Geographically Weighted Regression:

An Overview

Marie Claire Anderson
26633157
GEOB479: Lab 2
Dr. Klinkenberg
February 29, 2020
Abstract

Regression analysis examines the relationship between a dependant variable and one or more independent variables. Typical global regression does not account for spatial variation, and as such provides limited insight into spatial phenomena. Geographically Weighted Regression (GWR) is a regression technique that accounts for spatial variation. GWR is contextualized though a case study in Vancouver, British Columbia where the relationship between children’s language scores and a set of child specific and neighbourhood specific explanatory variables are accounted for. Additionally, the applications of GWR in various fields are investigated though considering the work of other scholars and why GWR provides better spatial insights into analysis.
Regression and Geographically Weighted Regression

Regression is a statistical technique that describes the spatial relationship between a dependant variable (outcome variable), and one or more independent variables (explanatory variable/predictor variable). There are two basic types of regression, if there is one independent variable it will be known as simple regression and if there are two or more independent variables it will be known as multiple regression. Multiple regression seems to occur most often. In a multiple regression only one independent variable is changing, while the others remain constant. This allows for a calculation to be made that indicates how much the dependant variable is changing as the independent variable continually advances by 1 unit. Regression will determine the strength of the variable’s relationship, how well the model fits, and weather the model’s direction is positive, negative, or zero.

There are many regression methods for spatial analyses, such as OLS (Ordinary Least Squares), GLR (Generalized Linear Regression), and GWR (Geographically Weighted Regression). OLS is the most common, because it produces a single equation and is a global model. Typical global regression models like OLS do not account for the modelling of processes that change over space and as such they are considered to be an aspatial form of analysis. For that reason, Geographically Weighted Regression becomes the best analysis option as it is a local regression model and is area specific and spatially significant. GWR allows for spatial patterns to be interpreted more accurately because it assesses how processes differ over space rather than remaining constant like in an OLS regression. There are multiple reasons why relationships will possibly differ over space, including but not limited to random sampling variations, attitudes and preferences, and misspecifications of reality. For a GWR to yield proper results the dataset should be very large. Sample variance becomes a factor when using large datasets as they are
representative of the population but have the potential to inaccurately represent it. However, GWR is only concerned with very large variations in parameter estimates and variations of this size are unlikely to be due to sample variance alone. Additionally, different areas are often associated with different values, opinions, attitudes and preferences. A good way to think about this would be to consider how politics may create varying opinions across a city. Furthermore, models may vary due to misspecification. This can perhaps be a result of using the wrong types of variables to discuss a relationship or leaving out important variables intentionally or non-intentionally. To attempt to mitigate this, conducting an exploratory regression to determine what variables are most significant can help results. A GWR will result in parameter estimates and a parameter surface for each variable which will allow for visualization of the phenomena across the study area.

**GWR Case Study: Exploration of a Child’s Language Skills in Vancouver, British Columbia**

To exemplify the use of Geographically Weighted Regression in context, the following discussion will investigate the relationship between a child’s language skills and several explanatory variables associated with that child and their determined neighbourhood in Vancouver, British Columbia. The data for this analysis was originally gathered though census Canada data and Canada’s Early Development Instrument (EDI), but for the purpose of this analysis it was obtained collectively from the University of British Columbia Department of Geography. The EDI is effectively a survey which teachers are given to fill out about their students, resulting in “help scores” that determine how well a child’s development meets the expectation for their age group (What is the EDI, 2019). The dataset used included a sample of 2,695 children across Vancouver, BC. It is important to note that this analysis incorporates two
types of variables: neighbourhood variables, and child specific variables (Table 1). The
following analysis was conducted in ArcPro by ESRI. To begin assessing this data an
exploratory regression was carried out in order to determine which explanatory variables are best
describing the dependant variable, which in this case is Lang_sc (language score). Through
running two exploratory regressions it was determined that the most important explanatory
variables associated with a child’s language score are ESL, Soc_c, Loneparent, Recimmig, and
Income1000.

The next method of analysis was a Generalized Linear Regression (GLR). Using the help
scores as input features, language scores as the dependant variable and the most important
explanatory variables from the exploratory regression, the GLR was carried out. Note that he
model type used for the GLR was continuous (Gaussian), this means that the GLR is set to
conduct an Ordinary Least Squares (OLS) regression. Using the same explanatory variables,
dependant variable and input features with 250 neighbours and a bisquare weighting scheme the
GWR was carried out. Furthermore, a grouping analysis was preformed to decipher the
neighbourhood groupings. To do this the Spatially Constrained Multivariate Clustering tool was
used to assess the following variables: Childcare, Fam4, Loneparent, RecImmig, and
Income1000.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lang_sc</td>
<td>1 – 100 value reflecting a child’s language abilities</td>
<td>Child specific</td>
</tr>
<tr>
<td>Soc_c</td>
<td>1 – 100 value reflecting a child’s social abilities</td>
<td>Child specific</td>
</tr>
<tr>
<td>ESL</td>
<td>Is English the child’s second language (1 - Yes or 0 -No [Binary])</td>
<td>Child specific</td>
</tr>
<tr>
<td>Gender</td>
<td>The child’s gender (1 – female or 0 – male [Binary])</td>
<td>Child Specific</td>
</tr>
<tr>
<td>Fam4</td>
<td>% of families with 4+ members</td>
<td>Neighbourhood Specific</td>
</tr>
<tr>
<td>Income1000</td>
<td>Average income – divided by 1000</td>
<td>Neighbourhood Specific</td>
</tr>
<tr>
<td>Loneparent</td>
<td>% of families that are lone parent</td>
<td>Neighbourhood Specific</td>
</tr>
<tr>
<td>RecImmig</td>
<td>% of families that are recent immigrants to Canada having been her less then 5 years</td>
<td>Neighbourhood Specific</td>
</tr>
</tbody>
</table>

Table 1 - Child specific and neighbourhood specific variables, derived from EDI results
The grouping analysis yielded five different neighbourhood groups that have vastly different socio-economic characteristics (Table 2). It is clear through the grouping analysis that there are lots of families of 4 and lots of childcare, of which fall mainly within the blue neighbourhood (East Vancouver and surrounding area). Childcare also seems to be associated with higher income, dual income families that reside on in the orange neighbourhood (West-side of Vancouver and surrounding area). Upon investigation of the GLR (OLS) and GWR results in relation to these five neighbourhoods it is clear that using a GWR yields results that are spatially relevant. The GLR results show a slightly similar pattern to the GWR results around the Kingsway, however the GWR results are more effective than the GLR results in describing the area and show evidence of a good fit model (Table 3, Map 1, and Map 2). To determine where the GWR model worked well for each explanatory variable, the R² values were plotted onto to the parameter surfaces, with higher values indicating a better model fit (Figure 1). Areas with low R² values likely indicate that there are more variables that should be accounted for. The results show that in the areas of West Point Grey, Oakridge, Hastings-Sunrise, and Riley Park being a lone parent has a strong relationship with a child’s language scores (Map 3). Additionally, income has a strong relationship with a child’s language scores in the areas of Oakridge, Riley Park, Mount Pleasant, Strathcona and the Downtown Eastside (Map 4). Through these results it can be speculated that income plays a bigger role in the lives of people living on the east side of Vancouver, and that lone parents are pretty well dispersed throughout the city. For additional results on ESL and social skills see Maps 5 and 6.
### Table 2 - Neighbourhood groupings

<table>
<thead>
<tr>
<th>Neighbourhood Grouping</th>
<th>Childcare</th>
<th>Families of 4</th>
<th>Income</th>
<th>Lone Parent</th>
<th>Recent Immigrant</th>
<th>Neighbourhoods Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Above average</td>
<td>Above average</td>
<td>Below Average</td>
<td>Above Average</td>
<td>Average</td>
<td>Strathcona, Grandview-woodland, Hastings Sunrise, Renfrew Collingwood, Kensington Cedar Cottage, Riley Park, Mount Pleasant, Sunset, Victoria Fraserview, Killarney, South Cambie, Shaughnessy, Marpole, Oakridge</td>
</tr>
<tr>
<td>2</td>
<td>Below average</td>
<td>Very below average</td>
<td>Very very above average</td>
<td>Extremely above average</td>
<td>Very below average</td>
<td>Granville Island</td>
</tr>
<tr>
<td>3</td>
<td>Below average</td>
<td>Very below average</td>
<td>Average</td>
<td>Below average</td>
<td>Average</td>
<td>Downtown, West-End, Stanley Park, Olympic Village</td>
</tr>
<tr>
<td>4</td>
<td>Above average</td>
<td>Above average</td>
<td>Very above average</td>
<td>Below average</td>
<td>Average</td>
<td>West Point Grey, UBC, Dunbar Southlands, Kerrisdale, Arbutus Ridge</td>
</tr>
<tr>
<td>5</td>
<td>Below average</td>
<td>Very below average</td>
<td>Above average</td>
<td>Below average</td>
<td>Very below average</td>
<td>Kitsilano, Fairview</td>
</tr>
</tbody>
</table>

### Table 3 - GLR (OLS) parameters and GWR parameter ranges by variable

<table>
<thead>
<tr>
<th>GLR (OLS) Parameter Value</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \leq -2.497195 )</td>
<td>(-2.5) Std. Dev.</td>
</tr>
<tr>
<td>( \leq -1.498317 )</td>
<td>(-1.5) - (-0.50) Std. Dev.</td>
</tr>
<tr>
<td>( \leq 0.499439 )</td>
<td>(-0.50) - (0.50) Std. Dev.</td>
</tr>
<tr>
<td>( \leq 1.498317 )</td>
<td>(0.50) - (1.5) Std. Dev.</td>
</tr>
<tr>
<td>( \leq 4.353324 )</td>
<td>(&gt;1.5) Std. Dev.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>GWR Parameter Value Ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESL</td>
<td>High: 10.7879</td>
</tr>
<tr>
<td>SOC_C</td>
<td>High: 0.879077</td>
</tr>
<tr>
<td>Loneparent</td>
<td>High: 2.02893</td>
</tr>
<tr>
<td>Income1000</td>
<td>High: 1.79355</td>
</tr>
<tr>
<td></td>
<td>\leq 0.349782</td>
</tr>
<tr>
<td>------------</td>
<td>--------------</td>
</tr>
<tr>
<td></td>
<td>\leq 0.451146</td>
</tr>
<tr>
<td></td>
<td>\leq 0.651341</td>
</tr>
</tbody>
</table>

*Figure 1 - GWR $R^2$ values*
Neighbourhood Grouping Analysis and GLR Results
for Children's Language Skills in Vancouver, British Columbia

GLR Results
- < -2.5 Std. Dev.
- -2.5 - -1.5 Std. Dev.
- -1.5 - -0.50 Std. Dev.
- -0.50 - 0.50 Std. Dev.
- 0.50 - 1.5 Std. Dev.
- > 1.5 Std. Dev.

Neighbourhood Groupings
1
2
3
4
5

Note: Neighborhood groupings are based on census enumeration areas.

Map 1 - Neighbourhood grouping analysis and GLR results, see Table 1, 2 and 3 for additional information
Neighbourhood Grouping Analysis and GWR Results
for Children's Language Skills in Vancouver, British Columbia

Spatially Constrained Multivariate Clustering Box-Plots

Analysis Fields

GWR Results
- ≤0.349782
- ≤0.451146
- ≤0.651341

Neighbourhood Groupings

Note: Neighborhood groupings are based on census enumeration areas.

Map 2 - Neighbourhood grouping analysis and GWR results, see Table 1, 2 and 3 for additional information.
Map 3 - Lone Parent Analysis and GWR Results, see Figure 1 and Tables 1, 2, 3 for additional information
Map 4 - Income Analysis and GWR Results, see Figure 1 and Tables 1, 2, 3 for additional information.
Map 5 - Social Skills Analysis and GWR Results, see Figure 1 and Tables 1, 2, 3 for additional information
Map 6 - ESL Analysis and GWR Results, see Figure 1 and Tables 1, 2, 3 for additional information
GWR Applications

Geographically Weighted Regression is a useful tool for spatial analysis across several contexts. This is because often particular phenomena are being investigated at the local level, rather than the global, and as such require varying spatial considerations to be accounted for. A 2017 study by Inti Nashwari looks at the impact of the agricultural sector on food security in Indonesia. It was determined that the use of GWR would be more suitable for their study location and set of variables. Through the GWR results the authors concluded that the effect of their variables on poverty is not general, but actually very unique and varied significantly across space (Nashwari, 2017). Similarly, a 2018 study in Guizhou Province, China was conducted by Xu et. al that investigated what variables are correlated to poverty in their study area. Due to the local scale of the research and the expectation that poverty levels likely change over space depending on a number of circumstances, a GWR was identified as the best tool for analysis. The authors concluded that through the GWR there is an indication that “spatial nonstationarity occurs for each geographical indicator in the different administrative towns” (Xu et. al., 2018 p. 967).

In addition to socio-economic applications, GWR also has ecological applications. A 2005 article by Wang, Ni and Tenhunen investigated relationships between Net Primary Production (NPP) of forest ecosystems in China and environmental variables. The authors conducted a number of analysis methods including OLS and GWR and determined that GWR preformed best as it best displayed the spatially non-stationary relationships between the dependant and independent variables (Wang, Ni, Tenhunen, 2005). Likewise, a 2012 study by Pearsall and Christman explored the spatially varying relationship between urban greenness and socio-economic conditions in Philadelphia, USA. In this study a GWR was used to decipher the
strength of the relationships among variables that can only be determined at the local scale and concluded that relationships within the city rely on location-specific contexts (Pearsall and Christman, 2012).

Furthermore, Geographically Weighted Regression can also be applied to health geography. A 2018 study by Purwaningsih and Noraprilla investigates the variables that have a relationship with cervical cancer in Indonesia through a GWR. The authors found this method to be useful as rates of cervical cancer seem to change significantly across the country (Purwaningsih & Noraprilla, 2018).

Spatial phenomena are complex and likely always has multiple reasons for its variance. When global models do not fit, local models such as GWR prove to provide accurate insight into what variables matter and why. Socio-economic applications like poverty, and children’s language abilities, in addition to ecological and health applications are just a few of the many things that Geographically Weighted Regression can be applied to in order to attempt to understand spatial relationships.
References


