LAB 2: Geographically-Weighted Regression

Alex Briault

GEOB 479
February 15, 2019

Abstract

This report examines geographically-weighted regression (GWR) by examining the method, applying it to a case study of children's social scores in Vancouver, BC, and discussing other possible applications. GWR accounts for location in its regression which allows for spatial patterns resulting from spatial variations of variables to be visible. The case study examines children's social scores in Vancouver in relation to three variables: gender, income, and language score. Results indicate that the language score model is the best fitting of the three but that there are still other variables at work influencing social scores that were not included in this analysis. Other applications of GWR are included to show its broad range of usefulness across disciplines compared to other regression methods.

Geographically-weighted Regression, a non-technical explanation

When trying to explain or predict where a particular phenomenon occurs, a regression analysis can be helpful to determine which of the variables present has the greatest effect on the phenomenon. Regression analysis can be used to both explain and predict spatial relationships and patterns. To perform a regression analysis, a dependent variable and at least one independent variable are needed. The dependent variable is the variable your analysis is attempting to explain while the independent (explanatory) variable(s) is what may potentially explain the dependent variable. In this lab, the child's social score was the dependent variable and gender, income, and language score were the independent variables. If there are a large number of explanatory variables, it is useful to first do an explanatory regression analysis as this will narrow down the number of variables to only those that are most important. The "most important variables" can then be used to do a regression analysis like Ordinary Least Squares (OLS) which is a linear regression that predicts or models a dependent variable based on its relationship to explanatory variables. OLS produces a global model of the process being tested because it does not consider location in its analysis. A global model, like OLS, assumes that variables are constant across the analysis area and that location does not cause variations in the data being analysed. Despite this issue with OLS, it is useful to perform an OLS regression before a geographically-weighted regression because comparing the results of both will show if the variables are spatially influenced best.

In contrast to global models like OLS, a geographically-weighted regression (GWR) analysis does consider location and produces a local model which accounts for spatial variation in the data. GWR takes neighbouring values into consideration when creating the model so if variables are spatially autocorrelated or nonstationary then the resulting analysis will reflect this. Following Tobler's first law of geography, near things have a greater affect than far things do so a GWR assumes that variables spaced more closely together will be affected to a greater degree than those that are more dispersed. While GWR can be used to better model nonstationary variables than an OLS model, it is not without problems. If one variable, or a combination of variables, is redundant or too spatially similar in variation then the problem of multicollinearity will result. Multicollinearity will cause the model to be biased as too much weight is ascribed to one variable.

GWR is useful in a variety of contexts where location may be influential. GWR can be employed in health geography, landscape ecology, real estate analysis, crime studies, and more. In health geography, a dependent variable like number of heart attacks can be examined in relation to other health factors like age, gender, smoking rates, or alcohol consumption to determine where public health interventions may be useful. In landscape ecology, GWR can be used to determine where a particular plant species may be found based on elevation, slope aspect, precipitation, proximity to human development. Real estate analysis is very spatially dependent and GWR can be used to explain why house prices are high in one area of the city but low in another by using variables like crime rates, proximity to schools, average incomes, and average age of residents.

As Tobler's first law of geography states: everything is related but near things are more related than distant things. GWR accounts for nearby variables that may affect the dependent variable in a way that global linear regression models cannot and this makes it an invaluable method of analysis when working with spatial data.

GWR Results Discussion

This analysis explored several variables and their affects on children's social scores. To determine which variables were to be used, an explanatory regression analysis was carried out first. This regression analysis resulted in income[1], gender, and language being the most important variables influencing social scores. With these variables, an OLS regression analysis and a GWR analysis were done to model the influence of these variables. As the OLS produces a global model where location is not accounted for, its result do differ from those of the GWR, however, they share some similarities. On Vancouver's east-side, the results of both the OLS and GWR were very similar indicating that children's social scores in this area of the city are influenced by geography. The results of the OLS and GWR vary more widely on the westside of Vancouver, in the Downtown core, and on the northern east-side which indicates that variations in variables in these areas are less influenced by geography.

The GWR using gender reveals no correlation between the effect of gender and children's social scores. High effects of gender are concentrated predominantly in the eastern portion of Vancouver with lower effects present across southwest and central southern Vancouver. Despite the highest R^2 values also being present in the eastern portion of Vancouver, the R^2 values span across areas with high and low effects of gender. Because of these results, gender is not a viable variable to explain the distribution of children's social scores across Vancouver.

Using income as the GWR explanatory variable produces a map displaying a greater degree of correlation between income and social scores than gender. While the correlation is stronger using income, there is still a large variance in areas where high R^2 values are present

[1] Income calculated as the average neighbourhood income divided by 1000

compared with high effects of income. There is a low effect of income on social scores in east Vancouver but there are pockets adjacent to the low-defect area where income has a high effect on social scores.

The language score variable produces the strongest correlation with the local R^2 values of the GWR. As visible in the map below, areas of Vancouver where language score has a high effect, there is also a high local R^2 value. This is the best fitting explanatory variable but it does not perfectly explain the distribution of children's social scores as the GWR model only accounts for 61% of the variation present in the results. The effects of language on social scores is highest in the downtown core, Kitsilano, and east Vancouver areas but, as visible in the map, there are areas of east Vancouver that are not explained by the GWR model.

The grouping analysis shows a fairly clear east-west division where Vancouver's east-side is dominated by neighbourhoods with the lowest average incomes, highest percentages of families spending 30 or more hours on childcare, and the highest percentage of lone parents. The westside is dominated by the highest average neighbourhood incomes and highest percentage of families of four. When the local R^2 coefficients of the GWR are displayed on this grouping map, the strength of the model is apparent. Despite the strength of the GWR model on the east-side of Vancouver, it is important to note that the highest R^2 value is 0.612776 which, while closer to 1.0, can only account for 61% of the variation in social scores which indicates that there is another variable at play that has not been accounted for. Downtown Vancouver and Kitsilano do not fit this model which reinforces the premise that there is another variable affecting social scores that was not accounted for in this analysis.

Other potential uses of GWR

In this analysis, GWR was used to determine which variables best explained a child's social score but GWR can be used in a variety of other applications like health, real estate, and crime analysis. The 2011 study by Liu et al. looked to investigate the factors contributing to the spread of drug-resistance tuberculosis (DR-TB). Using data on ecological factors from the World Health Organization/International Union Against Tuberculosis and Lung Disease, OLS regression and GWR were conducted (Liu et al., 2011). The OLS regression indicated a global linear spatial relationship between DR-TB and its latent synthetic risk factors (Liu et al., 2011). The risk factors identified by Liu et al. (2011) were Annual Precipitation, Annual Atmospheric Temperature, Temperature Climate Zone, Geography Climatic Zone, and Geography Latitude and from these five factors, two latent synthetic risk factors were extracted, "Temperature" and "Humidity" which explain 87.17% of the total variance among the five original factors. The results of the GWR showed a stronger relationship between the latent synthetic risk factors and DR-TB and this spatial variability led to the use of GWR for local estimations (Liu et al., 2011). The results of the Liu et al.'s (2011) study recommend that due to the spatial variability of factors influencing DR-TB that planning, prevention, and control should be done according to the local relationship between it and the latent synthetic risk factors. In this case study, relying only on the regression produced by the OLS regression would have resulted in planning, prevention, and control strategies being applied universally across affected regions which would have resulted in some regions benefitting while others would have experienced no benefit as inappropriate methods would have been employed.

In an example of using real estate data, Legg and Bowe (2009) applied GWR to list prices of single family houses in Marquette, Michigan. List prices were the dependent variable while house square footage and lot size were the explanatory variables (Legg & Bowe, 2009). Originally, the number of bedrooms was included as an explanatory variable but as it was closely linked to house square footage, it was excluded as its inclusion would have resulted in multicollinearity (Legg & Bowe, 2009). GWL was introduced in this case study because previous use of linear models resulted in inaccurate predictions and GWR allowed for this to be corrected for (Legg & Bowe, 2009). The results of this analysis showed that lots located nearer the urban core had higher lot square footage prices while also suggesting that larger houses on these (closer to urban centre) lots contributed less to the list cost (Legg & Bowe, 2009). This analysis also suggested that this list price pattern was indicative of the age of homes between urban and rural settings. Rural subdivisions, per Legg and Bowe (2009), had newer homes and lower land values and these newer homes contributed more towards list prices per square foot. Using GWR in this way could be beneficial to developers looking to build new subdivisions or medium- to high-density housing as land costs relative to urban core as well as current list prices could indicate areas where the most profits stand to be made.

Crime analysis benefits from GWR because it allows for a variety of variables to be tested to determine effects on crime levels and as a result can offer local crime-reduction suggestions that would otherwise not be visible through an OLS regression. In Drum's (2016) article, lead levels in gasoline were linked to crime rates across the United States, Canada, Australia, France, Great Britain, Finland, Italy, New Zealand, and Germany with the same results visible in each country: as lead emission rose so did crime rates and as lead emissions dropped, so did crime rates. In New Orleans, high levels of lead in soils were shown to correspond with lower income neighbourhoods which also corresponded with crime maps of the city (Drum, 2016). As lead levels vary across a city, a linear regression would miss this spatial variation and its results would not be useful whereas including spatial variation by using a GWR, the spatial patterns become apparent which can lead to focused management and removal of lead in areas of high concentration.

References

Drum, K. (2016). Lead: America’s Real Criminal Element. Retrieved from https://www.motherjones.com/environment/2016/02/lead-exposure-gasoline-crime-increase-children-health/

Legg, R., & Bowe, T. (2009). Applying Geographically Weighted Regression to a Real Estate Problem. *ArcUser, Spring*, 44-45. https://www.esri.com/news/arcuser/0309/files/re_gwr.pdf

Liu, Y., Jiang, S., Liu, Y., Wang, R., Li, X., Yuan, Z., … Xue, F. (2011) Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis. *International Journal of Health Geographics 10*(50), 1-10. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3173290/pdf/1476-072X-10-50.pdf