[AM-03-034] The Geographically Weighted Regression Framework

Abstract

Local multivariate statistical models are increasingly encountered in geographical research to estimate spatially varying relationships between a response variable and its associated predictor variables. In geography and many other disciplines, such models have been largely embedded within the framework of regression and can reveal significantly more information about the determinants of observed spatial distribution of the dependent variable than their global regression model counterparts. This section introduces one type of local statistical modeling framework: Geographically Weighted Regression (GWR). Models within this framework estimate location-specific parameter estimates for each covariate, local diagnostic statistics, and bandwidth parameters as indicators of the spatial scales at which the modeled processes operate. These models provide an effective means to estimate how the same factors may evoke different responses across locations and by so doing, bring to the fore the role of geographical context on human preferences and behavior.

Keywords: bandwidths, geographically weighted regression, local spatial models, multiscale, spatial context, spatial process heterogeneity, spatial process scale

Author & citation

Sachdeva, M., and Fotheringham, A. S. (2020). The Geographically Weighted Regression Framework. The Geographic Information Science & Technology Body of Knowledge (4th Quarter 2020 Edition), John P. Wilson (Ed.).
DOI: 10.22224/gistbok/2020.4.7.

This Topic is also available in the following editions:

DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., and Wentz, E. (2006). Spatial Expansion and Geographically Weighted Regression. The Geographic Information Science & Technology Body of Knowledge. Washington, DC: Association of American Geographers.

Explanation

1. Local Statistical Modeling Frameworks
2. Spatially Varying Processes and the Concept of Data-borrowing
3. Geographically Weighted Regression Model Specification
4. Multiscale Geographically Weighted Regression
5. MGWR 2.2 Software
6. Concluding Remarks

1. Local Statistical Modeling Frameworks

A fundamental aspect of geographical research is to understand processes operating between people, objects and events through an examination of their observed spatial patterns (Harvey, 1968). The fact that an attribute value varies across space is taken for granted within most studies - it is extremely rare to uncover some measurement of our physical and social environments that is constant across space. Various methods have been developed to investigate aspects of spatial variation of an observed variable such as fractal analysis, variograms to analyze spatial structure, local models of spatial autocorrelation and wavelets (Lloyd, 2006). While it is interesting to examine spatial variation in data, it is often of more value to ask why such variation exists. Exploring this question leads to a need to understand the underlying processes causing the observed spatial pattern. Traditional global statistical models assume that these processes are constant over space so that variations in y are assumed to be caused solely by variations in one or more of the covariates which affect y. Local models relax this assumption and allow the possibility that these processes vary over space. Such models hence quantify not only how each covariate can affect y but also if this effect varies spatially. In this sense, as shown in Figure 1, traditional regression models such as the Ordinary Least Squares (OLS) regression, which estimate a single, average coefficient for each relationship are special cases of the more general local model formulation and only apply when relationships are stationary over space.

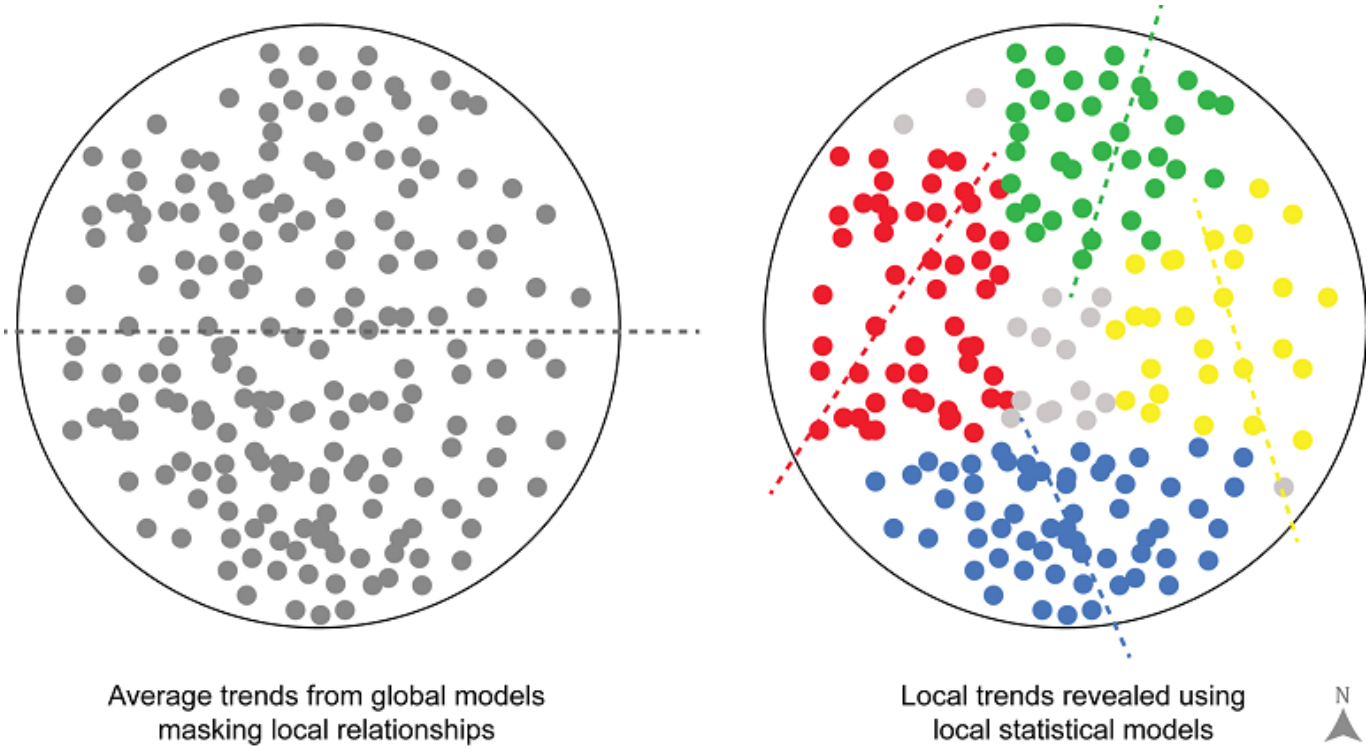


Figure 1. Local trends revealed using local statistical models (right) which are otherwise masked by ‘averaged’ global trends (left). Source: authors.

This section of the GIS&T Book of Knowledge focuses on the Geographically Weighted Regression (GWR) framework - the GWR model and its recent multiscale extension, Multiscale Geographically Weighted Regression (MGWR) (Brunsdon et al., 1996;

Fotheringham et al., 1996; Fotheringham et al., 2003; Fotheringham et al., 2017; LeSage, 2004). Other local modeling frameworks such as eigenvector spatial filter-based local regression (SFLR) and Bayesian spatially-varying coefficient (SVC) models (Banerjee et al., 2014; Gelfand et al., 2003; Murakami et al., 2017) are not discussed but comparisons between GWR and these other frameworks can be found in (Oshan & Fotheringham, 2018; Wolf et al., 2018).

2. Spatially Varying Processes and the Concept of Data-borrowing

In the context of the GWR framework, we define processes as the conditional relationships between the dependent variable (y) and independent variables (x). Where global regression models, such as OLS and some spatial regression models, assume these relationships to be constant over space, this assumption is relaxed in the GWR framework. For example, a traditional global regression model linking variations in crime rates across an urban area to variations in income levels would produce a single parameter estimate to establish the effect of income on crime rates. It would be assumed that the effect of income on crime rates is constant across the entire study area. However, many social processes, particularly those involving human decisions and behaviors, might vary over space so that in such cases the global model will be misspecified. It might be, for example, that in some parts of the urban area, income variations have a stronger effect on crime rates than in other parts. Models within the GWR framework hence estimate location-dependent relationships between the independent and dependent variables, in effect allowing the processes being modeled to vary across space. For this reason, GWR and some other techniques, such as Anselin's local indices of spatial associations (Anselin, 1995), have been classified as place-based analytic techniques which respond to the reality of complex sciences (Goodchild, 2009). The insights gained from such techniques can highlight the important role place plays in affecting people's beliefs, behaviours and decision-making. For example, many studies suggest that preference for type of housing, political affinity and choice of mode of transport depend significantly on location (Chandola et al., 2005; Enos, 2017; Escobar, 2001; Panter et al., 2016; Walker & Li, 2007). Models within the GWR framework account for and quantify such spatial variations in processes.

In order to calibrate a separate model for each location when only one observation of y and each covariate is typically recorded, GWR models borrow data from neighboring observations and weight these data according to a smooth decay function based on either a physical distance or the number of nearest neighbors. Data from nearby locations are weighted (1-0) more highly than are data from more distant locations with the rate of decrease in the weighting determined by a bandwidth parameter which controls the distance, or the nth nearest neighbor, at which weights fall to zero (or approximately zero). Small bandwidths denote more local processes; large bandwidths indicate regional or global processes. As long as an optimal bandwidth is determined in the calibration of the GWR model and some continuous smooth function of distance is used, the specific kernel function chosen is not critical. Several kernel functions have been employed in GWR with the most common being bisquare, Gaussian and exponential functions.

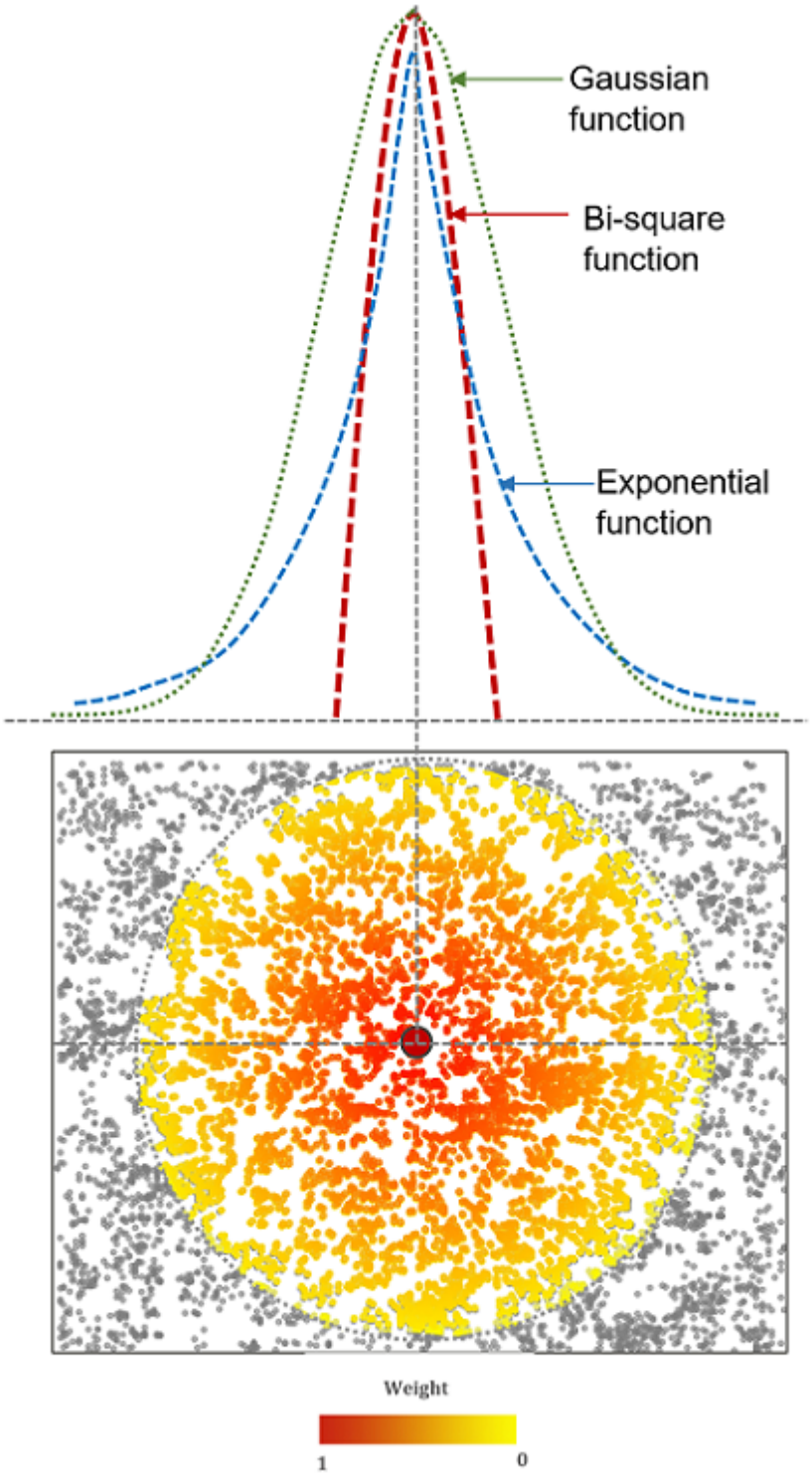


Figure 2: Weighting scheme in GWR with different distance-decay kernel functions (top) and a generic example of the weights at each point (bottom). Source: authors.

Spatial weighting kernels can be defined as ‘fixed’ or ‘adaptive’ (Figure 3 left and right, respectively). Fixed kernels have the same rate of distance-decay for all locations whereas adaptive kernels have different rates of distance-decay depending on the density of data points in the vicinity of the regression point. Depending on the choice of the kernel, the bandwidth is typically defined as either the distance at which weights fall below a certain

value (fixed) or the number of nearest neighbors from the regression point which receive a non-zero weight in the local regressions (adaptive). A large bandwidth allows data from locations further from the regression point to be included in the local regressions whereas a small bandwidth restricts data in the local model calibrations to those recorded at locations in close proximity to the regression point.

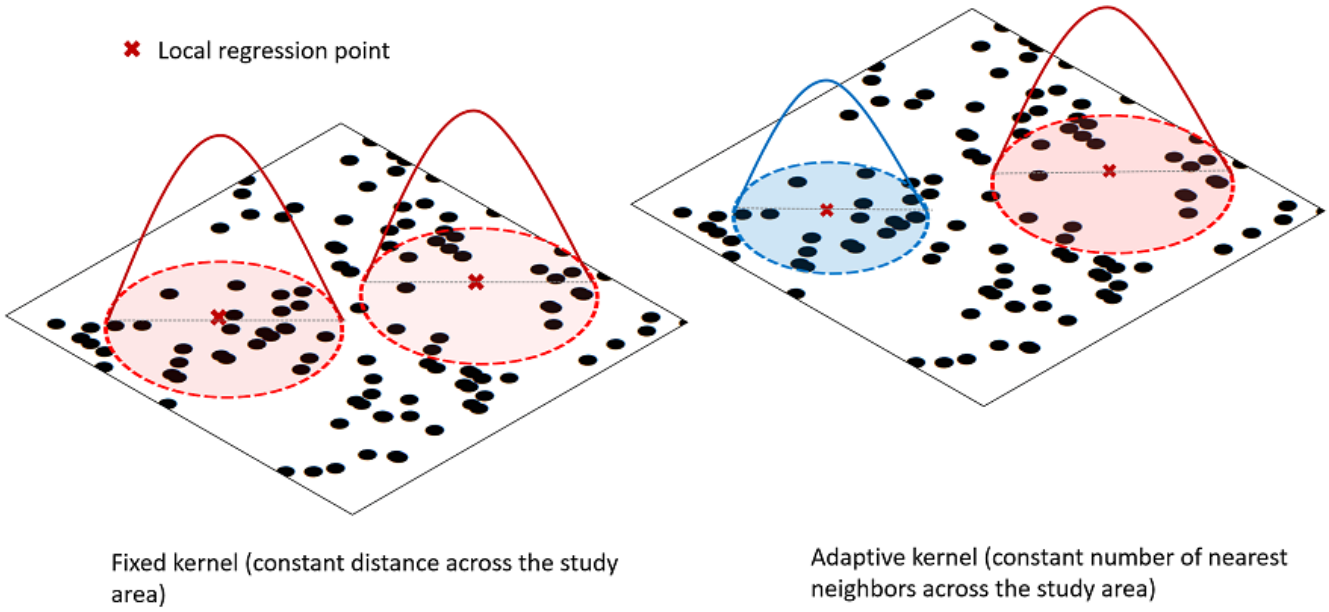


Figure 3. Conceptual diagram explaining fixed (left) and adaptive weighting (right) schemes. Source: authors.

Conceptually, the optimal bandwidth selection is a trade-off between bias and variance. If the relationships being modeled are spatially varying, greater bias will be introduced into the local parameter estimates as the bandwidth increases because the data borrowed from locations farther from the local regression point will have been produced by increasingly different processes. However, the smaller the bandwidth, the greater will be the uncertainty regarding the local parameter estimates because the local models will have been calibrated with fewer data points. The selection of the bandwidth parameter is based on a statistical optimization criterion that includes a trade-off between model fit and model complexity, such as a corrected Akaike Information Criteria (AICc) (Fotheringham et al., 2003; Yu et al., 2020(a)). Most GWR software packages include several options for the goodness-of-fit criterion used in bandwidth selection but as Yu et al. (2020(b)) demonstrate, minimizing AICc is a very good approximation to finding the optimal trade-off between bias and variance.

The bandwidth parameter is a useful output in GWR calibration because it is an indicator of the spatial scale over which processes operate. A small optimal bandwidth indicates local processes that vary significantly across the study region while a large optimal bandwidth suggests a stable relationship that does not vary much over space.

3. Geographically Weighted Regression Model Specification

A GWR model is calibrated as an ensemble of ordinary least squares regressions, individually estimated at each location where data are observed using the neighboring weighted data. By allowing coefficient estimates to be derived for each location, GWR explicitly incorporates spatial context. The model can be written as:

$$y_i = \beta_0(u_i, v_i) + \sum_j \beta_j(u_i, v_i) X_{ji} + \varepsilon_i$$

where yi is the dependent variable, Xji is the jth independent variable, βj (ui,vi) is the jth coefficient at location (ui,vi), and εi is the random error term. Unlike OLS, the parameters are allowed to vary by location (ui,vi).

Since the GWR technique estimates a surface of local parameters for each modeled relationship using overlapping subsets of data, a classic t-test to examine the significance of each local estimate would lead to an excess of false positives being recorded at a given significance level due to multiple hypothesis testing (Williams et al., 1999). The severity of the multiple hypothesis testing issue for GWR is dependent on the optimized bandwidth obtained in the model calibration as shown by da Silva and Fotheringham (2016). To obtain an α-value that accounts for this issue, a correction is applied such that:

$$\alpha = \frac{\xi}{\frac{ENP}{p}}$$

where ξ is the expected type I error rate before correction, ENP is the effective number of parameters in GWR which is a function of the optimal bandwidth parameter and p is the number of independent variables in the model (da Silva & Fotheringham 2016). It is important to use this corrected α value to assess the significance of local parameter estimates to avoid the proportion of false positives exceeding α.

A Monte Carlo test can be employed for parameter estimates from MGWR (see below) and GWR to distinguish situations where the observed spatial variation in the parameter surfaces may only be due to noise. This test checks the variation in the estimated parameter surface against the variations in parameter surfaces constructed from randomly arranged data points to examine whether the spatial variation in the estimated surface is significantly higher than that in the randomly constructed surfaces. Such a test is implemented in the software described below.

4. Multiscale Geographically Weighted Regression

Since different relationships might operate at different spatial scales, an emerging trend in local models is to allow the bandwidth parameter to be covariate-specific rather than having a single bandwidth represent the spatial scale at which every relationship being

modeled varies. For instance, the recently developed multi-scale extension to GWR, Multiscale GWR (MGWR), allows for a unique bandwidth parameter to be derived for each relationship within a model:

$$y_i = \beta(u_i, v_i) + \sum_j \beta_{bwj}(u_i, v_i) X_{ji} + \epsilon_i$$

where the subscript bwj indicates the bandwidth used for the calibration of the jth conditional relationship (Fotheringham et al., 2017; Yang, 2014). The MGWR model calibration employs a more complex backfitting algorithm and is therefore more time-consuming to converge but recent advances have led to the model being calibrated on even very large data sets in a manageable time (Li et al., 2019; Li & Fotheringham, 2020). For example, in a model explaining house prices using multiple independent variables such as income, age of a structure, living area and access to public amenities, each affecting the response variable at a unique spatial scale, MGWR can estimate a unique scale for each covariate-specific association (Figure 4).

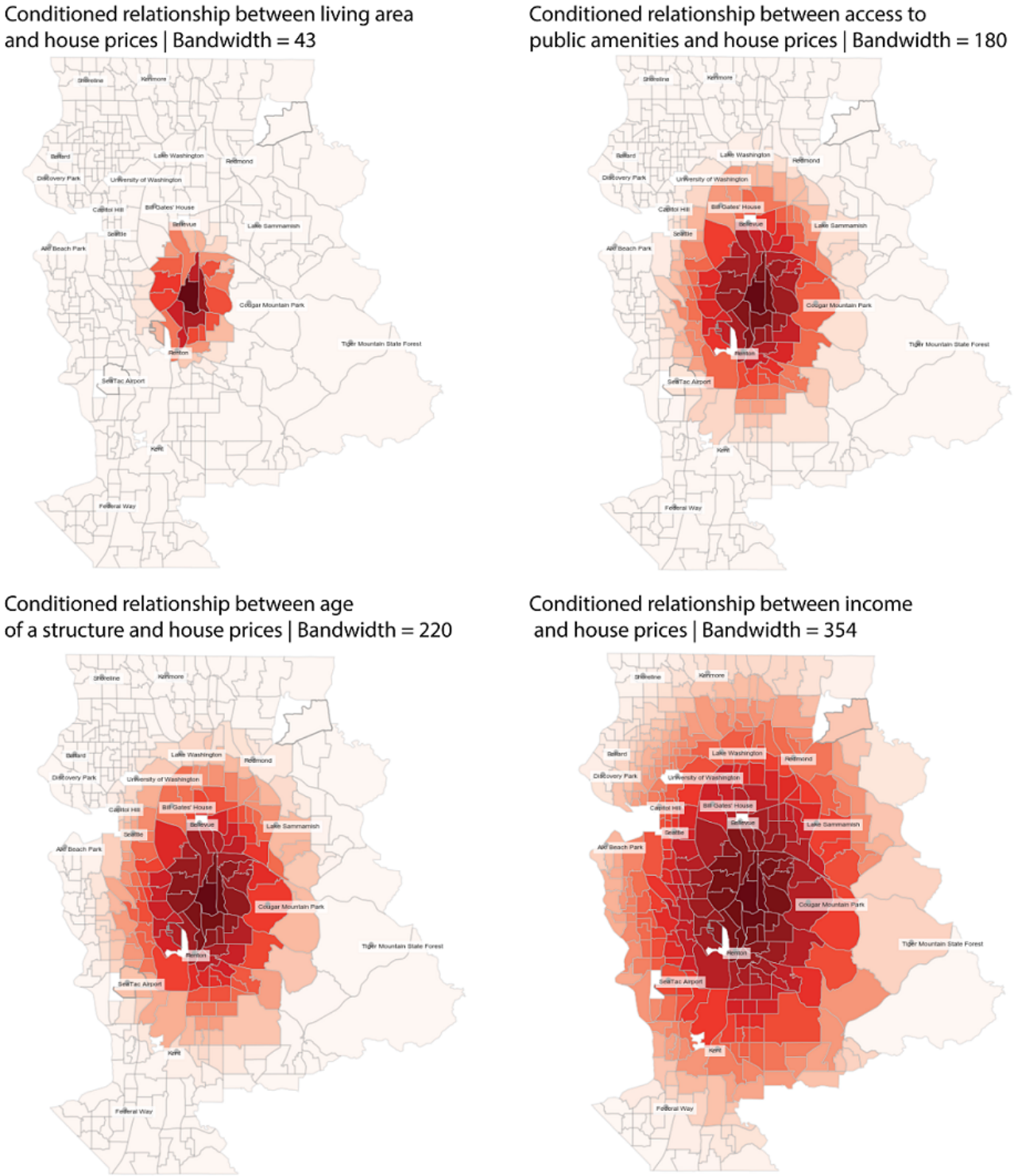


Figure 4. Example of multiple estimated bandwidths interpreted as the unique scale at which different processes operate. In this example, the data are for house prices in King County, Washington, USA (377 total Census tracts). Source: authors.

Similar to the inference in GWR, the location specific parameter estimates produced in the MGWR model also require a correction to account for multiple dependent hypothesis tests. In this case, the correction in equation (2) is still applied but it is covariate-specific because the value of ENP will vary across covariates as it is controlled by the optimized bandwidth.

The software described below outputs a covariate-specific value of ENP as well as a corrected t value for each covariate.

The selection of optimal covariate-specific bandwidths in MGWR is based on a model selection score such as the corrected Akaike Information Criterion (AICc), similar to the selection criterion of a single bandwidth in GWR. As discussed, the bandwidth parameters are either given in number of nearest neighbors or a distance metric and can be compared to assess the relative operational scale of processes across covariates. The MGWR calibration process also produces confidence intervals for covariate specific bandwidths using Akaike weights (Li et al. 2020). MGWR marks a significant advancement over GWR and is described further in Fotheringham et al. (2017), Wolf et al. (2018), and Yu et al. (2020(a)).

5. MGWR 2.2 Software

MGWR 2.2 software provides a user-friendly, graphical interface for calibrating the GWR and MGWR models and can be downloaded from https://sgsup.asu.edu/sparc/mgwr for both Windows and Mac OS platforms (Figure 5). The types of models that can be fit using the software consist of the Gaussian, Poisson or Logit GWR models and a Gaussian MGWR model. Standard file formats such as comma separated, dbase IV and excel can be used to input data into the software. The software reads the first row of any dataset as variable names by default and populates the variable list. The location variables can be either projected or spherical and are required inputs from the user in the software. The spatial kernel types can be chosen from adaptive bisquare or fixed Gaussian and the optimization criterion has AIC, AICc, Bayesian Information Criterion (BIC) and Cross Validation (CV) as available options. After choosing the dependent and independent variables from the variable list, the user can select the folder for storing the resulting output files. The MGWR 2.2 software produces two output files once a model is successfully run: a summary text file and a .csv file containing georeferenced parameter estimates, their associated t-values, standard errors and local goodness of fit measures. The .csv file can be subsequently joined to the associated shapefile and the results can be mapped using any software such as ArcMap or QGIS. Some advanced options are also available in the software including Monte Carlo tests for spatial variation, tests for local collinearity, confidence intervals for estimated bandwidths and options to change the threshold for convergence of the MGWR backfitting algorithm.

An open-source Python package is also available for the software (https://github.com/pysal/mgwr) that has some added functionality and provides the option to directly map and analyze the results from the model in Python or R. A comprehensive route map to appropriately choosing (M)GWR in any research is provided by Comber et al. (2020) and an in-depth guide to using the open-source package is available by Oshan et al. (2019).

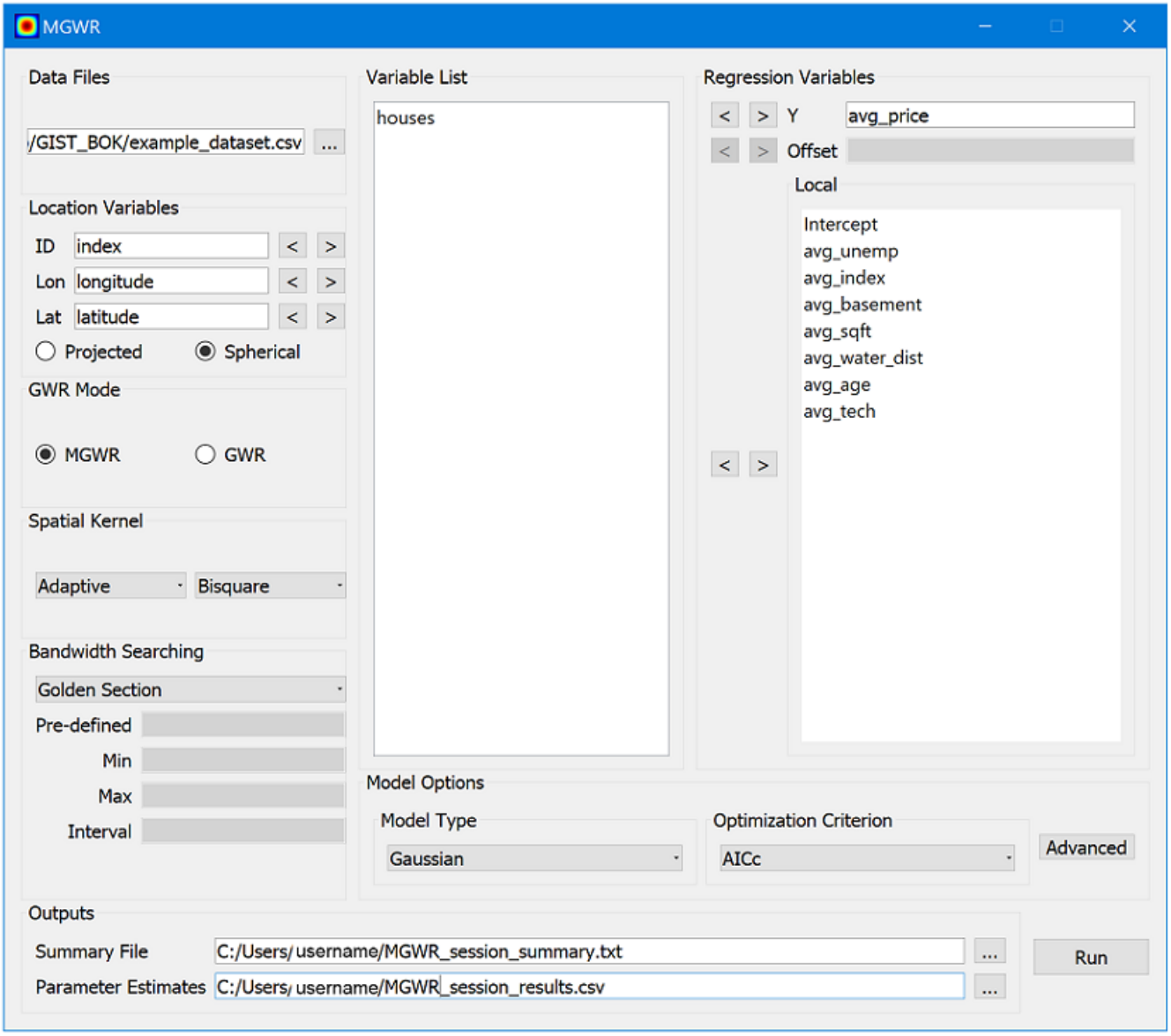


Figure 5. MGWR 2.2 main interface with variables selected for model run. Source: authors.

6. Concluding Remarks

Models within the GWR framework provide an effective method of capturing spatially varying processes and estimating the spatial scale at which those processes operate. By estimating location-specific parameters, such models are able to provide more information about the processes that lead to an observed pattern of data than traditional global regression models. Readers interested in applying the GWR framework of models are referred to Comber et al., 2020; Cupido et al., 2020; Fotheringham et al., 2019; Oshan et al., 2020; Fotheringham et al., 2020; Gu et al., 2020; Li, 2020; and Wu et al., 2019 for best practices and examples of recent applications.

References

Anselin, L. (1995). Local Indicators of Spatial Association -- LISA. Geographical Analysis,

27(2):93-115.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2014). Hierarchical Modeling and Analysis for Spatial Data, 2nd edition. Chapman and Hall/CRC Press Monographs on Statistics and Applied Probability, Boca Raton, FL.

Brunsdon, C., Fotheringham, A.S., & Charlton, M.E. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis, 28(4), 281-298.

Chandola, T., Clarke, P., Wiggins, R. D., & Bartley, M. (2005). Who you live with and where you live: Setting the context for health using multiple membership multilevel models. Journal of Epidemiology and Community Health, 59(2), 170–175.

Comber, A., Brunsdon, C., Charlton, M., Dong, G., Harris, R., Lu, B., Lü, Y., Murakami, D., Nakaya, T., Wang, Y., & Harris, P. (2020). The GWR route map: A guide to the informed application of Geographically Weighted Regression. ArXiv:2004.06070 [Stat].

Cupido, K., Fotheringham, A. S., & Jevtic, P. (2020). Local modelling of U.S. mortality rates: A multiscale geographically weighted regression approach. Population, Space and Place. 27(1).

Enos, R. D. (2017). The Space between Us: Social Geography and Politics. Cambridge University Press.

Escobar, A. (2001). Culture sits in places: Reflections on globalism and subaltern strategies of localization. Political Geography, 20(2), 139–174.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons.

Fotheringham, A. S., Charlton, M., & Brunsdon, C. (1996). The geography of parameter space: An investigation of spatial non-stationarity. International Journal of Geographical Information Systems, 10(5), 605–627.

Fotheringham, A. S., Yang, W., & Kang, W. (2017). Multiscale geographically weighted regression (MGWR). Annals of the American Association of Geographers, 107(6), 1247-1265.

Fotheringham, A. S., Yue, H., & Li, Z. (2019). Examining the influences of air quality in China's cities using multi-scale geographically weighted regression. Transactions in GIS, 23(6), 1444–1464.

Gelfand, A. E., Kim, H. J., Sirmans, C. F., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. Journal of the American Statistical Association, 98(462), 387-396.

Goodchild, M. F. (2009). What Problem? Spatial Autocorrelation and Geographic Information

Science. Geographical Analysis, 41(4), 411–417.

Gu, H., Yu, H., Sachdeva, M., & Liu, Y. (2020). Analyzing the distribution of researchers in China: An approach using multiscale geographically weighted regression. Growth and Change.

Harvey, D. W. (1968). Pattern, Process, and the Scale Problem in Geographical Research. Transactions of the Institute of British Geographers, 45, 71–78.

LeSage, J. P. (2004). A Family of Geographically Weighted Regression Models. In L. Anselin, R. J. G. M. Florax, & S. J. Rey (Eds.), Advances in Spatial Econometrics: Methodology, Tools and Applications (pp. 241–264). Springer.

Li, Z. (2020). Multiscale Geographically Weighted Regression Computation, Inference, and Application [Ph.D. Dissertation, Arizona State University].

Li, Z., & Fotheringham, A. S. (2020). Computational improvements to multi-scale geographically weighted regression. International Journal of Geographical Information Science, 34(7), 1378–1397.

Li, Z., Fotheringham, A. S., Li, W., & Oshan, T. (2019). Fast Geographically Weighted Regression (FastGWR): A scalable algorithm to investigate spatial process heterogeneity in millions of observations. International Journal of Geographical Information Science, 33(1), 155–175.

Li, Z., Fotheringham, A. S., Oshan, T. M., & Wolf, L. J. (2020). Measuring Bandwidth Uncertainty in Multiscale Geographically Weighted Regression Using Akaike Weights. Annals of the American Association of Geographers, 110(5), 1500–1520.

Lloyd, C. D. (2006). Local Models for Spatial Analysis, 1st Edition. CRC Press.

Murakami, D., Yoshida, T., Seya, H., Griffith, D. A., & Yamagata, Y. (2017). A Moran coefficient-based mixed effects approach to investigate spatially varying relationships. Spatial Statistics, 19, 68–89.

Oshan, T. M., & Fotheringham, A. S. (2018). A Comparison of Spatially Varying Regression Coefficient Estimates Using Geographically Weighted and Spatial-Filter-Based Techniques. Geographical Analysis, 50(1), 53–75.

Oshan, T. M., Li, Z., Kang, W., Wolf, L. J., & Fotheringham, A. S. (2019). mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. ISPRS International Journal of Geo-Information, 8(6), 269.

Oshan, T. M., Smith, J. P., & Fotheringham, A. S. (2020). Targeting the spatial context of obesity determinants via multiscale geographically weighted regression. International Journal of Health Geographics, 19(1), 11.

Panter, J., Heinen, E., Mackett, R., & Ogilvie, D. (2016). Impact of New Transport

Infrastructure on Walking, Cycling, and Physical Activity. American Journal of Preventive Medicine, 50(2), e45–e53.

Silva, A. R. da, & Fotheringham, A. S. (2016). The Multiple Testing Issue in Geographically Weighted Regression. Geographical Analysis, 48(3), 233–247.

Walker, J. L., & Li, J. (2007). Latent lifestyle preferences and household location decisions. Journal of Geographical Systems, 9(1), 77–101.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling Error in Multiple Comparisons, with Examples from State-to-State Differences in Educational Achievement. Journal of Educational and Behavioral Statistics, 24(1), 42–69.

Wolf, L. J., Oshan, T. M., & Fotheringham, A. S. (2018). Single and Multiscale Models of Process Spatial Heterogeneity. Geographical Analysis, 50(3), 223–246.

Wu, C., Ren, F., Hu, W., & Du, Q. (2019). Multiscale geographically and temporally weighted regression: Exploring the spatiotemporal determinants of housing prices. International Journal of Geographical Information Science, 33(3), 489–511.

Yang, W. (2014). An extension of geographically weighted regression with flexible bandwidths [Thesis, University of St Andrews].

Yu, H., Fotheringham, A. S., Li, Z., Oshan, T., & Wolf, L. J. (2020(b)). On the measurement of bias in geographically weighted regression models. Spatial Statistics, 38, 100453.

Yu, H., Fotheringham, A. S., Li, Z., Oshan, T., Kang, W., & Wolf, L. J. (2020(a)). Inference in Multiscale Geographically Weighted Regression. Geographical Analysis, 52(1), 87–106.