Investigating the Social Skills of Children in the City of Vancouver

An Exploration of Geographically Weighted Regression

By: Grihalakshmi Soundarapandian
37082138
GEOB 479 L2A
February 8, 2018

About Geographically Weighted Regression

When thinking about spatial analysis, a common theme is the use of rasters or polygons. These elements have been used extensively when exploring the patterns, trends, and variables across an area. They are used instead of aggregated point data, as that brings rise to the issue of the Modifiable Areal Unit Problem (MAUP) (Klinkenberg). The MAUP occurs when the same data undergoes the same analysis, but the results differ if the aggregation scale or zoning is changed (Esri). Therefore, an uncertainty in the results of an analysis is created, along with issues in the inference and conclusion of the results. In order to avoid aggregation of point data and the MAUP, but still conduct viable analysis, this lab introduced the idea of Geographically Weighted Regression (GWR), a tool that allows individual analysis of multiple variables across a surface, creating a regression equation for each, and encouraging the comparison of and the relationship between variables to show spatial variability, which is usually not present in other regression models (Legg and Bowe). Additionally, the GWR provides a platform to analyze non-stationary data such as weather, climate, temperature, and human and population data (Columbia University). Importantly, it is a method which allows the patterns created by the point data to be observed and analyzed with greater certainty, than analysis conducted via aggregation methods that introduce the MAUP.

Regression analysis is a method of testing the relationship between variables, and observing whether or not they influence each other (Wikipedia). Heavily used in statistics and geographical information science, regression is a key concept in understanding present patterns between variables and predicting future responses, and in determining dependent and independent or explanatory variables (Wikipedia). In that manner, GWR extends this concept into a geographical context, taking into account local regression (Klinkenberg). GWR is an extension to the Ordinary Least Squares (OLS) method of regression, and is used when one believes that the variables will vary depending on location (Columbia University). As mentioned within it's name, the GWR uses a weighted matrix, similar to that of a Multicriteria Evalulation (MCE), wherein the strength of a variable is determined by how close or far away from a location it is, with the weight decreasing with increased distance (Columbia University). In order to conduct a GWR, the OLS tool, followed by the GWR tool in ArcMap must be used, along with a kernel type, bandwidth method, and the number of neighbours specified (Klinkenberg). Once conducted, the GWR will provide meaningful results, showcasing the influence of location on the variables of the analysis.

In this lab, we wish to see the local relationships between a child's social skills (Klinkenberg), and variables such as language abilities (Lang_sc), gender (Gender), and the average income within the neighbourhood (income1000). Some variables (Table 1) pertain to the child only, and others to the whole neighbourhood, allowing one to see which neighbourhood or personal attributes may have the greatest influence on social skills of a child (Klinkenberg). GWR allows us to conduct local spatial regression analysis, and focus on the local scale (Klinkenberg), which is crucial in understanding specific patterns and attributes of neighbourhoods, and can be used in a much more useful manner to create proper solutions, than a global scale analysis for the same problem. Here, the OLS tool was used, with the GWR tool afterward to render the overall results of the analysis. The results, ie. the impacts of the neighbourhood on the child's social skills, and the inference of these results, are explained below, along with Maps 1-5.

Lab Results and GWR & OLS Comparison

As mentioned previously, GWR analysis is suitable for local regression. However, OLS, or the Ordinary Least Squares tool serves global regression, and for measuring the relationship of a dependent variable with its explanatory variables (Esri 2017). In this lab, OLS and GWR were both used, using the variables indicated in Table 2, and the results were plotted over the Enumeration Area Grouping Analysis (EAGA).

The EAGA was created using the variables indicated in Table 3. The resulting groups and their variable values are described in Table 4. The OLS generated standard residual results with a Koenker (BP) Statistic of 48.331, a Jarque-Bera Satistic of 891.896, and a Chi-Squared value of 0.000 with 3 degrees of freedom. Looking at Map 1, the OLS results were scattered, showing high variability across the study area. High correlation points (between social skills and the Groups) are present everywhere, but slightly more concentrated on the right side of the map, suggesting that the strength of the variables in all groups can significantly influence social skills, which may not be true. Low and mid-correlation points are also present across the map, without distinct concentration areas. Also, as indicated in Figure1 from the OLS report, the results rendered a positive relationship between Social Score, and Language Score and Income1000, meaning that a higher Language Score and Income results in greater social skills. Gender did not influence social skills significantly. Whereas, the GWR (Map 2) used Local R2, and generated results of greater accuracy. This can be see by how precise and concentrated each of the high, mid, and low correlation points are across the map area. For example, on the right side of the study area, within Group 1, high correlation occurs heavily, along a slanted shape. Surrounding this are distinct areas of mid and low correlation. This replicates higher spatial autocorrelation and clustering, and less spatial variability than the OLS results. High correlation also occurs distinctly within Group 3 areas at the top of the map, and low correlation also occurs in Group 2 in the middle of the study area.

GWR and Income1000 (Map 3): The spatial distribution of the Income1000 variable is such that low income is quite distinct in the right side of Vancouver, between areas of high income. Mid income dominates the rest of the study area. The GWR results match the areas of low and high income, with high and mid correlation points respectively. Therefore, distinctly, higher income neighbourhoods result in greater social skills of a child, and vice versa. In areas of mid income, income is not a strong driver of social skills.

GWR and Language Score (Map 4): Similar to Income1000, high language scores are distributed predominantly along the right side and top of the study area, and slightly in the left of the map. Once again, the GWR results match, showing high correlation in the high language score areas. This means the social skills are positively influenced by language scores, such that a higher language score results in greater social skills of a child, and suggests that these neighbourhoods may have better language-orientated education.

GWR and Gender (Map 5): The spatial distribution of Gender is such that certain areas have more females, mainly on the right and lower middle areas of the map. These are then surrounded by areas of more males, specifically at the top of the study area, and then with equal-gender areas in-between. The GWR results here show that there is a no specific correlation between social skills and gender, as both high and mid-correlation points exist in areas of both high male and female concentration, with low correlation points sharing the genders as well.

The Versatility of Geographically Weighted Regression

In this lab, the GWR tool was used in a social context - child development. But, it is incredibly versatile and can also be used in a variety of areas where space is an important aspect, ranging from epidemiology to real estate, and more.

Within epidemiology, GWR is used to investigate the existence and spread of various diseases (Columbia University). A study conducted in Keur Soce, Senegal identified hotspots for malaria, and assessed the related risk factors using a GWR analysis. Various statistical and geostatistical modeling methods were used, including GWR coupled with spatial autocorrelation tools (Ndiath et al.). As Ndiath et al. states, the dependent variable was the reported cases of malaria, and the explanatory variables were various environmental and socioeconomic factors as listed in Table 5, all of which were monitored over a 6-month period. As a result of this analysis by (Ndiath et al.), a choropleth map was produced to show the number of households with reported malaria cases, under 3 map classifications: “households with no confirmed cases, households with one to three confirmed cases, and households with more than three cases” (Figure 3). Then, using “an empirical Bayesian kriging model with log empirical data transformation”, a continuous map of malaria hotspots was created (Figure 4). The technical flow chart for the project is shown in Figure 2. The results of an epidemiology study like this one can be incredibly useful to public health servants and doctors. In this case, doctors can be sent to the hotspot locations to administer medical services, and more awareness initiatives can be conducted within the surrounding areas of middle malaria case confirmations (ie. listed as 1-3 malaria cases in Figure 3), to educate the residents about malaria prevention and safeguarding.

Aside from epidemiology, GWR can also be used for real estate. Using location and other related variables, students of Northern Michigan University used GWR to investigate the listing prices of single family homes across Marquette, Michigan (Legg and Bowe). Using the GWR tool within ArcMap allowed them to model this real estate behavior in higher accuracy and certainty than before, when only linear regression models were used and resulted in under- or overestimated prices (Legg and Bowe). Dating back to 2008, Legg and Bowe chose 93 homes for sale, and the GWR tool rendered individual regression equations for every home. From these results, a raster surface was created, showing spatial variability (Legg and Bowe). As seen in Figure 5, the most expensive homes are located on the top left, with the prices slowly decreasing across the study area, with the most affordable homes near the middle right of the study area. Then, prices slightly increase once again in the right-most of the study area. The results from this study are quite interesting, and now can be used by home buyers and real estate agents alike to see where their respective markets will be within Marquette, Michigan.

As explained, the use of GWR extends simply past a strong spatial analyst tool within ArcGIS. It is a tool which can make any kind of study possible, and helps to generate results that will aid in discovering and implement solutions for real world problems, creating substance and meaning for the investigation as a whole. In this regard, geographically weighted regression is crucial to modern-day GIS analysis, within all fields.

APPENDIX

Variable	Description
Phys_sc	A child's physical abilities, a value between 0 – 100
Soc_sc (Social Score)	A child's social skills, a value between 0-100
Lang_sc (Language Score)	A child's language abilities, a value between 0-100
LoneParent	% of the neighbourhood families that are a single parent
Fam4	% of the neighbourhood families with 4 or more members
Immigrant	% of neighbourhood families that are immigrants
RecImmig	% of neighbourhood families that are recent immigrants
ESL_EA	% of the neighbourhood that does not have English or French as their first language
Childcare	% of families that spend 30 or more hours on childcare
VisMinority	% of neighbourhood that are visible minorities

Income	Average income for the neighbourhood
Income1000	Average income for the neighbourhood divided by 1000
Gender	Child’s gender: 1 = Female, 0 = Male
ESL	Is English the child’s first language? Yes = 1, No = 0

Table 1. All Variables Within the Data from Early Development Instrument

Dependent Variable	Explanatory Variables
Soc_sc	Lang_sc, Gender, Income1000

Table 2. The Dependent and Explanatory Variables chosen for the GWR and OLS Regression Analyses. The Explanatory Variables were determined to be the “most important” variables in determining the social skills of a child based on having the highest AdjR^2 and the lowest AICc values. *Source: Klinkenerg*

Variable
Childcare, Fam4, LoneParent, RecImmig

Table 3. The variables used in the Ennumeration Area Grouping Analysis. The Resultant groups (Table 4), incorporate these variables at varying values

Group	Variable Values
1	High Fam4, Childcare, and LoneParent; low Income; average RecImmig
2	High Fam4, Income, Childcare; low LoneParent; average RecImmig
3	High Income; low Fam4, RecImmig, Childcare, and LoneParent
4	High RecImmig, and Income; low Fam4, Childcare, LoneParent

Table 4. Grouping Analysis Results: Groups and Variable Values

Variables	Values
Age	0–69
Gender	0–100 % for both males and females
Household size	1–35
Village size	10–3698
Sleeping rooms	1–9
Bed net use	1–25
Distance to health Facilities	1–15
Temperature	18.1–41.7
Raining days	15–39
Housing materials	Material used for walls, roof, and floor
Cow	2–125
Goat	1–66
Sheep	0–99
Donkey	0–58
Horses	0–15
Distance to breeding site	10–300

Table 5. Environmental and Socioeconomic Explanatory Variables Used in Ndiath et al. *Source: Ndiath et al.*

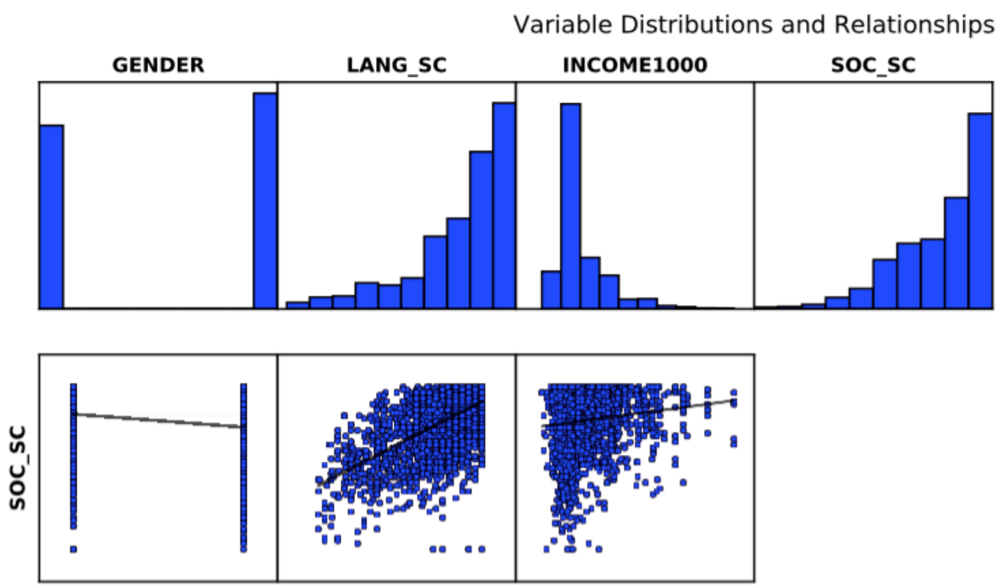

Figure 1. OLS Relationships Between Dependent and Explanatory Variables. *Source: OLS Results Report PDF*

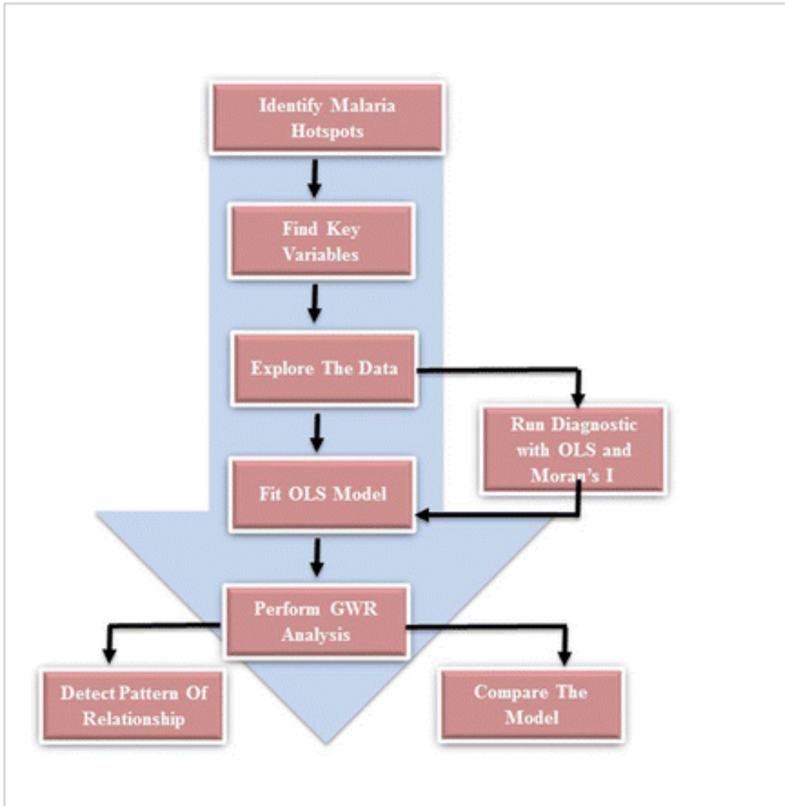

Figure 2. Technical Flowchart of Malaria Analysis. *Source: Ndiath et al.*

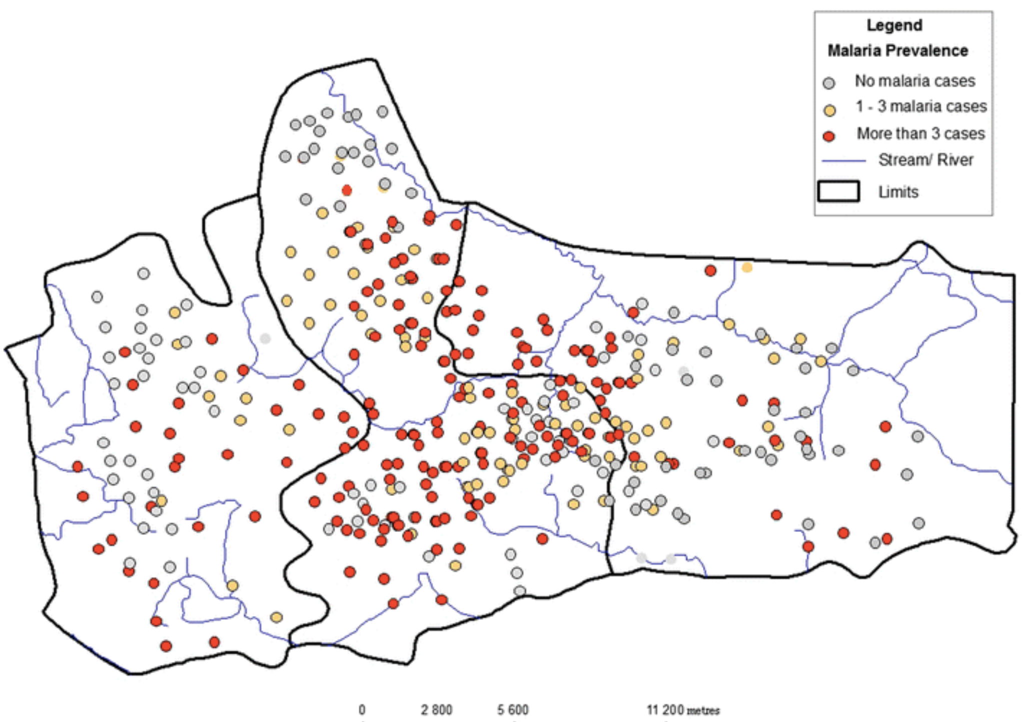

Figure 3. "Spatial distribution of household with and without RDT positive". *Source: Ndiath et al.*

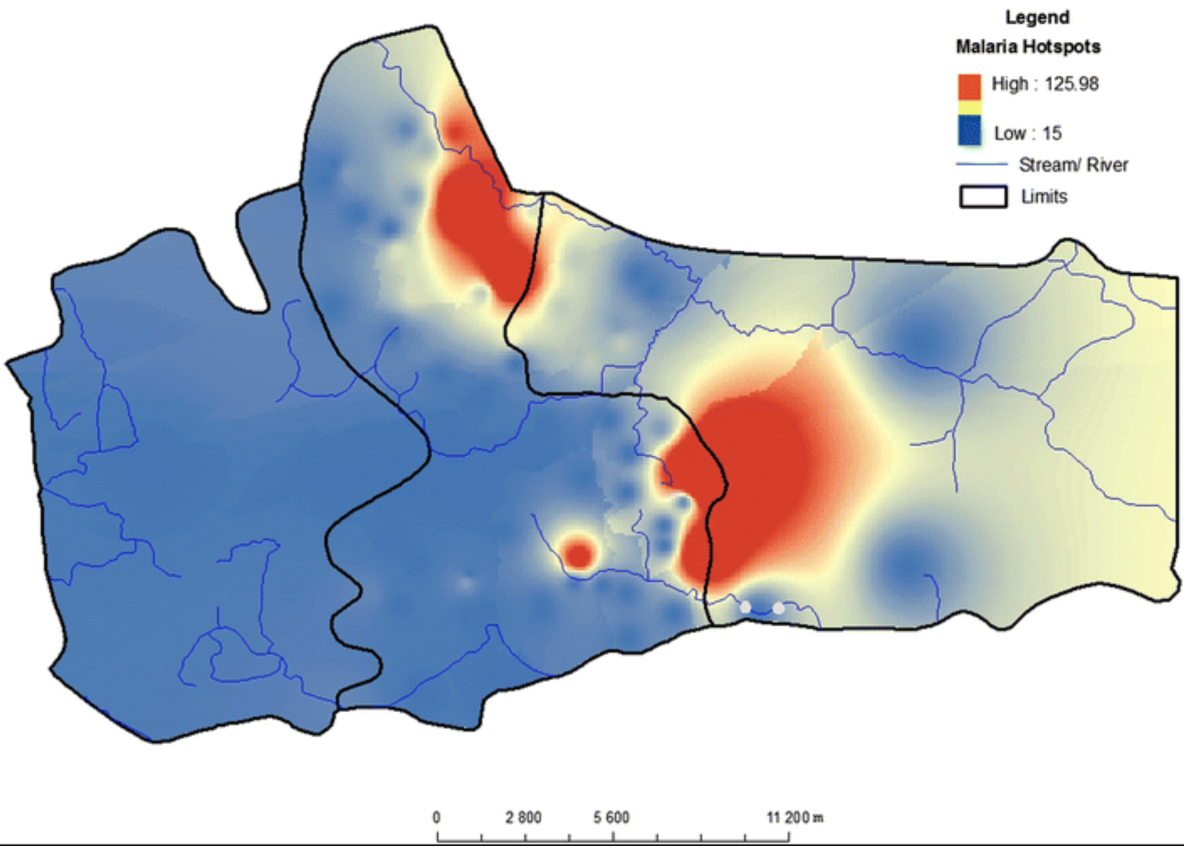

Figure 4. "Malaria hotspots in Keur Soce HDSS". *Source: Ndiath et al.*

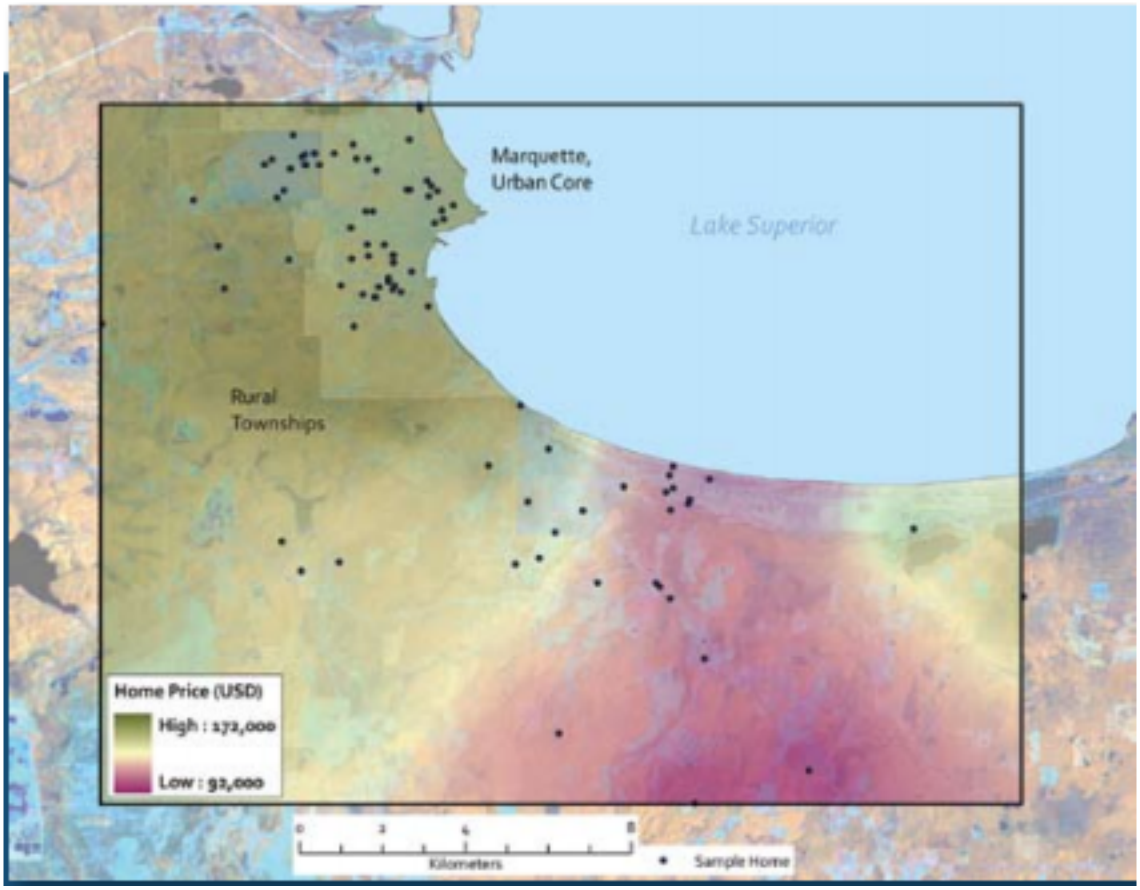

Figure 5. "The listing price of a typical house modeled based on spatially varying regression coefficients generated using the GWR tools in ArcGIS." *Source: Ndiath et al.*

Works Cited

Esri. “Maup GIS Dictionary Technical Support.” *Esri*. n.d. Web. 08 Feb. 2018. <https://support.esri.com/en/other-resources/gis-dictionary/term/maup>

Esri. “Ordinary Least Squares.” *ArcGIS Desktop ArcMap*. 2017. Web. 08 Feb. 2018.<http://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/ordinary-least-squares.htm>

“Geographically Weighted Regression.” *Population Health Methods.* Columbia University, n.d. Web. 08 Feb. 2018.

Klinkenberg, Brian. “Lab 3: Geographically Weighted Regression.” *Geob 479: GIScience in Researach*. Dept. of Geography. U of British Columbia, Vancouver, 2018. Web. 08 Feb. 2018.

Legg, Robert, and Bowe, Tia. “Applying Geographically Weighted Regression. An Example from Marquette, Michigan.” *Esri ArcUser Online Spring* 2009. Web. 08 Feb. 2018.

Ndiath, Mansour M. et al. “Application of geographically weighted regression analysis to assess the risk factors for malaria hotspots in Keur Soce health and demographic surveillance site.” *Malaria Journal* 14: 463 (2015). *BMC.* Web. 08 Feb. 2018.

“Regression Analysis.” *Wikipedia*. n.d. Web. 08 Feb. 2018. <https://en.wikipedia.org/wiki/Regression_analysis>