Applying and Comparing a Geographically Weighted Regression Analysis to Predict a Child's Abilities

Jack Irwin
2-29-2020
Geob 479
Lab 2

Abstract

This analysis used geographically weighted regression (GWR) modelling to determine how a child's language abilities in Vancouver are affected by different variables measured by the Human Early Learning Partnership. The analysis compares the results from the GWR analysis to an ordinary least square (OLS) regression analysis of the same variables to explore the differences in results from each approach. The GWR model for the relationship between a child's social skills score and their language abilities score was the only GWR model to not provide conflicting results with the OLS regression model. All other variables were shown to have locally significant relationships in different parts of Vancouver that appeared to be both positively and negatively related.

Introduction to Geographically Weighted Regression (GWR) Models

According to Tobler's First Law of Geography everything is related in space, but near things are more related than distant things. A GWR model effectively applies this law in spatial analyses. While an OLS or linear regression model in spatial analyses strictly focuses on determining the relationship between a dependant variable and some independent variable(s) *across* space, which renders a model for a global relationship that fails to account for Tobler's stipulation of near observations being more related. In the case of linear regression models used in spatial analyses, all observations across space are equally influential in determining the best fitted model for the dependent variable. A GWR model is a type of regression model that better accounts for the differences in variance across a heterogenous space or landscape by excluding or weighting the influence of observations measured further away.

There are a few ways to define the weighting scheme to deter the influence of further observations in a GWR model. When every measured point in the dataset is given a location a GWR model can exclude observations from a specified distance away. This is a discrete method of spatial weighting that can be used by a GWR model, but Martin Charlton, Stewart Fotheringham, and Chris Brundson (2003) suggest this method fails to represent actual geographical processes because of potential discontinuity. Another method of spatial weighting applies a Gaussian curve to weight the observations' influence according to their distance away from an area being modeled. So, the closer an observation is to areas being modeled, the greater the influence that observation has on the projected measurements to areas formed by the GWR model. However, if all observed points are distant from certain spaces being modelled or observed points are densely clustered in parts of the landscape, then the GWR model may require spatially adaptive weighting. This applies small bandwidths to the areas being modeled near densely clustered observed data and larger bandwidths to the areas being modeled in areas further away from observed data. If a large bandwidth is applied in a GWR model, results likely will not differ significantly from a linear regression model's results, conversely using too small of bandwidths can limit the statistical significance of results. Furthermore, deciding which variables to use as dependent variables in a GWR model should take careful consideration, and different statistical tools such as the Akaike information criterion can be applied in consideration. So, determining the variables, bandwidth, and method of weighting are all essential to running an appropriate GWR model.

After determining the variables and a suitable weighting scheme and finally running a GWR model, the results may yield differing relationships within the data. Where a linear regression model might determine a moderately strong positive relationship between a dependent and some independent variable(s), a GWR model could show that there is actually an extremely strong positive relationship in one region of the study area but a relatively weak relationship throughout the rest of the study area. Furthermore, GWR models can find contradicting relationships between variables in a study area. For example, one part of the study area may show the variables are negatively related, while another part shows the variables are positively related. Making sense of the results from a GWR model requires some knowledge of the local statistical significance of results within the study area. Due to how the observed data is distributed and the weighting scheme, some relationships determined by the GWR may prove less significant in areas.

Results

These results show how a GWR may be applied in determining the degree to which different factors in parts of Vancouver can affect a child’s language abilities. The Early Development Instrument employed by the Human Early Learning Partnership measures scores of language abilities based on whether the child is literate and can recognize numbers and count (EDI webpage, 2020). The independent variables involved in the following comparison between OLS regression and GWR are the child’s social score, household income, and whether English is a second language, the child has a lone parent, and is part of a recent immigrant family. Figure 1.3 depicts how these variables generally show up in loosely defined clusters throughout Vancouver after using the spatially constrained multivariate clustering tool in ArcGIS Pro. The labels of each category in Figure 2.1 were determined from the data of the spatially constrained multivariate cluster boxplots produced from ArcGIS Pro. These categorized areas are referenced in the comparison between GWR and OLS regression results.

Many of the GWR results for these independent variables dispute the global relationships put forth by the OLS results. Table 1.1 shows the linear regression coefficients for each independent variable that characterize the OLS regression's analysis of the relationship between the independent variable and the child's language abilities. Given the OLS regression results in Table 1.1, one might presume a child's language abilities are not affected by a household's income or whether the child is a member of a recent immigrant family. Instead, they might presume there is a stronger positive relationship between a child's language abilities and whether English is a second language for a child and a moderately negative relationship between associated with whether the child has a lone parent. However, the GWR results show high and low coefficients ranging between positive and negative values for these four independent variables. This suggests these explanatory variables may have opposing effects on a child's language abilities in separate parts of Vancouver and contest the OLS regression results that show either no significant effect or one conclusive positive or negative relationship in the model.

Figure 1.1 examines the GWR case for income. The color of the raster cells surrounding the white dots are more statistically significant than the smaller cells surrounding grey or black dots. The income map notably shows about every range (every color of raster cells) of predicted GWR coefficients in some area of the map surrounding multiple white dots. This confirms the OLS regression results which calculated a small coefficient representing no significant relationship between income and a child's language abilities *for all* of Vancouver. However, Figure 1.2 provides a nuanced view showing different relationships in pockets of Vancouver to be significantly positive, miniscule, and even slightly negative in the Low Income spatially constrained cluster of Vancouver. This variability could be due to the fact that smaller changes between observed income disproportionally affect the model in the Low Income cluster more than in the rest of Vancouver. It could also be due to differences between neighborhood schools within Vancouver, which is a major confounding variable in this study.

Figure 1.2 examines the GWR case for a child's social skills score as the explanatory variable. First, according to Table 1.1 the range for social skills' GWR high and low coefficients is relatively small, and the low coefficient stays above zero. So, the OLS regression model's coefficient for social skills is actually confirmed by the GWR results. However, the GWR still provides better insight as to where the positive relationship is stronger versus weaker. Figure 1.2 shows a choropleth map indicating a stronger positive relationship between a child's language abilities and social skills in the Lower Income section of Vancouver that is also more statistically significant given the distribution of higher R-squared valued measurements.

The GWR results described in this study help explain why OLS regression models are often less applicable in predicting outcomes at local levels. Furthermore, the GWR results have shown where relationships are found to be stronger versus weaker within a study area, which can allow researchers to find nuanced relationships and also better recommend areas to focus future work on to uncover potential confounding variables associated with particular regions.

Discussion and Other Applications for GWR

Beyond analyzing what variables may affect a child's language abilities in a city, GWR has a variety of applications and is commonly employed in research devoted to urban planning, health geography, and economics. For instance, Osvaldo Daniel Cardozo, Juan Carlos García-Palomares and Javier Gutiérrez (2012) used GWR modeling to forecast transit ridership along twelve lines in the Madrid Metro network. Cardozo et al. (2012) found out of nine possible explanatory variables that the number of suburban bus lines, total bus lines, places of employment, and number of workers in an 800m radius of transit line station stops significantly correlated with ridership. Through their GWR model they forecasted stronger positive relationships in the number of suburban bus lines, total bus lines, and places of employment in central and northern Madrid, while the number of workers showed a stronger positive relationship with transit ridership at stops in southern Madrid (Cardozo et al., 2012). Urban planners could use their GWR model to forecast how different types of urban developments across Madrid might affect transit ridership at specific train stops, which could then help the city's public transit sector schedule trains more accordingly.

GWR may also be useful in health geography research. As part of a study on chronic obstructive pulmonary disease (COPD) in Germany, Boris Kauhl, Werner Maier, Ju¨rgen Schweikart, Andrea Keste, and Marita Moskwyn (2018) examined the possible associations between different demographic and socioeconomic factors and COPD using a GWR model. The variables deemed risk factors used by Kauhl et al. (2018) included household size, insurants aged 65 and older, insurants with migration background, and area deprivation which characterizes the economic deprivation. Through their GWR model Kauhl et al. (2018) found a strong association between COPD instances and areas with higher proportions of elderly populations and economic deprivation. They determined elderly populations in disadvantaged areas are at a higher risk of getting COPD regardless of individual socioeconomic traits (Kauhl et al., 2018). In this case, the application of GWR helped researchers find commonalities in the spatial pattern of association between potential risk factors associated with a deadly lung disease.

One final example of the diversity of GWR application includes its utility for economists and policymakers to study human consumption patterns. Selima Sultana, Nastaran Pourebrahim, and Hyojin Kim (2018) used GWR to explore how economic, demographic, and spatial variables relate to energy consumption for households in fourteen different North Carolina metropolitan statistical areas (MSAs). First Sultana et al. (2018) divided energy consumption into two categories of utility expenditure and transportation and determined the variables that would affect each category of energy consumption. For example, income measures and physical household characteristic variables were used for utility expenditure variables, while spatial variables such as distance from city center were used for transportation (Sultana et al., 2018). The GWR results revealed which MSAs experienced heightened energy consumption devoted to transportation due to sprawl, while the OLS regression results only showed a somewhat positive relationship between transportation energy use and a house further located from city center amongst all MSAs (Sultana et al., 2018). This research also points out that policymakers strategizing ways to reduce utility energy consumption can begin by analyzing which GWR coefficients associated with detached versus attached or multifamily housing are highest and create local policies more applicable to each MSA.

These recent studies show how GWR can be used in different contexts from research in health to economics. Sultana et al. (2018) mentioned the GWR model producing better results compared to the OLS regression, but the OLS regression remains useful for its simplicity in interpretation. All studies mentioned effectively displayed how a GWR model provides a nuanced analysis, but simultaneously increases the complexity for interpretation of the results.

Appendix

Table 1.1

Variable	GWR High Coefficient	GWR Low Coefficient	GWR Coefficient Range	Linear Regression Coefficient
English Second Language	10.9939	-2.18108	13.17498	5.4519
Social Score	0.87506	0.44034	0.43472	0.6219
Lone Parent	1.99519	-1.79068	3.78587	-0.3162
Recent Immigrant	0.369193	-0.21265	0.581846	0.0674
Income	2.14084	-0.37184	2.51268	0.0867

Fig. 1.1

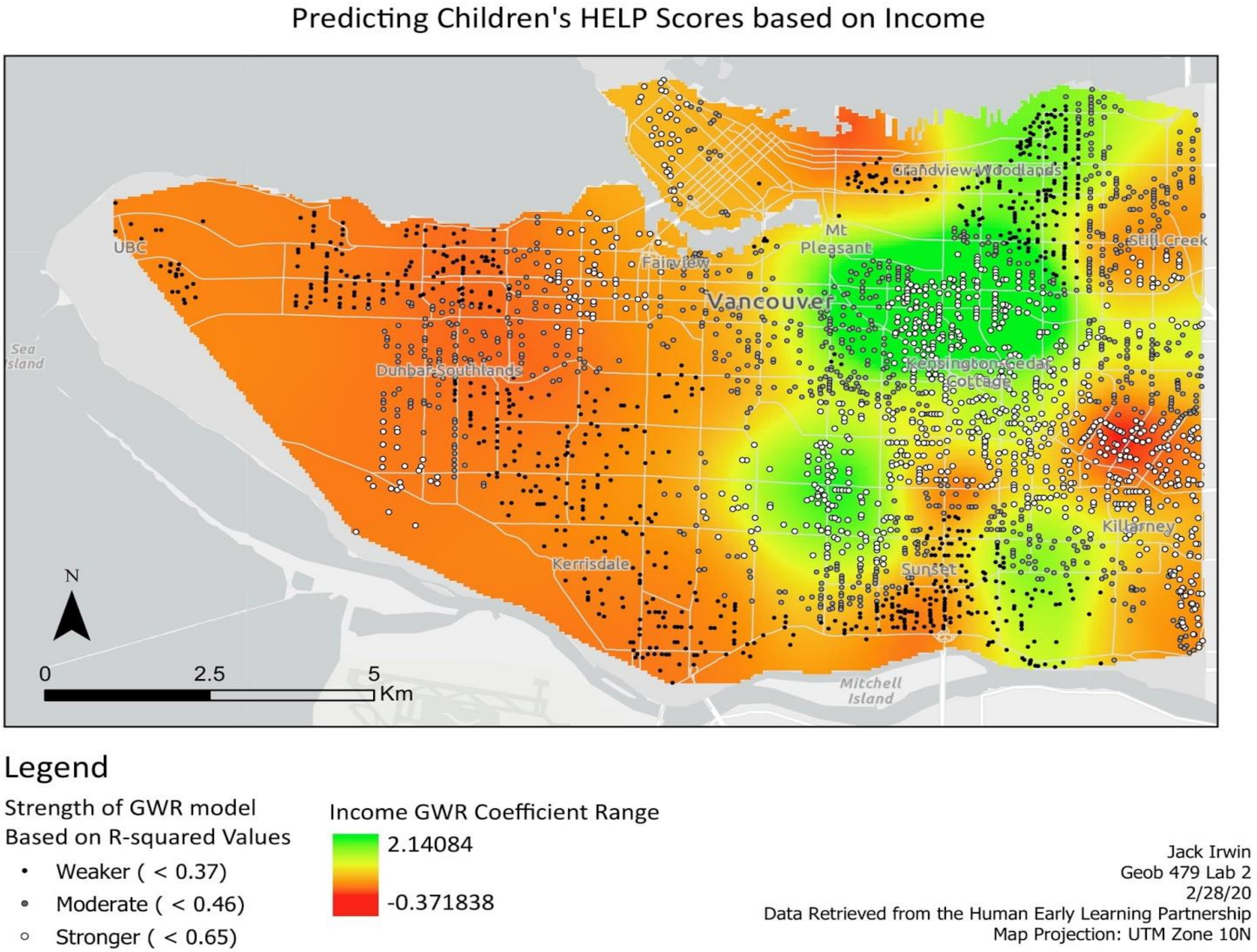
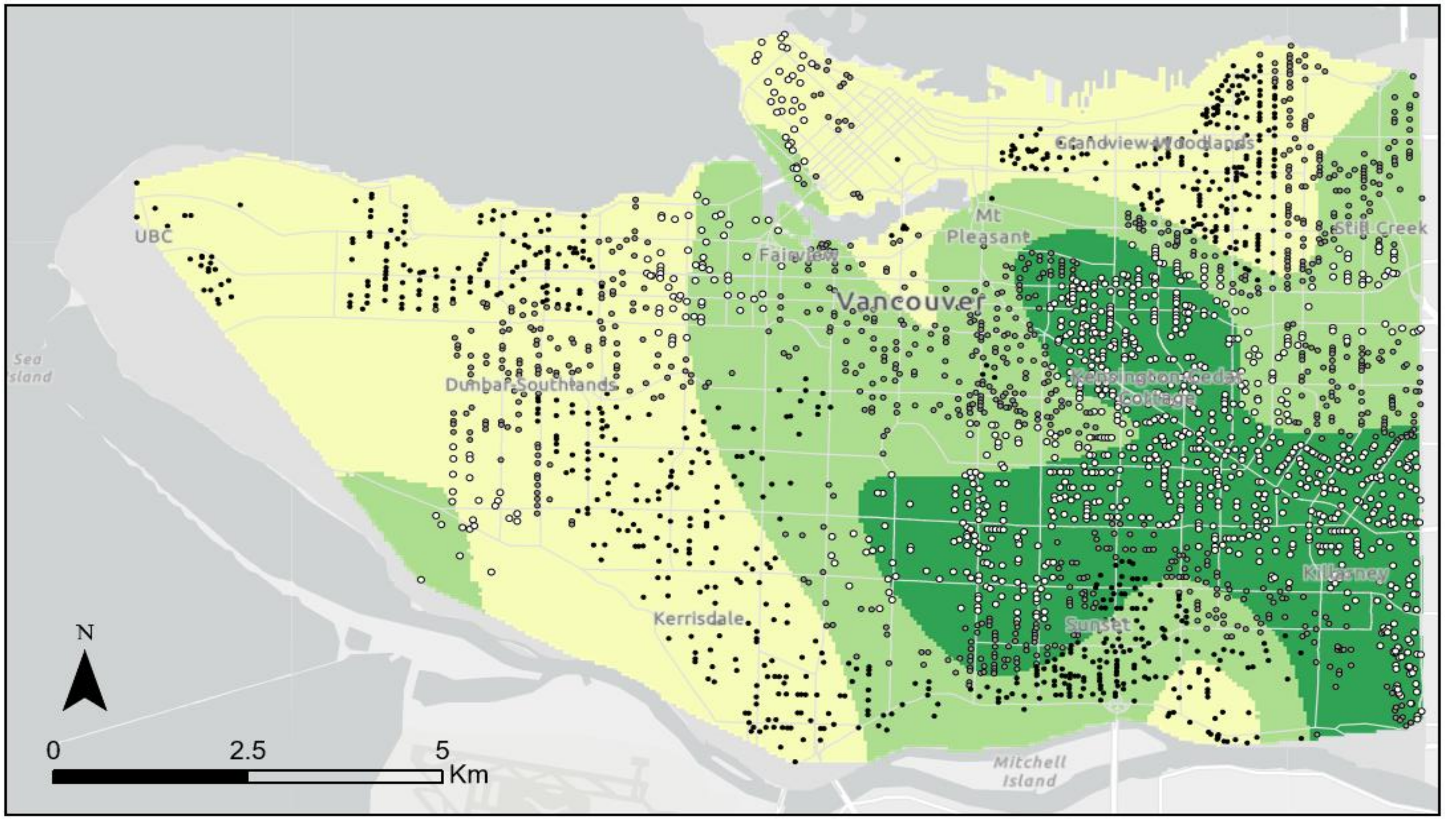

Predicting Children's HELP Scores based on Income
UBC
Fairview
Mt Pleasant
Grandview-Woodlands
Vancouver
Still Creek
Dunbar-Southlands
Kensington-Cedar Cottage
Sea Island
Killarney
Kerrisdale
Sunset
Mitchell Island
N
0
2.5
5
Km
Legend
Strength of GWR model Based on R-squared Values
• Weaker (< 0.37)
● Moderate (< 0.46)
○ Stronger (< 0.65)
Income GWR Coefficient Range
🟩 2.14084
🟥 -0.371838
Jack Irwin
Geob 479 Lab 2
2/28/20
Data Retrieved from the Human Early Learning Partnership
Map Projection: UTM Zone 10N

Fig. 1.2

Predicting Children's HELP Scores based on their Social Skills Criterion



Legend

Strength of GWR model Based on R-squared Values

- • Weaker (< 0.37)
- ● Moderate (< 0.46)
- ○ Stronger (< 0.65)

Social Skills GWR Coefficient Range

- 🟩 ~(0.67 - 0.875)
- 🟩 ~(0.55 - 0.67)
- 🟨 ~(0.44 - 0.55)

Jack Irwin
Geob 479 Lab 2
2/28/20
Data Retrieved from the Human Early Learning Partnership
Map Projection: UTM Zone 10N

Fig. 2.1

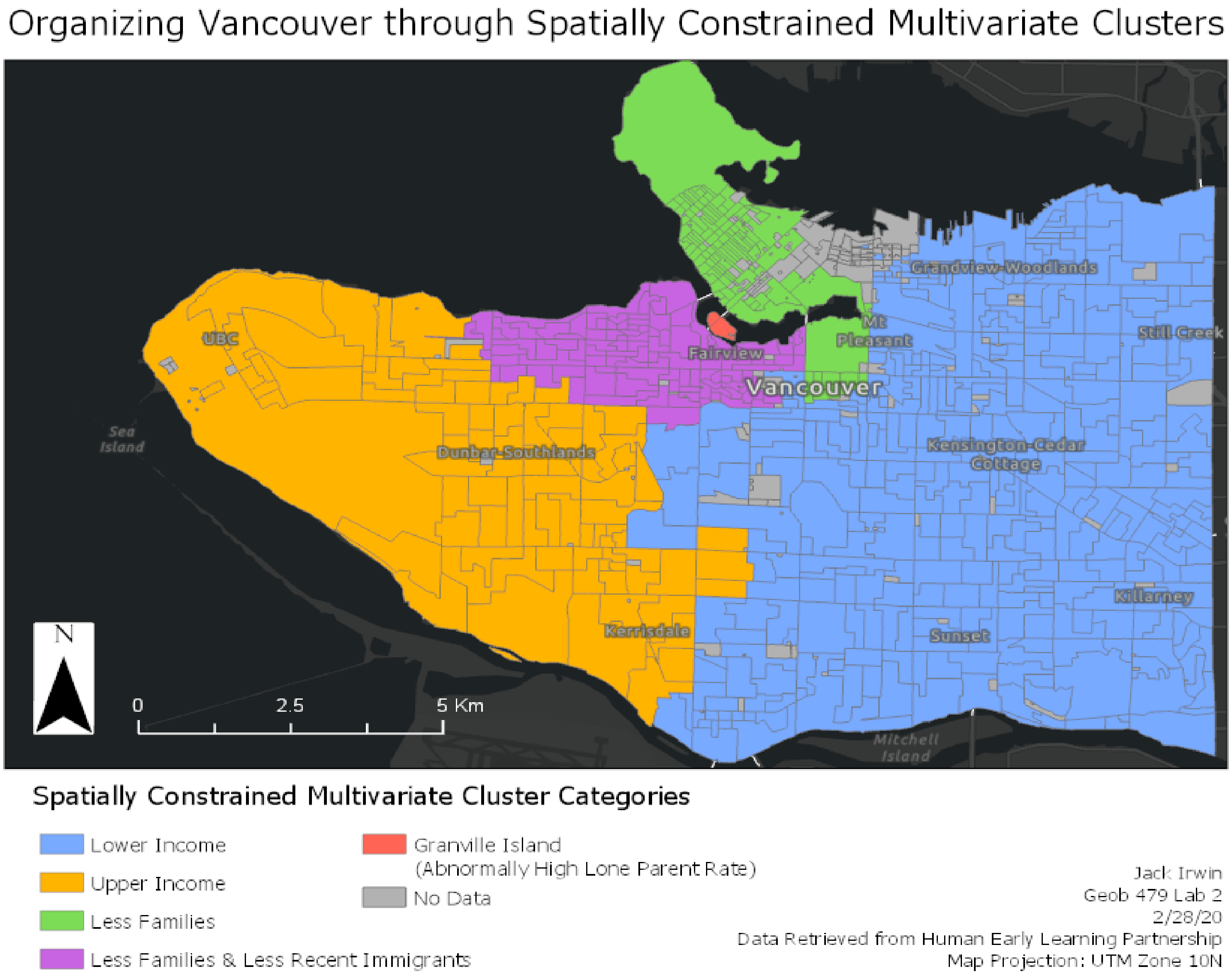

Organizing Vancouver through Spatially Constrained Multivariate Clusters
Grandview-Woodlands
Mt Pleasant
Still Creek
UBC
Fairview
Vancouver
Sea Island
Dunbar-Southlands
Kensington-Cedar Cottage
Killarney
Kerrisdale
Sunset
Mitchell Island
N
0
2.5
5 Km
Spatially Constrained Multivariate Cluster Categories
Lower Income
Upper Income
Less Families
Less Families & Less Recent Immigrants
Granville Island (Abnormally High Lone Parent Rate)
No Data
Jack Irwin
Geob 479 Lab 2
2/28/20
Data Retrieved from Human Early Learning Partnership
Map Projection: UTM Zone 10N

Fig. 2.2

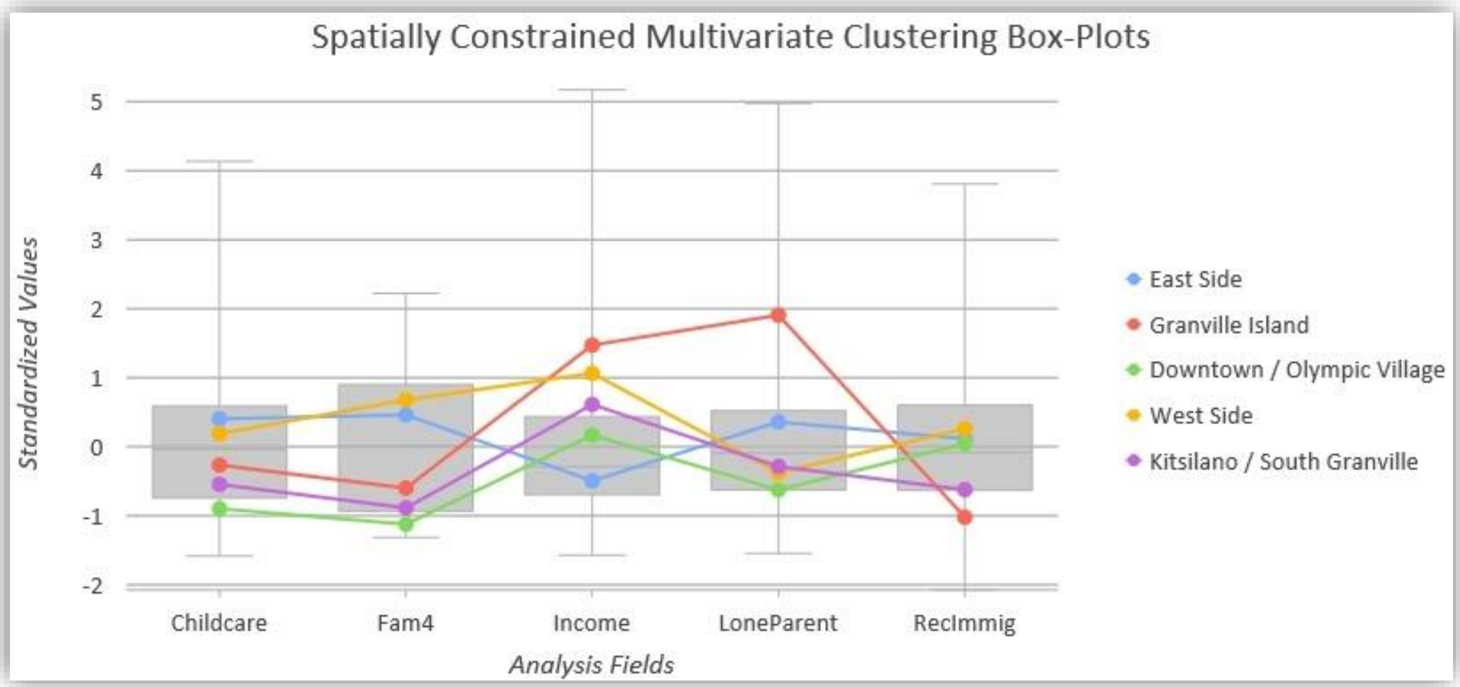


*Figure 2.2 helps show how the spatially constrained multivariate clustering tool organized Vancouver on the map. The legend categories for Figure 2.1 were termed accordingly with the characteristics that stood out in this box plot, rather than termed by geographic region as this box plot originally shows.

Works Cited

Cardozo, O. D., García-Palomares, J. C., & Gutiérrez, J. (2012). Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, *34*, 548-558. doi:10.1016/j.apgeog.2012.01.005

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

Kauhl, B., Maier, W., Schweikart, J., Keste, A., & Moskwyn, M. (2018). Who is where at risk for chronic obstructive pulmonary disease? A spatial epidemiological analysis of health insurance claims for COPD in northeastern germany. *Plos One*, *13*(2), e0190865. doi:10.1371/journal.pone.0190865

Sultana, S., Pourebrahim, N., & Kim, H. (2018). Household energy expenditures in north carolina: A geographically weighted regression approach. *Sustainability*, *10*(5), 1511. doi:http://dx.doi.org/10.3390/su10051511