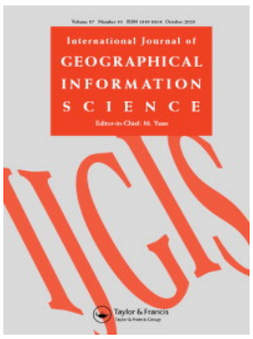

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

On the local modeling of count data: multiscale geographically weighted Poisson regression

Mehak Sachdeva, A. Stewart Fotheringham, Ziqi Li & Hanchen Yu

To cite this article: Mehak Sachdeva, A. Stewart Fotheringham, Ziqi Li & Hanchen Yu (2023) On the local modeling of count data: multiscale geographically weighted Poisson regression, International Journal of Geographical Information Science, 37:10, 2238-2261, DOI: 10.1080/13658816.2023.2250838

To link to this article: https://doi.org/10.1080/13658816.2023.2250838

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2023, VOL. 37, NO. 10, 2238–2261
https://doi.org/10.1080/13658816.2023.2250838

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

On the local modeling of count data: multiscale geographically weighted Poisson regression

Mehak Sachdeva[a], A. Stewart Fotheringham[b], Ziqi Li[c] and Hanchen Yu[d]

[a]Center for Urban Science and Progress, Tandon School of Engineering, New York University, New York, NY, USA; [b]School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA; [c]Department of Geography, Florida State University, Tallahassee, FL, USA; [d]School of Urban Governance and Design, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China

ABSTRACT

A recent addition to the suite of techniques for local statistical modeling is the implementation of the multiscale geographically weighted regression (MGWR), a multiscale extension to geographically weighted regression (GWR). Using a back-fitting algorithm, MGWR relaxes the restrictive assumption in GWR that all processes being modeled operate at the same spatial scale and allows the estimation of a unique indicator of scale, the bandwidth, for each process. However, the current MGWR framework is limited to use with continuous data making it unsuitable for modeling data that do not typically exhibit a Gaussian distribution. This study expands the application of the MGWR framework to scenarios involving discrete response outcomes (count data following a Poisson's distribution). Use of this new MGWR Poisson regression (MGWPR) model is demonstrated with a simulated data set and then with COVID-19 case counts within New York City at the zip code level. The results from the simulated data underscore the superiority of the MGWPR model in effectively capturing spatial processes that influence count data patterns, particularly those operating across diverse spatial scales. For empirical data, the results reveal significant spatial variations in relationships between socio-ecological factors and COVID-19 cases – variations often missed by traditional 'global' models.

ARTICLE HISTORY
Received 16 December 2022
Accepted 18 August 2023

KEYWORDS
Local Poisson regression; spatial process scale; multiscale geographically weighted Poisson model; COVID-19; local scoring algorithm

1. Introduction

Regression modeling is a popular ensemble of tools employed to unearth plausible explanations for the spatial and aspatial variations often inherent in observed phenomena. Many formulations within the broader regression framework have been developed to model the hypothesized relationships between a dependent variable (y) and single or multiple independent variables (x). 'Global' models, a term often used to refer to the traditional techniques that postulate spatially stationary processes, have been extensively deployed in the analysis of observed phenomena within the natural,

CONTACT Mehak Sachdeva mehaksachdeva@nyu.edu

social and physical sciences. However, a perspective that is gaining considerable traction within spatial analysis is that conditioned relationships might vary across space, suggesting geographical variations in how covariates influence the dependent variable based on context (Fotheringham 2020, Fotheringham and Sachdeva 2021, 2022). For example, although certain demographic variables might correlate with such preferences for certain types of music, an important determinant is that of cultural heritage, or what geographers refer to as 'spatial context' (Hauser 1970, Relph 1976, Agnew 1996, King 1996, Golledge 1997, Goodchild 2011, Agnew 2014). The contextual background to many decisions is widely recognized but is exceedingly difficult to measure. Sometimes, geographical interaction terms (indicator variables for example in spatial regime models) are introduced into models as a quick fix to solve the contextual problem but the addition of such terms demands *a priori* knowledge of the regions in which contextual effects occur. It also assumes that such contextual effects are uniform within these regions and change in a discontinuous manner at the boundaries of these regions. Both assumptions are highly questionable in most applications. Local models, such as multiscale geographically weighted regression (MGWR), provide a more effective alternative of identifying the range and intensity of contextual effects by allowing the intercept and the conditioned associations within a model to vary across space. Local models often convincingly outperform their global counterparts, as evident in the abundant empirical literature comparing the two frameworks (*inter alia*, Zhang *et al.* 2004, Malczewski and Poetz 2005, Maroko *et al.* 2009, Cardozo *et al.* 2012, Wang *et al.* 2018, Zhu *et al.* 2020).

The current implementation of the MGWR model assumes a Gaussian modeling framework. Fotheringham *et al.* (2017) applied the Gaussian MGWR model to population change data in Ireland from 1841 to 1851, the period of the Great Famine, and found spatially varying associations operational at unique spatial scales. MGWR has since been used in modeling house prices, air quality, obesity rates, voting behavior and mortality rates among many other spatial phenomena and appears to provide more accurate estimates of spatially varying associations than its uniscale counterpart (Fotheringham *et al.* 2019, Oshan *et al.* 2020, Cupido *et al.* 2021, Fotheringham *et al.* 2021, Sachdeva *et al.* 2022). While a Gaussian modeling framework is useful in modeling observed phenomena that follow a normal distribution, many empirical data, especially in epidemiology, transportation and ecological analyses, exist in the form of integer counts, such as traffic crash incidents, and disease contraction counts, which often follow a Poisson distribution. It is well-known that an implementation of a Gaussian model on data that follow a Poisson distribution can lead to erroneous predictions and misspecification problems. A Poisson random variable is often used to model counts with a minimum value of zero, a theoretically unbound maximum value and assuming only integer values, unlike normal data that are continuous and have a theoretical unbounded maxima and minima. Moreover, the variance of a variable following a Poisson distribution is assumed to be equal to its mean. If a Gaussian linear regression model is used to model data that are Poisson distributed, the predictions could result in erroneous negative estimations and the constant variance assumption for normal linear regression inference would be violated.[1]

Consequently, since Poisson regression provides a more appropriate framework to analyze discrete data (especially for low numbers) than conventional Gaussian regression, it would be useful to extend the current Gaussian-based MGWR framework to incorporate Poisson distributed data. Moreover, the predecessor model to MGWR, geographically weighted regression (GWR), extends the Gaussian model to accommodate a Poisson distribution for the dependent variable and this extension has been widely used in literature (Nakaya *et al.* 2005).[2] Here, we propose and develop a new statistical model, that of multiscale geographically weighted Poisson regression (MGWPR), to achieve this. This is not straightforward because the existing MGWR framework, which employs a backfitting algorithm used in the calibration of generalized additive models (GAMs) (Hastie and Tibshirani 1986, Buja *et al.* 1989, Everitt 2005), needs to be expanded and a general local scoring algorithm (LSA) has to be employed to allow the definition of different distributions of y (the response variable) and their associated link functions, before operationalizing the backfitting algorithm. The definition of different distributions and their associated link functions has further ramifications for the inference calculations for the model, which are also developed and described in this paper. Finally, this model is tested using simulated data and by expanding an existing COVID-19 study in New York City.

The remainder of this paper is organized as follows. In Section 2, we describe the framework of MGWPR and its calibration procedure. The inference procedure for the model is described in Section 3. A simulation experiment is constructed and tested in Section 4 where comparisons between global Poisson regression (Poisson GLM; Agresti 2002), GWPR and MGWPR are made. Finally, an empirical application of MGWPR using COVID-19 positive case counts in New York City at the zipcode level is described in Section 5, followed by conclusions and discussion in Section 6.

2. Specification of multiscale geographically weighted Poisson regression

Poisson regression falls within the umbrella of *generalized linear models*, a framework which generalizes OLS regression to enable its use with different distributions of response variables (e.g. binary, count, categorical, etc.). To do so, a transformation of the response variable is first applied using a *link function* specific to the distribution being modeled. This enables a linear estimation of the association between the response and predictor variables. In a Poisson regression model, the link function is the natural logarithm and the predicted response variable following the transformation is measured in the natural logarithms of the original counts. Since the Poisson distribution only allows discrete, non-negative integers, a Poisson regression model is a common choice to model count data.

2.1. Poisson regression specification

A typical Poisson regression to model the expected count value of y_i denoted by $E(y_i|x_{k,i})$ can be specified as follows:

$$E\left(y_i|x_{k,i}\right) \sim Poisson\left[\exp\left(\sum_{k=0}^{K}\beta_k x_{k,i}\right)\right] \tag{1}$$

where $x_{k,i}$ is the kth predictor covariate at location i, β_k is the parameter representing the conditioned relationship between the response variable and predictor variable k and *Poisson* $[\lambda]$ indicates a Poisson distribution with mean and variance $= \lambda$. This specification includes β_0, the intercept, with x_{0i} representing an array of all 1s. The specification in Equation (1) represents a global model where a single parameter represents the conditioned association between each covariate and the response variable across the entire study region. A local equivalent of this model, from the geographically weighted Poisson regression specification (Nakaya *et al.* 2005), can be expressed as follows:

$$E(y_i|x_{k,i}) \sim Poisson \left[\exp\left(\sum_{k=0}^{K} \beta_{k,i} x_{k,i}\right)\right] \quad (2)$$

where the response variable at each location i is modeled using weighted data from neighboring locations to estimate covariate and location-specific parameter estimates, $\widehat{\beta_{k,i}}$. A spatial kernel is used to develop a weighting matrix (with weights in the range 0–1) that assigns larger weights to neighboring locations and smaller weights to more distant locations following an optimized distance-decay function that is estimated from the data. The amount of neighboring data used in a local model is governed by either the number of locations used (adaptive bandwidth) or the radius used to select the locations (fixed bandwidth). The bandwidth parameter has a maximum of n, the number of data points (or a distance parameter), which represents a global relationship, with fewer data points (or smaller distance radii) representing more local processes. The inherent assumption in, and drawback of, GWPR is that the bandwidth parameter is the same for all relationships within a model, which could result in severe misspecification since different associations may vary over different spatial scales. This assumption is relaxed in the MGWPR model described below.

2.2. Multiscale geographically weighted Poisson regression

A multiscale version of the geographically weighted Poisson regression model is specified as follows:

$$E(y_i|x_{k,i}) \sim Poisson \left[\exp\left(\sum_{k=0}^{K} \beta_{bwk,k,i} x_{k,i}\right)\right] \quad (3)$$

where *bwk* as a subscript to the beta estimates represents the covariate (k)-specific bandwidths. Similar to the calibration procedure for MGWR, the challenge of estimating covariate-specific bandwidths is solved in MGWPR by using a backfitting algorithm. To calibrate the covariate-specific bandwidth, the term $\beta_{bwk,k,i}\ x_{i,k}$ from Equation (3) is defined as the kth additive term in a Gaussian additive model equivalent specification as follows:

$$y = \sum_{k=0}^{K} f_k + \epsilon \quad (4)$$

Since such a backfitting algorithm fits only normal additive models, further transformations are required to estimate unique bandwidths using the backfitting algorithm for a Poisson model. To do this, an adjusted response variable as described by

McCullagh and Nelder (2019) is constructed and then used within the backfitting algorithm. This calibration procedure for MGWPR draws from the LSA defined by Hastie and Tibshirani (1986) from the GAM literature.

First, the additive terms ${f_0}^0$, ${f_1}^0$, ${f_2}^0$, ${f_3}^0$, ..., ${f_k}^0$ are initialized. There are multiple options for the initialization including GLM estimates, GWPR estimates or zeroes. As Fotheringham *et al.* (2017) note, there is negligible difference in the results regardless of which initialization option is used. We describe the calibration procedure here with an initialization using GWPR estimates. Once a GWPR model is calibrated, a predictor term η_{ji} (Equation (5)), weights w_i (Equation (6)) and an adjusted dependent variable z_i (Equation (7)) are constructed. This is the first iteration of the LSA calibration loop and hence a superscript '0' is added as below.

$$\eta_i^{(0)} = \sum_{k=0}^{K} \beta_{i,k}^{(0)} x_{i,k} \tag{5}$$

where $\beta_{i,k}^{(0)}$ is estimated using GWPR for each covariate k. Next, predictions from the estimated GWPR parameters are calculated as follows:

$$\widehat{O_i}^{(0)} = E_i * e^{\eta_i} \tag{6}$$

where E_i is the expected count or the offset term in the model. Then, an adjusted dependent variable is constructed as:

$$z_i^{(0)} = \eta_i^{(0)} + \frac{O_i - \hat{O}_i^{(0)}}{\hat{O}_i^{(0)}} \tag{7}$$

The predictions from the GWPR estimates in Equation (6) are also used to construct a weighting matrix, as shown in Equation (8).

$$A_i = \begin{bmatrix} \hat{O}_1^{(0)} & 0 & 0 & 0 \\ 0 & \hat{O}_2^{(0)} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \hat{O}_n^{(0)} \end{bmatrix} \tag{8}$$

Finally, the adjusted dependent variable and the constructed weights (A_i) are used to fit the backfitting algorithm in a similar manner to that described by Fotheringham *et al.* (2017). Within the backfitting algorithm, the weights from LSA (A_i) and the spatial weights from the bandwidth optimization are both used until the convergence criterion is met. The overall weights in the backfitting algorithm are:

$$W^*{}_i^{(0)} = W_i^{(0)} * (A_i)^{(0)} \tag{9}$$

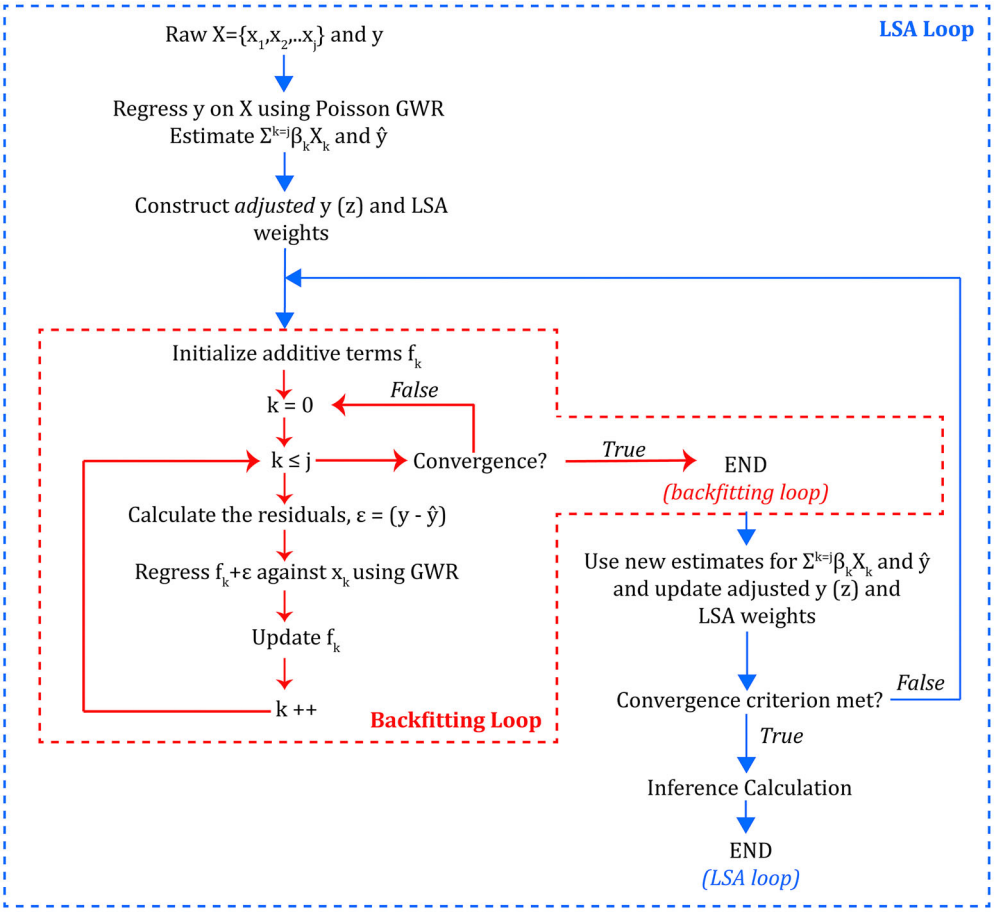

Figure 1. LSA loop for multiscale geographically weighted Poisson regression.

After convergence is achieved within the backfitting loop, a second iteration of the LSA loop starts after updating values from Equations (5)–(9). This continues until the convergence criterion for the LSA loop is satisfied. The convergence criterion for the *m*th iteration in LSA is defined as follows:

$$\delta^{m}{}_{numerator} = \sum_{i=1}^{n}\sum_{k=1}^{K}\frac{(\beta_{ik}{}^{(m)} - \beta_{ik}{}^{(m-1)})^2}{n};$$

$$\delta^{m}{}_{denominator} = \sum_{i=1}^{n}\sum_{k=1}^{K}(\beta_{ik}{}^{(m)})^2;$$

$$\delta^{m} = \sqrt{\frac{\delta^{m}{}_{numerator}}{\delta^{m}{}_{denominator}}} \tag{10}$$

We set the tolerance for δ^m at 10^{-8} by default.

The estimation procedure for MGWPR hence consists of two loops. This is further clarified through Figure 1. The outer loop (in blue) is the LSA that transforms the dependent variable using a link function and estimates weights. Inside each outer LSA loop is a weighted backfitting inner loop (in red) that runs until convergence. The

new estimates from the backfitting loop are then used to calculate a new adjusted dependent variable and new weights, which are again run until convergence within the backfitting loop. The LSA stops when the change in the parameter estimates is below the tolerance. Because of the added complexity, the computational time requirements to calibrate a MGWPR model are much greater than for GWPR and Gaussian MGWR as we demonstrate below. As noted by Li *et al.* (2019), the time complexity for a GWR model is $O(k^3 n^2 \log\ n)$ where k is the number of covariates, and n is the number of data points in a model. The time complexity for MGWR is given by $O(kdn^2 \log\ n)$ where d is the number of iterations required for convergence in the backfitting algorithm. The calibration of MGWPR requires an outer LSA loop that increases the time complexity to $O(mkdn^2 \log\ n)$, where m is the number of iterations of the LSA loop before convergence. Since the memory allocation for each LSA loop remains constant across the m iterations, the memory complexity for MGWPR is expected to be similar to that for MGWR $O(kn)$ (Li and Fotheringham 2020). Similarly, the time complexity for inference calculation in an MGWPR model is expected to increase from $O(kdn^3)$ for MGWR, to $O(mkdn^3)$ owing to the m LSA iterations while not affecting memory allocation requirements.

3. Inference for MGWPR

Calculations for MGWPR inference closely follow those for GWPR (Nakaya *et al.* 2005) and MGWR (Yu *et al*. 2020). At convergence, the parameter estimates at each regression point are given in the following equation:

$$\hat{\beta}_{ki} = C_k z_i \tag{11}$$

where

$$C_k = (X_k{}^t W_i\ A_i X_k)^{-1} * \ X_k{}^t\ W_i\ A_i \tag{12}$$

The LSA weights and adjusted dependent variable z_i on the right-hand side of Equation (11) are calculated based on the converged parameter estimates. The variance–covariance matrix of the estimated local parameter estimates is then given by:

$$cov\ (\hat{\beta}_{ki}) \ = \ C_k {A_i}^{-1} C_k{}' \tag{13}$$

and the standard error of the kth parameter estimate is given by:

$$SE(\hat{\beta}_{ki}) \ = \ \sqrt{cov\ (\hat{\beta}_i)_k} \tag{14}$$

where $cov\ (\hat{\beta}_i)_k$ is the kth diagonal element of the variance–covariance matrix defined in Equation (14). Consequently, the local pseudo t statistic for the kth parameter at location i is computed by:

$$t\left(\widehat{\beta_{ki}}\right) \ = \ \frac{\widehat{\beta_{ki}}}{SE(\widehat{\beta_{ki}})} \tag{15}$$

Given the assumption that the true regression parameter equals zero, the distribution of values as specified in Equation (15) will tend to follow a standard normal distribution. The covariate-specific hat matrix S_k and the model hat matrix S ($S = \sum_{k=1}^{k=K} S_k$)

can be calculated following the iterative process described in Yu *et al.* (2020) so that the effective number of parameters can be obtained to calculate model goodness-of-fit scores such as AICc:

$$AICc = \sum_{i=1}^{N}\left(\frac{O_i \log\left(\hat{O}(\beta_i)\right)}{O_i} + O_i - \hat{O}(\beta_i)\right) + 2 * v_1 + 2 * \frac{v_1 * (v_1 + 1)}{N - v_1 - 1} \tag{16}$$

where $v_1 = trace(S)$ is the model effective number of parameters.

4. A simulation experiment

To demonstrate the performance of MGWPR, and its superiority to GWPR and global generalized linear models, a simulation experiment was designed to assess the following five aspects of model performance:

1. Estimation of the scale across which processes vary;
2. Reproduction accuracy of the spatial heterogeneity of different processes;
3. Model performance (flexibility and goodness of fit);
4. Replication of the response variable; and
5. Computational overhead.

4.1. Simulation design

Three surfaces of parameters representing processes with varying degrees of spatial heterogeneity were constructed. Local surfaces for β_0 and β_1 were simulated using a two-dimensional spatial random field (SRF) with a Gaussian covariance model on a 25 by 25 grid ($n = 625$). The Gaussian variogram employed to construct the processes is:

$$\gamma(h, \lambda) = \sigma^2 \left[1 - e^{-\frac{\pi}{4}\left(\frac{h}{\lambda}\right)^2}\right] \tag{17}$$

where length-scale (h), the distance parameter (λ) controlling the range of spatial autocorrelation and variance (σ^2) control the amount of spatial heterogeneity in the surfaces with larger values of h corresponding to lower spatial heterogeneity and higher values of σ^2 resulting in greater variation in the estimates. To construct the surface of β_0 values, $h = 5$ and $\sigma^2 = 2$; for the surface of β_1, $h = 50$ and $\sigma^2 = 1$; and β_2 is constant over space with magnitude $= 0.5$. The three simulated parameter surfaces are shown in Figure 2.

The covariates x_1 and x_2 are drawn from random normal distributions with mean $=$ 0 and standard deviation $= 1$. To ensure a Poisson distribution for the response variable, an expected number of values, μ, is constructed using the following equation:

$$\mu = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} \tag{18}$$

and y is drawn at random from this distribution. The distribution and spatial variation of the constructed response variable are shown in Figure 3. One thousand such datasets were constructed to ensure robustness of the results given the randomness in the response variable.

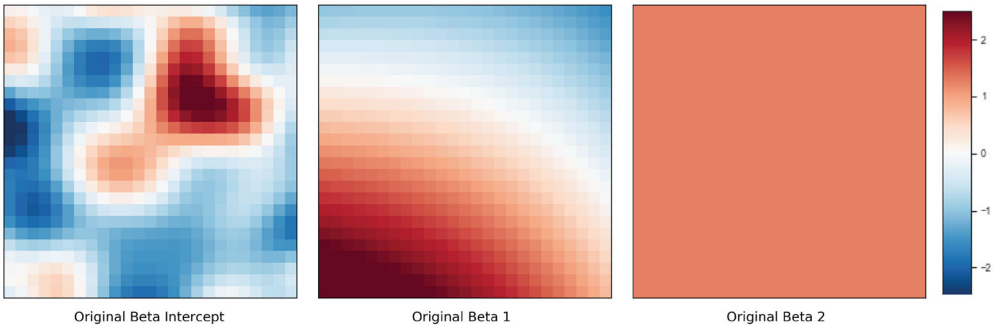

Figure 2. Simulated parameter estimates for β_0, β_1 and β_2 (left to right).

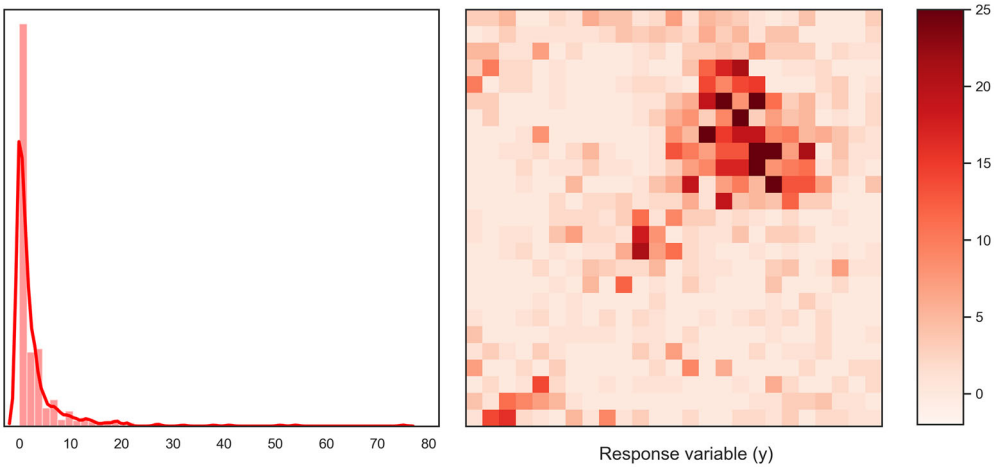

Figure 3. Simulated response variable distribution (left) and spatial distribution (right).

4.2. Simulation results

We now describe the results of calibrating MGWPR, GWPR and GLM models on the constructed synthetic data described above. Within the LSA we use a tolerance of 10^{-8} and GWPR estimates as the initialization values for the additive terms, for each of the 1000 iterations of the data construction and model runs. For the simulation experiments, we employ an adaptive kernel following a bisquare decay function, which optimizes and returns the number of nearest neighbors employed in each localized regression, also termed as the *bandwidth* parameter. We restrict the minimal value for the bandwidth optimization search algorithm to 43, following the current MGWR Gaussian implementation.

4.2.1. Bandwidth estimation

The estimated optimal bandwidth(s) from GWPR and MGWPR from each of the 1000 runs are shown in Figure 4, which demonstrates the sensitivity of the bandwidth across the realizations. The distribution for the single bandwidth estimated in a GWPR model ranges from 50 to 70. The bandwidth estimated for the local intercept using

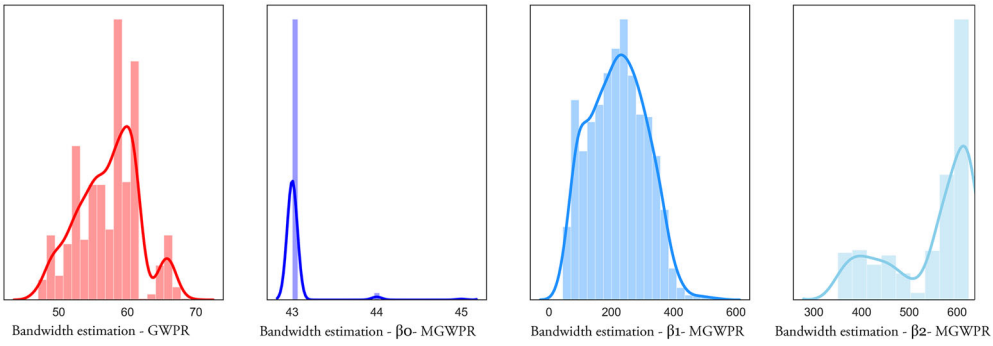

Figure 4. Bandwidths estimated from GWPR (leftmost) and from MGWPR for β_0, β_1 and β_2 (left to right, respectively).

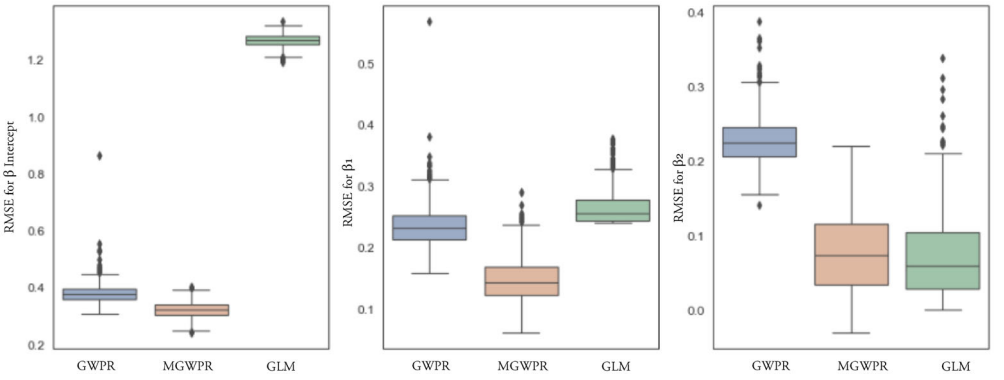

Figure 5. $RMSE_k$ comparisons for GLM, GWPR and MGWPR.

MGWPR is almost always 43 (the most local bandwidth estimation possible). The bandwidth for β_1 and β_2 ranges from about 100 to 400 (with a mean of 225) and from 400 to 624 (with a mean of 556), respectively. It is clear from these distributions in the context of the constructed process surfaces in Figure 2 that MGWPR accurately estimates the spatial scale at which the surfaces vary. GWPR, on the other hand, estimates a rather local bandwidth representing all the three parameter surfaces despite the differences in their degrees of spatial heterogeneity.

4.2.2. Local parameter estimation accuracy

The accuracy with which both MGWPR and GWPR replicate the three known parameter surfaces, $\beta_{i,0}$, $\beta_{i,1}$ and $\beta_{i,2}$, for each of the 1000 simulations is measured in two ways. First, the root mean squared error (RMSE) for each parameter estimate $\beta_{i,k}$ from both MGWPR and GWPR is calculated as follows:

$$RMSE_{i,k} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\beta_{i,k} - \widehat{\beta_{i,k}})^2} \tag{19}$$

Smaller values of RMSE represent better replication of the known parameter estimates and the calculated values for GWPR and MGWPR are shown in Figure 5.

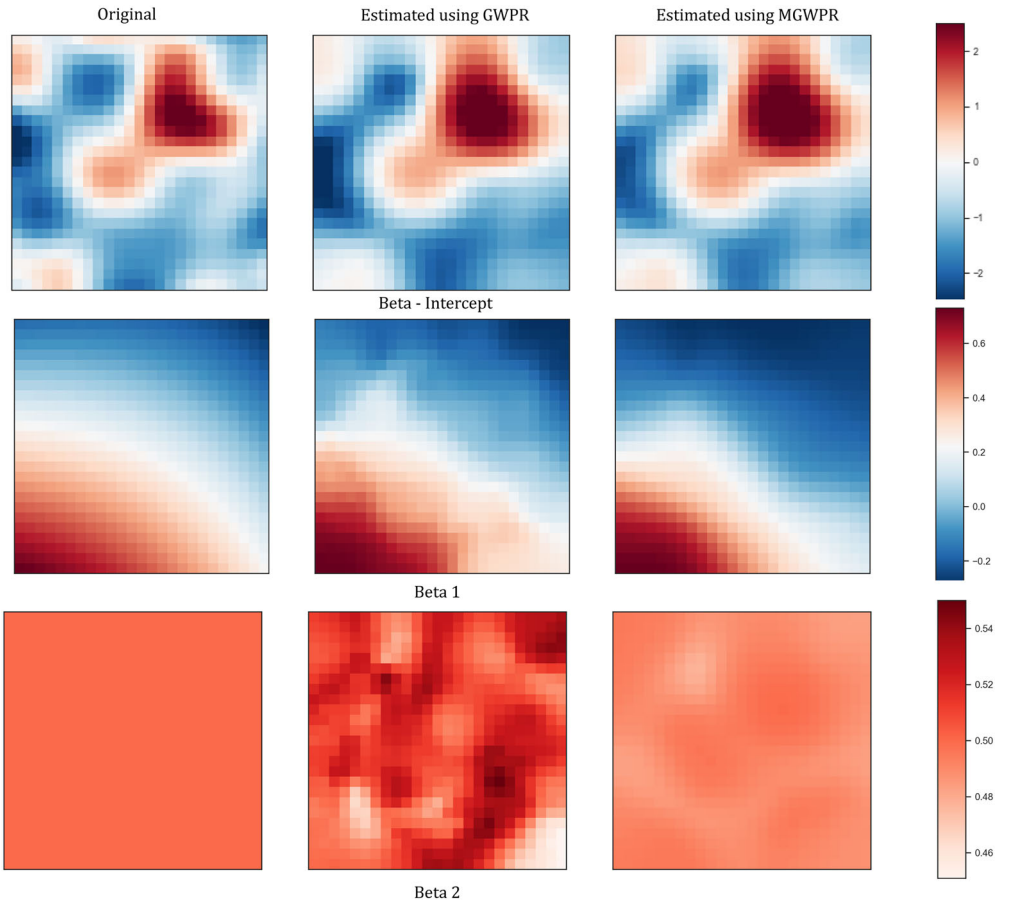

Figure 6. Estimated local parameters from GWPR and MGWPR.

Second, the estimated parameter surfaces from both MGWPR and GWPR are averaged across the 1000 simulations and mapped to visually inspect the local parameter estimation accuracy.

In Figure 5, the overall lower values and tighter fit around the mean for MGWPR indicate that the model is able to replicate the three parameter surfaces more accurately than GWPR. This is supported by the visual inspection of the plotted surfaces shown in Figure 6.

MGWPR closely replicates the local surfaces for all three parameters. While GWPR does a good job in replicating the more local parameter surfaces (β_0 and β_1), it is not able to replicate the global surface (β_2) correctly.

4.2.3. Model performance (flexibility and goodness of fit)

Figure 7(a) shows the comparison of the AICc values from the three models in each of the 1000 iterations. GLM clearly, and as expected, is not able to accurately replicate the data owing to the spatial nonstationarity in the simulated parameter surfaces. Figure 7(b) highlights the superiority of MGWPR over GWPR by rescaling without the GLM results.

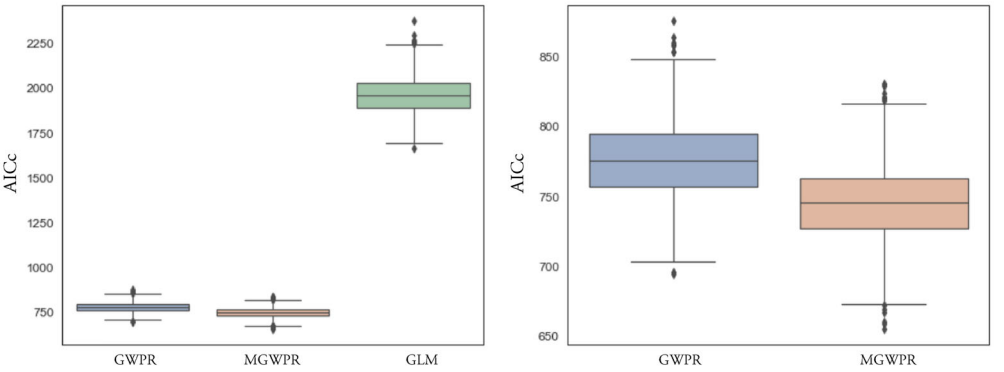

Figure 7. Comparison of AICc estimates from (a) GLM, GWPR and MGWPR and (b) GWPR and MGWPR.

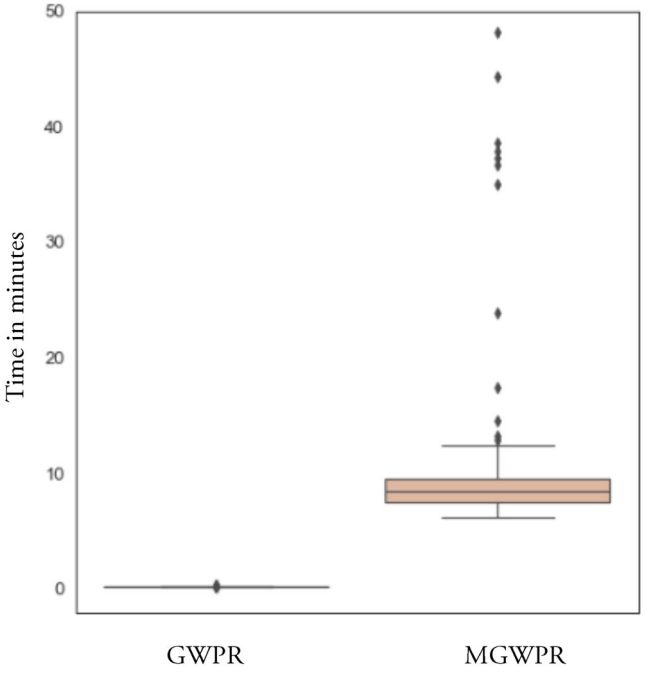

Figure 8. Average time taken to convergence for both the GWPR and MGWPR models.

4.2.4. Computational efficiency

MGWPR is a more complex model than GWPR and takes longer to reach convergence. The external LSA loop has an internal backfitting loop which adds to the computational time. Figure 8 shows the time in minutes required to run the two models in each of the 1000 data sets. For this data set ($n = 625$; $k = 3$), the calibration of MGWPR takes around 10 minutes on average on a 10 core system[3] while GWPR takes only 0.4 minutes on average.

In summary, in a controlled simulation experiment, MGWPR is better able to estimate both parameter spatial heterogeneity and the spatial scale of this heterogeneity

than GWPR, at the expense of additional computational time. We now examine the use of MGWPR with real-world data on COVID-19 positive tests in New York City at the zipcode level.

5. An empirical example of the use of MGWPR in analyzing the spatial distribution of COVID-19 counts

This section contains an empirical example of the MGWPR model using COVID-19 positive cases data in the early phase of the pandemic in New York City at the Zip Code Tabulation Area (ZCTA) level. In order to demonstrate the value of local modeling with count data, this example expands an existing study undertaken by DiMaggio *et al.* (2020) using a global Poisson model. Using similar data sources, model specification and discussions from DiMaggio *et al.* (2020), we can test the MGWPR model and showcase the importance of estimating covariate-specific indicators of scale of the spatial heterogeneity of the different processes being modeled. It is worth noting that there are two limitations of the data that restrict our findings from this analysis. First, the data are collected for only 183 areal units, which is too few to properly showcase the outputs from local models to the fullest extent but serves the purpose here of demonstrating how the new local modeling framework works. Second, the temporal span for which data are collected (from 3 April to 22 April 2020, representing the first phase of the pandemic), is extremely narrow and is not expected to represent the true variation in the phenomenon.

While a number of recent studies have investigated the ecological risk factors associated with COVID-19 at the county level in the United States (Gao *et al.* 2020, Khazanchi *et al.* 2020, Hughes *et al.* 2021), few studies have focused on the ZCTA or more granular spatial scales (Liu *et al.* 2021, Kedron *et al.* 2022). DiMaggio *et al.* (2020) provide an exception to this and make their data sources and code available. Additionally, they use a Poisson Bayesian regression model with a spatially structured term to analyze COVID-19 case counts. This presents an opportunity to analyze the same data using MGWPR. While DiMaggio *et al.* (2020) implement and discuss various univariate and multivariate models, we focus on the following model, using the notations in their paper:

$$Y \sim Pois(\lambda_i = E_i \cdot \theta_i);\ \log \theta_i = \beta_0 + \beta_j x_{ij} + v_i + \eta_i + (offset) \tag{20}$$

where θ_i is the response variable measuring positive COVID-19 test cases, $\beta_j x_{ij}$ represents the various predictors employed in the model and their corresponding parameter estimates (note that these are global estimates with a single value estimated for the covariate across the region), v_i is a spatially unstructured random effects term and η_i is the spatially structured random effects term. The primary conclusion from the calibration of this model in DiMaggio *et al.* (2020) is that areas with higher percentages of African American population were at higher risk of contracting COVID-19. The universe of predictors considered in their models and the source for those data are presented in Table 1. Figure 9 presents the number of total COVID-19 tests and the number of positive COVID-19 cases in NYC at the ZCTA level from 3 April 2020 to 22 April 2020. These closely match the data presented in DiMaggio *et al.* (2020).

Table 1. The universe of all variables explored in DiMaggio *et al.* (2020) with data sources.

Variable	Source (DiMaggio et al.)	Source for this study
COVID-19 test result data	NYC DOHMH Github page	NYC DOHMH Github page
Total Population	US Census 2010	US Census 2010
Proportion of people older than 65 years	US Census 2010	US Census 2010
% of African American pop.	US Census 2010	US Census 2010
% pop. speaking one language other than English	US Census 2010	US Census 2010
No. of people per sq. mile	US Census 2010	US Census 2010
No. of schools per sq. mile	US Census 2010	US Census 2010
No. of houses per sq. mile	US Census 2010	US Census 2010
% pop. receiving public assistance	US Census 2010	US Census 2010
Social fragmentation index (Congdon)	Combination of 4 variables calculated from US Census 2010	Combination of 4 variables calculated from US Census 2010
% of people with heart disease or congestive failure	Simply Analytics (paid data)	CDC 500 Cities data
% of people with chronic obstructive pulmonary disease (COPD)	Simply Analytics (paid data)	CDC 500 Cities data
Shapefiles of NYC ZCTAs	NYC Dept. of City Planning	NYC Dept. of City Planning

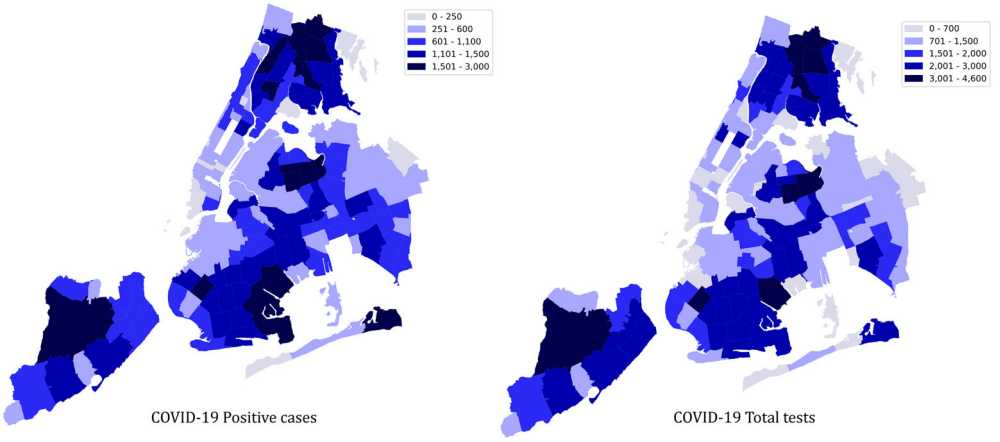

Figure 9. Total COVID-19 tests and positive cases in NYC at the ZCTA level (3 April–22 April 2020).

The multivariate model constructed and discussed by DiMaggio *et al.* (2020) consists of the following predictors: % of people with chronic obstructive pulmonary disease (COPD); % of pop. with heart disease; % of African American pop.; housing density; and % of pop. older than 65. Only two of these variables, % of African American population and % of the population older than 65, had significant associations (both positive) with COVID-19 cases. However, after the extraction, transformation and testing of the data employed by DiMaggio *et al.* (2020), we found multicollinearity issues in their predictors as shown in Table 2. In Bayesian models, unbiased estimates of the parameters can be obtained in the presence of multicollinearity by using an informative prior and this may have been the case in DiMaggio *et al.* (2020), but here we need to deviate from a direct reproduction of their model because we use a frequentist model, which would be prone to bias because of the

Table 2. Variance inflation factors for the predictors used by DiMaggio *et al.* (2020).

Variable	VIF
% of people with COPD	14.46
% of people with heart disease or congestive failure	15.33
% of African American pop.	1.35
No. of houses per sq. mile	1.07
Proportion of people older than 65 years	1.8

A value of VIF <10 is generally considered acceptable.

Table 3. Poisson GLM results.

Variable	Est.	SE	*t*(Est/SE)	*p* Value
Intercept	−0.661	0.003	−203.9	.00
% African American	0.051	0.003	16.4	.00
% Pop. with heart disease	0.056	0.005	11.5	.00
Pop. density (pop/sq.mile)	0.157	0.013	12.1	.00
No. of schools per sq. mile	−0.197	0.014	−14.1	.00
% Pop. receiving public asst.	0.015	0.005	3.2	.001
% Hispanic	0.057	0.004	14.8	.00

multicollinearity in the data. The covariates used in the model here are selected using an optimization algorithm designed to select the set with the best model performance based on AICc.[4] The model is shown in Equation (21), which follows the MGWPR model described in Equation (3).

$$\begin{aligned} \mathrm{E}\ (\mathit{Positive\ cases}) \sim \mathit{Poisson}[\mathit{Offset}_i\ (\mathit{Total\ tests})\ \exp\ (\beta_{0,i,bw0} \\ +\ \beta_{1,i,bw1}\ (\%\ \mathit{Afr.\ American}) \\ +\beta_{2,i,bw2}\ (\mathit{pop\ density})+\beta_{3,i,bw3}\ (\%\ \mathit{heart}) \\ +\ \beta_{4,i,bw4}\ (\%\ \mathit{Hispanic})+\ \beta_{5,i,bw5}\ (\%\ \mathit{pub\ asst}) \\ +\ \beta_{6,i,bw6}\ (\mathit{schools\ per\ sq.\ mile}))] \end{aligned} \tag{21}$$

5.1. Poisson GLM results

The global model estimated using a Poisson GLM has an explained deviance of 44.2% and an AICc value of 1660. All the predictors are significant at the 1% confidence level and the estimated parameters and their standard errors are shown in Table 3. The residuals from the model are mapped in Figure 10.

Except for the school density parameter, which is significantly negative, all the predictors affect the response variable in a significant positive manner. Since the link for the Poisson regression model is the natural log, holding all other predictors constant, an increase in one unit for x multiplies the rate of y by the exponent of β. Since the predictors in this model are also standardized, the raw parameter estimates correspond to a one standard deviation increase in x leading to y being increased or decreased by a factor of e^{β}. To simplify interpretation of coefficients hereafter, we transform the parameter estimates as follows:

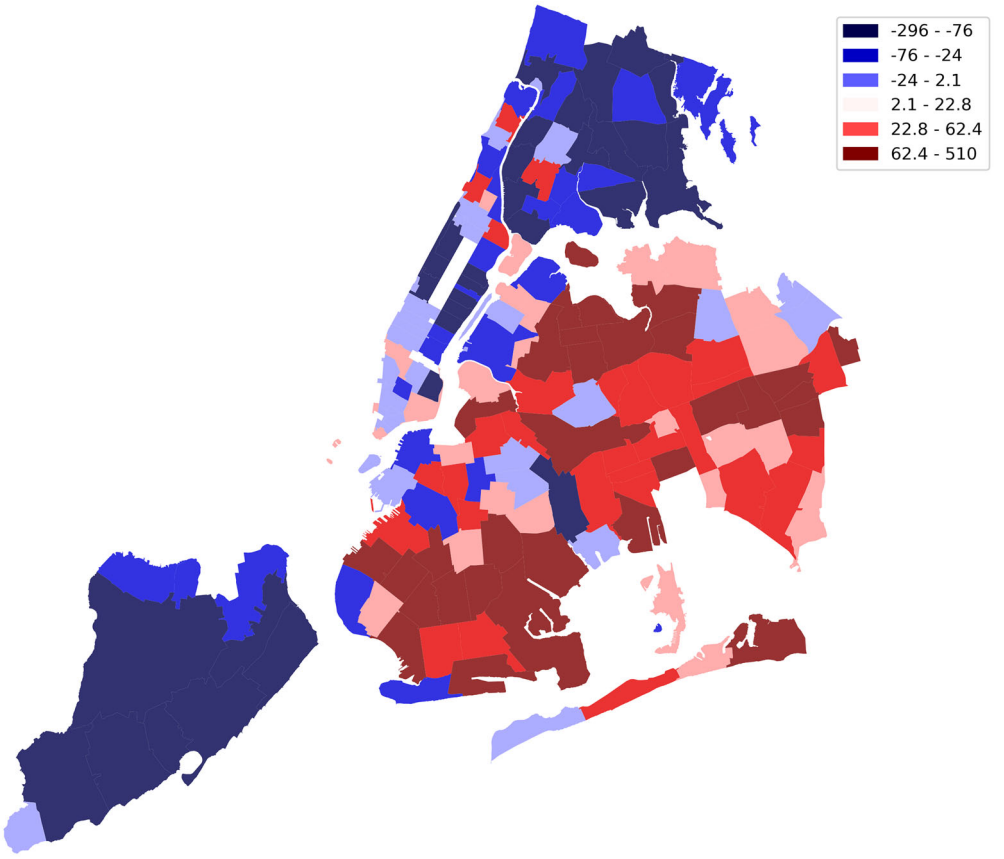

Figure 10. Global model (Poisson GLM) residuals.

$$\beta^* = (e^{\left(\frac{\beta_{raw}}{\sigma_x}\right)} - 1) * 100 \tag{22}$$

This transformation results in coefficient values that are interpreted as follows: a one unit increase in x would affect y by β^* percentage.

As shown in Figure 10, the global residuals from the Poisson GLM are strongly positively spatially correlated with a Moran's I value of 0.45 and p value of .0024. This suggests the global model is severely misspecified and that a local model might be more appropriate. Consequently, we calibrate the GWPR and MGWPR models using the predictors, response and offset variables as described in Equation (21).

5.2. GWPR and MGWPR results

The AICc for the GWPR model is 511 and that for MGWPR is 480 (compared to 1600 for the GLM model). The deviance explained for the two models is 85.4% and 86.7%, respectively (compared to 44.2% for the GLM model). While the difference in model performance between GWPR and MGWPR is not large, the MGWPR model provides more information by allowing the estimation of covariate-specific bandwidths, which in turn should lead to superior prediction of the local parameters. The single

Table 4. Bandwidths estimated using MGWPR with their 95% confidence intervals.

Variable	Bandwidth	Bandwidth CI
Intercept	43	(43.0, 47.0)
% African American	92	(88.0, 96.0)
% Pop. with heart disease	45	(45.0, 75.0)
Pop. density (pop/sq.mile)	65	(63.0, 75.0)
No. of schools per sq. mile	49	(49.0, 63.0)
% Pop. receiving public asst.	111	(109.0, 129.0)
% Hispanic	80	(75.0, 96.0)

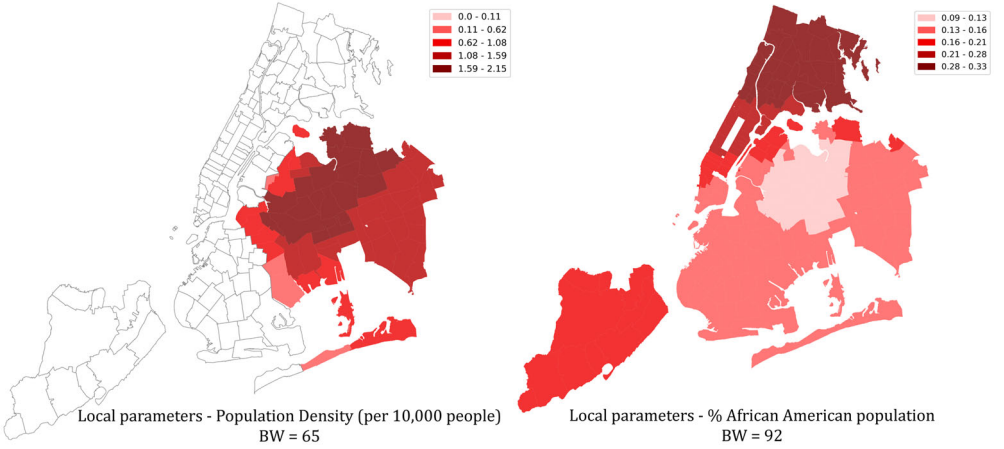

Figure 11. Local parameter estimates for population density and % of African American population (calibrated using MGWPR).

bandwidth estimated for GWPR is 59 (given a total of 183 ZCTAs). The covariate-specific bandwidths (along with their 95% confidence intervals followed by Li *et al.* 2020) estimated from MGWPR are shown in Table 4 and suggest a range of heterogeneity in the processes being modeled. For example, the bandwidth for the local association between % of the population receiving public assistance and the probability of testing positive for COVID-19 is 111 (109–129) representing a fairly global process while the local intercept represents a local process with an estimated bandwidth of 43. It bears repeating that the temporal and spatial scale of analysis for this study is extremely limited and hence the processes estimated using sophisticated regression models such as MGWPR are not expected to reveal plausible insights into the actual operational processes. For example, the most populous zip code in NYC has 108,661 residents (ZCTA 11368 in Brooklyn; US Census 2010) and the average population for all the zip codes is approximately 46,000. Additionally, the arbitrary and short timespan for which the data are collected and analyzed are hardly representative of the true phenomena and the factors associated with it in the city. For this reason, we describe only a limited number of local parameter maps resulting from MGWPR to showcase the implementation of MGWPR rather than to interpret the processes affecting the spread of COVID-19 in NYC.

Figure 11 shows the significant (at the 95% confidence level accounting for multiple hypothesis testing[5]) local parameter estimates for the predictors representing

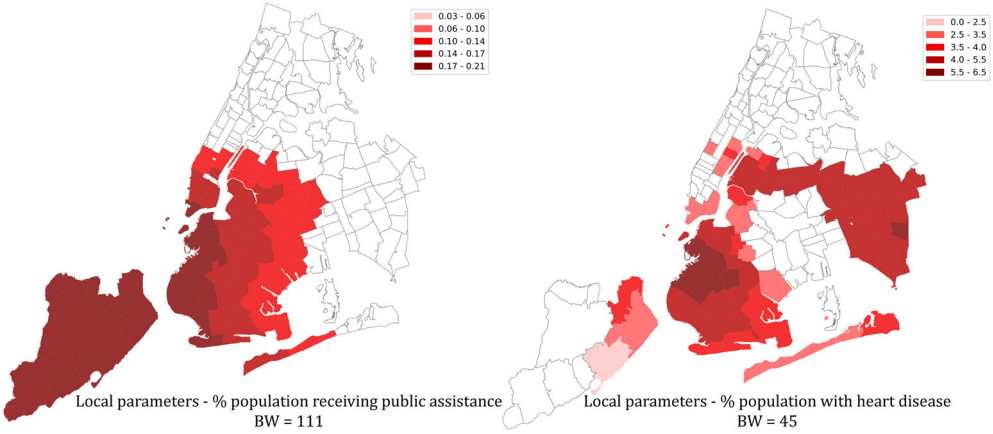

Figure 12. Local parameter estimates for % of population with heart disease and % of population receiving public assistance (calibrated using MGWPR).

population density and percentage of African American population. DiMaggio *et al.* (2020) report a significant positive effect of population density on positive COVID-19 cases across NYC using a global model. The local estimates in Figure 11(a) from MGWPR similarly suggest that holding all other covariates constant, an increase in population density would lead to an increase in positive COVID-19 cases but this association is only significant in Queens and in some parts of Brooklyn. Figure 11(b) shows the significant associations between the percentage of African American population and the % of COVID-19 cases, *ceteris paribus*. This association was reported as significantly positive across NYC by DiMaggio *et al.* (2020) but the local parameter estimates suggest a large variability in this relationship across the city. A 1% increase in African American population would lead to a 0.33% increase in COVID-19 positive cases in neighborhoods in the Bronx and in parts of Upper Manhattan, while in some parts of Queens this drop to just a 0.1% increase.

Figure 12 shows two other relationships with Covid cases estimated in this analysis, that of the percentage of the population with heart disease and the percentage of the population receiving any kind of public assistance. DiMaggio *et al.* (2020) note that in their multivariate model where racial predictors were accounted for, predictors measuring comorbidities such as heart disease or COPD were not significantly associated with positive COVID-19 cases. It is possible to estimate significant local relationships that might not be significant at the global level, as noted by Fotheringham and Sachdeva (2022) and Sachdeva and Fotheringham (2023). Using MGWR, the predictor measuring the association between heart disease and the probability of contracting the virus is seen to be significant in some parts of the city. This suggests that holding the racial predictors constant, not only the disease severity but also the acquisition of the disease might be significantly associated with comorbidities in some parts of the city. One explanation could be the more frequent hospital visits for patients suffering some conditions such as heart disease that could potentially bring them into contact with other infected patients. Finally, the association between the percentage of the population receiving public assistance and the percentage of positive COVID-19 cases

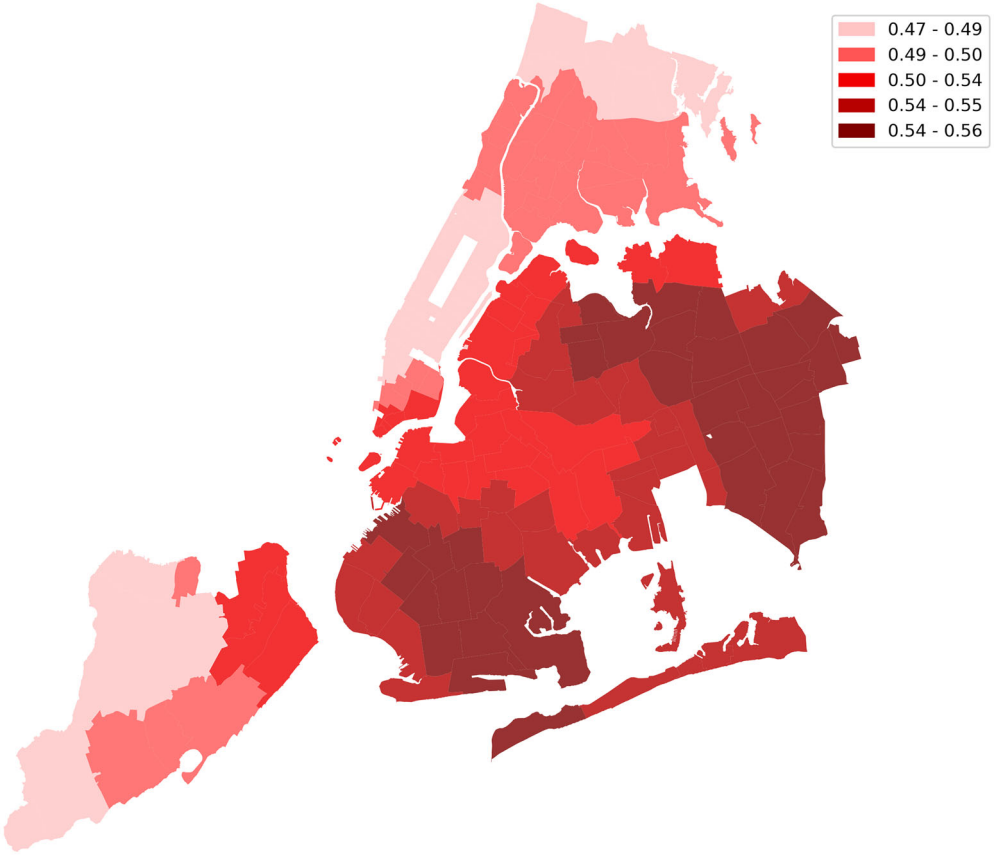

Figure 13. Local intercept estimated using MGWPR.

is positive and significant in areas mainly in the southern part of the city, especially in Staten Island. Public assistance in this context refers to any kind of assistance or benefits received in a household in either cash or in-kind from any governmental entity. This predictor could thus be capturing low-income residents that are otherwise not included in the model.

Finally, the local intercept estimated using MGWPR can be interpreted as the average probability of testing positive for COVID-19 in the areal unit of analysis (ZCTA in this case), holding all other predictors at their mean values. This interpretation follows from the transformation of the estimated β_{0i} values as below:

$$\frac{Positive\ cases\ (y_i)}{Total\ tests\ (offset_i)} = \exp\ (\beta_{0i}) * \exp\left(\sum_{k=1}^{K} \beta_k x_{i,k}\right) \tag{23}$$

These values are mapped in Figure 13 where it can be seen that for most of Manhattan and parts of the Bronx and Staten Island, a little less than half the total tests conducted were expected to return positive for COVID-19, assuming average population characteristics. In some parts of Queens and south Brooklyn, this probability increases to more than half. The variability can be interpreted as the impact of local effects not included in the model. Given the probable local nature of the

mechanisms leading to the spread of COVID-19, the coarse spatial unit of analysis is perhaps not appropriate to capture the actual variability of the COVID-19 spread.

6. Discussion and conclusions

Local regression models enable the estimation of spatially nonstationary processes by allowing the parameter estimates to vary across space. Additionally, local models estimate a bandwidth parameter that represents the spatial scale across which a process exhibits heterogeneity. Generalized extensions of a rudimentary local model, GWR, exist that allow the response variable to assume non-normal distributions such as Poisson and Binomial. However, GWR is restricted in that it allows the estimation of only a single bandwidth to represent the scale of all associations in a model. A multiscale extension of GWR, MGWR was recently developed that removes this restriction and allows the estimation of process specific bandwidths. However, MGWR is restricted to response variables following a normal distribution. The development and implementation of MGWPR as described here is thus a major advance allowing the MGWR framework to handle count data. MGWPR is shown to accurately represent the unique scales at which surface heterogeneity occurs in relationships using simulated and empirical data. The estimated unique bandwidths from MGWPR provide more information and valuable insights on the nature of the operational processes being estimated when compared to GWPR. As such, it expands the tool-base within the non-Bayesian regression modeling field by allowing the modeling of a spatial phenomenon following a Poisson distribution.

The LSA, employed in the calibration of MGWPR, is a natural extension of GAMs. Where GAMs are used to estimate unique functions between a predictor and response variable in the covariate space, MGW(P)R estimates unique relationships between a predictor and response variable in the geographical space. Through the use of MGWPR, we can now estimate spatial associations varying at unique scales between count data such as disease spread, traffic crash incidents, crime counts, etc. and their associated socio-economic and ecological predictors. This new model provides an alternative methodology to spatial scientists especially within the fields of epidemiology and health studies. The availability of open source and intuitive local models such as MGWPR that enable measurement of the unique scales at which behavioral, ecological and environmental processes affect health outcomes opens up new possibilities to understand the multiscale processes affecting such phenomena. Additionally, the calibration algorithm demonstrated here can be used to extend MGWR to a generalized multiscale geographically weighted framework. The LSA calibration and inference procedure can easily scale to include other distributions of the response variable such as binomial, negative-binomial, gamma distributions, etc., by expanding the link function options. This will further increase the kinds of analysis that could leverage the potential of local modeling tools and help remove further limitations from the framework. Finally, these new forms of the MGWR model, with flexible input data types, will be implemented in the open-source MGWR software[6] and made available for use publicly.

Notes

1. The constant variance assumption for a linear regression model states that the variance of the errors/residuals is assumed to be constant (Poole and O'Farrell 1971).
2. An up-to-date bibliography of all the peer-reviewed journal articles applying the geographically weighted regression framework and its extensions is available here: https://sgsup.asu.edu/sparc/multiscale-gwr
3. We used cores of Intel Xeon Processor E5 v4 Family (E5-2680V4) on the high-performance computing platform at ASU Core research facilities.
4. We used the commonly employed statistical variable selection techniques namely, best subset selection and forward selection (Marhuenda *et al.* 2014), using the AICc as the diagnostic criterion and both resulted in the same subset of variables as depicted in Equation (21).
5. We follow da Silva and Fotheringham (2016)'s effective correction criterion to maintain the expected family-wise error rate and to avoid false positives.
6. MGWR desktop software is available for open download at: https://sgsup.asu.edu/sparc/multiscale-gwr; the open-source Python implementation of MGWR is embedded within PySAL: https://github.com/pysal/mgwr

Author contributions

Mehak Sachdeva: project administration, conceptualization, software development, graphics production, analysis, writing original draft and editing subsequent drafts. A. Stewart Fotheringham: conceptualization, writing original draft and editing subsequent drafts. Ziqi Li: conceptualization, assistance with writing original draft, software development. Hanchen Yu: analytical development, editing original draft, software development.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Science Foundation (#2117455) awarded to Prof. A. Stewart Fotheringham.

Notes on contributors

Mehak Sachdeva is a Faculty Fellow at the Center for Urban Science and Progress within the Tandon School of Engineering at New York University. E-mail: mehaksachdeva@nyu.edu. Her research interests include developing and testing spatial analytical methods to model and understand urban processes and phenomena.

A. Stewart Fotheringham is Regents' Professor of Computational Spatial Science and Director of the Spatial Analysis Research Center in the School of Geographical Sciences and Urban Planning at Arizona State University. E-mail: stewart.fotheringham@asu.edu. His research interests include local spatial models, spatial processes, spatial analytics, and spatial interaction modeling.

Ziqi Li is an Assistant Professor of Quantitative Geography in the Department of Geography at Florida State University, Tallahassee, FL 32306. E-mail: Ziqi.Li@fsu.edu. His research interests include spatial statistical modeling, explainable geospatial artificial intelligence, and their applications in interdisciplinary fields.

Hanchen Yu is a visiting assistance professor in Urban Governance and Design Thrust, Society Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. His research interests include spatial analysis, geographic information science, and spatial interaction modeling.

Data and codes availability statement

The data and code used in the manuscript are openly available on Figshare: https://doi.org/10.6084/m9.figshare.21743021.v1. A local version of the MGWR repository from https://github.com/pysal/mgwr was used as a base code for the experiments. The simulation experiment code files are available in file 'Simulation_experiment_version-1_IJGIS.ipynb'. The code runs the experiment once – to obtain the results reported in the paper, the code was run 1000 times. The replication of the NYC Covid data study from DiMaggio *et al.* (2020) uses the data file named 'nyc_all_data.csv'. The data are compiled from the sources mentioned by DiMaggio *et al.* (2020) and are enumerated in Table 1 of the manuscript. The code for the NYC replication study is available in the file 'NYC_replication_code_submission-IJGIS.ipynb'.

References

Agnew, J., 1996. Mapping politics: how context counts in electoral geography. *Political Geography*, 15 (2), 129–146.

Agnew, J.A., 2014. *Place and politics: the geographical mediation of state and society*. London: Routledge.

Agresti, A. 2002. Categorical data analysis. 2nd ed. New York: John Wiley & Sons, Inc., 320–332. http://dx.doi.org/10.1002/0471249688

Buja, A., Hastie, T., and Tibshirani, R., 1989. Linear smoothers and additive models. *The Annals of Statistics*, 17 (2), 453–510.

Cardozo, O.D., García-Palomares, J.C., and Gutiérrez, J., 2012. Application of geographically weighted regression to the direct forecasting of transit ridership at station-level. *Applied Geography*, 34, 548–558.

Cupido, K., Fotheringham, A.S., and Jevtic, P., 2021. Local modelling of U.S. mortality rates: a multiscale geographically weighted regression approach. *Population, Space and Place*, 27 (1), e2379.

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

DiMaggio, C., *et al.*, 2020. Black/African American Communities are at highest risk of COVID-19: spatial modeling of New York City ZIP Code-level testing results. *Annals of Epidemiology*, 51, 7–13.

Everitt, B.S., 2005. Generalized additive model. *In*: *Encyclopedia of statistics in behavioral science*. John Wiley & Sons, Ltd.

Fotheringham, A.S., Brunsdon, C., and Charlton, M., 2002. *Geographically weighted regression: the analysis of spatially varying relationships*. Hoboken: Wiley.

Fotheringham, A.S. and Sachdeva, M., 2021. Modelling spatial processes in quantitative human geography. *Annals of GIS*, 28 (1), 5–14.

Fotheringham, A.S. and Sachdeva, M., 2022. On the importance of thinking locally for statistics and society. *Spatial Statistics*, 50, 100601.

Fotheringham, A.S., 2020. Local modelling: one size does not fit all. *Journal of Spatial Information Science*, 21 (21), 83–87.

Fotheringham, A.S., Li, Z., and Wolf, L.J., 2021. Scale, context, and heterogeneity: a spatial analytical perspective on the 2016 U.S. presidential election. *Annals of the American Association of Geographers*, 111 (6), 1–20.

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

Fotheringham, A.S., Yue, H., and 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.

Gao, S., *et al.*, 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special*, 12 (1), 16–26.

Golledge, R.G., 1997. *Spatial behavior: a geographic perspective*. New York: Guilford Press.

Goodchild, M.F., 2011. Formalizing place in geographic information systems. *In*: L.M. Burton, S. A. Matthews, *et al.*, eds. *Communities, neighborhoods, and health: expanding the boundaries of place*. New York: Springer, 21–33.

Hastie, T. and Tibshirani, R., 1986. Generalized additive models. *Statistical Science*, 1 (3), 297–310.

Hauser, R.M., 1970. Context and consex: a cautionary tale. *American Journal of Sociology*, 75 (4, Part 2), 645–664.

Hughes, M.M., *et al.*, 2021. County-level COVID-19 vaccination coverage and social vulnerability—United States, December 14, 2020–March 1, 2021. *MMWR. Morbidity and Mortality Weekly Report*, 70 (12), 431–436.

Kedron, P., *et al.*, 2022. A replication of DiMaggio et al. (2020) in Phoenix, AZ. *Annals of Epidemiology*, 74, 8–14.

Khazanchi, R., *et al.*, 2020. County-level association of social vulnerability with COVID-19 cases and deaths in the USA. *Journal of General Internal Medicine*, 35 (9), 2784–2787.

King, G., 1996. Why context should not count. *Political Geography*, 15 (2), 159–164.

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

Li, Z., *et al.*, 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., *et al.*, 2020. Measuring bandwidth uncertainty in multiscale geographically weighted regression using Akaike weights. *Annals of the American Association of Geographers*, 110 (5), 1500–1520.

Liu, C., Liu, Z., and Guan, C., 2021. The impacts of the built environment on the incidence rate of COVID-19: a case study of King County, Washington. *Sustainable Cities and Society*, 74, 103144.

Malczewski, J. and Poetz, A., 2005. Residential burglaries and neighborhood socioeconomic context in London, Ontario: global and local regression analysis. *The Professional Geographer*, 57 (4), 516–529.

Marhuenda, Y., Morales, D., and Pardo, M.C., 2014. Information criteria for Fay–Herriot model selection. *Computational Statistics & Data Analysis*, 70, 268–280.

Maroko, A.R., *et al.*, 2009. The complexities of measuring access to parks and physical activity sites in New York City: a quantitative and qualitative approach. *International Journal of Health Geographics*, 8 (1), 34.

McCullagh, P. and Nelder, J.A., 2019. *Generalized linear models*. 2nd ed. New York: Routledge.

Nakaya, T., *et al.*, 2005. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine*, 24 (17), 2695–2717.

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

Poole, M.A. and O'Farrell, P.N., 1971. The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 52 (52), 145–158.

Relph, E.C., 1976. *Place and placelessness*. London: Pion.

Sachdeva, M., and Fotheringham, A.S., 2023. A geographical perspective on Simpson's paradox. *Journal of Spatial Information Science*, (26), 1–25.

Sachdeva, M., Fotheringham, S., and Li, Z., 2022. Do places have value? Quantifying the intrinsic value of housing neighborhoods using MGWR. *Journal of Housing Research*, 31 (1), 24–52.

Wang, S., *et al.*, 2018. Spatial variations of PM2.5 in Chinese cities for the joint impacts of human activities and natural conditions: a global and local regression perspective. *Journal of Cleaner Production*, 203, 143–152.

Yu, H., *et al.*, 2020. Inference in multiscale geographically weighted regression. *Geographical Analysis*, 52 (1), 87–106.

Zhang, L., *et al.*, 2004. Modeling spatial variation in tree diameter–height relationships. *Forest Ecology and Management*, 189 (1–3), 317–329.

Zhu, C., *et al.*, 2020. Impacts of urbanization and landscape pattern on habitat quality using OLS and GWR models in Hangzhou, China. *Ecological Indicators*, 117, 106654.