**GGRC32 – Essential Spatial Analysis**

**Assignment 4: Local Statistics and Interpolation**

**Fall Semester 2017**

**By: Grisham Nathan**

**Due November 21st by Noon**

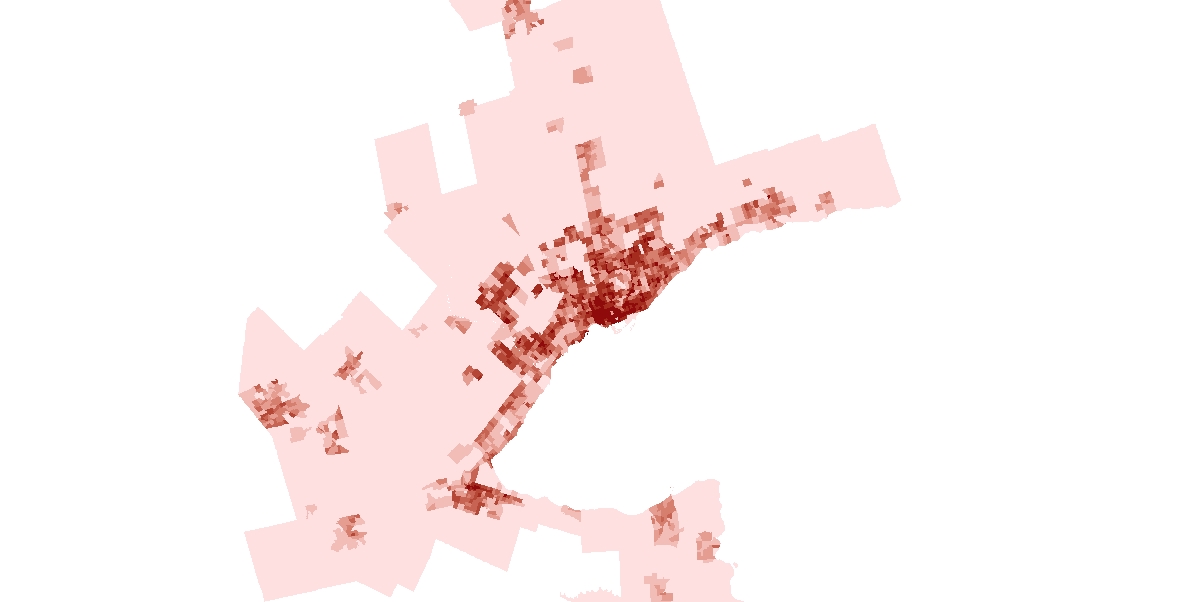
**Submit via Blackboard**

**Part I: Local Spatial Statistics**

You will be learning about Local Moran’s I (called Cluster and Outlier Analysis in ArcMap) and Getis and Ord’s statistic (called Hotspot Analysis in ArcMap) in this assignment. Both of these functions are available under “Mapping Clusters” in the Spatial Statistics Toolbox. You will be analyzing 2011 population density in the Greater Golden Horseshoe (GGH).

Make a choropleth map of population density. Choose a colour gradient appropriate for displaying low-to-high data values and make the polygon boundaries transparent so that spatial patterns in the core of the region are discernable.

1. **Provide a copy of your map. In a few sentences, describe the pattern of population density in the GGH. Where do you see areas of high (low) density? Where are the highest areas, where are the lowest? Is this pattern random, or do you see positive or negative autocorrelation?** (3 points)

****

I see most of the high density areas in downtown, midtown, Mississauga and Brampton. The lowest density areas are in Scarborough and also outside of Toronto. The pattern is positively autocorrelated, as most of the high density areas are highly clustered.

Compute the global Moran’s I of population density using queen contiguity.

1. **What are the results of the global Moran’s I calculations? What can you conclude about the spatial pattern of population density?** (3 points)

The results of the calculations are: Moran’s I = 0.418690, Z-Score = 30.753122, P-value = 0.

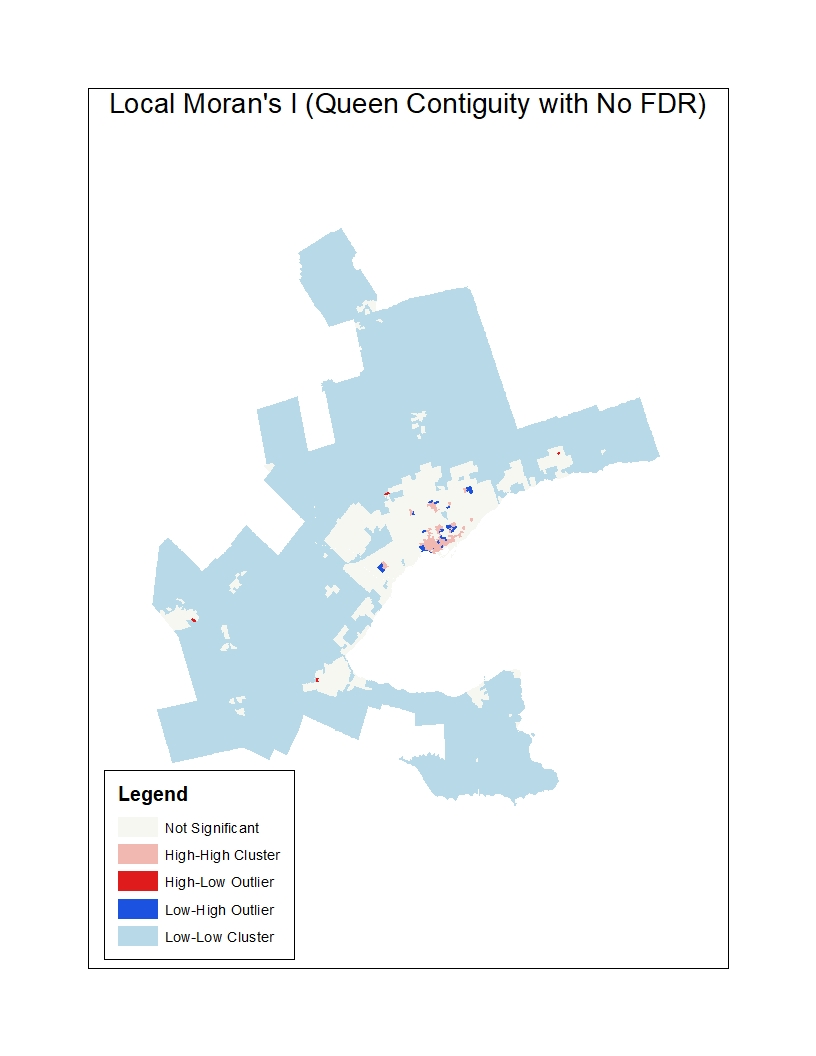
The population density of Toronto is positively autocorrelated and is more clustered than random.

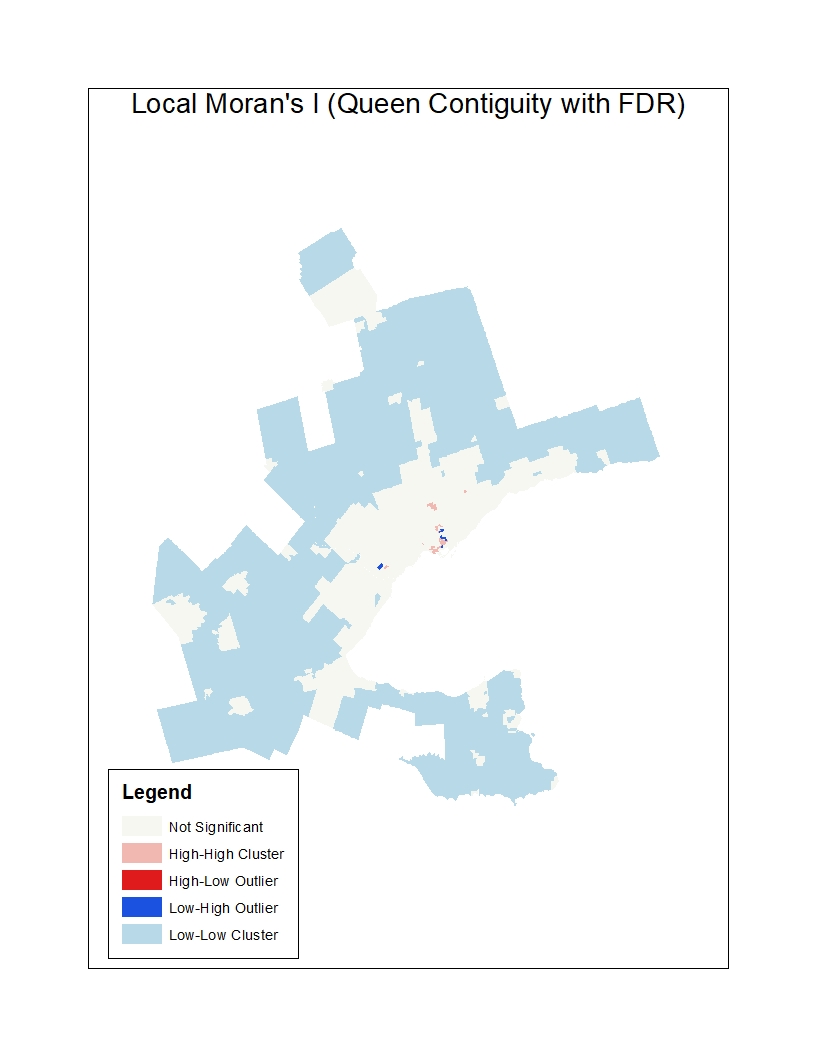
Compute Local Moran’s I using two different distance metrics: 1) queen contiguity and 2) inverse distance with a 15km bandwidth. Note that the bandwidth must be in the same units as the co-ordinate system of the data-frame. In both cases, compute the neighbour matrix with a row-standardization. Also, compute each of these statistics both with and without False Discovery Rate Correction. You should produce four sets of results in this step.

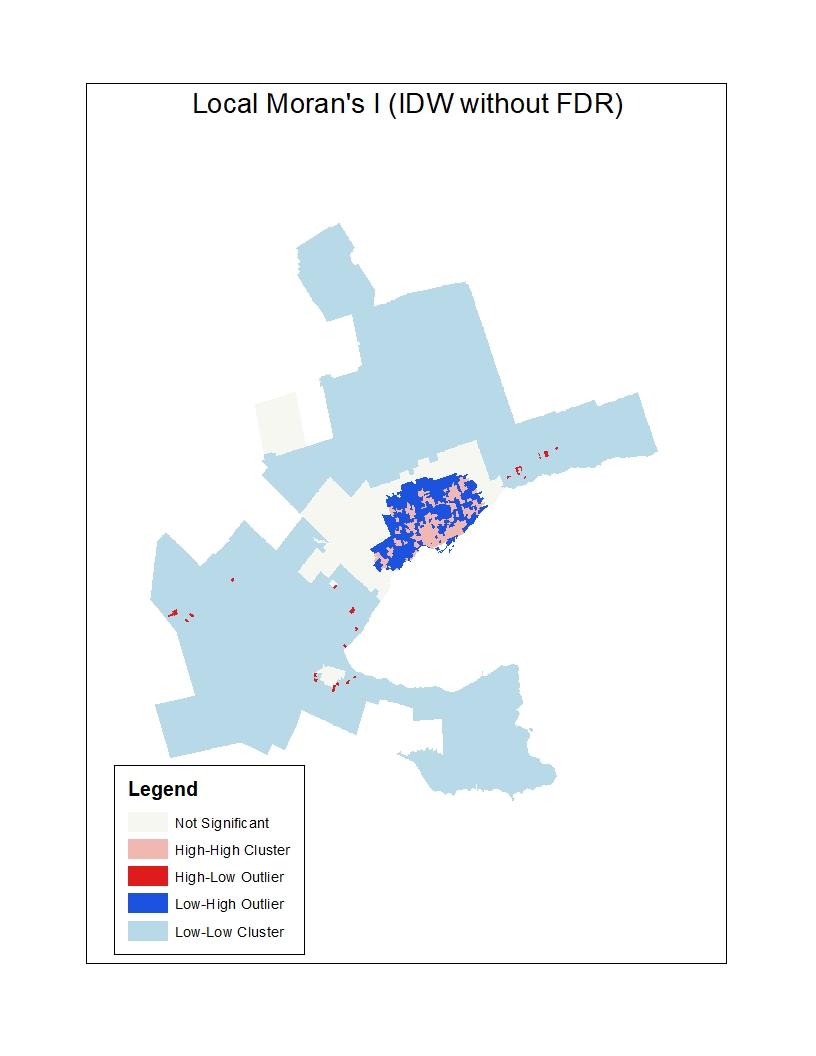
1. **In the output layers, describe what each of the new fields are: LMiIndex, LMiZScore, LMiPValue and COType. How are they related?** (5 points)

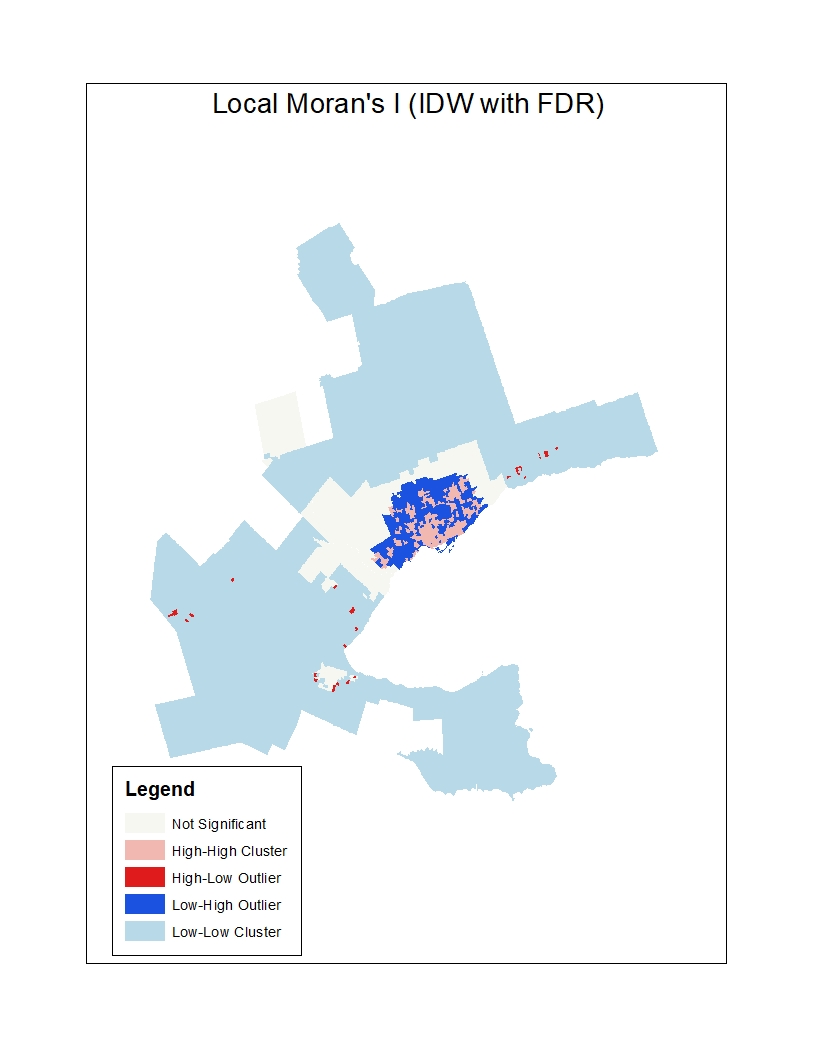
LMiIndex is the Local Moran’s Is. LMiZScore is the Z-Score of the Local Moran’s Is. LMiPValue is the p-value of the Local Moran’s Is. COType is the Cluster-Outlier Type (LL, HH,HL,LH). LMiIndex measure the spatial autocorrelation at a local level. LMiPValue and LMiZScore is calculated from the hypothesis testing of the Local Moran’s Is from the LMiIndex. Cluster-Outlier Type is the comparison of the local moran’s I of a census to the local moran’s Is of the surrounding censuses.

1. **Show the 4 output maps, and compare the results of the computations. Remember to turn off polygon boundaries so that polygon colours can be seen while zoomed out. Summarize relative numbers and locations of each type of spatial cluster.** (5 points)

****

****

****

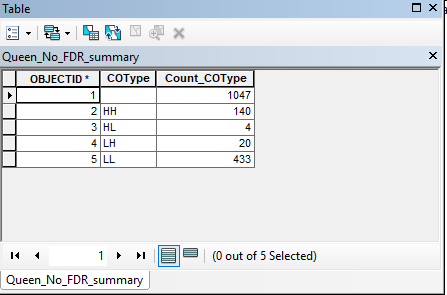
****

Queen: More high-high clustering in Queen (No FDR) than Queen (With FDR) and there is more clustering in Queen (With FDR) that is not significant.

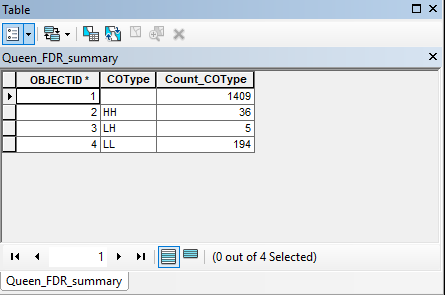
IDW: Slightly more clustering that is not significant in IDW (With FDR) than IDW (No FDR). Otherwise, the distribution of High-High clustering and Low-High clustering is mostly the same in both IDW methods.

Queen vs IDW: Significantly more High-High clustering and Low-High clustering shown in IDW than Queens. Queens (With FDR) has far more clustering that is not significant throughout the Golden Horseshoe than the other three methods.

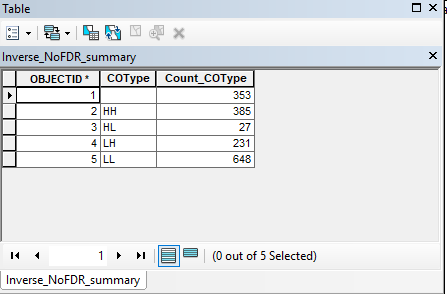
**Queens (No FDR)**

****

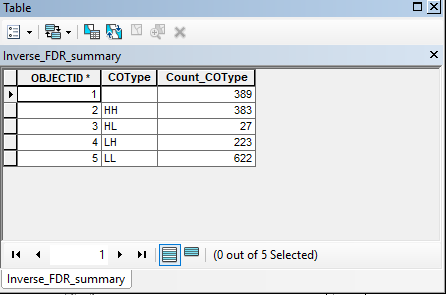
**Queens (With FDR)**

****

**Inverse Distance (No FDR)**

****

**Inverse Distance (With FDR)**

****

1. **What is false discovery rate (FDR) correction? How do the maps with FDR correction compare to those without**? **Why are they different?** (5 points)

**When making Local Moran’s I calculations, we must take into account of overlapping subsets of data. FDR corrects this problematic situation by finding a significance level based from the p-values of the Moran’s Is that does not make our hypothesis testing, too risky or too conservative. The maps with FDR correction tend to have more clustering that is not significant. The reason why they are different is because the maps with FDR correction uses lower significance levels than the maps without FDR correction.**

1. **If =0.05, what would the adjusted and adjusted be for this dataset using Sidek and Bonferroni corrections?** (4 points) **Using these adjustments, how many significant clusters of each type would you find in the queen-based and distance-based Local Moran’s I maps?** (4 points) **How does this compare to the number found without corrections, and the number found using FDR corrections?** (4 points) **(**It would probably be helpful to organize your results into a table, and then discuss the table.)

**Bonferroni Correction**

**Sidek Correction**

=

After adjustments, all the clusters are significant clusters for each type in the queen-based and distance-based Local Moran’s I maps.

Now we are going to work with the statistic. Compute for population density using queen contiguity and FDR Correction. Compute it again using fixed distance bands of 1km, 5km and 10km.

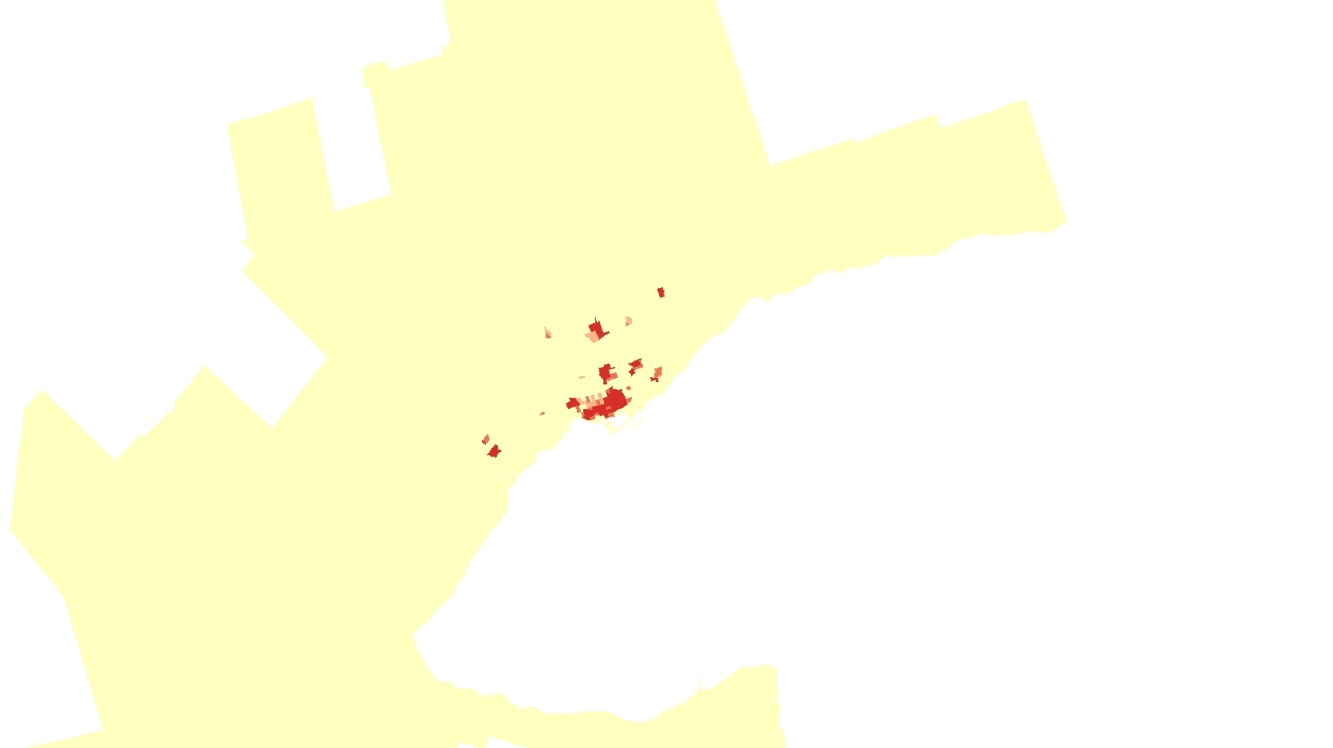
1. **In the output table, what is contained in the fields: GiZScore, GiPValue and Gi\_Bin?** (3 points)

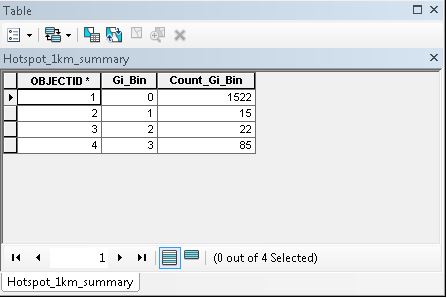
GiZscore is the Z Score of the Getis-Ord Gi Statistic. GiPValue is the p-value for Getis-Ord GiStatistic. Gi\_Bin identifies whether a hot spot or cold spot is statistically significant or not. In addition, for each of the hot spots or cold spots that are statistically significant, the Gi\_Bin will also identify at what confidence level (99%, 95%, 90%) they are each statistically significant at.

1. **Provide copies of the output maps. Remember to change the boundaries so that they are transparent. Describe the differences you see between the 3 maps, and back up your findings with a numerical summary (i.e. a table showing the number of significant hot and cold spots in each map).** (6 points) (Hint: Use the summarize tool on the Gi\_Bin field). **What is the relationship between bandwidth size and the number of hotspots/cold-spots detected?** (2 points) **Thinking about the formulas for how Gi and E(Gi) are calculated, why do you think so many cold-spots appear when using larger bandwidths, even in populated areas? Which spatial pitfall is this related to?** (2 points).

0 means hotspot/coldspot is not significant, 1-3 means hot spot has statistical significance and -1 – -3 means cold spot has statistical significance.

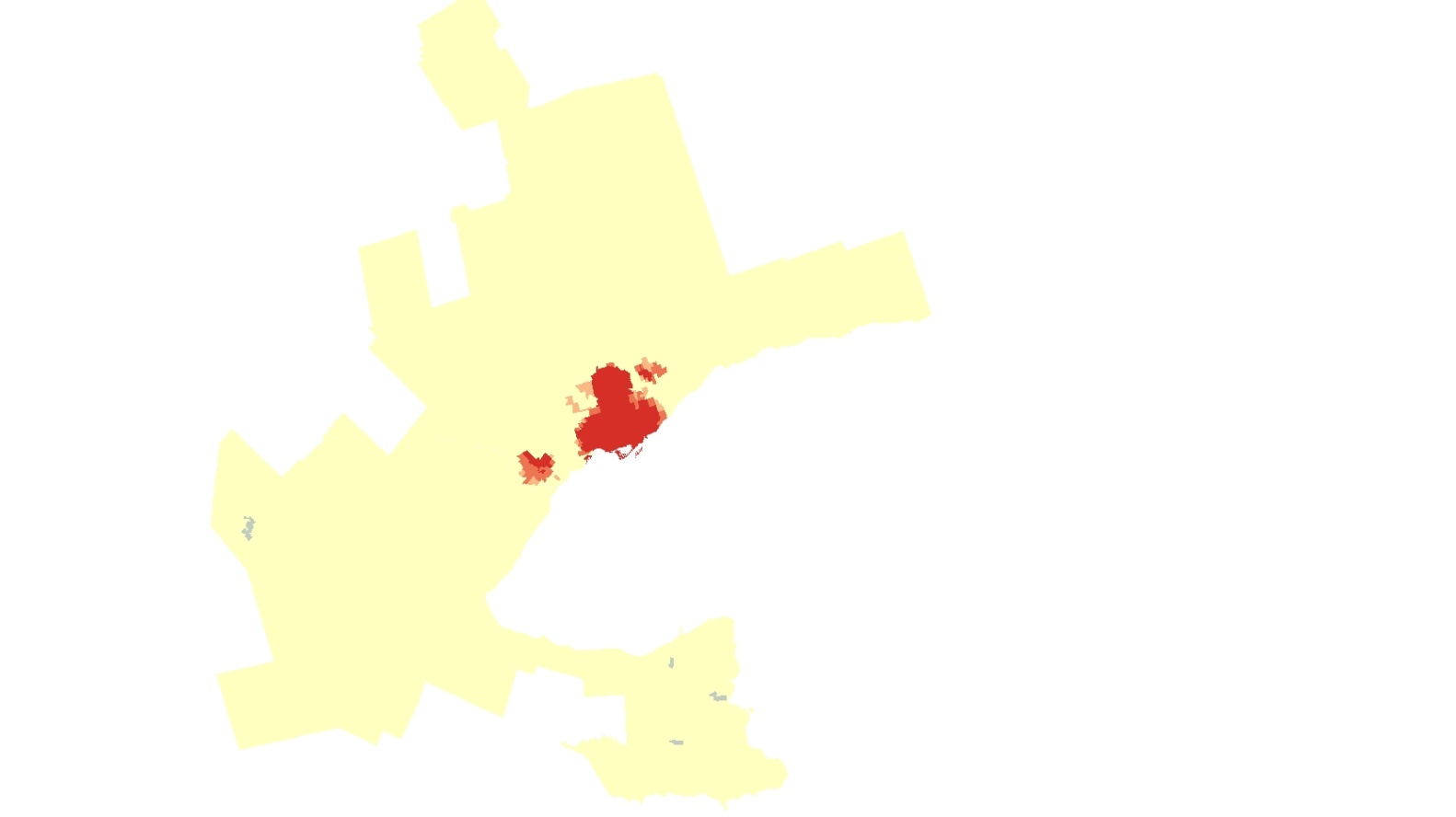
**Gi Statistic with 1 km fixed distance band**

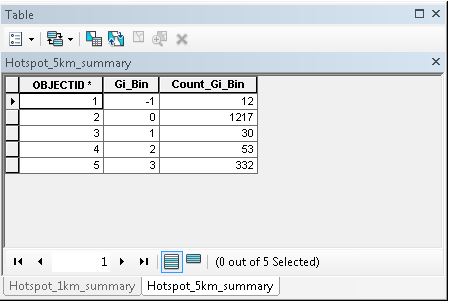




Compared to the larger fixed distance bands, this map has the most hotspots/coldspots that are not significant and has the least hotspots.

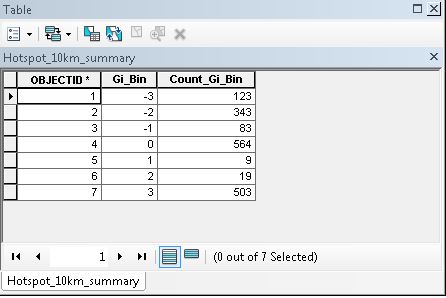
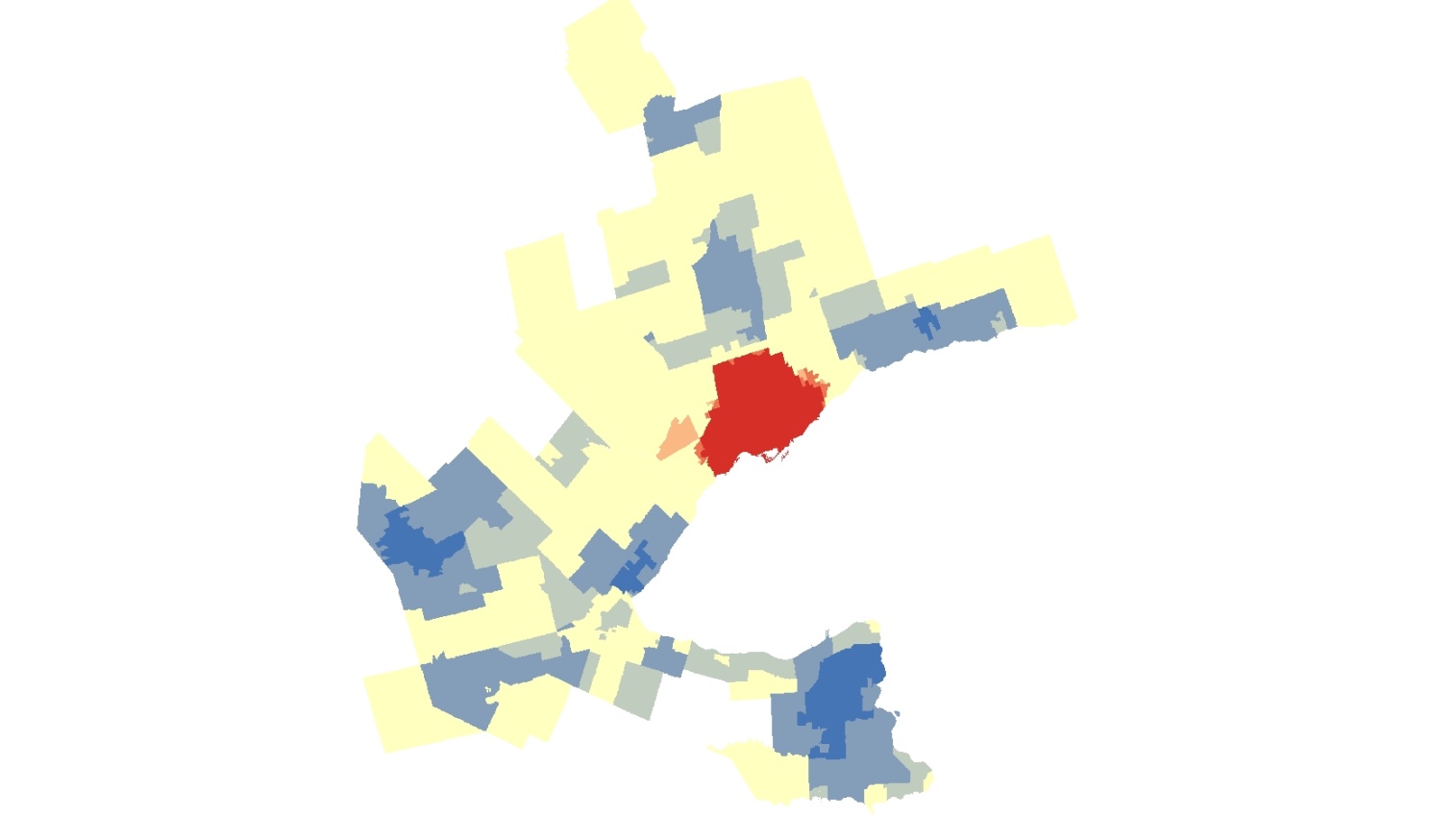
**Gi Statistic with 5 km fixed distance band**





This map has more statistically significant hotspots than the 1 km fixed distance band, and a few cold spots in Southern Ontario. This map has many not significant hotspots/coldspots just like the map with 1 km fixed distance band.

**Gi Statistic with 10 km fixed distance band**



This map has the most spots that are statistically significant hotspots and coldspots compared to the lower fixed distance bands. This map has the least spots that are not significant compared to the other maps.

As bandwidth increases, the number of hotspots and coldspots increases. I think many occurrence of cold spots when using larger bandwidths is because the weight matrix includes more neighbours in Gi statistic. I think the spatial pitfall that this is related to is scale.

Now we are going to compare Local Moran’s I to the statistic. Compute Local Moran’s I using 5km fixed distance bandwidth and FDR Correction. For the sake of comparison, don’t use row standardization in the Moran’s I calculations.

1. **Describe the differences seen between the Mi and Gi maps. What unique information does each statistic provide?** (5 points)

Anywhere there is Hotspots in Toronto in either the maps with 1km fixed distance or 5km fixed distance, the Local Moran’s I classifies as either Low-high cluster or high-high cluster. Anywhere, outside of the Hotspots of the 5 km fixed distance map, the Local Moran’s I map shows plenty of clustering that is not significant, but there are many Low-low clusters throughout the Golden Horseshoe. The Gi statistic calculates the percentage of the total sum of population density found in the neighbourhood of a census tract. The Local Moran’s I calculates the spatial autocorrelation between a census tract and the surrounding census tracts.

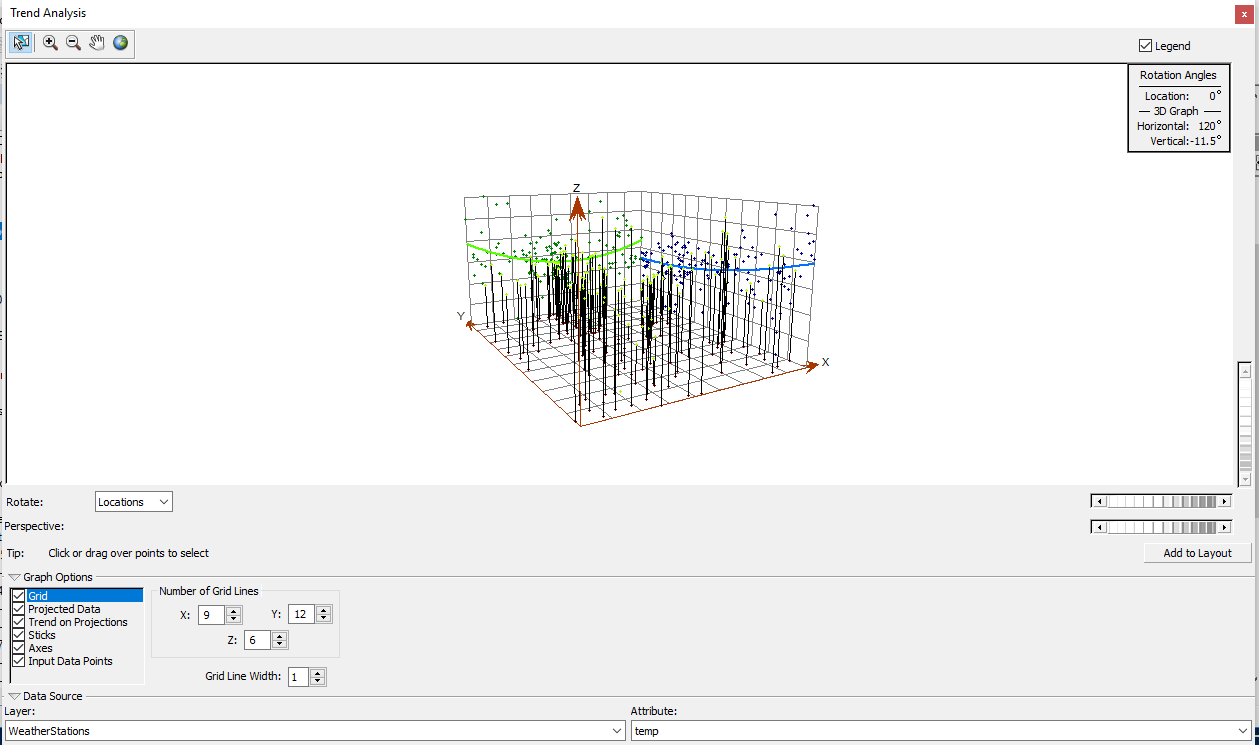
**Part 2: Interpolation**

In this part of the lab, we will be learning about interpolation using global trend surface analysis and inverse distance weighted methods.

Find the Demo folder and complete Tasks 1-3 in the demo instructions.

When completed with the demo, you can repeat these interpolation steps on the Utah temperature dataset provided. Make sure to find the best polynomial degree and IDW weights, and record the Root-mean-square-errors you retrieve.

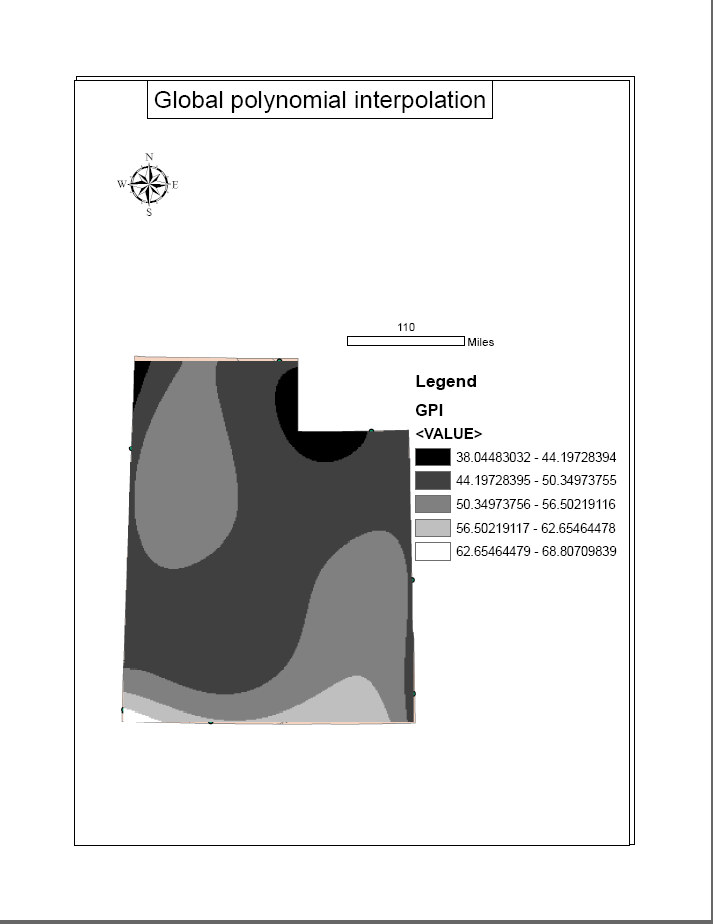
1. **Provide a screenshot of the trend analysis 3D scatterplots and describe the spatial trend in the data.** (5 points)

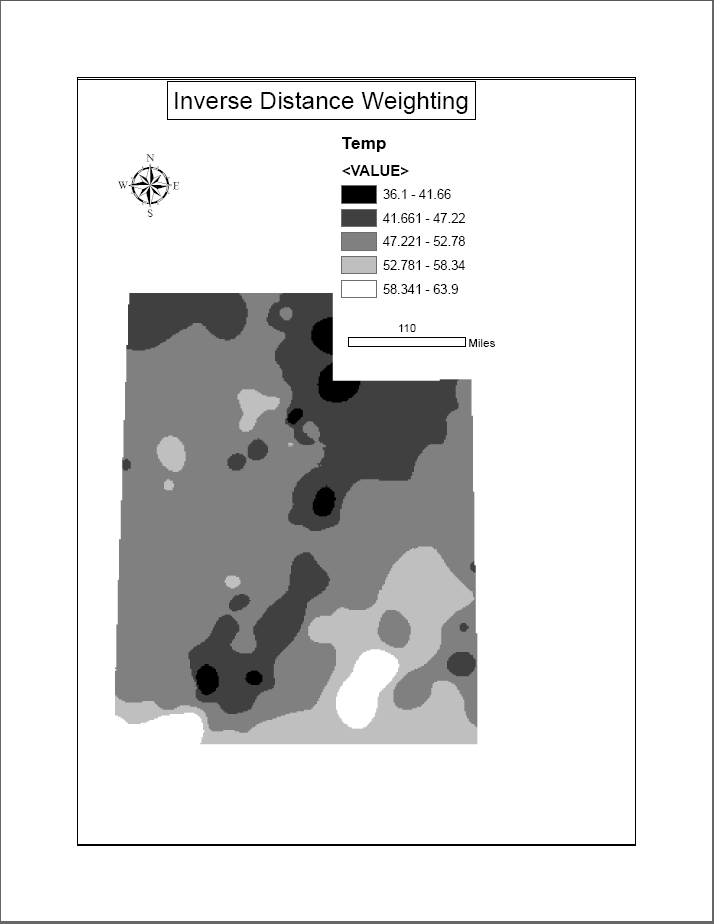


According to the trend projections, the YZ plane dips from the north until it reaches midway and then increases slightly as it keeps going to the south. The XZ plane decreases from the west until it reaches midway and then increases slightly as it keeps going to the east. The strength of the east-west trend is steeper than the north-south trend. However both trends are similar to each other (Both are parabolic), which leads me to believe that the general temperature pattern in Utah is distributed uniformly.

1. **Provide maps for interpolated Utah temperatures using global polynomial and IDW interpolation. Which method is doing a better job of interpolation? Where are the biggest differences between the two methods found? Why?** (10 points)

The root mean square error for IDW for 3.1379047413182692. The root mean square error for global polynomial is 3.816731602162265. IDW does a better job of interpolation than global polynomial, since root mean square for IDW is lower than root mean square for global polynomial. Root mean squared error measures “the standard deviation of the differences between predicted values and observed values.” In global polynomial method, the best order of polynomial is 4. For IDW method, the power is 3.568537249152612. The IDW weights are: 0.99722, 0.00076, 0.00069, 0.00044, 0.00035, 0.00021, 0.00006, 0.00006, 0.00005, 0.00004, 0.00003, 0.00003, 0.00003, 0.00003, 0.00002.



****