GEOB 479 LAB 2

Geographically Weighted Regression

Nigel Tan
17568163

GWR, or Geographically Weighted Regression, is a form of multiple regression that adds importance weights based on distance between objects. GWR can be used to answer a variety of questions where variables are non-stationary and processes and factors may change over drifts in space and changes in scale. GWR analysis of children's test scores in Vancouver shows that ESL may be the best predictor of test performance. GWR was also compared to OLS/GLR and found to be more accurate with less error for spatial relation modelling. GWR can also be applied to many other types of analyses, and is particularly suited for evaluating and modelling environmental relationships with many factors, such as disease geography.

What is GWR?

A regression analysis is an extremely important statistical function that describes the relationship between two different variables. Regression is used to identify to what extent, if any, one variable is correlated to another. That is, if the value of one variable is dependent upon the value of the other.

Geographically Weighted Regression (GWR) on the hand, adds a spatial dimension to the typical regression analysis. In GWR, variables are geographic in nature, corresponding to points on the map. These variables are often dynamic and constantly in motion. This can include people and other organisms, climate and weather, disease and health effects, or environmental variables.

For example, say one wanted to identify what the biggest human impact on salamander habitats would be. The options are numerous. It could be a coal power plant, or a bridge, or a road, or pollution. Furthermore, after conducting enough regression analyses to isolate the main cause, there is also a question of whether this holds true in different locations, making this a geographic question, which then calls for GWR. The geographic weighting adds an extra dimension to the typical regression analysis, making it more tailored for specific informational needs.

GWR goes beyond the normal regression analysis as it is multivariate, making this a form of multiple regression. The key difference here is that a typical regression/correlation analysis is bivariate, meaning that it tests the relation between two variables only. Normally, if one wanted to test more than two variables, one would first do a normal regression with two variables, then another using the residuals from the first analysis and a new set of variables. The process would then keep repeating until every variable has been tested against every residual, a highly tedious process. This is where multiple regression comes into play. Multiple regression is an automatic function that streamlines the process of conducting numerous regression analyses with many variables. GWR then takes this multiple regression a step even further by weighing the important of proximity. Geography is based on the inherent assumption that near things are more related than distant things. As such, GWR assigns higher weights to variables that sit closely together geographically, with the weight reducing as distances increase.

GWR also makes use of the Monte Carlo simulation test. In this process, the GWR algorithm will constantly resample data until the likelihood of a given event can be guessed or approximated. This is a method for using randomness to try and approach a deterministic value. In GWR, values in given locations are analyzed for patterns. Values are then randomly resampled over and over again, with the pattern compared each time for clustering. The fewer times the shuffled values are able to match the real data pattern, the higher the chance of clustering and spatial autocorrelation, meaning that values are likely to be geographically correlated.

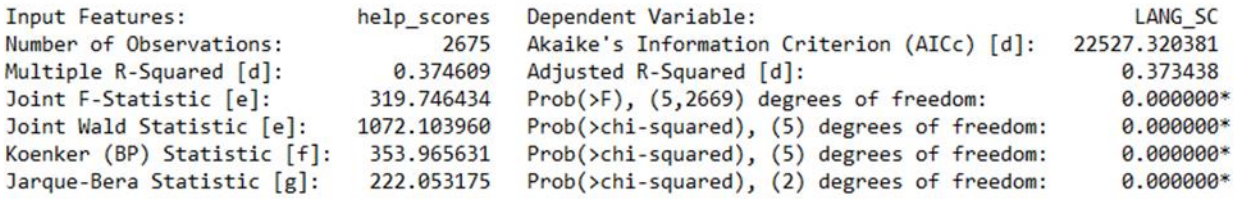
GWR is helpful as it analyzes relationships where dependency and therefore correlation may vary wildly across different spatial areas and scales. Through GWR, many key relationships that occur at a local scale can be made clear.

Analysis Results Discussion

In the results of the Generalized Linear Regression and Geographically Weighted Regression tests carried out, we found that the factors all had a statistically significant impact on children's test scores. It was found that the GWR analysis had an R^2 value of 0.48, much higher than the 0.37 given by the GLR/OLS test, meaning that the GWR analysis was better able to model the spatial clustering correlation effects. GWR also had a slightly lower Akaike's Information Criterion (AICc) score at 22361, against the GLR's 22567. This means that the GWR was able to predict with a lower error rate.

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	22.866067	2.053085	11.13742	0.000000*	2.468069	9.26476	0.000000*	--------
ESL	5.451903	0.67323	8.098127	0.000000*	0.682263	7.990911	0.000000*	1.139015
SOC_SC	0.621941	0.017114	36.34036	0.000000*	0.02233	27.85183	0.000000*	1.027846
LONEPARENT	0.316155	0.162472	1.9459	0.051766	0.164324	1.923977	0.05446	1.153368
RECIMMIG	0.067444	0.027998	2.408857	0.016055*	0.027404	2.461064	0.013903*	1.020833
INCOME1000	0.086723	0.035709	2.428607	0.015209*	0.033571	2.583294	0.009831*	1.27446

Table of OLS Results



```
Input Features:              help_scores   Dependent Variable:                          LANG_SC
Number of Observations:             2675   Akaike's Information Criterion (AICc) [d]:   22527.320381
Multiple R-Squared [d]:         0.374609   Adjusted R-Squared [d]:                          0.373438
Joint F-Statistic [e]:        319.746434   Prob(>F), (5,2669) degrees of freedom:           0.000000*
Joint Wald Statistic [e]:    1072.103960   Prob(>chi-squared), (5) degrees of freedom:      0.000000*
Koenker (BP) Statistic [f]:   353.965631   Prob(>chi-squared), (5) degrees of freedom:      0.000000*
Jarque-Bera Statistic [g]:    222.053175   Prob(>chi-squared), (2) degrees of freedom:      0.000000*
```

Looking at the tables, it seems ESL was by far the best predictor of children's test scores.

To find how each area of Vancouver differs in terms of results, we can look to the spatial clustering map and the associated clustering box plots. The map shows each of the five identified clusters and their locations across the city, while the box plots show the characteristics of each analysis. The clusters have been identified as follows:

1. Blue area. There is a relatively high usage of childcare which likely means younger children. The most significant aspect of this group is that they have the lowest average income by far. The Income 1000 raster shows many areas here as being highly sensitive to changes in income, as a small increase in income would be proportionally larger due to the area's low average income.
2. Red Area. This group appears to be an outlier group as it consists of only Granville Island and has several extreme box plot values. It scores the highest for income of all regions and boasts by far the highest lone parent score, approaching two standard deviations above expected values.
3. Green area. This zone encompasses downtown and the False Creek Olympic Village areas. Here we see the lowest rates for childcare, four-person family rate, and lone parent rate. Income is average with a higher proportion of recent immigrants. This is indicative of an area with possibly very young couples and small families.
4. Orange area. There is a large proportion of recent immigrants with high income and the highest rate of Fam4. This is an affluent area, which when combined with the Income 100 raster, shows that the area seems to benefit the least in test scores when given a change in income.

5. Purple area. This area was generally quite average across the chart but was usually slightly under the prediction values.

Other Applications of GWR

GWR has a wide variety of possible applications. However, it is claimed that one key limitation of it is that GWR tends to be more useful for exploring information than it is for predicting values. This is, GWR provides important statistical measures for given data sets, but tends to fall short in future value prediction due to an over-reliance on assumptions.

Perhaps where GWR finds the most use is in health geography and epidemiology, due to its usefulness in identifying environmental and socioeconomic health determinants on a population. Since GWR is able to perform multiple regressions, it is preferred in situations where several factors must be compared to identify

One example of a good use of GWR would be Chen et al.'s (2018) analysis of socioeconomic determinants of PM2.5 exposure in China. The goal of this study was to identify the extent to which six different location-based factors contributed to population exposure. These factors were: industry share, urbanization, construction level, urban expansion, income disparity, and private vehicle use rate. The results found that in every single test city, statistically significant test results showed that each of the six factors had a large effect on population exposure to PM2.5 particles.

A study like this is highly useful as it provides evidence for policy makers and urban planners that urbanization and related development activities have tangible and measurable impacts on human health. GWR provides a numerical metric that can be used to display relational information in a visual manner.

An article by Szymanowski and Kryza (2011) compares the utility of GWR to that of a standard multiple regression analysis for modelling urban island heat effects, a subdiscipline of urban climatology and meteorology. They found that GWR was significantly better than multiple regression due to its ability to take non-stationarity into account. While both types of regression offered similar results, where multiple linear regression fell short was that it failed to be able to analyze for location-dependency and inherently assumes stationarity.

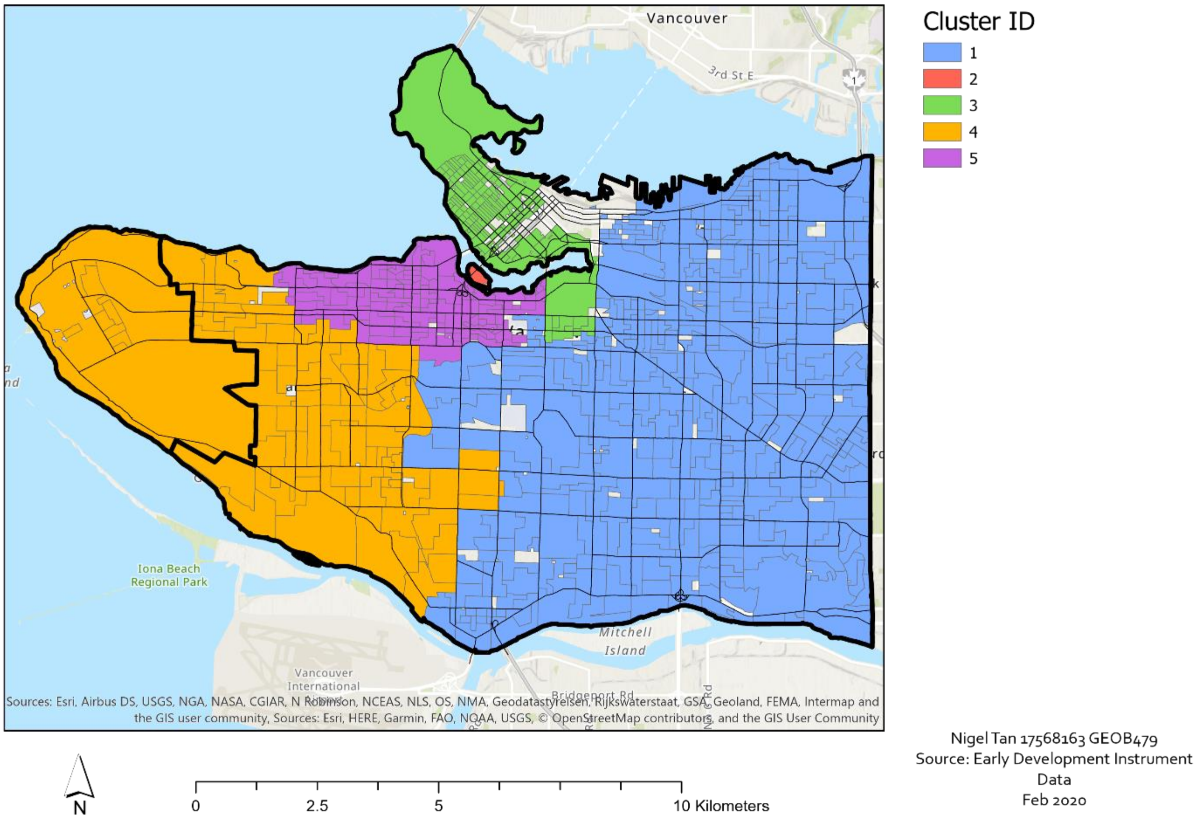
GWR is not limited to only being useful in social and environmental contexts. For example, GWR is highly useful for analyzing housing prices. Housing is a concept that is innately geographic; there is a greater desire to live in some areas over others and clusters of houses often have similar values due to surrounding amenities and features. As a result, data is heavily location dependent and does not lend itself to random patterns. GWR is then needed to identify patterns in prices and correlate them to outside factors in the neighbourhood.

To conclude, essentially any sort of analytical question where one seeks to find the relation between location-specific parameters can be answered using GWR, as GWR is able to weigh multiple variables and account for spatial clustering effects and spatial autocorrelation.

Figures

1. Spatial Cluster Analysis

Spatial Cluster Analysis of Children's Test Scores

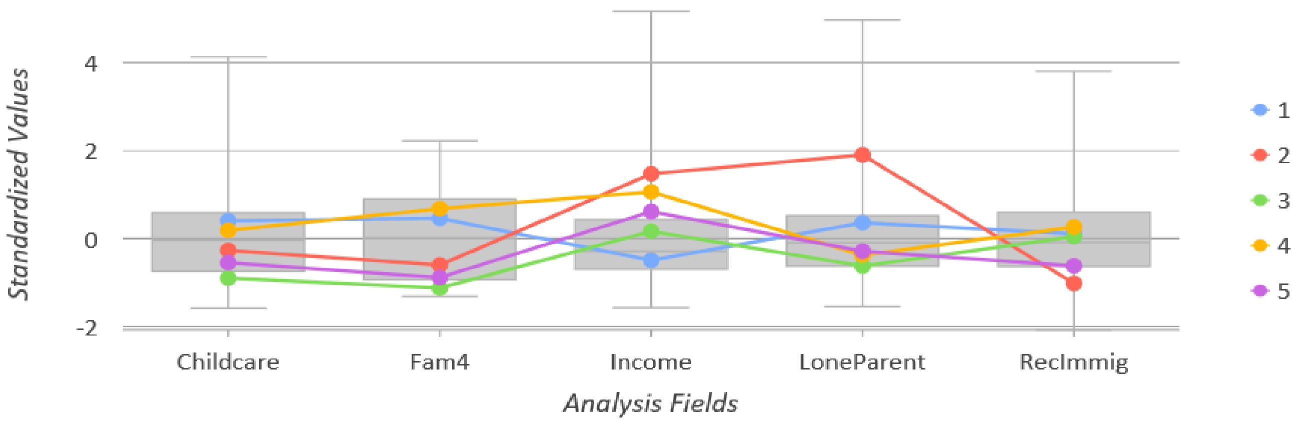


Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument
Data
Feb 2020

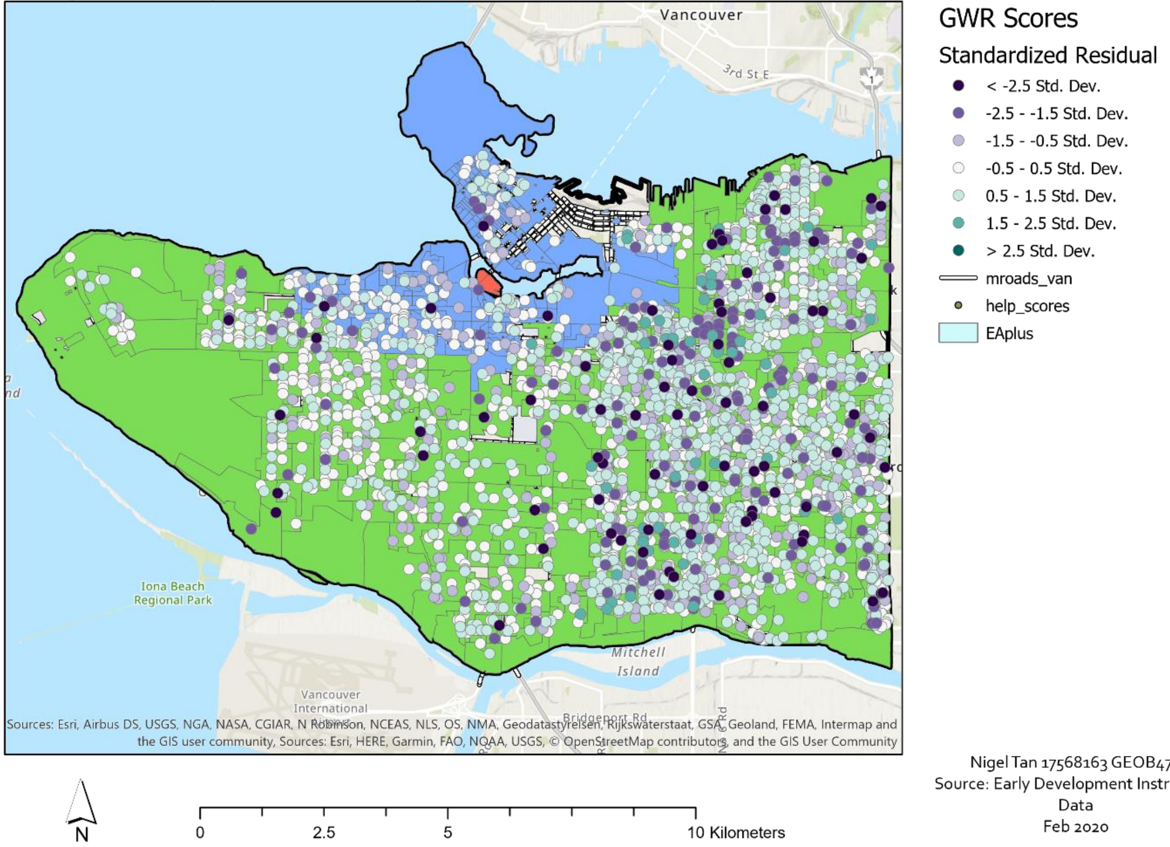
2. Spatial Cluster Box Plots

Spatially Constrained Multivariate Clustering Box-Plots



3. Geographically Weighted Regression Residual Scores

Generalized Linear Regression and Geographically Weighted Regression Analysis of Children's Test Scores

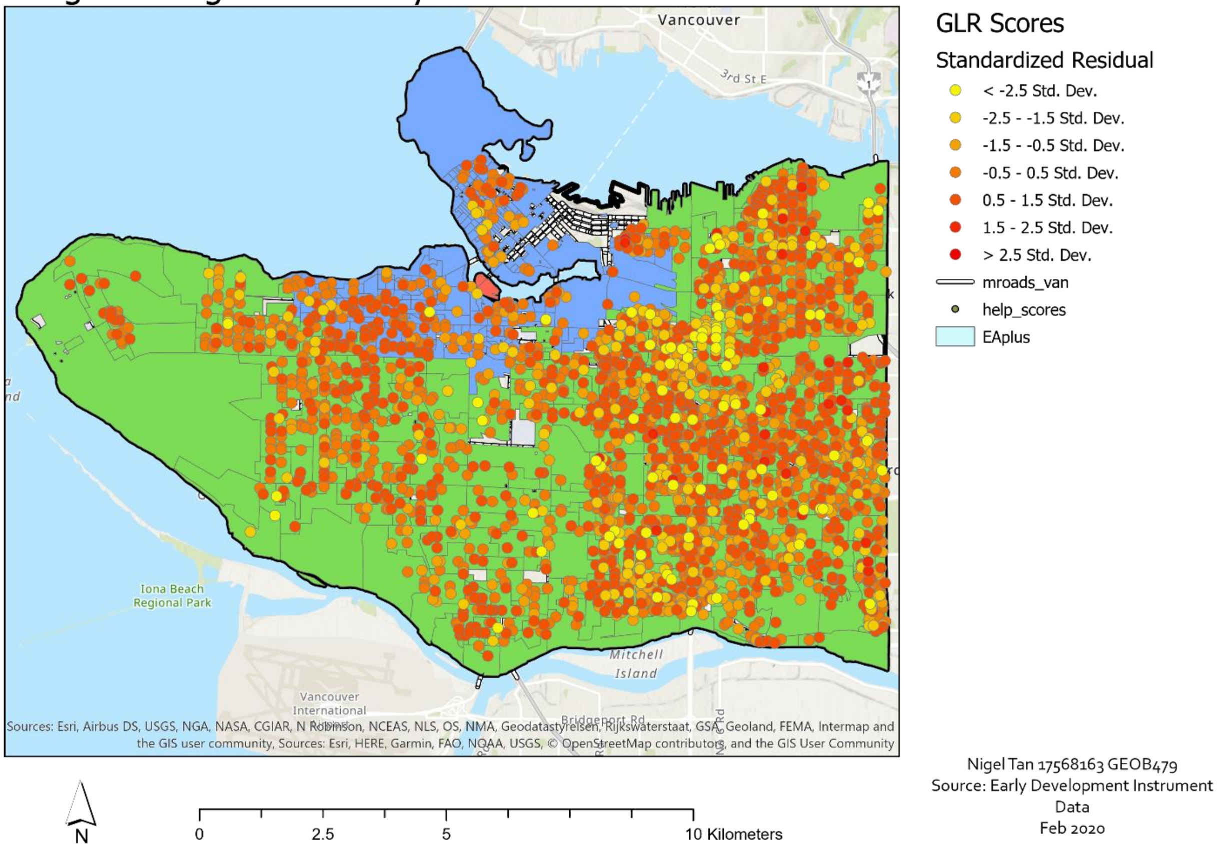


Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument Data
Feb 2020

4. Generalized Linear Regression Residual Scores

Generalized Linear Regression and Geographically Weighted Regression Analysis of Children's Test Scores

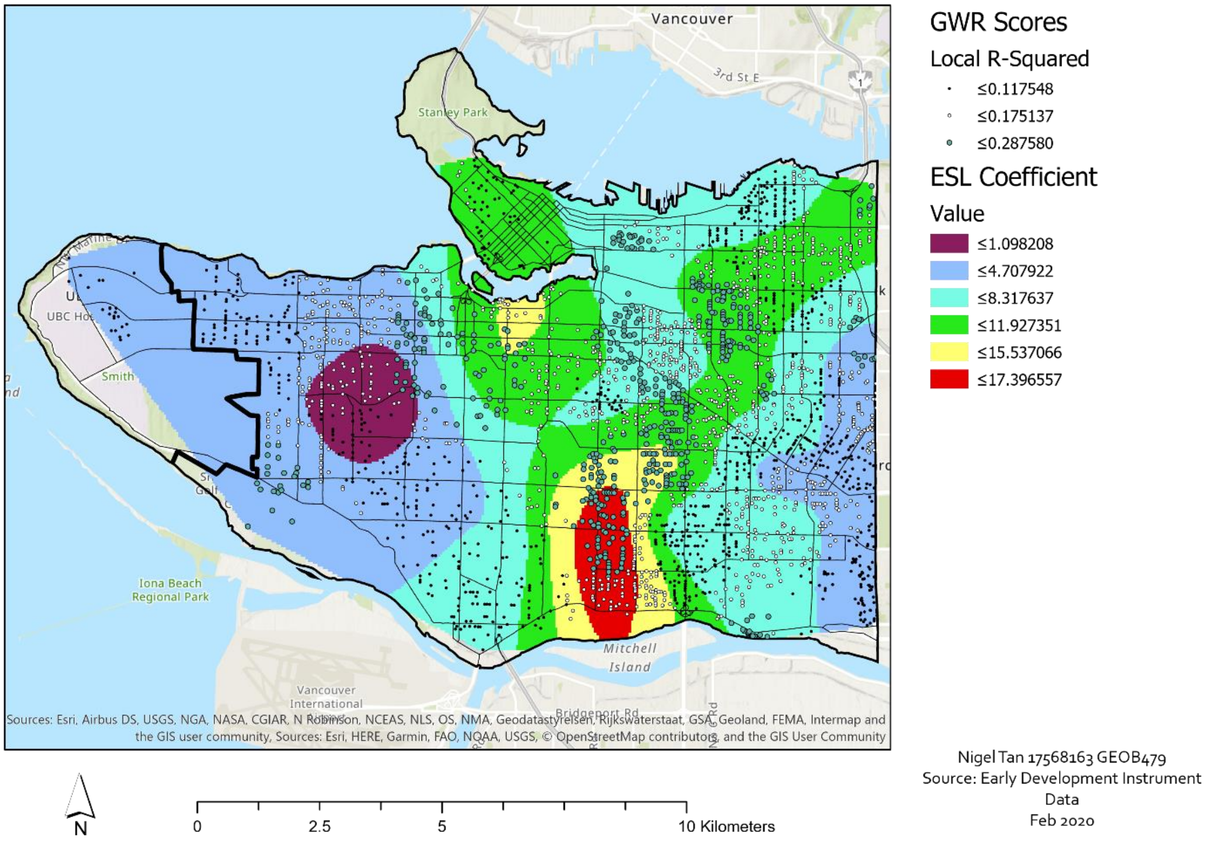


Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument Data
Feb 2020

5. Local R^2 and ESL Coefficient Raster

Spatial Cluster Analysis of Children's Test Scores

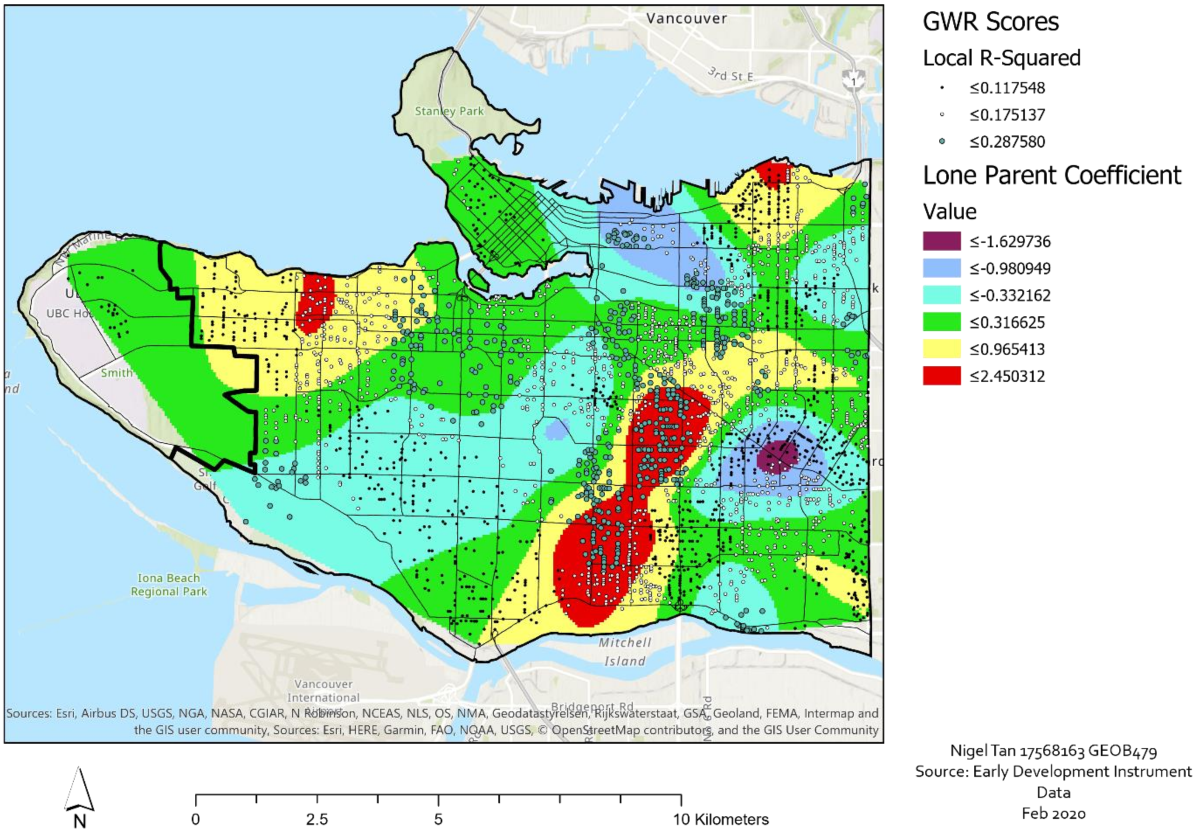


Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument Data
Feb 2020

6. Local R^2 and Lone Parent Coefficient Raster

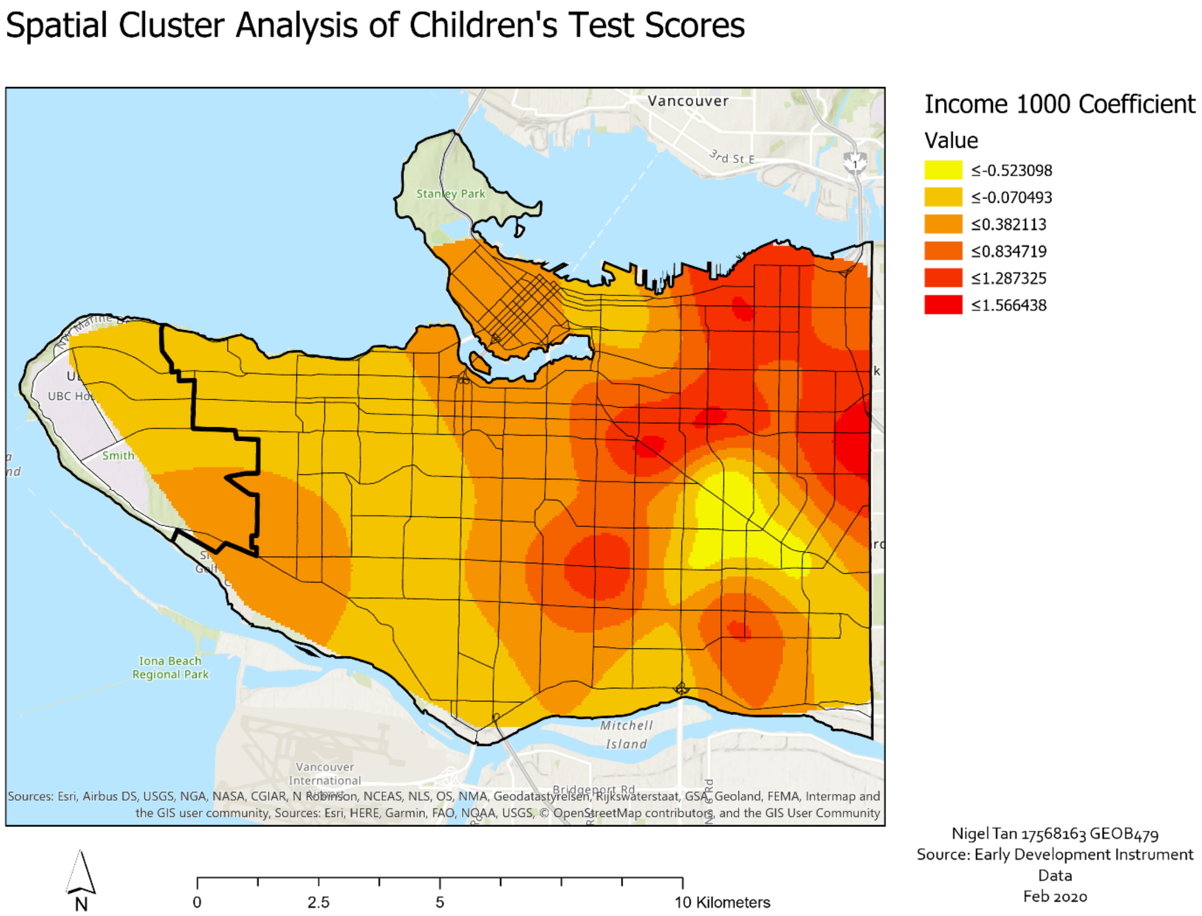
Spatial Cluster Analysis of Children's Test Scores



Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument Data
Feb 2020

7. Income 1000 Coefficient Raster



Sources: Esri, Airbus DS, USGS, NGA, NASA, CGIAR, N Robinson, NCEAS, NLS, OS, NMA, Geodatastyrelsen, Rijkswaterstaat, GSA, Geoland, FEMA, Intermap and the GIS user community, Sources: Esri, HERE, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community

Nigel Tan 17568163 GEOB479
Source: Early Development Instrument Data
Feb 2020

Bibliography

Chen, J., Zhou, C., Wang, S., & Hu, J. (2018). Identifying the socioeconomic determinants of population exposure to particulate matter (PM2.5) in China using geographically weighted regression modeling. Environmental Pollution, 241, 494–503. doi: 10.1016/j.envpol.2018.05.083

Geographically Weighted Regression. (n.d.). Retrieved from https://www.mailman.columbia.edu/research/population-health-methods/geographically-weighted-regression

Szymanowski, M., & Kryza, M. (2011). Application of geographically weighted regression for modelling the spatial structure of urban heat island in the city of Wroclaw (SW Poland). Procedia Environmental Sciences, 3, 87–92. doi: 10.1016/j.proenv.2011.02.016