GEOG 418 Lab 4: Spatial Regression and Geographically Weighted Regression

Introduction and Data:

The Canadian Arctic Archipelago is a group of islands, north of the Canadian mainland, and covering about 1,424,500 km^2. It is a group of 36,563 islands in the Arctic sea that comprises much of Northern Canada (Figure 1). This region is showing some effects of global warming and is estimated that the melting will contribute +2.4cm to the global sea level (Wayman, 2103). With that said, the monitoring of sea ice and glaciers melt duration is an important process to maintain the sea level budget (Sharp, 2011). This can be done through analyzing the relationship between the variables provided and to develop both linear and spatial models for understanding the melt duration (MD) in the Canadian Arctic Archipelago. Linear regression is an approach for modeling the relationship between dependent variable and explanatory variables. The two linear regression tools that will be interpolated in this lab are Ordinary Least Square (OLS) and Geographically Weighted Regression (GWR). On the other hand, spatial regression methods were developed to manage spatial data and to improve the model data relationship. The Exploratory Regression is a data mining tool that is used to evaluate all possible combinations of the input candidate explanatory variables to see which models pass all the necessary diagnostics. Through this process, users can greatly increase the chance of finding the variables that best explains the model (Braun, 2011). The GWR tool is a local form of linear regression that is used to model spatially varying relationships with a dependent variable and a set of explanatory variables (Charlton, 2009).

The dataset provided for this lab is a MD vector point file with the spatial projection of CIS_Lambert_NAD27. This MD shapefile contains several climate attributes for the Canadian Arctic Archipelago. The attribute table consists of Row Identifier (ID), Albedo, Surface Air Temperature (TS), Multi- Year Ice Concentration (MYI), Total Ice Concentration (NCT), MD, Net All- Wave Radiation (Qnet), Longitude (X), and Latitude (Y). Lastly, the study area layer was retrieved from the Lab 1 folder and has the same spatial projection as the MD file.

Goal:

The goal of this lab assignment is to explore the process of model selection, spatial regression, GWR and to compare and contrast the linear, spatial regression and GWR as spatial analysis techniques.

Methods:

Correlation Matrix and Exploratory Regression

To being the processes for this lab, users first need to analyze and understand the file/variables that are being interpolated. This can be done using two different methods: Correlation Coefficient Matrix and Exploratory Regression. The Correlation Coefficient Matrix chart was derived from exporting the given MD data file to Excel and by performing correlation

data analysis on the seven variables (TS, MYI, NCT, MD, Qnet, X, and Y). Through this chart we can gain a better understand the relationship between the data variables. Next, users can perform the “Exploratory Regression” tool from the “Spatial Statistics Tools” on the given MD vector point file to further evaluate all possible combination of the candidate variables. After completion of these two methods, users can derive the best suitable variables from analyzing these results. In this case, the most suitable variables are the Surface Air Temperature (TS) and the Total Ice Concentration (NCT).

Linear Regression: Ordinary Least Square & Geographically Weighted Regression

The results from Ordinary Least Square model are only trustworthy if the explanatory variables satisfy the assumptions required by the task. Therefore, once the suitable variables are determined, users can employed the OLS tool to model the relationship between the dependent variable (MD) and the explanatory variables (TS and NCT). The derived results of OLS are then used to understand the correlation of the given variables and to analyze the coefficient information: estimated value, observed value and the residuals. Afterward, the TS and NCT variables are once again used as the explanatory variables and MD as the dependent variable for the performance of GWR tool found in the Spatial Statistics Tools. The resulting outputs of GWR are used to examine the predicting values (Predicted, Residuals, Coefficient, and Coefficient Standard Error) or understanding the factors that contribute to dependent variable outcomes.

Spatial Interpolation: Ordinary Kriging & Spatial Autocorrelation: Moran’s I & Histogram

The resulting coefficient information from OLS and the prediction values from GWR can be employed to perform the ordinary kriging to map a parameter surface across the study area. The ordinary kriging is performed for the OLS predicted values, observed values, and the residual values to derive different surfaces. Following this step, the ordinary kriging are also performed for the GWR predicted values, residual values, coefficient and the coefficient standard error for TS and NCT. These derived ordinary kriging layers can be joined with the OLS and GWR vector points as the background surface for an easier visual analyzation and prediction. The spatial autocorrelation Moran’s I were performed on the OLS residual layer, GWR residual layer, and the GWR coefficient standard error for TS and NCT layers. The output of this performance can be used to evaluate the characteristic of the pattern and the trend. Lastly, the OLS and GWR residual layers are used to generate the Histograms.

Results:

The exploratory regression allows user to define suitable variables that can be used to interpolate the best model (Figure 2). Due to seven variables contain in the given MD file, the regression tool derived seven different models for user to select. The adjusted R-squared is a statistical measure of how well the model fits the observed dependent variable values from 0 to

100 percent. The higher the percentage indicates that there is more correlation between the variables. Thus, the best case of the adjusted R-squared value is generated from the TS and NCT variables, with an adjusted R-squared value of 0.65. Variance Inflation Factor (VIF) measures the redundancy among explanatory variables. If the VIF value exceeds 7.5, than the variables should not be acknowledged for further model processing. Resultantly, the TS and NCT model has a lower VIF value of 1.84 than most models.

The correlation matrix is used to investigate the dependence between multiple variables at the same time. The result is a table containing the correlation coefficients between each explanatory variable and can also be used to determine the best suitable variables (Figure 3). The closer the value of coefficient is to 1, the closer the relationship between the two variables is. With that said, the correlation coefficient between TS and NCT is -0.67517; meaning there is a close relationship between the variables and can be chosen to create the most suitable model. Moran's I is a way of measuring degrees of spatial autocorrelation and was interpolated for the OLS residual (Figure 4), GWR residual (Figure 5), and GWR standard error for TS (Figure 6) and NCT (Figure 7). The values can be observed below in the table.

Table 1 Spatial Autocorrelation Reports

Models:	Moran's Index	P-value	Z-score	Patterns
OLS residual	0.353549	0.0000	24.715286	clustered
GWR residual	0.275830	0.0000	19.313064	clustered
GWR standard error for TS	1.048920	0.0000	73.163419	clustered
GWR standard error for NCT	1.093691	0.0000	76.394777	clustered

The OLS and GWR tools were performed with the Surface Air Temperature variable (TS) and the Total Ice Concentration value (NCT). This was due to its nature of high R-squared value and its low VIF value. From the Observed OLS, the user can observe a significant melting towards Southwest/east and lower melt duration as it moves towards North (Figure 8). The Predicted OLS also presents a similar trend as the Observed OLS (Figure 9). However, the Residual OLS model indicates large residuals towards Southwest/east and at the center of the Canadian Archipelago (Figure 10). This indicates the model has a poor fit and should be taken into consideration. The Predicted GWR presents high melt duration pattern from Southwest to Southeast, which is almost identical to the OLS predicted model (Figure 11). The Residual GWR also presents a very alike residual trends as the Residual OLS model (Figure 12). Next, the GWR Coefficient for TS shows a significant clustering of high value towards the Southwest direction (Figure 13). On the other hand, the GWR Coefficient for NCT has the significant

clustering at the Southeast direction (Figure 14). Lastly, the GWR Standard Error Coefficient for TS (Figure 15) and NCT (Figure 16) presents a similar trend of low values towards the center of the study region.

Histograms are graphical representation of the distribution of numerical data, it is an estimate of the probability distribution of the continuous variables. The histogram for the OLS residual (Figure 17) is relative similar to the histogram for the GWR residual (Figure 18), both appeared to be symmetric. However, the GWR residual histogram indicates a more skewed left than the OLS residual.

Discussion and Conclusion:

Finding a properly specified regression model to answer a specific question can be challenging, especially when there are lots of potential explanatory variables that might be contributing factors to the model. Dependent variable is the variable representing the process that user are trying to predict or understand. In this case, the dependent variable is the Melt Duration. Explanatory variables are variables used to model or help explain the dependent variable values. As briefly mentioned above, to create a reliable OLS and GWR model, users needs to derive explanatory variables that satisfy all the assumptions required by the tasks. The suitable explanatory variables can be chosen from the exploratory regression by analyzing the variables' adjusted R-squared and the VIF values. The number of selected explanatory variables for the model should be as low as possible, as long as the R-squared and VIF values represent the same quality as the models with many variables. With that said, the model consists of TS and NCT is chosen as the best due to similar high adjusted R-squared (65%) and VIF (1.84) values as the model that contains three variables (adjusted R-squared: 67% and VIF 1.92). The OLS and GWR tools are both linear methods; consequently, if the relationship between explanatory variables and dependent variable is nonlinear the result model will perform poorly. Thus, it is important to select the proper explanatory variables that will derive a trustworthy model. Therefore, the selected TS and NCT variables indicates this model explained 65% of the model performance, which is considered as a trustworthy model.

The global Ordinary Least Squares (OLS) linear regression is used to generate predictions or to model a dependent variable in terms of its relationships to a set of explanatory variables (Chumney, 2005). The Geographically Weighted Regression is a local form of linear regression used to model spatially varying relationships (Charlton, 2009). A global interpolator derives a surface model using all available data provided for the study area by applying single function across the entire region. The resulting global output gives a best fit for the entire sample data, but may provide poor fit at some locations (Babish, 2006). In comparison, a local interpolator calculates prediction from the measured points within the neighborhoods to ensure that interpolated values are determined only by nearby points. Local interpolation applies an algorithm to small portion at a time (Babish, 2006).

After the performance of spatial autocorrelation test on the OLS residual, GWR residual, and GWR coefficient standard error for TS and NCT, the result indicates a clustering pattern for the study area. If the Moran's I statistic is close to 1, it indicates clustering in the geographic space (Babish, 2006). Table 1 in the result section shows that the clustering is true due to all values are close to 1. The GWR coefficient standard error for the surface air temperature indicate a high temperature at the Southwest direction and the GWR standard error for total ice concentration indicates the opposite. This phenomena of high air temperature at Southwest direction can be influenced by the latitude and ocean currents. The lower the latitude indicates a higher temperature due to closer to the equator and more sunlight receives. This can derive the assumption that the higher surface air temperature at Southeast direction has a correlation with the total ice concentration (Hund, 2014). The higher temperature will lead to higher melt duration rate, as observed in Figure 8, and higher melt duration rate will lead to a lower ice concentration due to significant melting.

The GWR tool consist of several 'BLUE' properties. It is a linear regression method and as a result, GWR has the property of 'L'. The property 'L' indicates that the estimator is linear because it estimations are weighted linear combinations of the selected variables. The geographically weighted regression is also an estimator that is used to generate local model to define the relationship of the available data. Therefore, GWR also consist of the 'E' properties.

In conclusion, the surface air temperature and the total ice concentration has a very strong correlation that governs the melt duration rate at the study area. Based on the result, areas with high surface air temperature also has a lower sea ice concentration due to the higher melt duration rate caused by the temperature. These assumptions are derived by modeling the best suitable explanatory variables found from the exploratory regression and correlation matrix. The selected explanatory variables will greatly influence the trustworthy and reliability of the model. Therefore, choosing the TS and NCT as the variables for the OLS and GWR tools allow a closer interpolation in the model to the real world. These models can also be used to help explaining other environmental phenomena that influences the melt duration in the Canadian Arctic Archipelago.