Lab 3 – Geographically-Weighted Regression

Effect of neighbourhood variables on social skills of children in Vancouver

Connor Guilherme

31581144

GEOB479

Feb. 9, 2018

Part 1 – Explanation of Geographically-Weighted Regression

A Geographically-Weighted Regression (GWR) is essentially a multiple regression that has spatial significance based on Tobler's Law that near things are more related than distant things. To explain this, the following discussion explains the difference between bivariate linear regression and multiple regression and the importance of the R^2 value and the residual.

Bivariate linear regression compares a dependent variable to a single independent variable (explanatory variable) and fits a straight line to the data to quantify a correlation. This is a simple relationship that can be used to predict the dependent variable based on the trend set by the explanatory data. The vertical distance between the straight line fit to the data and each data point is the residual (or the difference between the actual value and the predicted value that the line represents). The R^2 quantifies the variation in the explanatory variables and, as a result, is often called the "coefficient of determination". The closer the R^2 value is to 1, the better the fit of the model, therefore, if a model has a high R^2 value its usefulness as a predicting tool is higher than a model with a lower R^2 value.

Multiple regression compares a dependent variable to multiple independent variables (explanatory variables) and uses the residuals of each bivariate relationship between the dependent and independent variables in order to determine the best explanation for the dependent variable. There can be any number of explanatory variables, but the more there are the more complex the explanation for the dependent variable becomes as there are more and more residuals for each variable to take into account.

The GWR uses a multiple regression technique to explain the relationships between the dependent and independent variables and also associates these relations to geographic locations. By associating these relations to geographic locations, GWR adds another set of complexities to the attempt to explain the dependent variable. As Tobler suggests, the difference between near and far and being able to quantify this is integral to understanding the geographic relationships between variables. So, to conduct a GWR it is important and challenging to define what "near" and "far" are quantitatively with respect to the relations between variables. To deal with this challenge, the GWR uses the Monte Carlo method. The

Monte Carlo method determines the effect of spatial distribution by comparing the actual spatial arrangement to a random spatial arrangement to quantify the fit of the regression model.

Now that the premise of how GWR works has been articulated this discussion will introduce how to conduct a GWR. The first step in regression is to run an Ordinary Least Squares (OLS) regression for the important explanatory variables (excluding any unimportant variables). In order to determine the "important" explanatory variables, the explanatory regression tool should be used prior to conducting the OLS. This tool determines the AdjR^2 and the AICc for each variable – the most important variables have the highest AdjR^2 values and the lowest AICc values. The OLS provides a model for the dependent variable and creates a single regression equation to explain the dependent variable. It is important to run the GWR with the same input variables as the OLS as this will ensure that consistency in the relationships the model identifies. For example, if 4 variables were input into the OLS and only 3 variables were input into the GWR then the 4th variable in the OLS (that is not included in the GWR) can change influence the prediction of the OLS. The GWR provides a model for the dependent variable by fitting a unique regression equation to each of the explanatory variables. The R^2 determined by the GWR identifies the locations where the model fits best and can be used to help support or discredit the influence of a explanatory variable on the dependent variable.

In conclusion, the GWR is a powerful tool in understanding and quantifying the relationships between a subject of interest and explanatory variables through both determining statistical and geographical significance. However, as the GWR is based on linear regression the model produced can only represent linear relationships even when in reality the relationship between variables maybe non-linear. Another important issue with regression in general is misspecification. This is when an important explanatory variable is missing from the regression calculation and therefore an important piece to the puzzle in understanding the dependent variable is missing. Non-linear relationships and misspecification are two major issues that should be acknowledged when analyzing data through the use of GWR.

Part 2 – Discussion of GWR results

Step 1 - Explanatory regression analysis tool

The explanatory regression analysis tool was used to determine the most important variables based on the highest AdjR2 values and the lowest AICc values. Based on the dependent variable of social skill, the most important explanatory variables were found to be income, language ability and gender. This means that the other variables: ESL, fam4, loneparent, recent immigrant and physical ability were not included in the OLS regression or GWR.

Step 2 - Ordinary least squares tool

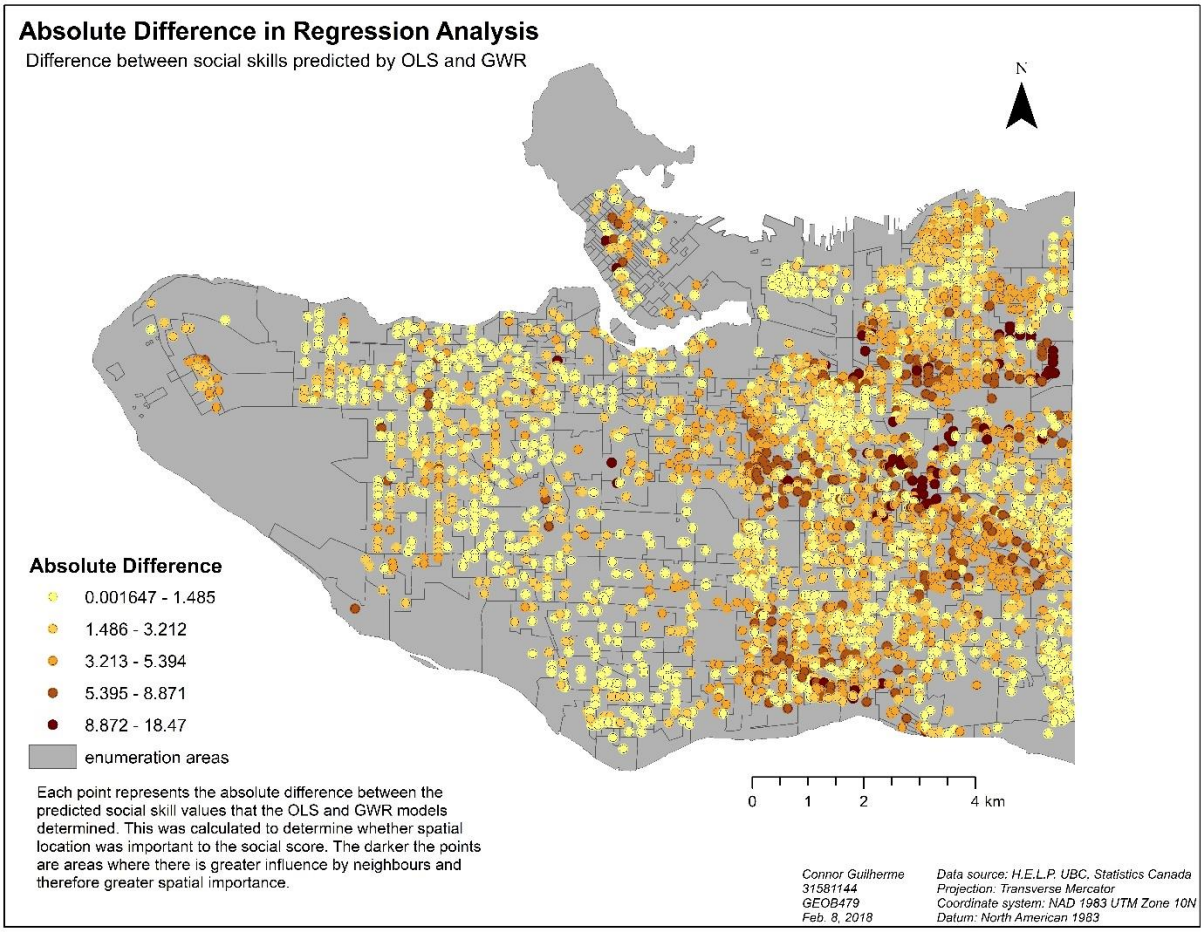
With the inputs of income, language ability and gender the OLS created a linear model based on a single regression equation. This model predicted the social skill of the child based on geographic location and the three important explanatory variables.

Step 3 - Geographically weighted regression tool

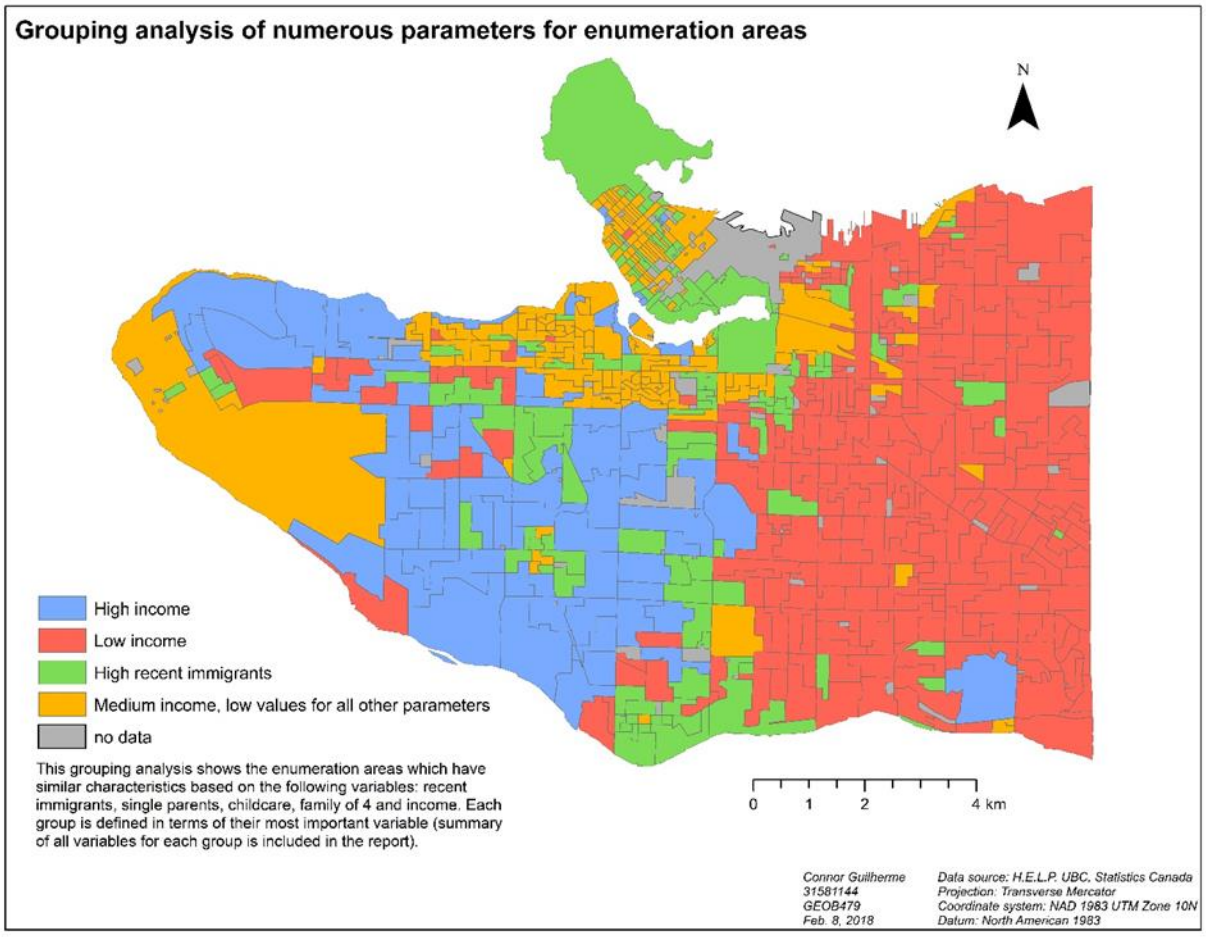
To visually represent the relationship between the explanatory variables and the dependent variable (social skill), a map of the absolute difference between the OLS and GWR predicted social skill values along with a map of the GWR for each explanatory variable is presented below. The GWR tool was used with the same inputs as described above and predicted the social skill of the child based on the geographic location with unique regression equations for each of the important explanatory variables. This produced 3 raster maps, one for each of the three explanatory variables (2 of these maps are included below).

Step 4 - Grouping analysis tool

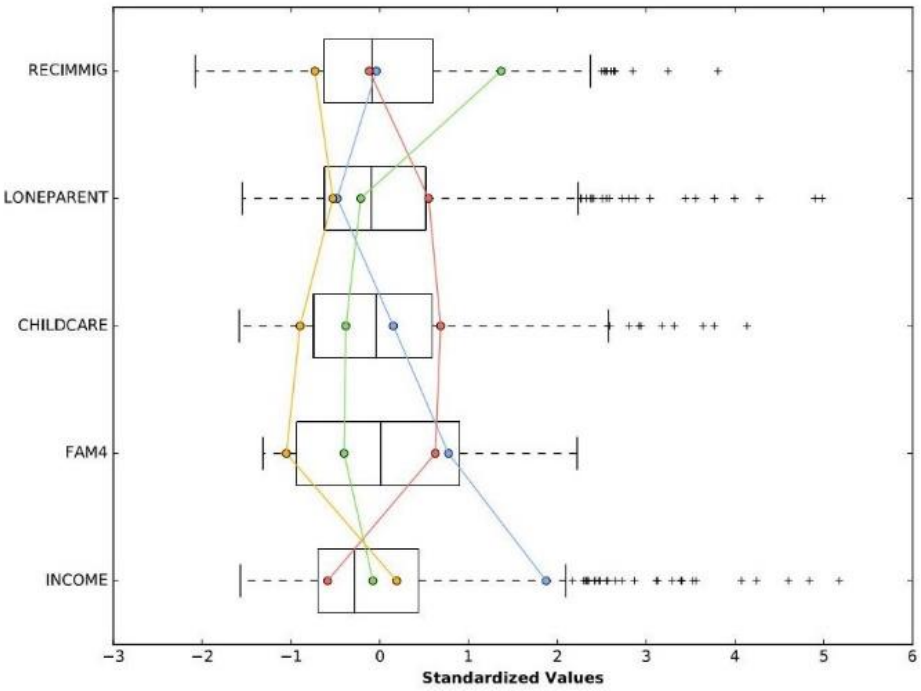
The grouping analysis tool was used to provide context to the GWR in order to better understand the context behind the relationship between each of the important explanatory variables, the dependent variable (social skill) and spatial distribution. In order to achieve this the group analysis took in 5 inputs (recent immigration, lone parent, family of 4, and income). The output report was used to interpret the meaning of each of the 4 output groups.



This map was created by subtracting the predicted social skills values determined through the OLS by the predicted social skill values determined through the GWR. The smaller the absolute difference the better the OLS fits the GWR. As the map above shows, for most of Vancouver the absolute difference for predicted social skill is small (absolute difference = 0.001647-1.485). This means that for most of Vancouver the OLS fits the GWR well and thus spatial distribution is not very important to social skills in children in Vancouver. However, in the few locations where the OLS does not fit the GWR well (darker spots), spatial distribution is important to the social skill. As it shows, there are two clusters of areas where spatial distribution can be seen as “very important” (absolute difference = 8.872-18.47) to social skill. There are a few other areas where spatial distribution is “important” (absolute difference = 5.395-8.871). Most of these “very important” and “important” clusters can be found East of Main Street.



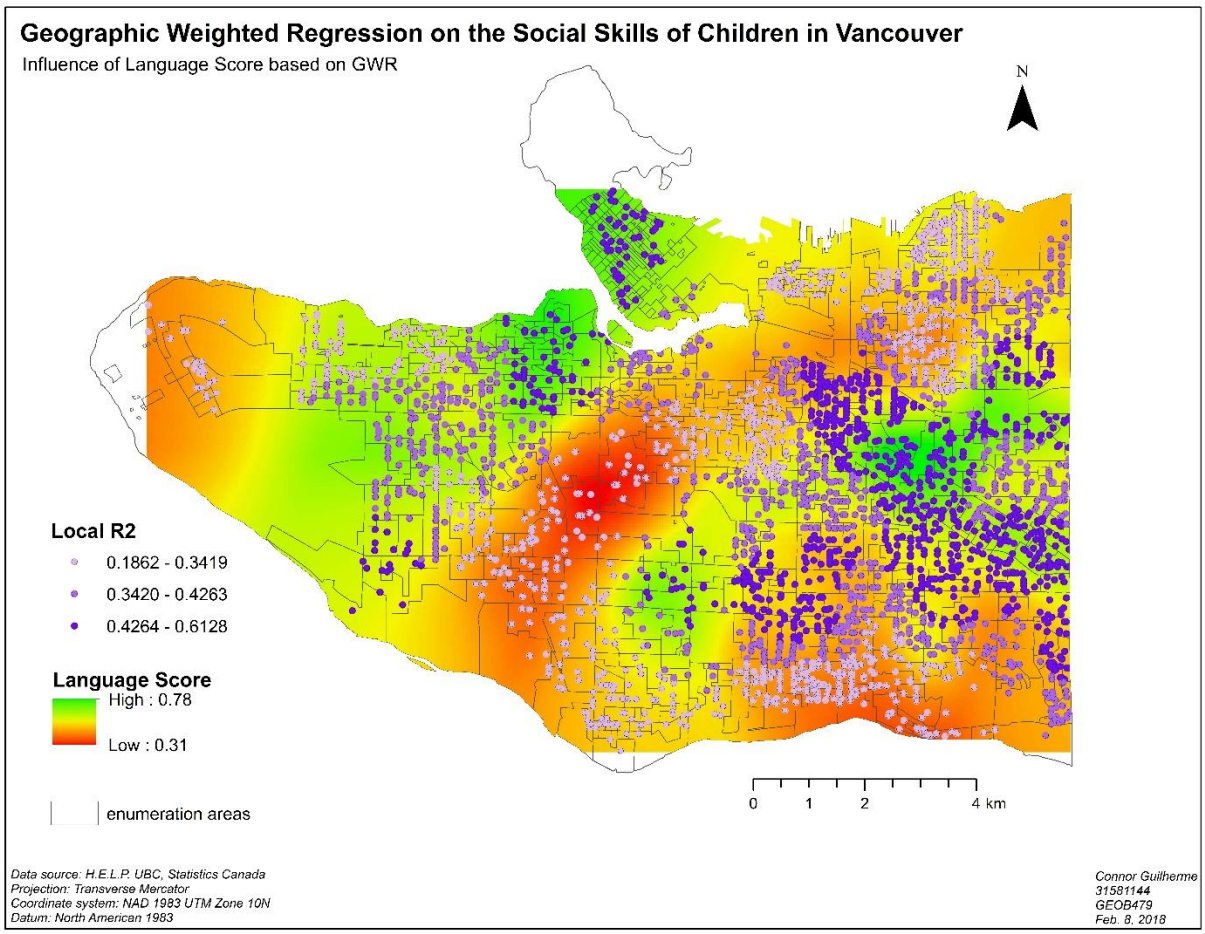
As the map above shows, the majority of the East Vancouver has low income (red), most of Kitsilano has medium income (orange), Point Grey, Shaugnessy and Kerrisdale have high income (blue), and high recent immigration is scattered across Vancouver and high in Marpole (green). It is also interesting to note that downtown is split between medium income (orange) and high recent immigration (green), although there is no data for a large portion of downtown.



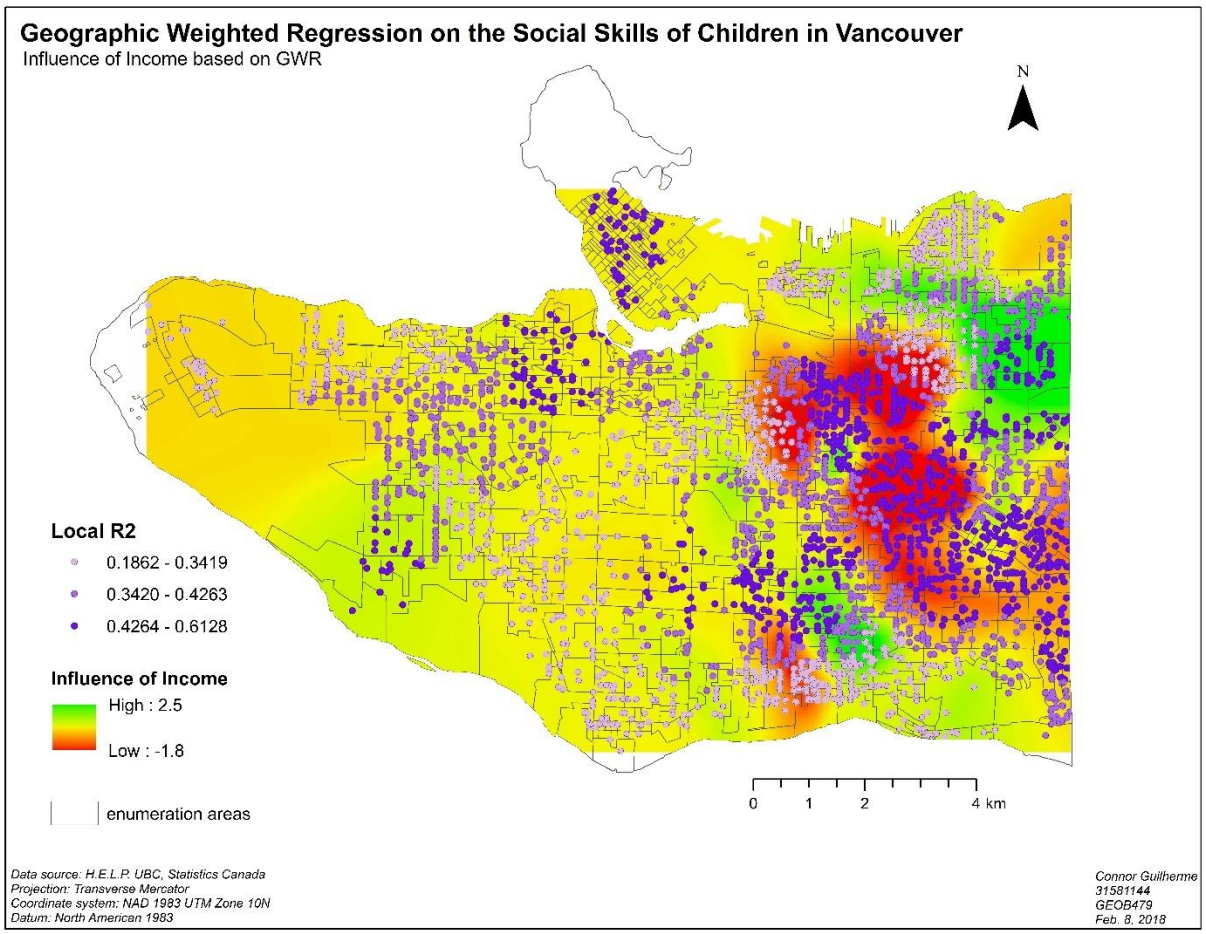
	Recent Immigrant	Lone Parent	Childcare	Family of 4	Income
Blue	Medium	Low	Medium	High	**High***
Red	Medium	High	High	High	**Low***
Green	**High***	Medium	Medium	Medium	Medium
Orange	Low	Low	Low	Low	**Medium***

***** = identifies most important variable rating as shown in the grouping analysis map above

The 4 groups determined by the grouping analysis map above were determined using the grouping analysis tool which calculates the group mean for each variable. The above cat and whisker boxplot from the group analysis output report shows the mean group values as dots in terms of standardized values for each variable, while the table above is an interpretation of the results in the boxplot that qualitatively assigns each group a rating of low, medium or high for each variable based on where the group mean falls in each of the boxplots. As you can see, the bolded ratings are the corresponding ratings for each respective colour as shown in the map legend – these are the most important variable ratings for each respective group.



This map shows the GWR results for relationship between language ability/score and social skills for children in Vancouver. Based on this map, it appears that language ability/score plays a significant role in determining social skill as the high R^2 values for the GWR are associated with high language score areas based on the model (for the most part). This is most evident in the West side of downtown and the East side of Kitsilano. However, the high R^2 values are not very close to a value of 1 so language score is not the entire cause for the social skills of children. For example, the area near Main Street and Broadway has high R^2 values but low language score values thus showing that the above described relationship is not found everywhere.



This map shows the GWR results for relationship between income and social skills for children in Vancouver. Based on this map it appears that there may be a relationship between low income areas and social skills as the large red area that identifies a low-income area is almost entirely covered by high R^2 values. However, in the other areas, where the R2 value is high (strong explanatory power), the income is high suggesting the opposite relationship that, contrasting, there is a relationship between high income areas and social skills. Clearly, the issue of non-stationarity is present. With this in-mind, although some non-stationarity issues can be found in the language map above too, it appears that the relationship between high language ability/score and social skills is greater than the relationship between low income and social skills.

Part 3 – How GWR could be used in a variety of contexts.

Using GWR to determine explanation factors for the social skills of children is just one of a limitless number of ways in which GWR can be applied. By understanding the reasons for using GWR, the great magnitude of possible applications of it becomes easy to imagine. As a form of regression analysis ESRi identifies the following three primary reasons to apply these types of methods:

i) To model some phenomenon to better understand it and possibly use that understanding to effect policy or make decisions about appropriate actions to take.

ii) To model some phenomenon to predict values at other places or other times.

iii) To explore hypotheses

The above reasons are specifically for the regression analysis portion of the GWR, but the true power of GWR as a regression analysis method is that GWR links statistical regression analysis of data to spatially dependent factors. This spatial component enables many applications of GWR in many different fields such as health, transportation, politics, education, resource management, environmental science, etc.

Three examples of GWR applications:

i) Increasing the accuracy of nitrogen dioxide pollution mapping

by Robinson, Llyod and McKinley

This report had a single input data set of current nitrogen dioxide for the GWR that was used to develop a predictive model for NO_2 pollution based on a GWR model.

ii) GIS-based analysis of obesity and the built environment in the US

by Xu and Wang

For the GWR in this report the independent variables were 3 built environment factors: street connectivity, walk score and fast-food/full-service restaurant ratio and two sociodemographic variables: race heterogeneity and poverty rate. The dependent variable was obesity and the through the use of the GWR model the

report found that walk score and street connectivity are negatively related to obesity, poverty rate and metro are positively related, and the fast-food/full-service restaurants ratio is not significant.

iii) Modelling malaria treatment practices in Bangladesh using spatial statistics *by Haque et al.*

In this study the dependent variable was "malaria treatment-seeking preferences" and 15 explanatory variables were run through the GWR model. Based on the GWR, the study found that several factors including tribal affiliation, housing materials, household densities, education levels, and proximity to the regional urban centre, were found to be effective predictors of malaria treatment-seeking preferences.

As these examples show, the GWR applications are diverse and the number of explanatory variables can vary based on the scope of the study. These examples and the use of GWR in this report show the great power and versatility of the GWR method. But, what cannot be forgotten is that the use of GWR can be challenging as there are a few issues that can arise. As described above, some such issues surrounding the use of GWR include: non-linear relationships between variables, misspecification and non-stationarity.

References:

Altman, Douglas G. Practical statistics for medical research. CRC press, 1990.

ESRI ArcGis. (2014). Regression Analysis Basics, How GWR Works, How OLS Works, Interpreting GWR Results, Interpreting OLS Results, Geographically Weighted Regression (GWR) (Spatial Statistics). Retrieved from ArcGIS Resources: http://resources.arcgis.com/en/help/main/10.2/index.html#//005p00000023000000

Haque, U., Scott, L. M., Hashizume, M., Fisher, E., Haque, R., Yamamoto, T., & Glass, G. E. (2012). Modelling malaria treatment practices in Bangladesh using spatial statistics. *Malaria journal*, 11(1), 63.

Klinkenberg, B. (2018). Introduction to Multiple Regression and GWR, Lab 3 GWR, GWR Considerations. Retrieved from GEOB 479 GIScience in Research: http://ibis.geog.ubc.ca/courses/geob479/labs/gwr_explanation.htm

Robinson, D. P., Lloyd, C. D., & McKinley, J. M. (2013). Increasing the accuracy of nitrogen dioxide (NO2) pollution mapping using geographically weighted regression (GWR) and geostatistics. *International Journal of Applied Earth Observation and Geoinformation*, 21, 374-383.

Xu, Y., & Wang, L. (2015). GIS-based analysis of obesity and the built environment in the US. *Cartography and Geographic Information Science*, 42(1), 9-21.