pagepress

Mastering geographically weighted regression: key considerations for building a robust model

Behzad Kiani,[1] Benn Sartorius,[1,2,3] Colleen L. Lau,[1] Robert Bergquist[4]

[1]UQ Centre for Clinical Research, The University of Queensland, Brisbane, Australia; [2]Centre for Tropical Medicine and Global Health, Faculty of Medicine, University of Oxford, Oxford, UK; [3]Department of Health Metric Sciences, Faculty of Medicine, University of Washington, Seattle, US; [4]Ingerod, Brastad, Sweden

Geographically weighted regression (GWR) takes a prominent role in spatial regression analysis, providing a nuanced perspective on the intricate interplay of variables within geographical landscapes (Brunsdon *et al.*, 1998). However, it is essential to have a strong rationale for employing GWR, either as an addition to, or a complementary analysis alongside, non-spatial (global) regression models (Kiani, Mamiya *et al.*, 2023). Moreover, the proper selection of bandwidth, weighting function or kernel types, and variable choices constitute the most critical configurations in GWR analysis (Wheeler, 2021).

Global regression models operate under the assumption that the relationships between independent and dependent variables are constant across the entire study area. However, GWR challenges this assumption by acknowledging that these relationships can vary spatially. This makes GWR particularly well-suited for situations where local variations in independent variables play a pivotal role in influencing the overall pattern of dependent variables, as well as when the relationship between the independent variables and the outcome could vary across space (Soroori *et al.*, 2023). Researchers employ various strategies, such as spatially distributing predictors based on domain knowledge or existing literature and measuring the spatial autocorrelation of predictors, to ascertain if they adhere to a spatial pattern (Mohammadi *et al.,* 2022). For example, although it is evident that the distribution of urban greenspaces is spatially heterogeneous, it is crucial to objectively assess their spatial autocorrelation in each setting since the distribution may be quite uniform in some of them (Kiani, Thierry *et al.*, 2023).

One of the pivotal aspects in configuring GWR lies in the judicious selection of bandwidth, a parameter that governs the spatial extent, over which neighbouring observations influence the estimation of local parameters. The bandwidth acts as a critical filter determining the degree of localization in the analysis. A narrow bandwidth may result in oversensitivity to local variations, potentially capturing noise in the data, while broad bandwidths may lead to oversmoothed representations masking subtle spatial patterns (Kiani, Mamiya *et al.*, 2023). Striking the right balance is thus essential to ensure that the GWR model captures the true spatial heterogeneity without being unduly influenced by distant observations. The challenge lies in avoiding the imposition of a uniform distance across the entire study area, something which can be done by choosing a bandwidth that is spatially adaptive and tailored to the intricacies of the underlying spatial processes. From the fixed and adaptive bandwidths that can be employed, we advocate for the use of the latter as it varies based on the size of each geographical area and that of its neighbours (Guo *et al.*, 2008). Using this approach, the model would opt for a narrower bandwidth in dense areas, such as central business districts, and a larger one for suburban areas (Guo *et al.*, 2008). Most spatial software applications can determine the optimal bandwidth based on the definition of the best model performance set by the user.

A critical facet in configuring GWR pertains to the choice of weighting functions, which govern the influence of neighbouring observations on the local parameter estimation. These functions define the spatial decay of influence that determines the degree at which observations closer to the target location contribute to the estimation process. Different weighting functions, such as Gaussian, exponential or bisquare, offer varying degrees of emphasis on nearby versus distant observations (Wheeler, 2021). For example, in the study of the spread of vector-borne diseases, e.g., leishmaniasis which involves vectors with limited mobility (Firouraghi *et al.*, 2023), a Gaussian weighting function that gives more weight to close observations might be appropriate as this emphasizes the capture of local transmission dynamics. On the other hand, for vectors with longer mobility ranges, or waterborne diseases like cholera, where the infectious agent can spread far through connected water sources, an exponential weighting function that considers a broad spatial range may better reflect the transmission patterns. The decision regarding which function to employ should align with the underlying spatial processes and the scale at which variations in relationships are expected as this underlines the importance of tailoring the approach to the transmission drivers of the infectious disease under investigation.

Correspondence: Behzad Kiani, UQ Centre for Clinical Research, The University of Queensland, Brisbane, Australia.
E-mail: B.Kiani@uq.edu.au

Key words: geographically weighted regression, spatial epidemiology, spatial regression.

Conflict of interest: the authors declare no potential conflict of interest, and all authors confirm accuracy.

Availability of data and materials: all data generated or analyzed during this study are included in this published article.

Received: 11 February 2024.
Accepted: 14 February 2024.

Geospatial Health 2024; 19:1271
doi:10.4081/gh.2024.1271

When configuring a GWR model, the judicious selection of variables is a fundamental consideration wielding significant influence over the model's capacity to capture spatial heterogeneity. Unlike global regression models that assume uniform relationships across all locations, GWR allows for the exploration of spatially varying coefficients for each variable. Thus, the selection of variables demands careful attention to ensure that the chosen set appropriately captures the nuances of the spatial processes under investigation. Selection of variables in GWR involves identifying those predictors that exhibit spatial non-stationarity, acknowledging that their impact varies across different geographic regions (Mohammadi *et al.*, 2023). Researchers must navigate the delicate balance between inclusivity and relevance, recognizing that overly complex models may introduce noise, while overly simplistic ones risk overlooking important spatial variations. GWR is sensitive to multicollinearity among independent variables, so researchers should carefully examine the spatial correlation between predictors to avoid inflated standard errors and unreliable coefficient estimates. Furthermore, it is critical to identify and handle outliers and influential observations as these data can disproportionately impact local parameter estimates and their presence may require exclusion or statistical adjustments such as differential weighting.

After having run a GWR analysis, assessment of spatial autocorrelation in the residuals is crucial. Spatial autocorrelation, which measures whether nearby locations have similar residuals, needs careful consideration to avoid biased parameter estimates. Diagnostic tools such as Moran's *I* or spatial autocorrelation maps are valuable for this assessment (Kiani, Fatima *et al.*, 2023). Additionally, it is essential to conduct thorough model validation to ensure the reliability of GWR results. Techniques like cross-validation or comparing predicted and observed values aid in assessing the model's predictive performance and generalizability. Furthermore, conducting sensitivity analysis allows researchers to explore the robustness of GWR results to changes in configuration parameters, ensuring the stability of findings. Finally, to derive meaningful insights from GWR, interpretation of local parameter estimates is paramount (Mohammadi *et al.*, 2022). Researchers must carefully interpret spatially varying coefficients in the context of the study area and underlying spatial processes.

Several available GWR software tools provide researchers with diverse options to explore spatial relationships at the local level. Geospatial Exploratory Data Analysis (GeoDa), a free, user-friendly and widely used platform, facilitates GWR implementation and offers visualization tools for spatial data exploration (Anselin *et al.*, 2022). Geographic Information Systems (GIS) software, developed by Environmental Systems Research Institute (ESRI) under the name ArcGIS, is a popular commercial tool that allows users to perform GWR analyses seamlessly through its Geostatistical Analyst extension (Scott & Janikas, 2009). R, a statistical programming language, provides the 'spgwr' package for GWR, which offers flexibility and customization for advanced users (Bivand *et al.*, 2017). Additionally, Python Spatial Analysis Library (PySAL) is a powerful open-source library that supports GWR analysis catering to researchers who prefer coding-based approaches (Rey *et al.*, 2022). Finally, the MGWR 2.2 free software is an alternative developed by Spatial Analysis Research Center (SPARC) at Arizona State University, USA (Li *et al.*, 2019). These software options vary in terms of user interface, capabilities and programming requirements that allow researchers to choose the tool that best fits their preferences and analytical needs when exploring local spatial relationships.

In conclusion, this paper has highlighted the nuanced nature of GWR analysis, emphasizing the importance of meticulous configuration to capture spatial intricacies. Each step, from variable selection to choices of bandwidth and weighting functions, influences the model's ability to reveal spatial heterogeneity. GWR's recognition of spatial non-stationarity provides a unique perspective beyond global regression models. Addressing challenges like spatial autocorrelation, conducting model validation and careful interpretation of local parameter estimates are crucial steps in ensuring the reliability of GWR analyses. Navigating this geospatial labyrinth and crafting confidence through precise configuration make GWR a valuable tool for unravelling spatial complexities across diverse landscapes. Finally, it is worth mentioning that GWR captures local spatial patterns and heterogeneity. However, it does not explicitly model spatial dependence. In contrast, geostatistical models are designed to handle spatial autocorrelation and provide a more accurate representation of spatial processes by explicitly incorporating spatial structures. Therefore, for data with strong spatial autocorrelation or complex spatial patterns, geostatistical models offer more comprehensive insights than GWR.

References

Anselin L, Li X, Koschinsky J, 2022. GeoDa, from the desktop to an ecosystem for exploring spatial data. Geogr Anal 54:439-466.

Bivand R, Yu D, Nakaya T, Garcia-Lopez M.-A, Bivand M.R, 2017. Package 'spgwr'. R software package.

Brunsdon C, Fotheringham S, Charlton M, 1998. Geographically Weighted Regression. J R Stat Soc: Series D (The Statistician) 47:431-43.

Firouraghi N, Bergquist R, Fatima M, Mohammadi A, Hamer DH, Shirzadi MN, Kiani B, 2023. High-risk spatiotemporal patterns of cutaneous leishmaniasis: a nationwide study in Iran from 2011 to 2020. Infect Dis Poverty 12:49.

Guo L, Ma Z, Zhang L, 2008. Comparison of bandwidth selection in application of geographically weighted regression: a case study. Can J For Res 38:2526-34.

Kiani B, Fatima M, Amin N.H, Hesami A, 2023. Comparing geospatial clustering methods to study spatial patterns of lung cancer rates in urban areas: A case study in Mashhad, Iran. GeoJournal 88:1659-69.

Kiani B, Mamiya H, Thierry B, Firth C, Fuller D, Winters M, Kestens Y. 2023. The temporal sequence between gentrification and cycling infrastructure expansions in Montreal, Canada. Habitat Int 139:102899.

Kiani B, Thierry B, Fuller D, Firth C, Winters M, Kestens Y. 2023. Gentrification, neighborhood socioeconomic factors and urban vegetation inequities: A study of greenspace and tree canopy increases in Montreal, Canada. Landsc. Urban Plan. 240:104871.

Mohammadi A, Bergquist R, Fathi G, Pishgar E, de Melo SN, Sharifi A, Kiani B, 2022. Homicide rates are spatially associated with built environment and socio-economic factors: a study in the neighbourhoods of Toronto, Canada. BMC Public Health 22:1482.

Mohammadi A, Pishgar E, Fatima M, Lotfata A, Fanni Z, Bergquist R, Kiani B, 2023. The COVID-19 Mortality Rate Is Associated with Illiteracy, Age, and Air Pollution in Urban Neighborhoods: A Spatiotemporal Cross-Sectional Analysis. Trop Med Infect 8:85.

Rey SJ, Anselin L, Amaral P, Arribas−Bel D, Cortes R.X, Gaboardi J.D, Kang W, Knaap E, Li Z, Lumnitz S, Oshan T, Shao H,

Wolf L.J. 2022. The PySAL ecosystem: Philosophy and implementation. Geogr Anal 54:467-87.
Scott LM, Janikas MV. 2009. Spatial statistics in ArcGIS. Handbook of applied spatial analysis: Software tools, methods and applications, Springer, pp. 27-41.
Soroori E, Kiani B, Ghasemi S, Mohammadi A, Shabanikiya H, Bergquist R, Kiani F, Tabatabaei-Jafari H. 2023. Spatial Association between urban neighbourhood characteristics and child pedestrian–motor vehicle collisions. Appl Spat Anal Policy 16:1443-62.
Wheeler D.C. 2021. Geographically weighted regression. Handbook of regional science, Springer: 1895-921.