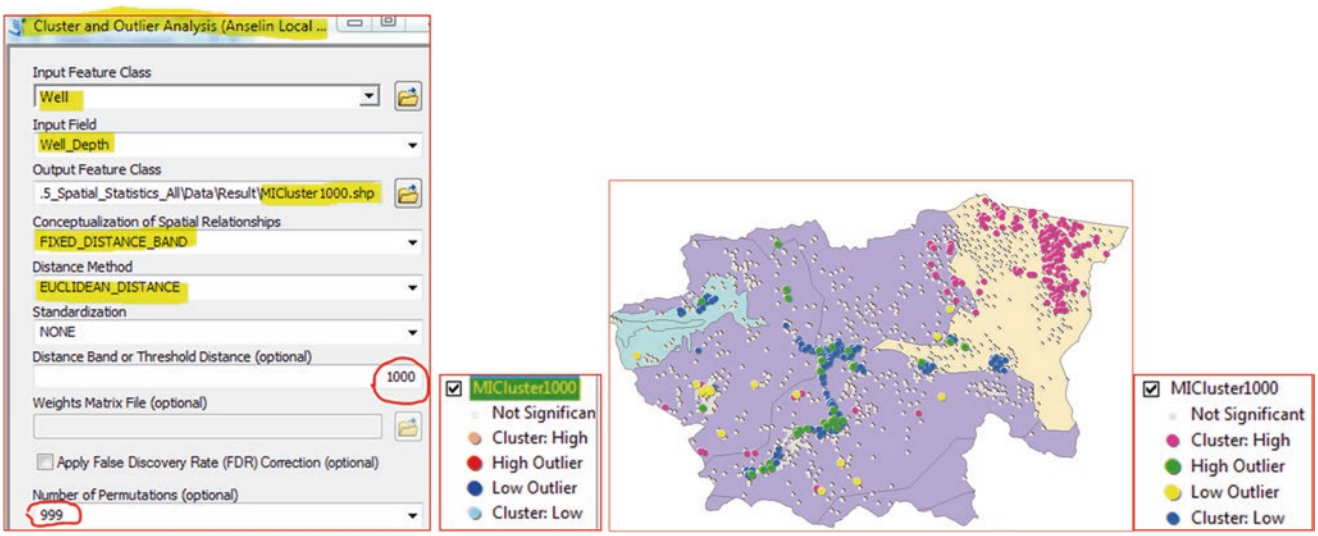
13. Distance Band: 1000
14. Number of Permutations: 999
15. OK

Result: The MICluster1000 layer will be displayed in the TOC with five classes.

16. R_click the symbol of "**Not significant**" and change its color to **Grey 10%**, the "**High-High Cluster**" to **Ginger Pink**, "**High-Low Outlier**" to **Leaf Green**, "**Low-High Outlier**" to **Solar Yellow**, and the "**Low-Low Cluster**" to **Cretan Blue.**

Interpretation: 192 wells have HH records, which means statistically significant cluster of deep wells and surrounding by deep wells. The depth of the wells in meters range from 230 to 675.

387 wells have LL records, which means statistically significant cluster of shallow wells and surrounding by shallow wells. The depth of the wells in meters range from 5 to 217.



Hot Spot Analysis (Getis-Ord GI*)

This is another method to identify statistically significant spatial clusters of wells of high depth (hot spots) and Shallow depth (cold spots) using the Getis-Ord Gi* statistic. The tool creates a new output layer with a z-score, p-value, and confidence level bin (Gi_Bin) for each well in the input layer. The z-scores and p-values are measures of statistical significance which tell the users whether to accept or reject the null hypothesis. The Gi_Bin field also identifies statistically significant hot and cold spots as below:

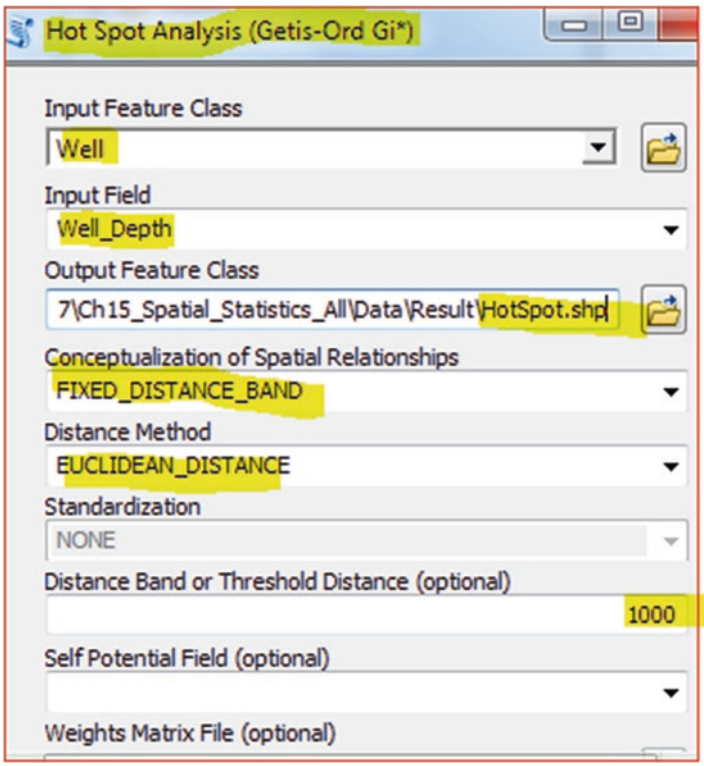
- Wells (+3 bins) reflect "Hot Spot" statistical significance with a 99% confidence level
- Wells (+2 bins) reflect "Hot Spot" statistical significance with a 95% confidence level
- Wells (+1 bins) reflect "Hot Spot" statistical significance with a 90% confidence level
- Wells (−3 bins) reflect "Cold Spot" statistical significance with a 99% confidence level
- Wells (−2 bins) reflect "Cold Spot" statistical significance with a 95% confidence level
- Wells (−1 bins) reflect "Cold Spot" statistical significance with a 90% confidence level
- Well with 0 bin indicates no apparent spatial clustering

A high z-score and small p-value for a well indicates a spatial clustering of deep well, while a low negative z-score and small p-value indicates a spatial clustering of shallow wells. The higher (or lower) the z-score, the more intense the clustering. A z-score near zero and with 0 bin indicates no clustering.

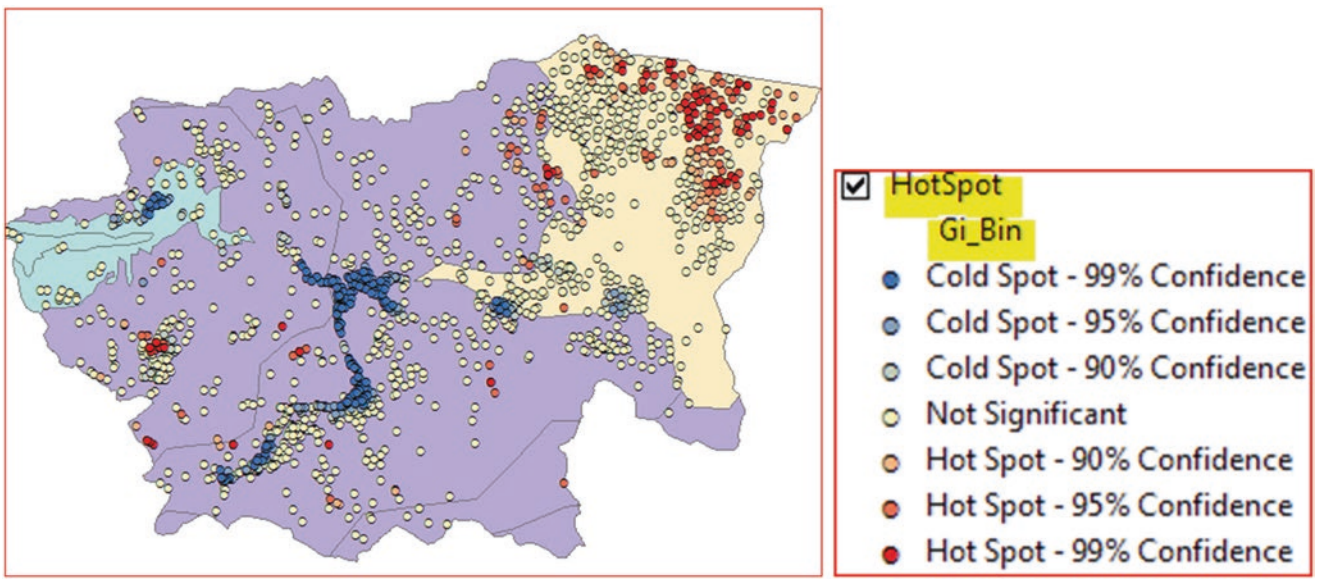
Scenario 7: Your advisor asked you to use the wells from the previous scenario to identify the hot and cold spot in Amman-Zarqa basin

1. Insert Data Frame and call it **Hot and Cold Spot**
2. Integrate the Well.shp and Geology.shp, from **Clustering** data frame

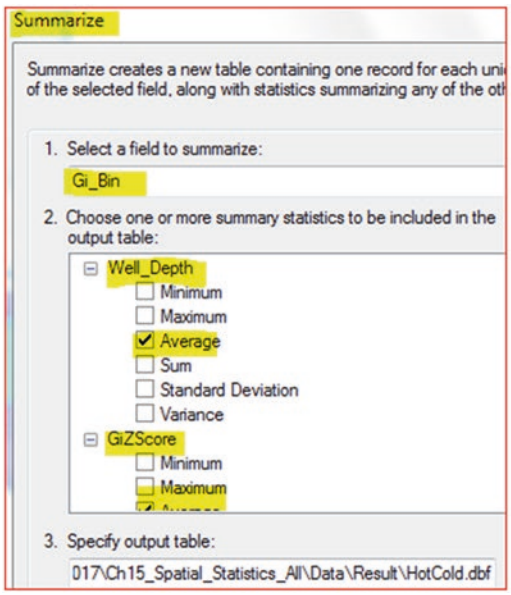
3. ArcToolbox/Spatial Statistics Tools/Mapping Clusters
4. D-click "Hot Spot Analysis (Getis-Ord Gi*)"
5. Input Feature Class: Well
6. Input Field: Well_Depth
7. Output Feature Class: \\Result\HotSpot.shp
8. Conceptualization of spatial relationship: FIXED_DISTANCE_BAND
9. Distance Method: EUCLIDEAN DISTANCE
10. Distance Band: 1000
11. OK



Result: The HotSpot layer will be displayed in the TOC with seven classes.



12. Open the attribute table of HotSpot Layer
13. R-click the Gi_Bin/Summarize
14. Select a field to summarize: Gi_Bin
15. Well_Depth: check Average
16. GiZScore: check Average
17. Specify output table: HotCold.dbf
18. OK and open the HotCold.dbf table



	OID	Gi_Bin	No of Wells	Average Well Depth	Average ZScore
	0	-3	305	82.74	-4.05
	1	-2	85	115.66	-2.27
	2	-1	76	119.34	-1.79
	3	0	1064	245.04	0
	4	1	65	380.42	1.79
	5	2	103	430.07	2.3
	6	3	89	459.55	3

Interpretation: The HotCold.dbf table includes seven records: Cold Spot, Hot Spot, and no clustering. The cold and hot spot each consists of three groups. A cold spot is depicted by a 99%−90% confidence interval which shows that they are shallow wells and have a negative average Gi_Bin and ZScore values. A Hot Spot is depicted by a 99%−90% confidence interval which shows that they are deep wells with a positive average ZScore and Gi_Bin values. Wells with ZScore close to 0 and 0 bin reflect no clustering.