Analysis Author: Kayla Aburida

Title: Lab 2: Hazardous Liquid Spills: South Central United States

Dataset source: ***GW Geography, Professor Hurley***

Dataset location: ***South Central United States.***

Date(s) of analysis: 9/25/2023- 10/11/2023.

Dataset time span: ***2010 - 2019***

Dataset Scale: ***State Level***

1. Layers

For this lab, the layers I had on my map were Study_Spills, Piplines_intersect, TESS_FINAL_Realistic, TESS_FINAL, TESS_POINTS_AND_LINES, TESS_PIPE_1, TESS_POINTS_1, TESS_TRIM_1, LAB_2, Study_States, and US_States.

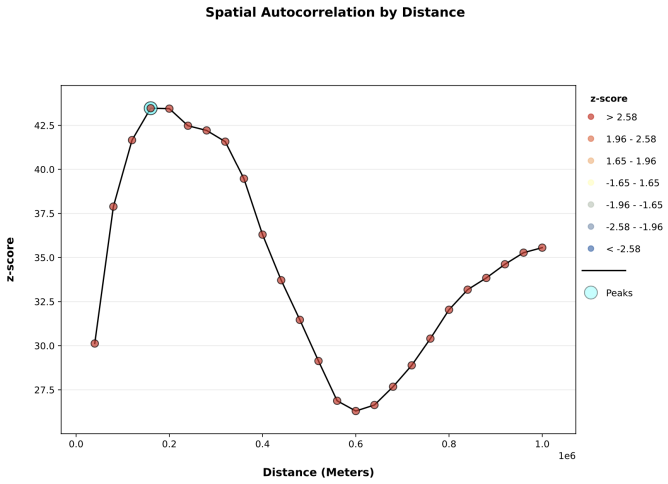
2. ISA

a. Explain what ISA is, making sure to detail how Moran's I is used.

ISA is the spatial autocorrelation function that examines spatial clustering at different scales or distance bands. ISA is the same as running the global moran's tool for a series of increasing distances, measuring the intensity of spatial clustering for each distance. This tool is especially helpful in determining distances where significant clustering occurs. This tool generates a graph with distance on the x-axis, and z-score on the y-axis. Peaks in the ISA graph show distances where there is significant spatial clustering.

b. What does your "SUM_Length" ISA graph show? Interpret the graph.

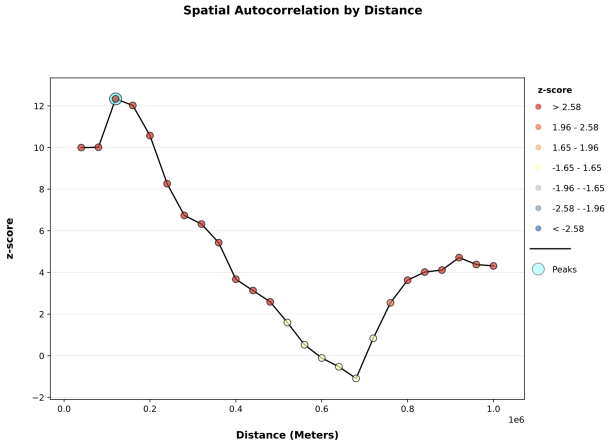
SUM_Length



Our SUM_Length ISA graph shows us what distances the highest amount of spatial clustering occurs. SUM_Length is the length of the pipeline in a tessellation hexagon. In this ISA graph, we can see that the distance we see the highest amount of spatial clustering is 43.473219 m.

c. What does your “Point_coun” Isa graph show? Tell me what you see.

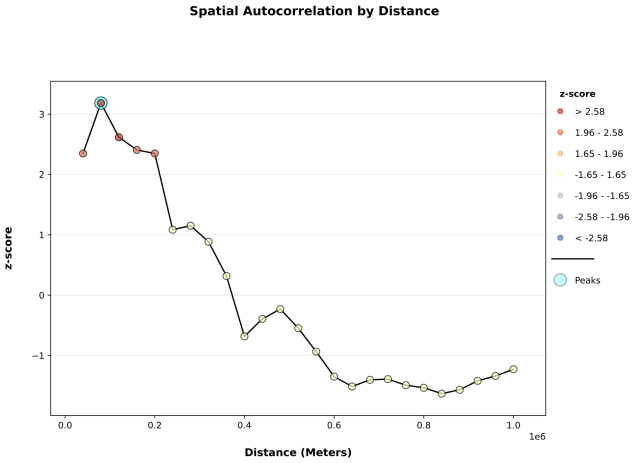
ISA_TESS_PointCount_Report



Our Point_coun ISA graph shows the spatial autocorrelation of the number of spills over 25 distance bands of 40,000 m with a 40,000 m increment. This ISA graph shows us the highest amount of spatial clustering is at 12.338755 m.

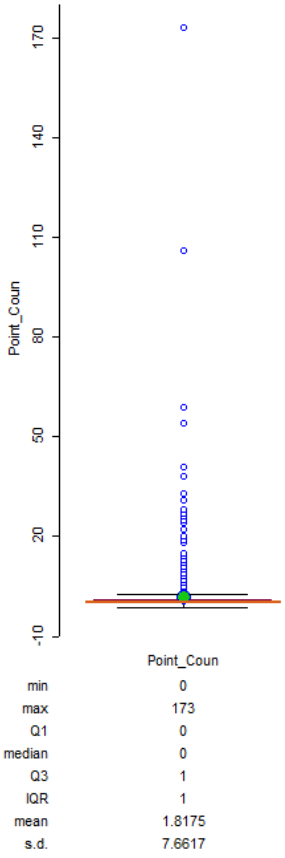
d. Compare your “Spills_km” ISA graph to your “Point_coun” graph. Did the peak distance change? Why or why not?

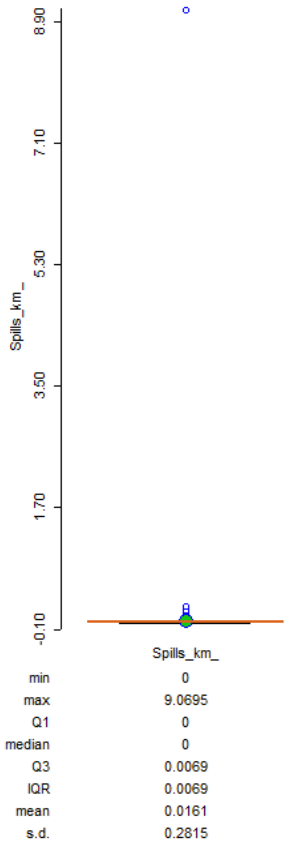
ISA_TESS_SPILLS_Report



The Spills_km ISA graph depicts the spatial autocorrelation of the proportion of the number of spills over the length of pipelines in each tessellation hexagon over 25 distance bands of 40,000 m with a 40,000. increment. The Spills_km ISA graph shows a peak of spatial autocorrelation at 3.183538 m. The peak of Spills_km vs. Point_coun differs by an increment of 9.155217. This is because the Spills_km ISA function shows spatial autocorrelation of the number of spills per length of pipeline, which differs from the spatial autocorrelation of the number of spills.

3. Box Plots





a. A boxplot of your "Point_coun" file. What does a boxplot do? Do you see any outliers?

A box plot is a method for graphically representing locality, spread, and skewness in groups of data. Additionally, through the Geoda application, the box plot function shows min, max, Q1, median, Q3, IQR, the mean, and standard deviation. The Point_coun box plot has the amount of spills per quadrant on the y-axis. All of the blue points on the box plot are outliers, there are quite a few of them. The maximum amount of spills in one quadrant is 173 and the minimum is 0. Because there are so many quadrants where there are no spills or even no pipeline, the data is skewed lower, with the mean being 1.8175.

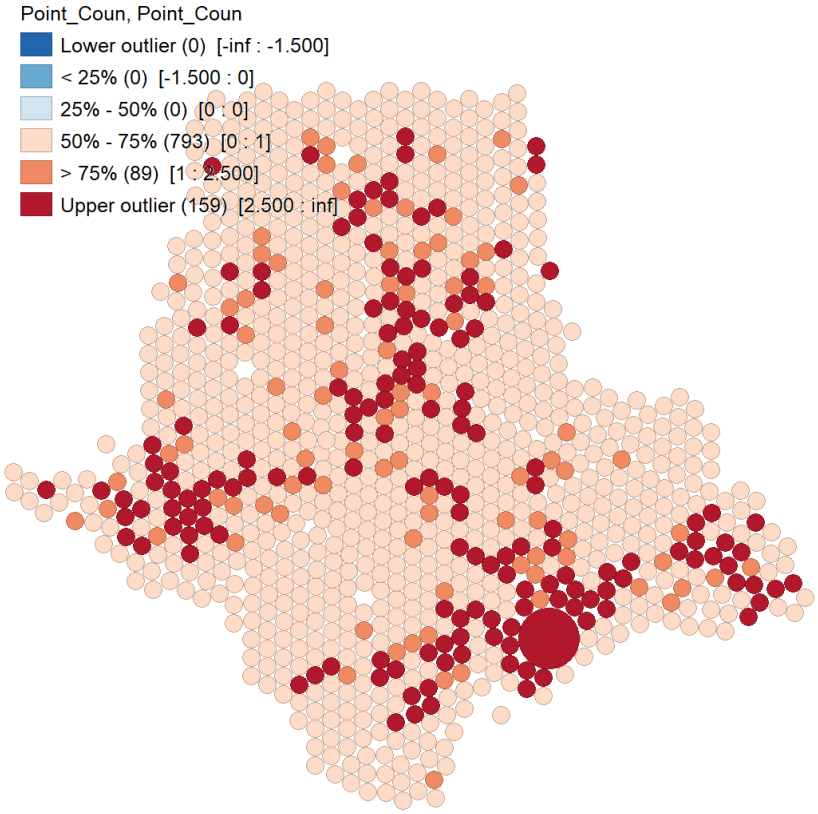
b. A boxplot of your “Spills_km” file. How does it differ from the “Point_Coun” file?

The spills_km data is a ratio of how many spills per km of pipeline, therefore we can assume results will be significantly lower. This data is hard to determine because we have one outlier that is so far from most of the data that we can see much of the box plot. That outlier lies around 9 spills/ km of pipeline. The reason we got this result is that the quadrant this data is coming from had only a fraction of the pipeline with one spill. Therefore 1/ 0.11026= 9.06947. This is an anomaly because it is rare to have that tiny fraction of pipeline and to create a large number in a proportion can throw off your data like it did here. Therefore there is one extreme outlier with a few other less extreme outliers we can barely see.

4. Cartograms

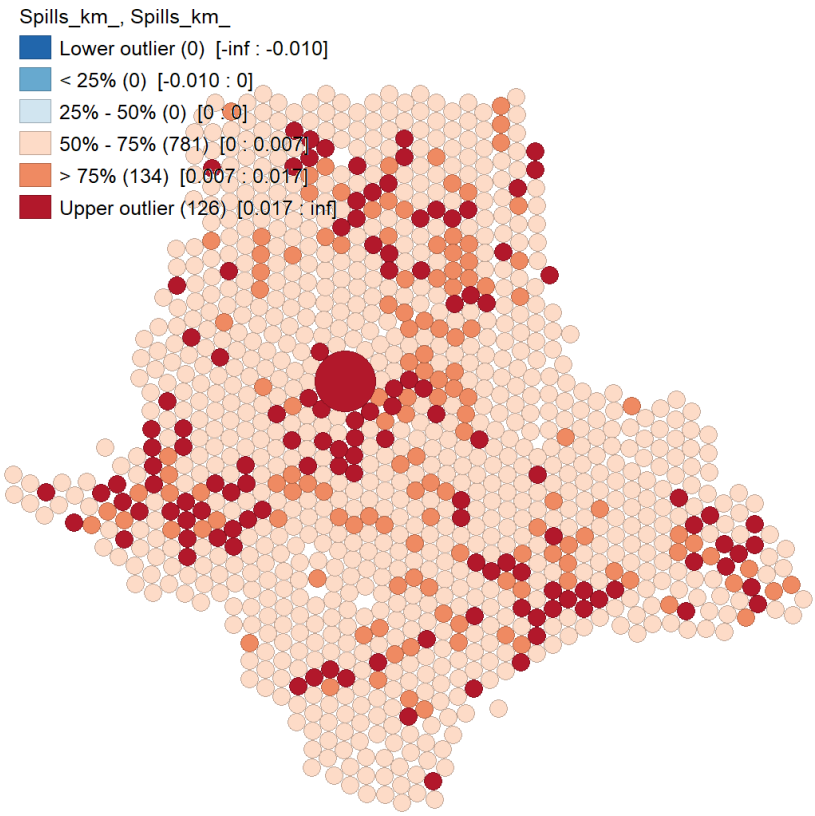
a. A cartogram of your “Point_Coun” file. What does a cartogram do? Do you see any patterns?

It’s a map type where the unit is expressed/replaced by a geometric form- a circle, rectangle, hexagon, etc. The proportion on the geometric form is equal to the value of the variable for the location. All in all. A cartogram is where geometric forms are used to reflect the spatial arrangement of locations based on their values. In the Point_Coun cartogram, I see a density of values and the upper outlier in the southeastern part of our study site. Another pattern I notice is that the darker values are rarely isolated from other dark values. I assume this is because when there are spills it generally is localized.



b. A cartogram of your "Spills_km" file. How does it differ from the "Point_Coun" file?

The cartogram of Spills_km looks a bit different although there seems to be some overlap. The spills_km cartogram has more isolated high values not near other high values. Additionally, the high outlier is that anomaly with the fraction of pipeline and one spill point. If we ignore that value the next highest value would be in Northern Texas where there were 106 spills and about 406 km of pipeline. The similarity I see between the two cartograms is in the southwest, southeast, and dead center of the study site; there seems to be a similar pattern of high-value circles.



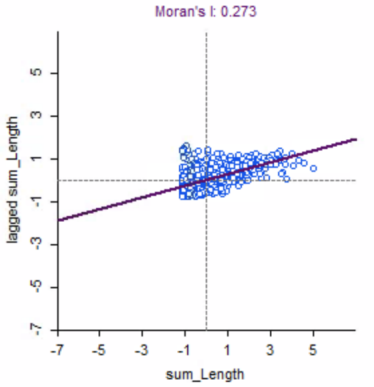
5. Moran's I

a. A Moran's I for "SUM_Length". Explain Moran's I, and interpret your output.

Moran's I is a tool that measures spatial autocorrelation based on both the feature's location and the feature's values at the same time. Based on a set of features and their attributes, it evaluated the pattern expressed as clustered, dispersed, or random. The Moran's I tool indexes z-score, and p-values to evaluate the significance of that index. The tool computes the mean and variance for the attribute being evaluated. Then, for each feature value, it subtracts the mean, creating a *deviation from the mean*. Deviation values for all neighboring features (features within the specified distance band, for example) are multiplied together to create a *cross-product*. We are specifically looking at the univariate Moran's I which specifically looks to be one variable of a unit's spatial correlation.

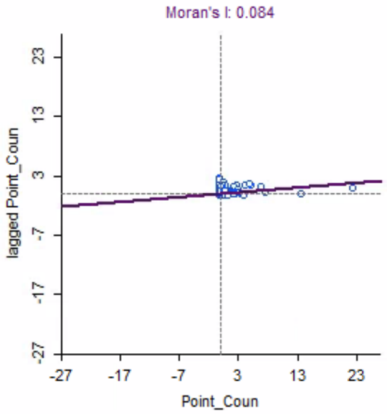
b.

The Moran's I scatter plot of **Sum_length** looks interesting, but to a spatial statistician, it's not very interesting. Notice how many of the points are centered in the middle, that means they are near the mean therefore they are not so many outliers and are uninteresting to us. However, we do see a few outliers essentially in the high-high positive correlation section (upper right) of our scatter plot. These will likely be the spatial outliers where there is high clustering and where there is a greater length of pipeline.



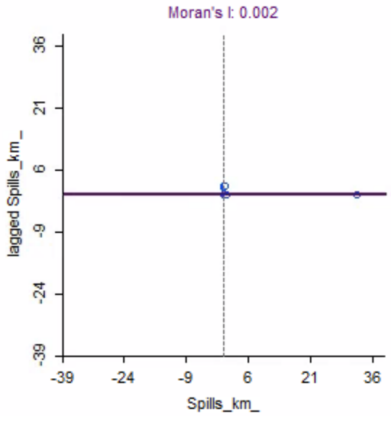
c.

Moran's I scatter plot for **Point_Coun** shows a few high-high positive spatial autocorrelation points, meaning that there are cases where there is a clustering of a high amount of spill points. I would assume this is around the southeast and southwest parts of our study section. If we look at our Point_coun cartogram we can see this visually.



d.

The Moran's I scatter plot for **Spills_km_** shows a pretty normal distribution of points… So it seems. Again we must remember that the Spills_km data set is the proportion of spills per km of pipeline, and the anomaly from the fraction of pipeline lives here. Therefore, if the outlier of this data set did not occur (the far right point), we might be able to see if there is more significant spatial autocorrelation in this data. However because of the anomaly Moran's I is at 0.002, which would have been higher if that anomaly had not occurred. Therefore without removing the data point, we don't have a way of knowing if there are outliers that indicate points of clustering/ spatial autocorrelation.

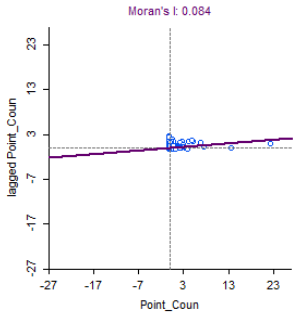
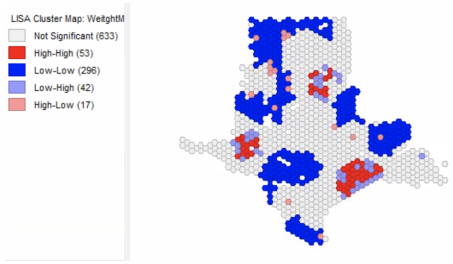
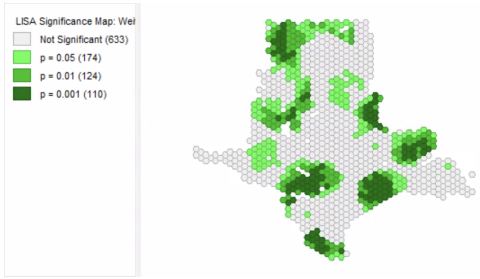
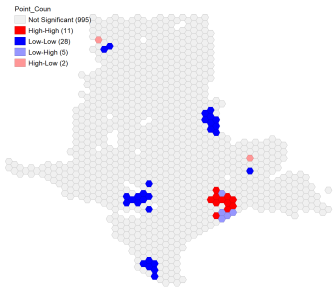
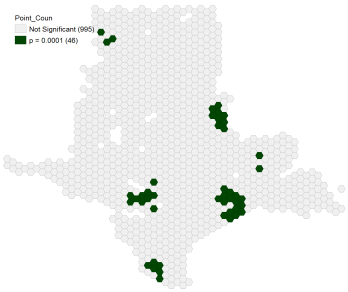


6. Local Moran's I

Local Moran's I runs a very similar function to that of global Moran's I. The Local Moran's I also have a principle called the local indicator of spatial association (LISA). LISA has two important characteristics, it provides statistics for each location with an assessment of significance, and it establishes a proportional relationship between the sum of the local statistics and corresponding global statistics. With this function you input a set of features and an analysis field the tool identifies locations of spatial clustering of features with high or low values as well as outliers, a local Moran's I value, a z-score, a pseudo p-value, and a code representing the cluster type for each statistically significant feature. The output is a positive or negative value of I. The positive I value indicates a feature that has neighboring features with similar high or low attribute values; the feature is part of a cluster. A negative I value means that the feature has neighboring features with dissimilar values; the feature is an outlier. The cluster/ outlier type is either a cluster of high values (HH), a cluster of low values (LL), an outlier of high value surrounded by low values (HL), or an outlier of low values surrounded by neighbors of high value.

To continue to confuse students, GIS statisticians added statistical significance, which is typically at a 95% confidence level. When no false discovery rate (FDR) correction is applied, features with p-values smaller than 0.05 are considered statistically significant. The FDR corrected this and reduced the p-value from 0.05 to a value that better reflects the 95% confidence level of your data set.

a. A Moran's I for "Point_Coun". Interpret your output.

		
After 9999 randomization, the significance filter at 0.0001		

Significance Map

These results indicate areas where there are HH, LL, HL, or LH values. I think these values for how the values of spills correlate with their neighbors make sense. We see a high amount of spill clustering located around other areas where there are high amounts of soil in the southwest and southeast parts of our study site. I interpreted this visually in the cartogram by noticing where there were clusters of a high amount of spills. The blue values indicate a low amount of spills neighboring other low amounts of spills areas. This also makes sense to me especially if we reference the Point_coun cartogram, the areas that have small and light-colored geometries align with the areas on the LISA significance map as dark blue.

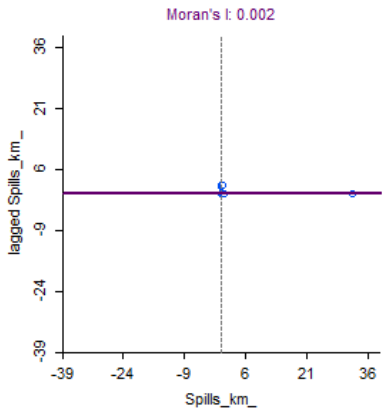
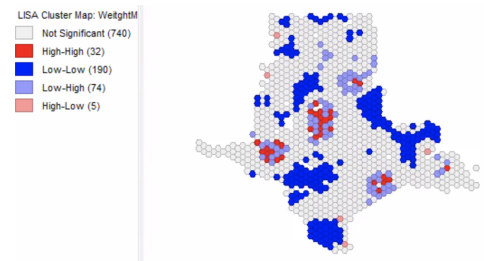
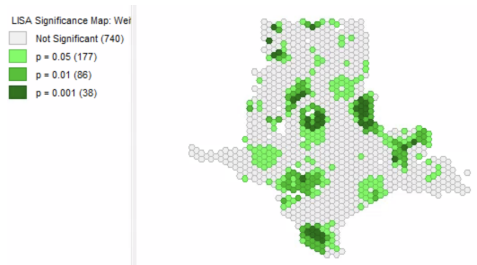
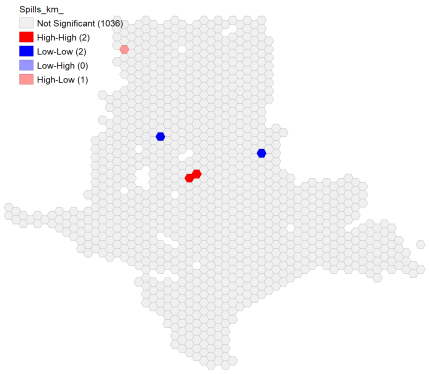
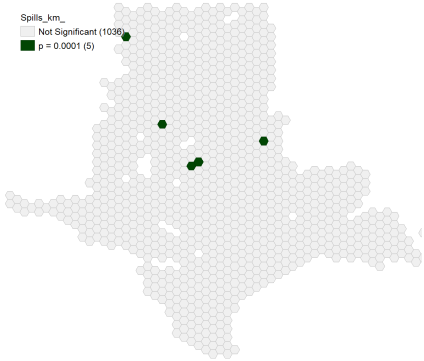
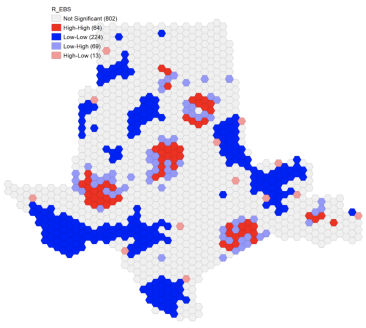
Cluster Map

This map indicates areas where there are statistically significant clusters. The dark values mean more clustering. The blank values determine that the clustering in these areas is not significant. We see sections of high clustering in the areas that are significant (HH, LL, HL, & LH) on the significance map. Although the cluster map does not tell us if there are high or low amounts of spills, but simplifies where the statistically significant values are significant concerning their neighbors. In this map, I notice high amounts of clustering throughout the study site with only the spatial correlation of where the points are on the significance map, not with correlation of the values of the significance map.

9999 randomization, significance value of 0.0001

Once we add 9999 randomization and change the significance filter to 0.0001 we see that there are a lot of points colored. We can interpret this by noting these values have one p-value, and the values that are shown are 99.99% probability against the null hypothesis. These results likely indicate where the most significant results are or where the densest clustering occurs

b. A Moran's I for "Spills_km". How has Local Moran's changed since we corrected for the length of the pipeline?

<table>
<tr>
<td> 
</td>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>After 9999 randomization, the significance filter at 0.0001</td>
<td> 
</td>
<td> 
</td>
</tr>
<tr>
<td>Empirical Bayes local univariate moran's I map for Spills_Km</td>
<td>  </td>
<td></td>
</tr>
</table>

The local Moran's I change in the Spills_km data set once we correct for the length of the anomaly pipeline after we run the empirical Bayes local univariate Moran's I in Geoda. That anomaly created skewed data as there was one outlier different from the rest, then after running empirical Bayes univariate Moran's I that anomaly no longer is significant and therefore the data becomes less skewed. In the map, we can see more values are significant with those H-H, L-L, H-L, and L-H values. They become more common because the data set is closer to the mean and therefore the function can see more outliers that are closer to the mean without the anomaly. In statistics jargon, we inflate the statistical significance of the dataset.

Question: In general, what is going on with the significance filter? What rules are we working under here (lab page 30, and lecture)?

Changing the significance filter to 0.0001 makes our results a 99.99% probability against the hypothesis. Therefore visually we only see the most statistically significant features of spatial autocorrelation on our map, which are two H-H values near the center, and two L-L values, one in the east and the other in the west of our study site. The rule we are working under here is False Discovery Rate (FDR), FDR correction estimates the number of false positives for a given confidence level and adjusts the critical p-value accordingly.