Addressing Multicollinearity in Local Modeling of Spatially Varying Relationship using GWR

Rizwan Shahid[1,2], and Stefania Bertazzon[2]

1 Alberta Health Services, rizwan.shahid@ahs.ca
2 Department of Geography, University of Calgary, Calgary, AB, bertazzs@ucalgary.ca

Abstract

Geographically weighted regression is a powerful and computationally intensive method to model varying spatial relationships, but it may introduce high local multicollinearity which, if not dealt with properly, leads to misleading inference and unreliable results. We introduced a novel solution to deal with multicollinearity by adopting a more localized and refined approach. This solution is demonstrated by modeling the varying local association of childhood obesity and its risk factors at neighborhood level to target only vulnerable neighborhoods for the prevention and control of obesity locally.

Background and Relevance

In any given study area, global models generally do not account for varying spatial relationship (spatial non-stationarity) and provide only one parameter estimate for each relationship in the model. Geographically weighted regression (GWR) is an extension of the global linear regression model: it addresses spatially varying relationships by yielding a set of local parameter estimates and their significance (Fotheringham et al., 2002; Chalkias et al., 2013). The technique has enjoyed increasing popularity over the last decade, and has been used in a variety of applications, beyond the original work in hedonic models (Fotheringham et al., 2002). However, in modeling local relationships, GWR may introduce high local multicollinearity in the local models even if the variables are not correlated globally (Wheeler & Tiefelsdorf, 2005). There have been efforts to deal with this issue, but the results have not been completely satisfactory. Brunsdon et al. (2012) suggested three possible strategies to deal with multicollinearity: a) omitting the variables causing high multicollinearity, b) working with large bandwidth (nearest neighbors or distance to account for in local model), and c) doing nothing and treating results with caution. Omitting a variable means excluding it from all the local models, including those where there is no multicollinearity and the variable would be significant. Working with large bandwidth would drive the GWR model towards the global model and would lead to more globalized parameter estimates rather than localized ones (Guo et al., 2008). Doing nothing is a dangerous approach, and may lead to a large portion of meaningless local models.

This research presents preliminary results of an innovative approach to handle multicollinearity while retaining all the significant independent variables and estimate the local GWR parameters coefficients, yet avoiding multicollinearity as suggested by Marill (2004), “The general goal should be the inclusion of all predictor variables that add substantial independent information while avoiding excessive collinearity”.

Our application is an analysis of childhood obesity at the neighbourhood level in the City of Calgary (Shahid and Bertazzon, 2015). Obesity is a global phenomenon. In spite of continuous efforts and spending millions of dollars, the obesity rates among Canadian adults and children is increasing at an alarming rate (Sharma, 2016). One of the reasons for this increase may be the one-size-fit-all approach in tackling the obesity epidemic ignoring the fact that each area (neighborhood or set of neighborhoods) has unique physical environmental and social characteristics that need to be addressed differently. Obesity may not be a problem in many neighborhoods; conversely, some of the obesity risk factors may not matter in some neighborhoods. We used GWR to analyze childhood obesity in the Calgary to uncover the risk factors that are important locally. We used four years (2005 to 2008) BMI (Body Mass Index) percentile data for age group 4.5 to 6 years old. Candidate explanatory variables were selected based on prior knowledge (known and suspected risk factors – supported by the theory). These are socioeconomic variables from census 2006 (immigrants, people with no certificate, diploma or degree, and median census family income); walkscore, an index of neighborhood walkability; location of parks and fast food restaurants; and pathway length in each neighborhood. The BMI percentile was used to classify children into underweight ($<5^{th}$ percentile), healthy weight ($\geq 5^{th}$ percentile and $<85^{th}$ percentile), overweight (($\geq 85^{th}$ percentile and $<95^{th}$ percentile) and obese ($\geq 95^{th}$ percentile). The location of each child was geo-coded using children's residential postal code and aggregated at the neighbourhood level. Neighbourhoods having zero obese children were removed, to avoid misleading results of the local analyses. Neighbourhoods with missing socioeconomic variables were also removed from the analysis. As a result, 174 neighbourhoods were retained. Upon combining the 4 years of BMI data, the total sample size became 37,460 children with an average of 215 children per neighbourhood (ranging from 8 to 1,232). The average number of obese children per neighbourhood is 21. Overall, the percentage of obesity for the City of Calgary is 9.7%; it varies spatially by neighbourhood ranging from 2.7 % to 21.4%. This suggests that the prevalence of obesity varies across the City of Calgary, which may be an indication of variations in obesogenic environment. Non-spatial models such as multivariate regression, ignoring local variability in the dependent as well as in the independent variables, may lead to model misspecification and fail to unmask local associations. Geographically weighted regression was, thus, used to capture the spatial dynamics by modelling the variation in local associations between dependent and independent variables across different neighbourhoods. Although the global model did not show any signs of multicollinearity, many of the local models did exhibit strong multicollinearity. If not handled properly, local multicollinearity would have yielded misleading results and inference. We handled it by employing a more localized approach.

Methods and Data

Prior to implementing regression models, some transformations were applied to some of the variables. The dependent variable (obese children) and independent variables (people with no certificate diploma or degree, proximity to fast food restaurants, proximity to parks) were standardized as suggested by Chalkias et al. (2013) and Preston et al. (2000) by taking the proportion of obese children to total children and

multiplying it by 1000. The median family census income was converted to $1000s (for example, 72,200 transformed to 72.2); pathway length was recorded in Km and walkscore was used as an index ranging from 0 to 100, with zero as not walkable at all to 100 as totally walkable.

Multicollinearity was measured using condition numbers. A condition number is the square root of the largest eigenvalue divided by the smallest eigenvalue of the cross product of geographically weighted matrices (Brunsdon et al., 2012) . A condition number above 30 is an indication of high multicollinearity (Belsley et al. 1980). We used the value of 30 as condition number threshold to identify the local models affected by multicollinearity. Once local models were flagged, based on their condition numbers, variable removal procedures were applied locally as follows. All independent variables 'X' were used initially to estimate the local parameters for all local models. Multicollinearity was assessed for all models and the models without multicollinearity problems were retained. Whenever multicollinearity was identified, the variable causing local multicollinearity was removed, resulting in a local model with 'X-1' variables. GWR was applied again to all neighborhoods. The models without multicollinearity were retained again but the models obtained from the first run using all variables were excluded. This strategy provided a new set of local models without multicollinearity. As condition numbers still gave evidence of some multicollinearity, a second variable was removed from the affected local models, resulting in 'X-2' variables and GWR was run again. This time none of the local models suffered from multicollinearity. Again, the models obtained from 'X' and 'X-1' were removed from the set of local models obtained from 'x-2' resulting in new set of models. The procedure is general and can be applied in other instances. If multicollinearity persists, the process can be repeated iteratively until no multicollinearity is detected. The final set of local models contained all the significant variables and was free of multicollinearity. This more localized approach can be referred to as Locally Refined GWR (LRGWR).

Results

Mapping the GWR parameter estimates and the t-value for each variable provides important insights of the relationship between dependent and independent variables. The t-values were mapped to visualize the changing relationship and the strength of the associations between dependent and independent variables across the city. The t-values were categorized with commonly used significance thresholds that are 90%, 95% and 99% (Mennis, 2006). Table 1 shows the number of neighbourhoods where the association is significant at 99%, 95% and 90% significance level. Positive association (positive beta coefficient) is shown by (+) and negative association is shown by (-). Results suggest that childhood obesity does not remain constant across the neighbourhoods, nor do its associated factors. The variation across neighbourhoods indicated the need of a local modeling approach that could deal with multicollinearity.

LRGWR Model			
Variable	**99%**	**95%**	**90%**
Walkscore	20 (-)	4 (-)	1 (+)
people with no certificate, diploma or degree	15 (+)	19 (+)	5 (+) 8 (-)
median census family income	8 (-)	29 (-)	5 (-)
immigrants	9 (+)	51 (+)	12 (+)
proximity to fast food restaurants	0	8 (+)	5 (+)
pathway length	0	13 (-)	3 (-)
proximity to parks	27 (-)	11 (-)	2 (-)

Table 1: variables of the LRGWR model and their significance

Figure 1 illustrates the extent of multicollinearity in the standard GWR model vs. our LRGWR model.

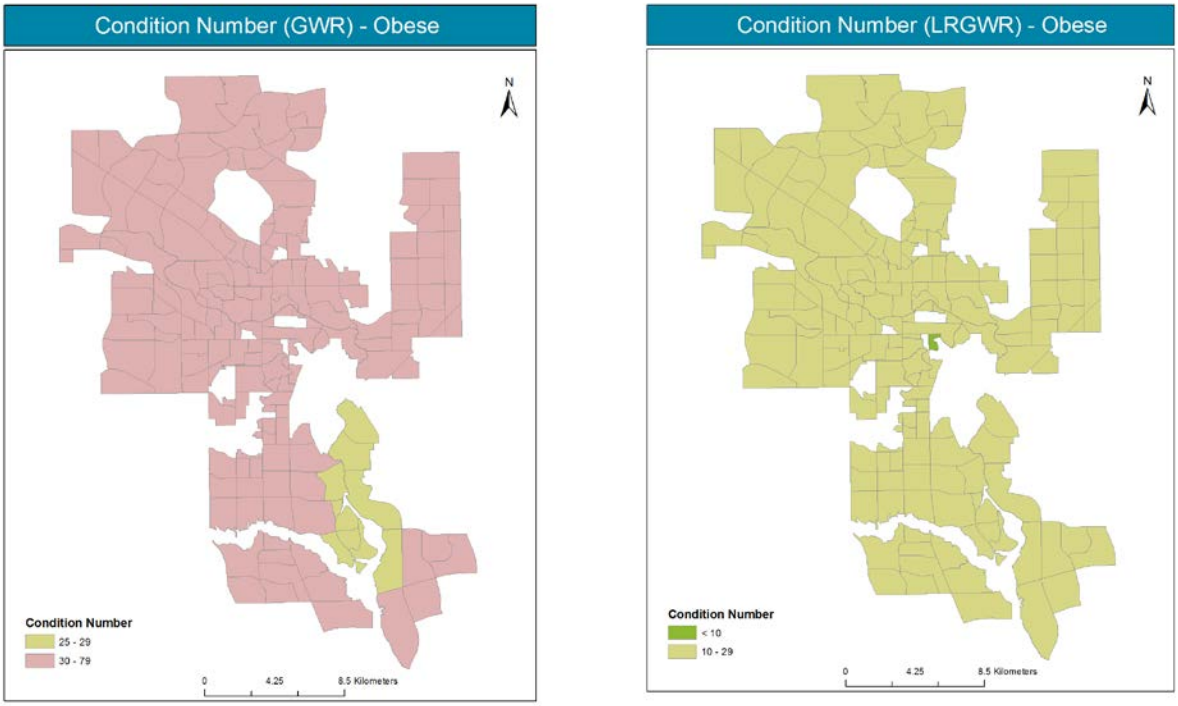

Figure 1: Condition numbers standard GWR and LRGWR

It is evident from the above figure that a very large number of local models suffer from multicollinearity. Without dealing with local models exhibiting high multicollinearity, inferences and the interpretation of results from GWR models would be misleading and unreliable. Conversely, when LRGWR was implemented by removing variable causing high multicollinearity locally, none of the condition numbers exceeded the threshold value of 30, thus improving the overall model and confidence on the results.

Conclusions

GWR is very useful in understanding and modeling varying local spatial relationships, but despite its merits and increasing popularity, GWR coefficients may exhibit local multicollinearity that may produce unreliable estimates and misleading inferences. Our method takes the emerging GWR technique a step further by addressing this limitation in a novel way that makes local models trustworthy and more reliable. The approach may decrease the power of the model, but this cannot be concluded until more analysis is performed. Further research is required to make this process robust and to embed it in the available software packages offering GWR.

References

Belsley DA, Kuh E, Welsch RE (1980) Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons, New York, NY.

Brunsdon, C. Charlton, M. & P. Harris, (2012), Living with collinearity in local regression models, Proceedings of the *10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Science*, Florianopolis SC, Brazil.

Chalkias, C. Papadopoulos, A.G. Kalogeropoulos, K. Tambalis, K. Psarra, G. & L. Sidossis, (2013). Geographical heterogeneity of the relationship between childhood obesity and socio-environmental status: empirical evidence from Athens, *Greece, Applied Geography 37*, 34-43.

Fotheringham, A.S. Brunsdon, C. & M. Charlton, (2002). Geographically Weighted Regression: The analysis of spatially varying relationships, Wiley, West Sussex.

Guo, L. Ma, Z. & L. Zhang, (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study, *Canadian Journal of Forest Research, 38,* 2526-2534.

Marill, K.A. (2004). Advanced statistics: linear regression, Part II: multiple linear regression, *Academic Emergency Medicine, 11(1),* 94-102.

Mennis, J. (2006). Mapping the results of geographically weighted regression, *The Cartographic Journal, 43(2),* 171-179.

Shahid, R., Bertazzon, S. (2015). Local spatial analysis and dynamic simulation of childhood obesity and neighbourhood walkability in a major Canadian city. *AIMS*

Public Health, special issue: “Spatial Aspects of Health: Methods and Applications”. **2**(4):616-637.

Sharma, A.M. (2016). Critical fat studies and obesity in Canada, The Lancet Diabetes & Endocrinology, In press.

Preston, S. Heuveline, P. & M. Guillot, (2000). Demography: Measuring and Modeling Population Processes, *Blackwell Publishing, Oxford*.

Wheeler, D. & M. Tiefelsdorf, (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression, *Journal of Geographical Systems, 7*, 161-187.