October 4th 2016 By: Jordan (Yu-Lin) Wang
TA: Steeve Deschenes V00786970

GEOG 418 Lab2- Interpolations and Trend Surface Analysis

Introduction:

Interpolation is a process that manipulates spatial information in order to derive new information and meaning from the original data. It processes points with known values to estimate values at other unknown points (QGIS, 2016). In this lab assignment, we are using ArcGIS tools to create an interpolated surface using the IDW of the average incoming for Calgary city in 2010 (Figure 1). Inverse Distance Weighted (IDW) is an analysis that assumes that each coordinate, (X, Y), and surface value, (Z), has a local influence that reduces with distance. The weight is a function of inverse distance. Points that are closer to the study area have more weight than those farther away (Badish.G, 2006). Trend surface analysis (TSA) can be used to filter out large scale spatial trends to help focus on the smaller scale residuals. It is also be used for fitting a surface to the sample points when the surface varies gradually from one region to another; in this case, we are interested in the average income throughout the city (ArcGIS, 2007). The IDW, TSA, and Ordinary Least Square (OLS) are used together in this assignment to interpolate the overall average income of the City of Calgary based on the provided data. OLS is a linear regression used to generate prediction or to model the relationships for a set of variables (ArcGIS, 2013). In order to determine if the model is significant or not, we used the ANOVA F statistic test to determine the results.

The data sets provided for this lab are in a Geo Database format that contains two data sets for the City of Calgary. The Census Tract centers (point vector) and the City of Calgary Boundary (polygon vector) are included in the Feature Classes. The Census Tracts are provided in a polygon file that contains the area regions of the City of Calgary. The CTCentres point data provide crucial information such as the average income, Northings (X), and Easting (Y). All the data sets provided are projected in the same UTM spatial projection of NAD_1983_3TM_114.

Goal:

The goal of this assignment is to learn how to interpolate a surface using ArcGIS to perform a trend surface analysis and to evaluate model improvement.

Methods:

The lab process begins by importing the two data sets provided for the City of Calgary; the Census Tract Centers (CTCentres) point vector file and the City of Calgary Boundary polygon vector file, these two files help derive crucial information required for this assignment. In the attribute table for the CTCentres, we are required to manually add four new data field columns. They are consisted of X^2, Y^2, XY, and ID; the ID column's data are copied from the ObjectID column. The values of X^2, Y^2, and XY are calculated by using the "Field Calculator" tool and by selecting the appropriate calculating fields [For X^2: multiply column x with column x. For Y^2: multiply column y with column y. For XY: multiply column x with column y].

Inverse Distance Weighting (IDW) (Figure 2)

The Inverse Distance Weighting is a Spatial Statistic tool that is provided on the ArcMap, it interpolates a raster surface from points using a linearly weighted combination technique. In this case, the input point features for the CTCentres file are: the Input Z value is the Average Income, the Output cell size is 89.116m^2 (default setting) and the power selected is 2. The output value for a cell using IDW is limited to the range of the values used to interpolate. The output IDW value of pixel range min is 20391.5957m and the max is 114720.2109m, meaning that the average result cannot exceed this range.

Trend Surface Analysis

After completing the performance of IDW, the next process is to perform a Global Polynomial Interpolation from the Spatial Statistics Tool to determine the trend surface of average income. Within this tool, the selected input file is the CTCentres with the Z value of Average Income and the generated output is a raster file. We are required to complete this process twice with different inputs for the Order of Polynomial (1st order and 2nd order) to understand the relationship between the different variables (Figure 3 & 4). The generated output raster files are then classified into 10 groups that show the estimating average income area based on our inputs.

Ordinary Least Square (OLS)

Next, we used the Ordinary Least Square tool to analyze the Average Income to determine the most suitable model for this assignment. The input value chosen is the CTCentre with the Unique ID of ID field (generated column). The dependent variable is the Average Income and it determines the relationships to the Explanatory Variable [X, Y, X^2, Y^2, and XY]. The output results are displayed in point vector files and the relationship results can be found at the pop up messages. The values shown in these pop up messages are then copied onto an excel file. The first OLS being performed has the explanatory variables of only X and Y to test the regression relationship of the average income. The second OLS performed has the explanatory variables of X, Y, X^2, Y^2, and XY. Lastly, different combinations of explanatory variables were selected and compared to determine which model is the most suitable representation of the relationship. After running multiple models and comparing their results, the Ordinary Least Squares with explanatory variables of X^2, Y^2, and XY derived the best suitable results (Figure 5).

Following this, we can generate multiple statistic report for the models by using the Regression tool in the Data Analysis tool. Then we can also compare these statistics values to determine the best model. Another tool used in Excel is the Correlation tool, it takes all the Explanatory Variables (X, Y, X^2, Y^2, and XY) as inputs and output a correlation chart consisting of each variable's relationship. After the Excel processes, we return to ArcMap to derive our final residual output map for City of Calgary. This is done by spatially joining the attribute table of CTCentres layer and the best suitable OLS models based on the spatial location and selecting that each point will be given all the attributes of the point in the layer being joined that is closest to it. The derived output is the residual maps of the best suitable OLS model.

Last of all, we performed the spatial autocorrelation analysis from the ArcGIS Spatial Statistics Tools. The input feature class is the best OLS model and with the input field of the residual. This generated an output that presents the Spatial Autocorrelation Report (Figure 6).

Results:

As mentioned in the introduction, the points that are closer to the study area have more influence than those points further away. The generated IDW (Figure 2) presents the highest average income spread out at the middle of the city from Northwest to Southeast. Based on the IDW result, most of the high average income areas are towards the Northwest direction of the city and the Northeast direction tend to have a lower average income. This information will be further discussed in the discussion section.

Global Polynomial Interpolation (GPI) fits a smooth surface that is defined by a mathematical function to the input sample points. The Global Polynomial surface changes gradually and captures the coarse-scale pattern in the data (ArcGIS. 2007). Based on our results, the generated GPI results are not able to present the changes of the Average Income as well as the IDW result. The first and second orders of polynomials are used to generate the GPI surface for the City of Calgary. However, these generated GPI surfaces does not represent the average income relationship as well as the IDW. The first order polynomial generated an output that classified the study area into groups of linear polygon (Figure 3). The second order polynomial generated an output that classified the study area into groups of curve linear polygons (Figure 4). In comparison, the second order polynomial is a closer representation than the first order polynomial. This is due to the second order presents a higher average income region from Northwest direction to Southeast, very similar to the results of the IDW.

The three OLS models output are listed in the Table below (Table 1, 2, & 3). These three models are consist of the OLS 1st order polynomial, OLS 2nd order polynomial, and the best model. The Multiple R-Squared is a statistical measure of how well the model fits the observed dependent variable values. The value of R-squared ranges from 0 to 100 percent, and the higher the percentage indicates that there are more regression exists between the variables. As observed in the Table section, the OLS 2nd model has the highest R-squared values in comparisons to the other two models. However, this model contains several variables that has VIF values larger than a thousand, not a good representation. Variance Inflation Factor (VIF) measures the redundancy among explanatory variables (ArcMap. 2016). If the VIF value exceeds 7.5, than the variables should not be used in the next model run. Using these outputs provided in the Table section, we can compare the results and derived with the conclusion that the regression model of X^2, Y^2, and XY are the best model to represent the relationship between the variables and the average income. Then we combined the best model with the CTCentres attribute table to generate the best suitable residual map for this assignment (Figure 5).

Next, we used the best suitable regression model to run the Spatial Autocorrelation tool. The generated report tells us several crucial information such as the Moran's I, Z-score and the P-value of 0.002425 (Figure 6). The Moran's I value is 0.079277, and will be further discussed in the discussion section. The Z-score is 3.032561, and this indicates that there is a less than 1% likelihood that the clustered pattern is the result of random change. Lastly, we used all the variables of X, Y, X2, Y2, and XY to generate a correlation chart to determine whether or not the variables are correlated (Table 4).

Discussion and Conclusion:

Although IDW are a populate interpolation approaches, they are not perfect. IDW does not work well with phenomena that has complex variables because IDW can only account for the effects of distance. When the sample points are dense and evenly distributed, the IDW generated will be the best representation. However, the IDW can be easily influenced by the uneven distributed data points because equal weight is assigned to each data points even if the data are clustered (Babish, G. 2006). The main problems associated with IDW are: does not generate assessment of the prediction errors and Bulls-eye effect. IDW has a tendency to generate patterns of concentric contours around actual data points (Bull's-eye effect) and also tends to generate flat areas/average out trends. This can often mess with the output of the final elevation surface, such as flattening peaks and valley (Babish, G. 2006).

October 4th 2016 By: Jordan (Yu-Lin) Wang
TA: Steeve Deschenes V00786970

The trend surface analysis (TSA) we conducted in this assignment is in a local scale that estimate surface using only a selection of nearest data points. In the assignment, we used the CTCentres points provided for the City of Calgary to estimate the entire surface of the city. Both of the 1st and 2nd degree polynomial functions were interpolated. Ideally, the larger the degree of polynomial functions used, the output will present a more detailed interpolated surface for the average income. Since only two different degrees of polynomial functions were run, the closest representation will be the 2nd degree polynomial interpolation. The IDW and 2nd degree polynomial presents a higher average income towards the center of map which matches with the real world Calgary map. The higher average income tends to cluster towards downtown area of the City of Calgary.Based on our results generated using the Ordinary Least Square, we can compare the outputs to determine the best suitable outcome. The OLS outcome for the 1st order (X, Y) has a high R-squared value of 24.78% and a low VIF value. Ideally, this should be the best model. However, it only contains two explanatory variables (too little to be good). The OLS result for the 2nd order (X, Y, X^2, Y^2, and XY) has a high R-squared value of 26.82% and also has VIF value larger than 1000. Therefore, not a good model. Lastly, the OLS output for the 2nd order (X^2, Y^2, and XY) has an R-square value of 24.79% and the VIF values are lower than 7.5. Therefore, the OLS 2nd order (X^2, Y^2, and XY) is the best representation.

After, we run the spatial autocorrelation test on the best mode (X^2, Y^2, and XY) the generated are: Z-score of 3.3032561, Moran's I of 0.079277, and P-value of 0.002425. If the Z-score falls outside the range of -1.96 and +1.96, the observed spatial pattern is unusual to be result of random chance, and p-value will be small. In this case, it is possible to reject the null hypothesis (ArcGIS Pro. 2016). Since our Z-score is above +1.96 and the P-value is less than 0.05, therefore, we can reject our null hypothesis because the result is not likely to be the result of random change. If the Moran's I statistic is close to 1, tend to cluster in geographic space. Therefore, the Moran's I of 0.079277 demonstrates that there are clustering happening. Lastly, this result of the best model (X^2, Y^2, and XY) are combined together with CTCentres to generate the best suitable residual result. The output shows an over and under estimation of average income at some locations. For example, there are a lot of under estimation occurring towards the center of the city.

In this assignment, several spatial interpolation and trend surface analysis were performed to derive the best suitable residual maps for the average income of City of Calgary. Tools such as IDW, OLS and TSA were used along the steps to derive the result. Based on our table, we can see that the best suitable model is the 2nd order OLS (X^2, Y^2, and XY). It shows that there are clustering occurring at the study area and based on the Z/P- value we can reject the null hypothesis. In conclusion, due to the nature of our data the results we derived is still not a great representation of the real world. However, based on the given data, the residual map for the OLS 2nd (X^2, Y^2, and XY) joined to CTCentres presents the best suitable result.

References:

ArcGIS (August, 2007) How Global Polynomial interpolation works. ArcGIS 9.2 Desktop Help. Retrieved on October 4th 2016, from http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=How_Glo bal_Polynomial_interpolation_works

ArcGIS (2013) Ordinary Least Squares (OLS) (Spatial Statistics). ArcGIS Help 10.1. Retrieved on October 4th 2016, from http://resources.arcgis.com/en/help/main/10.1/index.html#//005p00000022000000

ArcGIS (May, 2007) Performing trend analysis. ArcGIS 9.2 Desktop Help. Retrieved on October 4th 2016, from http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Performing_trend_analysis

ArcGIS Pro (2016) What is a z-score? What is a p-value? Environment Systems Research Institute. Retrieved on October 4th 2016, from http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

ArcMap (2016) Interpreting OLS result. Environment Systems Research Institute. Retrieved on October 4th 2016, from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/interpreting-ols-results.htm

Babish. G (July 2006) Geostatistics Without Tears. Environment Canada. Retrieved on October 4th 2016, from www.ibrarian.net/navon/paper/Geostatistics_Without_Tears.pdf?paperid=15417085

QGIS (2016) Spatial Analysis (Interpolation). Documentation for QGIS 2.2. Retrieved on October 4th 2016, from https://docs.qgis.org/2.2/en/docs/gentle_gis_introduction/spatial_analysis_interpolatio n.html

Figures:

October 4th 2016 By: Jordan (Yu-Lin) Wang
TA: Steeve Deschenes V00786970

The City of Calgary Study Area

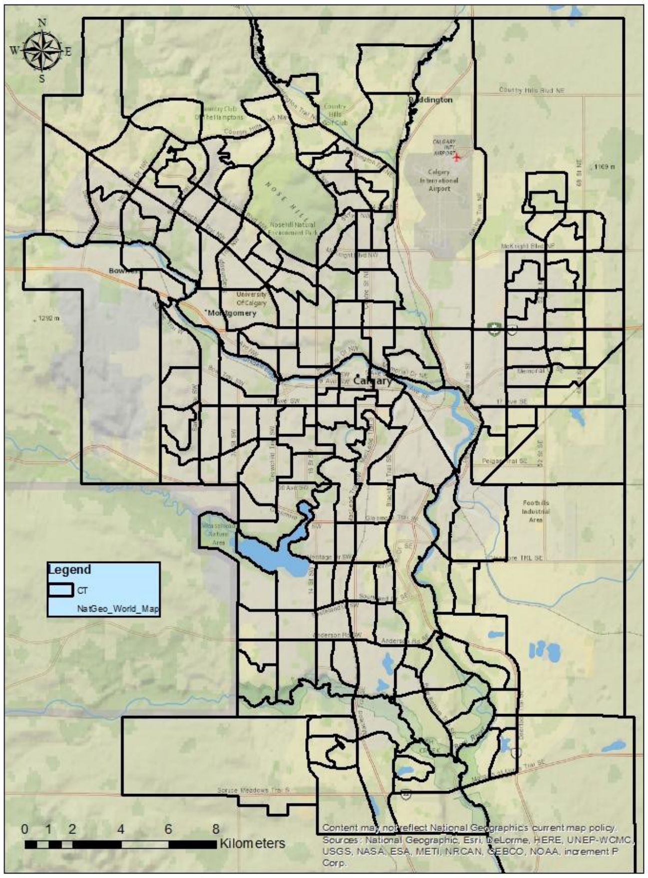


Figure 1. Study Map of The City of Calgary

October 4th 2016
TA: Steeve Deschenes

By: Jordan (Yu-Lin) Wang
V00786970

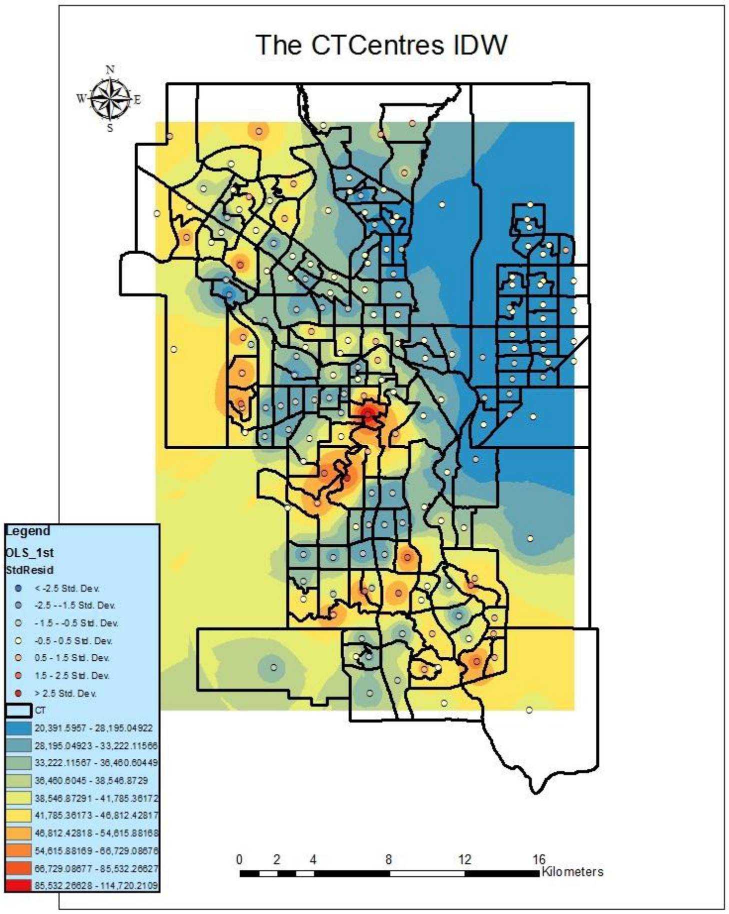


Figure 2. The CTCentres Inverse Distance Weight

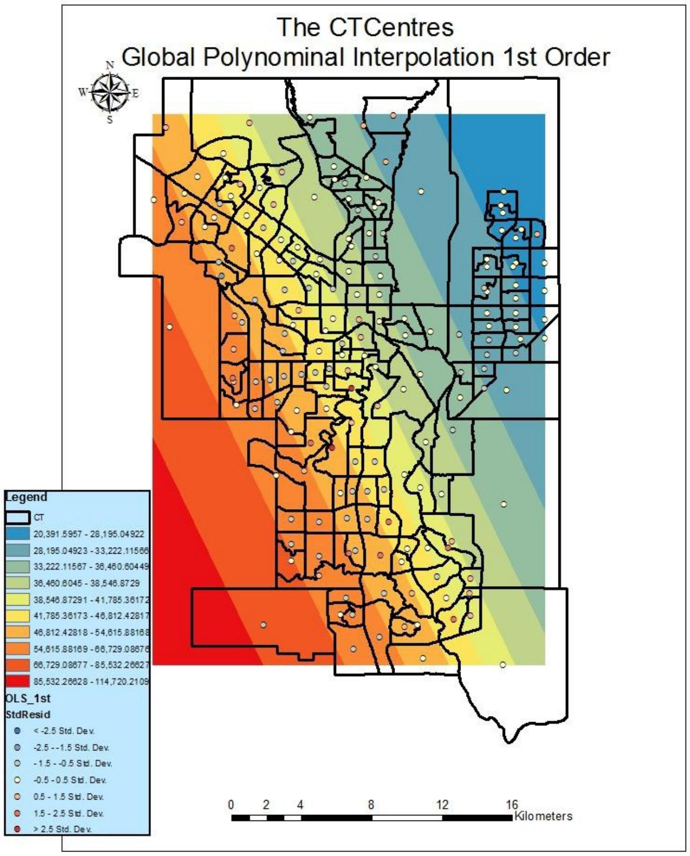


Figure 3. Trend surface Global Polynomial Interpolation 1st Order

By: Jordan (Yu-Lin) Wang
V00786970

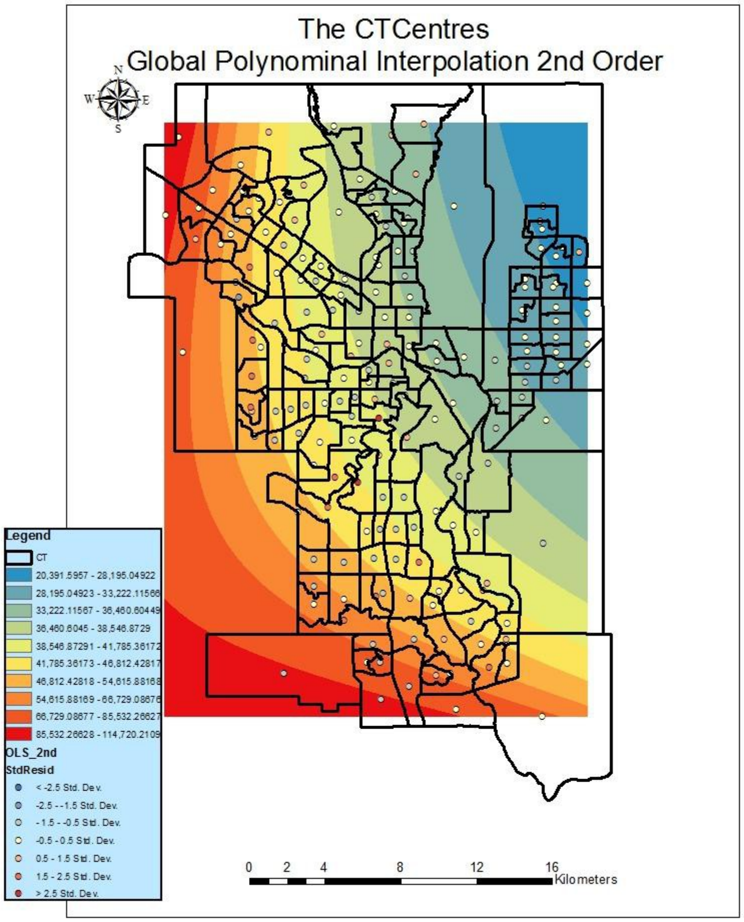


Figure 4. Trend surface Global Polynomial Interpolation 2nd Order

By: Jordan (Yu-Lin) Wang
V00786970

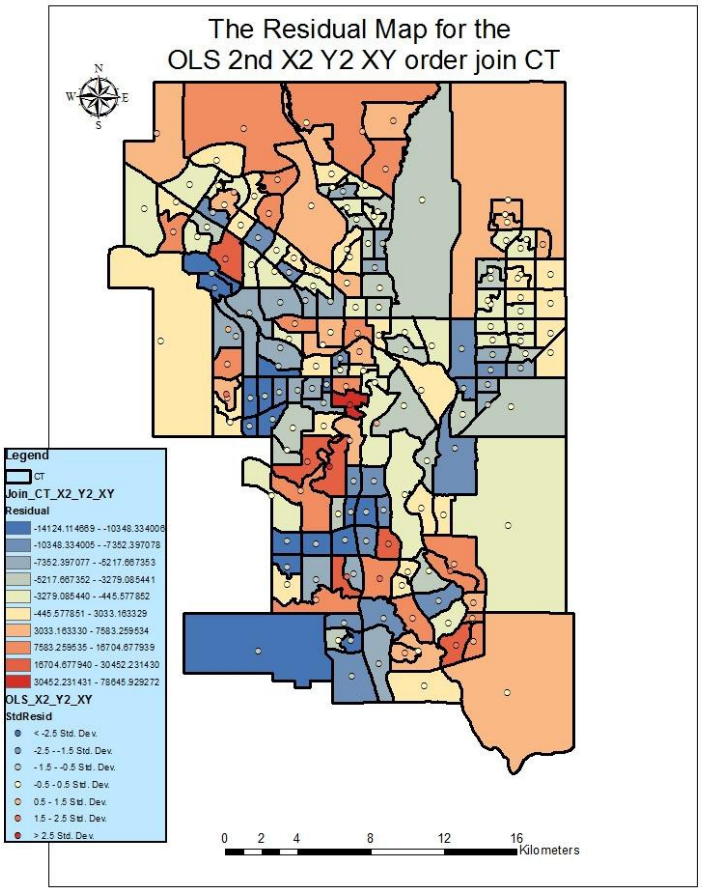
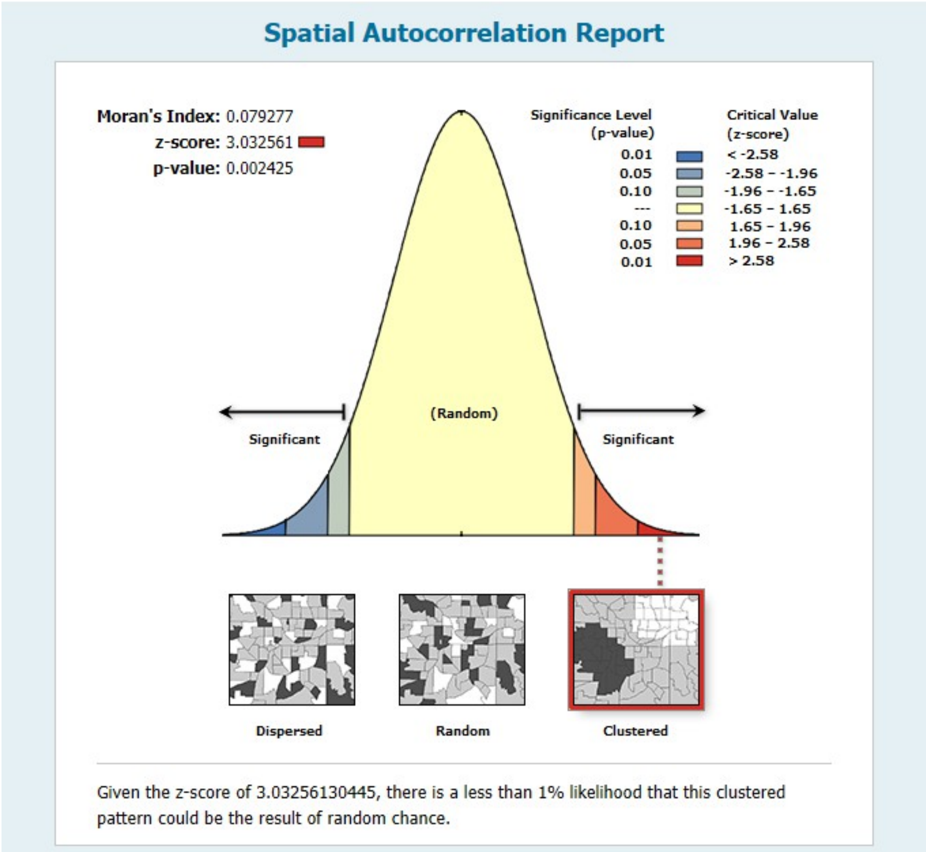


Figure 5. The final Residual Map for the OLS 2nd order (X^2, Y^2, and XY) join with CTCentres

October 4th 2016
TA: Steeve Deschenes
By: Jordan (Yu-Lin) Wang
V00786970



Global Moran's I Summary

Moran's Index:	0.079277
Expected Index:	-0.005556
Variance:	0.000783
z-score:	3.032561
p-value:	0.002425

Figure 6. Spatial Autocorrelation Report

October 4th 2016
TA: Steeve Deschenes

By: Jordan (Yu-Lin) Wang
V00786970

Tables:

Table 1. OLS1_X_Y

Variable	Coefficient	StdError	T-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF
Intercept	2651682.226994	543703.689277	4.877072	0.000003	472761.907238	5.608917	0.00000	--------
X	-0.956389	0.141043	-6.780826	0.000000	0.096756	-9.884516	0.00000	1.040160
Y	-0.463410	0.096154	-4.819437	0.000004	0.083536	-5.547401	0.00000	1.040160
Multiple R-Squared	0.247763							
Joint F-Statistic	29.313737							

Table 2. OLS2_X_Y_X2_Y2_XY

Variable	Coefficient	StdError	T-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF
Intercept	308025722.679688	406707123.179123	0.757365	0.449842	393058532.736614	0.783664	0.434288	------------
X	213.841900	132.713504	1.611305	0.108926	100.855095	2.120289	0.035383	>1000.0
Y	-108.318312	143.901873	-0.752723	0.452620	139.041594	-0.779035	0.437003	>1000.0
X2	-0.000009	0.000026	-0.354149	0.723664	0.000020	-0.443384	0.658046	4.314213
Y2	0.000010	0.000013	0.748163	0.455359	0.000012	0.774368	0.439750	>1000.0
XY	-0.000038	0.000023	-1.618295	0.107410	0.000018	-2.129525	0.034600	>1000.0
Multiple R-Squared	0.268205							
Joint F-Statistic	12.827618							

Table 3. OLS2_X2_Y2_XY

Variable	Coefficient	StdError	T-Statistic	Probability	Robust_SE	Robust_t	Robust_Pr	VIF
Intercept	1350443.260659	290491.290346	4.648825	0.000008	237874.204692	5.677132	0.00000	------------
X2	0.000002	0.000025	0.067890	0.945940	0.000019	0.087534	0.930336	4.05609

								3
Y2	-0.000000	0.000000	-4.540072	0.000013	0.000000	-5.551421	0.000000	1.181546
XY	-0.000000	0.000000	-3.523577	0.000552	0.000000	-4.963405	0.000002	3.713284
Multiple R-Squared	0.247875							
Joint F-Statistic	19.444443							

Table 4. Correlation

	X	Y	X2	Y2	XY
X	1				
Y	-0.19649	1			
X2	-0.84823	0.346789	1		
Y2	-0.19656	0.999999975	0.346847	1	
XY	0.99999	-0.1976473	-0.84858	-0.19771	1