Exploring Influential Variables on Children’s Social Skills Using Geographically Weighted Regression in Vancouver, BC

Alexander Coster

Introduction

Geographically Weighted Regression (GWR) is a regression technique for exploratory spatial data analysis. It is a form of linear regression analyses: it has predictor (independent) variables and a target (dependent) variable. The overall idea of regression is to answer two things: do the predictor variables adequately predict the dependent variable, and which ones are significant predictors. The ultimate objective is to find the best straight line relationship to explain how the variation in the outcome depends on variation of a predictor. For example, one could look at housing values as a dependent variable, and investigate age, distance to amenities, or local crime statistics as independent variables. Software such as ArcMap allows users to conduct the regression, and interprets the regression assumptions and output. This leads to greater understanding of the strength of the predictors, and grants the ability to forecast effects and trends, and explain complex phenomena.

However, a linear regression model assumes that the relationship in the formula is constant everywhere in the study area (regression parameters are “whole map” statistics). It produces a single parameter estimate for all relationships specified in the model; parameter estimates do not vary over space. This is what staticians would call a ‘global’ model: the predictor variables would have the same effect on the dependent variable regardless of where they are. It provides useful statistics, but it cannot be mapped. One may find that a regression analysis has a low coefficient of determination (model does not account for a significant variable), and that variable may be due to a missing spatial element.

Geographically Weighted Regression is like regression but with the spatial element added. The outcome is a set of estimated parameter values for each explanatory variable at every location obtained via weighted least squares (Fotheringham & Oshan, 2016). It includes distance, number of neighbours, bandwidth method (how great an influence surrounding areas have on the locally specific parameter estimates) and kernel type (fixed distance or varying as a function of feature density). It is like regression that is also sensitive to geographic relations among the variables: nearby value locations will influence the regression more than far away ones. This allows parameters to vary over space in a traditional OLS regression framework and specifies a separate regression model at every observation point, thus enabling unique coefficients to be estimated at each observation location (Bitter et al, 2006).

GWR emphasizes non-stationarity (relationships vary across the study area) (Liu et al, 2016). It simultaneously addresses spatial non-stationarity issues and demonstrates the spatial relationship between different variables. These location-specific parameters are mappable, so the user may visually investigate potentially spatially varying processes and examine the results using ArcGIS. The optional raster coefficient surface generated will show how relationships between the dependent variable and each explanatory variable fluctuate across a heterogeneous landscape. This can provide clues about potentially missing explanatory variables, depending on the pattern of coefficients of determination across the landscape. For example, the independent variables may better explain why children suffer from asthma in areas near freeways than in rural areas. GWR can also be used for prediction by supplying prediction explanatory variables on a one-to-one correspondence with the explanatory variables used to calibrate the model. Once your explanatory model is finished, you can tweak the predictor variables to predict the outcome of the newly calibrated predictor variables.

Data & Methods

This analysis was performed on data collected from the Human Early Learning Partnership: Early Development Instrument. 2815 individual observation points (exact locations modified to preserve anonymity) were used, which included a variety of standardized variables: physical abilities, social skills value, % of neighbourhood that are lone parents, % that are recent immigrants, % that are visible minorities, gender, ESL, % that are families of four or more, average income divided by 1000, language abilities, % that are immigrants, and % that spend 30 hours or more on childcare.

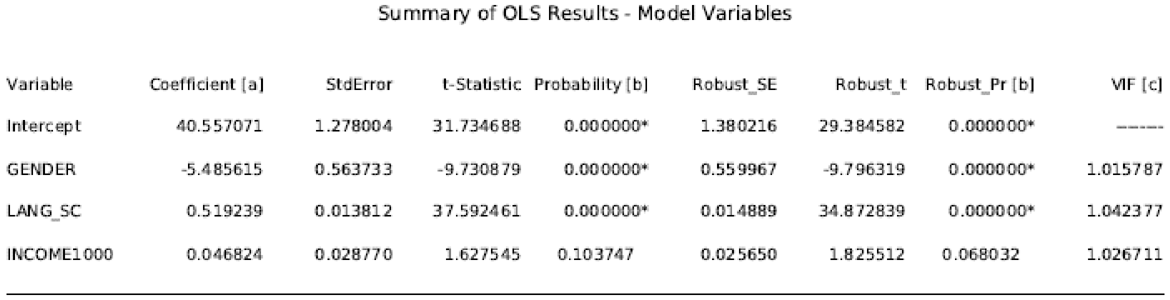
Using ArcGIS v10 software, two hierarchical explanatory regression analyses were performed on selected variables. Average income, gender, and language abilities were found to be the most significant independent variables (highest Adj.R2 and lowest AIC ratings). An Ordinary Least Squares (OLS) tool is used to determine the statistics associated with the best set of variables, make an output feature class, and produce a coefficient output table. The

Geographically Weighted Regression (GWR) tool is used to explore the spatial nature of relations using the same variables. A Grouping Analysis tool was used to create enumeration area ‘neighbourhoods’ using childcare, family of four, lone parent, recent immigrant, and income variables. Now, the analysis has produced global statistics (Figure 1), local statistics for each observation, four enumeration area classes (Table 1), and coefficient rasters for income, gender, and language skills (Maps 3, 4, & 5).

OLS Results

The OLS tool used three variables: gender, language skills, and income (neighbourhood average divided by 1000). OLS is aspatial: only one regression model is used to model the relationship, and it assumes the relationship between each variable is constant across space. According to the OLS statistics, gender has a strong negative correlation with social skills while language skills has a positive correlation (Figure 1). Given its low t-statistic (indicates the coefficient is not significant), low coefficient score, and high b probability, the income variable is deemed an insignificant variable in social skills scores. Mapping OLS standard deviation residual scores beneath GWR R2 scores gives an idea where the analyses found different results, and where spatial relationships caused a split from global statistics. In areas like Shaughnessy and South Granville, the OLS residual scores are closer to the mean, while GWR R2 values are low: OLS model predicts more accurately in these areas than GWR. In eastern Vancouver (Renfrew Heights, Collingwood, Victoria, Grandview) where R2 values are high, OLS residuals remain around 1 standard deviation away from the mean: the GWR performs better in these areas, suggesting a stronger spatial element to the statistics (Map 1).

Table 1



Summary of OLS Results - Model Variables

Variable	Coefficient [a]	StdError	t-Statistic	Probability [b]	Robust_SE	Robust_t	Robust_Pr [b]	VIF [c]
Intercept	40.557071	1.278004	31.734688	0.000000*	1.380216	29.384582	0.000000*	--------
GENDER	-5.485615	0.563733	-9.730879	0.000000*	0.559967	-9.796319	0.000000*	1.015787
LANG_SC	0.519239	0.013812	37.592461	0.000000*	0.014889	34.872839	0.000000*	1.042377
INCOME1000	0.046824	0.028770	1.627545	0.103747	0.025650	1.825512	0.068032	1.026711

Grouping Analysis Results

The Grouping Analysis tool created four groups where all the features in each group are as similar as possible, and all the groups are different as possible (Table 1). This was done using five variables: % of households that spend 30 hours or more on childcare, % of households that have families of four or more, % of households that have a single parent, % of households that are recent immigrants, and average income divided by 1000.

Table 1

Group Numbe r	Name	Description
One	**Small, middle class, established families**	Low Family/Four, Above Average Income, Low % Recent Immigrant, Low Childcare, Low Lone Parent
Two	**Larger, lower class, established, single-parent families**	Above Average Family/Four, Low Income, Below Average % Recent Immigrant, High Childcare, High Lone Parent
Three	**Small, middle class, immigrant families**	Below Average Family/Four, Above Average Income, High % Recent Immigrant, Below Average Childcare, Below Average Lone Parent
Four	**Larger, upper class, immigrant families**	High Family/Four, High Income, Above Average % Recent Immigrant, Above Average Childcare, Below Average Lone Parent

Group 2 (see Table 1) covers much of east and south Vancouver (Renfrew Heights, Victoria, Killarney, Fraser, Main, Renfrew, Hastings East, Grandview). GWR local R2 values are fairly consistently high in areas covered by Group 2: GWR models these areas well, suggesting there is a stronger spatial relationship between variables here (Map 2). Group 4 covers Kerrisdale, Shaughnessy, South Granville, Mackenzie Heights, and Dunbar. There is a pattern of lower local R2 values generated by the GWR tool in these areas; the GWR did not perform as efficiently here. Group 3 areas (Marpole, False Creek, Yaletown) are fairly accurate, but not great.

Geographically Weighted Regression (GWR) Results

Based on the spatially differing coefficients of the GWR, there is disagreement with the OLS based on the nature of relationship (positive or negative) and on the degree of influence of each variable (different coefficient values). There is obviously some heterogeneity in the model. The GWR analysis produced three raster maps based on social skills as a function of gender

(Map 3), income (Map 4) and language skills (Map 5). These rasters represent the fluctuating coefficients of each variable in conjunction with social skills across space. GWR has produced different coefficient scores than OLS in all three variables: where the OLS simply stated that gender, for example, has a coefficient of -5.49, the GWR fluctuates from 2.25 in some areas (South Vancouver, Hastings, Knight, Victoria) to -17.17 (Grandview, West End) (Map 3). GWR R2 values show a pattern of low scores in the more positive coefficient areas for gender, though this pattern is fairly inconsistent. This suggests a reduced significance of gender in these areas. High R2 scores exist both in negative coefficient areas (Grandview, West End) and positive areas (Victoria, Collingwood). Gender probably has an influence in these areas, given their significant negative and positive correlation values.

The income coefficient raster has a fairly discernable pattern where high R2 values are most prominent in areas where income coefficients are lowest, close to -1.79 (Grandview, Victoria, Knight & Collingwood) (Map 4). In these areas, income is expected to have a significant negative influence on social skills. However, R2 values are also high in positive coefficient areas (Renfrew), meaning income probably has a positive influence in these areas. Lower R2 values cover areas of average coefficient levels (University, Point Grey, Kerrisdale, South Granville, DT Eastside), suggesting the low influence of income in these areas. The income raster is somewhat less useful due to the fact that average income levels cover most of the city, and R2 values of all three classes exist in these areas. These inconsistent patterns reflect the degree that nonstationarity and spatial factors have on the dependent variable, as opposed to the global statistics supplied by OLS.

The language skills coefficient raster produced perhaps the most discernible results. There is a distinct pattern of low R2 values overtop areas of low language skills coefficients (Shaughnessy, South Granville, University, Marpole, South Vancouver, Grandview) (Map 5). Language is unlikely to have an influence on social skills in these areas. High R2 values consistently cover areas of higher coefficient scores (Victoria, Renfrew Heights, Kitsilano, False Creek, West End, Oakridge): language skills are likely to have a positive influence on social skills in these areas.

Geographically Weighted Analysis in Different Contexts

GWR can be used in conjunction with OLS in a variety of contexts to identify apparent spatial patterns, or simply explore the use of GWR to attain a higher coefficient of determination.

Environment & health, identification of poverty factors, and relationships between alcohol outlet densities and violence are three distinct contexts where GWR can help find spatial patterns and formulate hypotheses.

Environment & Health

Geographically Weighted Regression can be used in the study of pollutant dispersal in urban settings. Directly monitoring and gauging ground-level fine particle dispersion for air quality assessments is difficult due to the limited number of ambient monitoring sites. However, pollutant dispersal research using remote sensing has lead to a greater understanding of relationships between ground-level particulate rates and aerosol optical depth, meteorological fields (boundary layer height, relative humidity, air temperature, and wind speed), and land use information (road length, land use types, and population density) (Hu et al, 2013). Since these relationships are non-stationary, GWR was used to examine the spatial variability and nonstationarity by producing local regression results (2013). A particulate surface was generated to predict ground-level particulate concentrations of study area covering parts of Georgia, Tennessee and Alabama (2013). GWR generated a higher correlation coefficient than other non-spatial regression analyses, but found that model accuracy did not change significantly with the removal of the aerosol optical depth variable. Humidity, lower boundary layer height, wind speed, forest cover, and air temperature were the significant predictor variables. Given similar weather conditions, researchers can predict ground-level fine particle dispersion for urban settings where there are nearby polluters.

Identification of Poverty Factors

Poverty, and the role that the agricultural sector has on it, can be studied with geographically weighted regression. The objective of a study in Jambi Province, Indonesia, was to formulate strategies for poverty alleviation. Previous models attempting to pinpoint factors influencing poverty used the global approach, which were inappropriate given Indonesia’s different typology and geographical locations of Indonesian islands; poverty has a spatial dimension due to the highly heterogeneous landscape (Nashwari et al, 2017). The spatial approach does not relate only to distance, location and situation, access, correlation and pattern, but also to spatial pattern, correlation between variables, and spatial process (2017). Independent factors included road systems, location in upland area, and rainfall; the dependent variable was the percentage of poor food crop farmers. The study found that upland areas with poor access and low rainfall had the greatest percentage of poverty stricken farmers, though road access had both positive and negative impacts on poverty. Interestingly, the more villages near roads the higher the poverty index; there is a spatial autocorrelation element as well. Overall, the GWR model improved the coefficient determination, and lowered the AIC and Mean Square Error over global regression methods.

Alcohol Outlets & Violence

Spatial relationships between alcohol outlet density and violence (measure by police activity) can be investigated using GWR analyses. In the analysis, police attended incidents are the dependent variable, outlets are the independent variables. While violence varies significantly across space for off-licences and restaurants/cafes, bars and nightclubs do not show significant spatial variation. However, bar and nightclub density have significant positive relationships with violence (increase in number of police attended incidents) (Cameron et al, 2015). The analysis was modeled in New Zealand, using police data and census area unit data. Studies such as these can help city planners who are issuing liquor licenses: this sort of information can help guide planners to avoid grouping bars and clubs together in order to prevent violence.

Works Cited

Bitter C, Mulligan GF, Dall'erba S (2007) Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems,* 9:7-27. Retrieved from https://search-proquest-com.ezproxy.library.ubc.ca/docview/230098949?pq-origsite=summon&accountid=14656

Cameron, M. P., Cochrane, W., Gordon, C., & Livingston, M. (2015). Alcohol outlet density and violence: a geographically weighted regression approach. *Drug and Alcohol Review,* 35: 180-288. DOI: 10.1111/dar.12295

Fotheringham, A. S., & Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: dispelling the myth. *Journal of Geographical Systems*, 18: 303-329. Retrieved from https://search-proquest-com.ezproxy.library.ubc.ca/docview/1822609189?pq-origsite=summon&accountid=14656

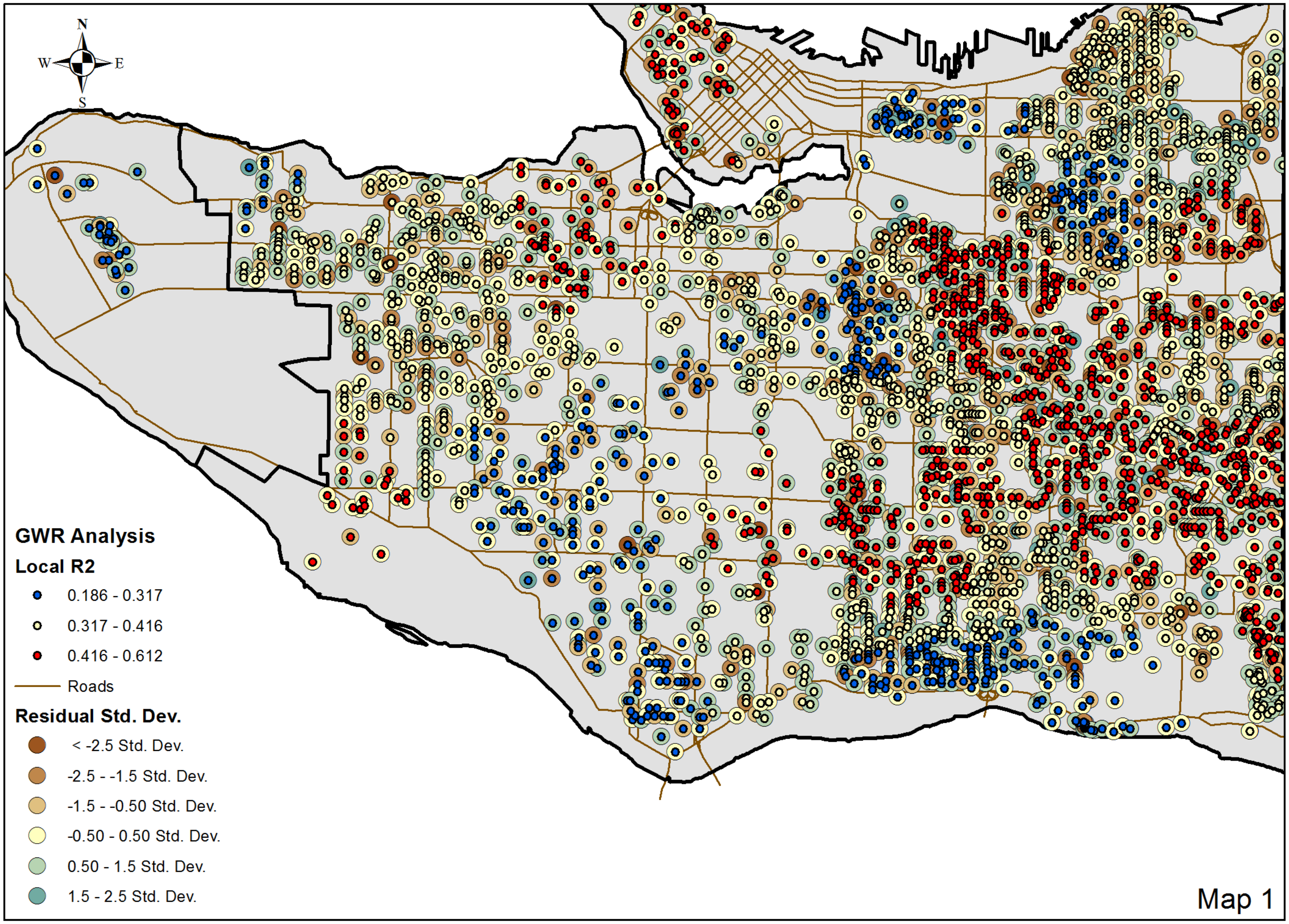
Hu, X., Waller, L. A., Al-Hamdan, M. Z., Crosson, W. L., Estes Jr., M. G., Estes, S. M., Quattrochi, D. A., Sarnat, J. A., & Liu, Y. (2013). Estimating ground-level PM2.5 concentrations in the southeastern U.S. using geographically weighted regression. *Environmental Research*, 121: 1-10. https://doi.org/10.1016/j.envres.2012.11.003.

Liu, J., Zhao, Y., Tang, Y., Xu, S., Zhang, F., Zhang, X., Shi, L., & Qiu, A. (2017). A mixed geographically and temporally weighted regression: exploring spatial-temporal variations from global and local perspectives. *Entropy*, 19: 1-20. Retrieved from http://www.mdpi.com/1099-4300/19/2/53

Nashware, I. P., Rustiadi, E., Siregar, H., & Juanda, B. (2017). Geographically weighted regression model for poverty analysis in Jambi Province. *Indonesian Journal of Geography,* 49: 42-50. DOI: http://dx.doi.org/10.22146/ijg.10571

Geographically Weighted Regression Analysis of Children's Social Skills, Vancouver, BC

R2 Results overtop OLS Standard Deviation Residual Values



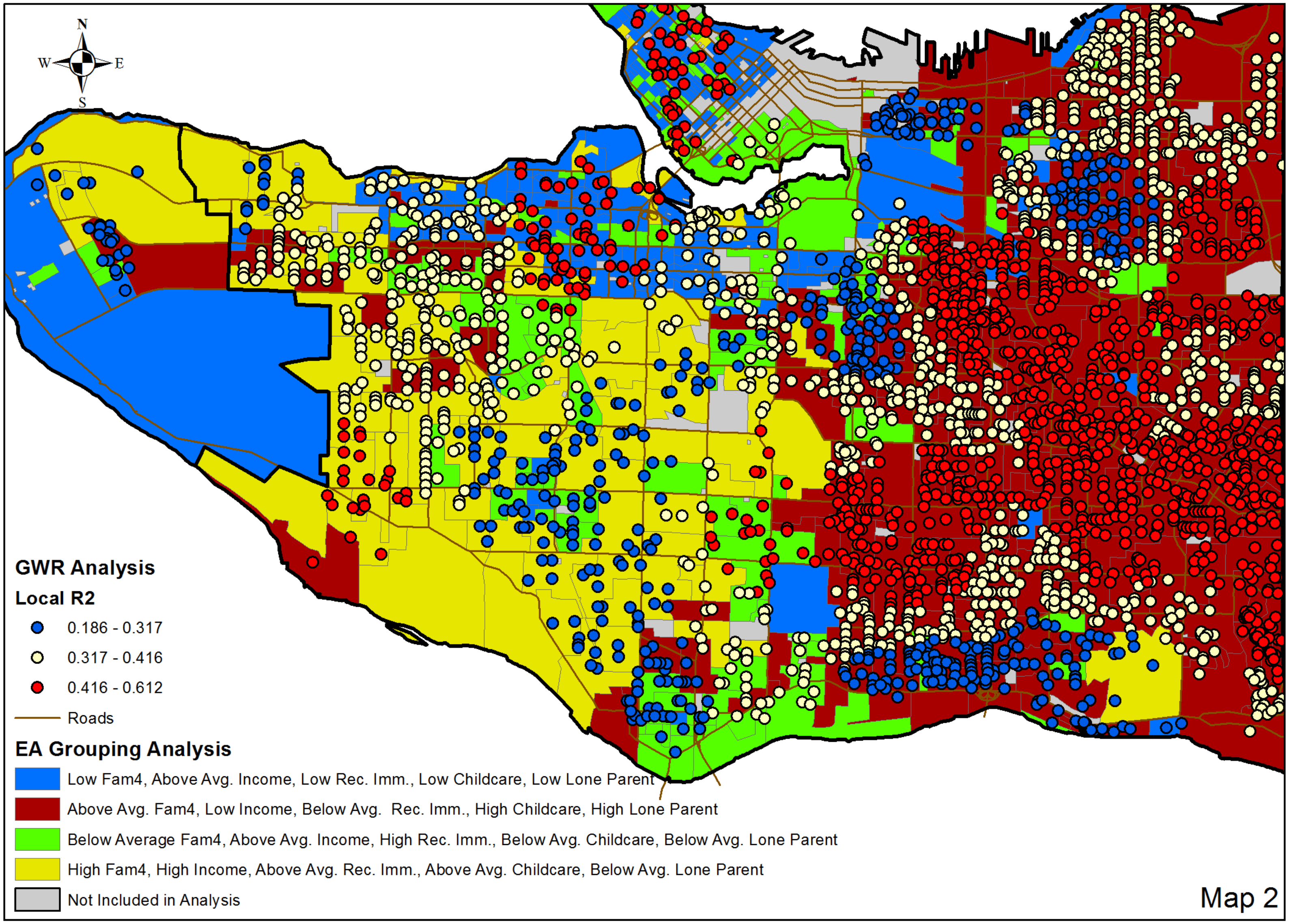
Data retrieved from Human Early Learning Partnership:
Early Development Instrument

NAD 1983
UTM Zone 10 N

Alexander Coster
February 9, 2018
GEOB 479: Research in GIS

Geographically Weighted Regression Analysis of Children's Social Skills, Vancouver, BC

R2 Results overtop Enumeration Area Groups


N
W
E
S
GWR Analysis
Local R2
● 0.186 - 0.317
○ 0.317 - 0.416
● 0.416 - 0.612
Roads
EA Grouping Analysis
🟦 Low Fam4, Above Avg. Income, Low Rec. Imm., Low Childcare, Low Lone Parent
🟥 Above Avg. Fam4, Low Income, Below Avg. Rec. Imm., High Childcare, High Lone Parent
🟩 Below Average Fam4, Above Avg. Income, High Rec. Imm., Below Avg. Childcare, Below Avg. Lone Parent
🟨 High Fam4, High Income, Above Avg. Rec. Imm., Above Avg. Childcare, Below Avg. Lone Parent
⬜ Not Included in Analysis
Map 2
0
1.25
2.5
5
Kilometers

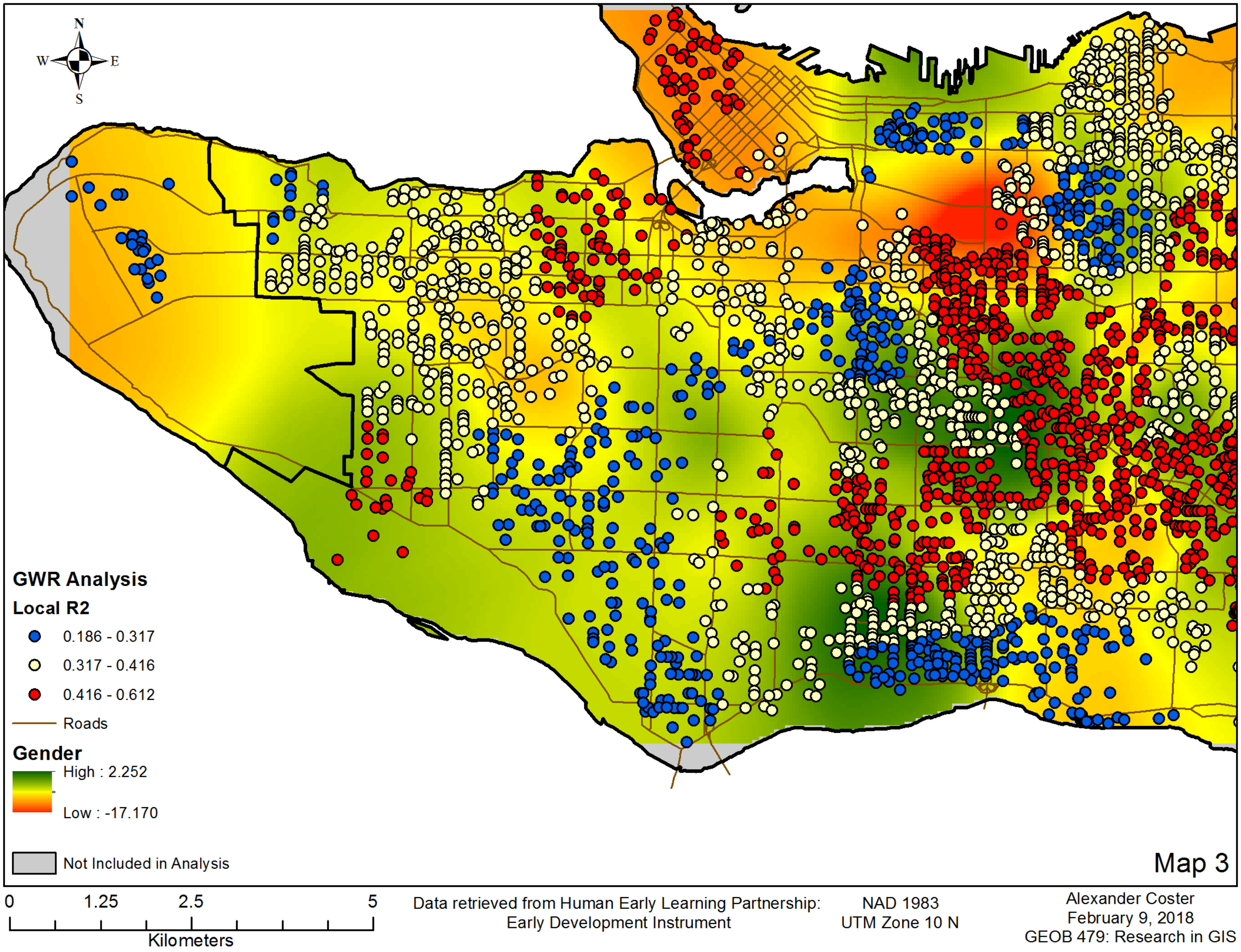
Data retrieved from Human Early Learning Partnership:
Early Development Instrument

NAD 1983
UTM Zone 10 N

Alexander Coster
February 9, 2018
GEOB 479: Research in GIS

Geographically Weighted Regression Analysis of Children's Social Skills, Vancouver, BC

R2 Results Overtop Social Skills as a function of
Gender Coefficient Raster



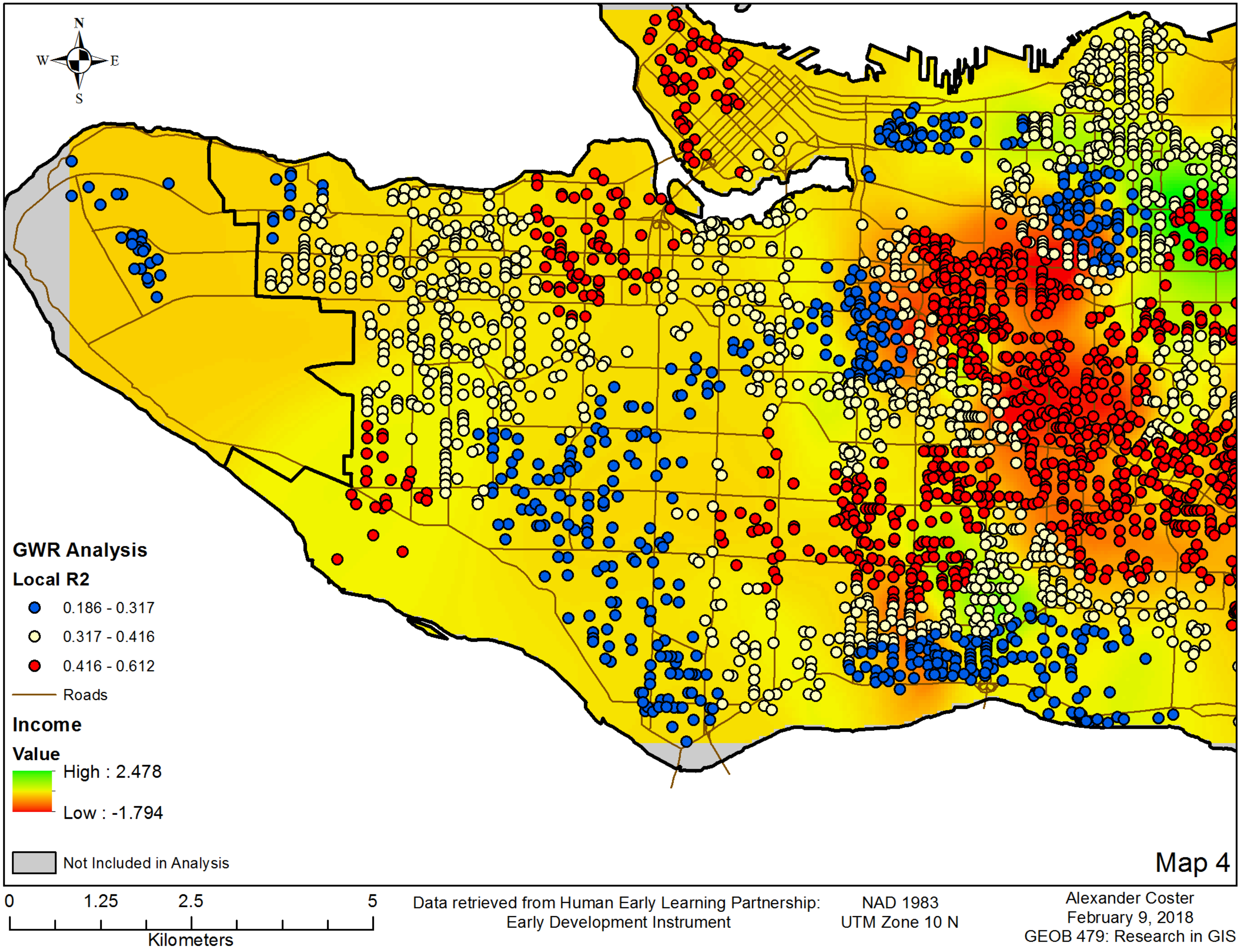
Data retrieved from Human Early Learning Partnership:
Early Development Instrument

NAD 1983
UTM Zone 10 N

Alexander Coster
February 9, 2018
GEOB 479: Research in GIS

Geographically Weighted Regression Analysis of Children's Social Skills, Vancouver, BC

R2 Results Overtop Social Skills as a
function of Income Coefficient Raster



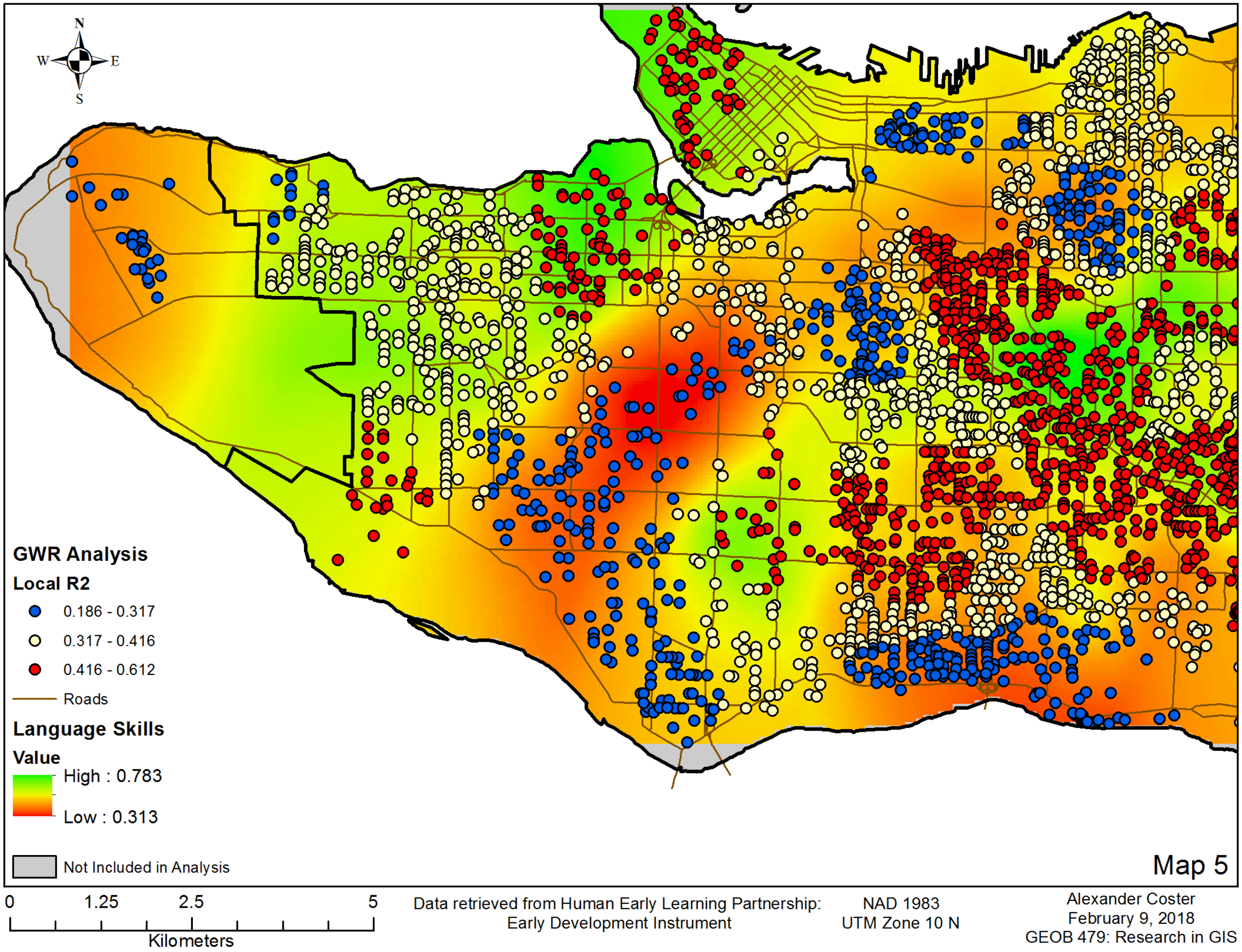
Data retrieved from Human Early Learning Partnership:
Early Development Instrument

NAD 1983
UTM Zone 10 N

Alexander Coster
February 9, 2018
GEOB 479: Research in GIS

Geographically Weighted Regression Analysis of Children's Social Skills, Vancouver, BC

R2 Results Overtop Social Skills as a
function of Language Skills



Data retrieved from Human Early Learning Partnership:
Early Development Instrument

NAD 1983
UTM Zone 10 N

Alexander Coster
February 9, 2018
GEOB 479: Research in GIS