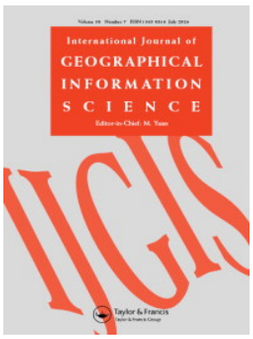

International Journal of Geographical Information Science

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

SGWR: similarity and geographically weighted regression

M. Naser Lessani & Zhenlong Li

To cite this article: M. Naser Lessani & Zhenlong Li (2024) SGWR: similarity and geographically weighted regression, International Journal of Geographical Information Science, 38:7, 1232-1255, DOI: 10.1080/13658816.2024.2342319

To link to this article: https://doi.org/10.1080/13658816.2024.2342319

Published online: 17 Apr 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2024, VOL. 38, NO. 7, 1232–1255
https://doi.org/10.1080/13658816.2024.2342319

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

SGWR: similarity and geographically weighted regression

M. Naser Lessani and Zhenlong Li

Geoinformation and Big Data Research Laboratory, Department of Geography, The Pennsylvania State University, University Park, PA, USA

ABSTRACT

Geographically weighted regression (GWR) offers a local approach to modeling spatial data, considering geographical location and spatial relationships between observations. A salient feature of GWR is the emphasis on geographical proximity, in accordance with Tobler's First Law of Geography, which assumes that closer entities have a greater influence on the target location. Traditional GWR models have been augmented to consider various forms of physical distances aimed at enhancing model performance, and they often disregarded the potential influence of other data attributes, a shortcoming that extends to most GWR extensions. In this study, we introduce a novel weight matrix construction, which integrates data attribute similarity alongside the conventional geographically weighted matrix. The two weights are integrated in a manner that results in improved model performance. The proposed model, called Similarity and Geographically Weighted Regression or SGWR, was applied to five distinct datasets: housing prices, crime rates, and three health outcomes including mental health, depression, and HIV. Results show that SGWR significantly improved model performance based on several statistical measures, outperforming the global regression model and the traditional GWR.

ARTICLE HISTORY
Received 5 July 2023
Accepted 9 April 2024

KEYWORDS
Attribute similarity; similarity weight matrix; GWR; regression; spatial relationship

1. Introduction

The geographically weighted regression (GWR) is a local regression method that enables the modeling of spatially varying relationships (Fotheringham *et al*. 1997, Brunsdon *et al*. 2010). It does this by allowing regression coefficients to vary over space rather than assuming a constant global effect as in the traditional regression model (Stapleton 2009). This local approach is particularly useful for analyzing complex spatial patterns that may not be easily detected with global models (Fotheringham *et al*. 2003). The power of GWR lies in its ability to capture such local patterns and improve predictive performance by accounting for spatial heterogeneity in data (Brunsdon *et al*. 2010). GWR is widely used across a variety of research domains due

CONTACT Zhenlong Li zhenlong@psu.edu

to its ability. In environmental science, for instance, GWR has been utilized to investigate the effects of urban expansion and alterations in landscape structure on the quality of wildlife habitats and land urbanization (Zhu *et al*. 2020, Zhu *et al*. 2022). In urban planning, GWR has increasingly been used to identify and interpret the spatial heterogeneity of relationships between various urban variables. This method allows urban planners to gain more granular insights into the spatial variability of relationships, which can inform more comprehensive and effective urban policy and intervention strategies (Hu and Xu 2019, Li *et al*. 2019b, Cui *et al*. 2021). Similarly, it has been extensively used in housing price modeling to take into account the spatial variation in the relationships between house prices and their determinants (Zhang *et al*. 2019, Tomal 2020). The GWR has also been used substantially in public health research to account for spatial variability in health outcomes and risk factors, and it can model how associations between a health outcome and its determinants vary spatially, providing a local rather than a global parameter estimate (Wende 2019, Jiao *et al*. 2021, Shi *et al*. 2023). In essence, accurately capturing the spatial patterns while studying today's phenomenon is critically important.

In scholarly discourse, the shortcomings of GWR are well recognized, and several refinements have been suggested. Multiscale geographically weighted regression (MGWR) took spatial scale into account (Fotheringham *et al*. 2017) to further enhance GWR performance, and geographically neural network weighted regression (GNNWR) was proposed aiming to more accurately estimate spatial non-stationary and improve its prediction capability (Du *et al*. 2020). Then regionally geographically weighted regression (RGWR) was introduced to optimize kernel function (Wang *et al*. 2022). The kernel function is an essential component of GWR because it determines how weights are assigned to the data points based on their spatial proximity to the location for which a prediction is being made. Another critical component of regression is the computation time; to address this challenge, researchers proposed a fast geographically weighted regression (FastGWR) and fast multiscale geographically weighted regression (FastMGWR) (Li *et al*. 2019a, Li and Fotheringham 2020), expanded the volume of data that could be processed in parallel by using Message Passing Interface (MPI) protocols. Following this, a four-dimensional GWR (4DGWR) was proposed aiming to improve the spatial dimensionality by considering spatial, altitudinal, and temporal non-stationary in the dataset (Tasyurek and Celik 2022). However, as a previous study expressed, the aspect of distance measurement, a significant yet independent facet of the GWR model, has been largely overlooked until recently (Lu *et al*. 2014). This issue is rooted in Tobler's first law of geography, which postulates that 'everything is interconnected, but things that are closer together are more strongly linked than things that are further apart' (Tobler 1970). The interpretation of 'near' and 'distant', however, has been a long-standing subject of academic contention.

In the realm of quantitative spatial analysis, the GWR model typically employs linear or Euclidean distance as the primary metric to measure spatial proximity. This spatial relationship forms the backbone of the weight assignment process during regression computations. Essentially, the proximity of a data point to the location of interest within the regression model inversely correlates with the weight it is accorded: the closer the data point is to the regression location, the larger its weight and, thus, the

more significant its influence on the parameter estimates of the local regression model. This approach ensures that estimates are locally tailored and sensitive to spatial variations, highlighting the potential geographical heterogeneity in relationships among variables. Capturing the complexity of geographic space merely through Euclidean distance presents a considerable challenge. Some researchers have endeavored to address this issue by enhancing the conceptual understanding of distance within the GWR model. Acknowledging the inherent non-stationarity of spatiotemporal data both in space and time, Huang *et al*. (2010) incorporated the time dimension into the GWR model. This inclusion led to the development of innovative spatiotemporal models such as geographical and temporal weighted regression (GTWR) (Fotheringham *et al*. 2015). Consequently, the notion of spatial proximity evolved from a single geographical distance to a more complex, multi-faceted temporal-spatial distance (Huang *et al*. 2010). To further refine the GWR model, scholars introduced non-Euclidean distance metrics, such as road network distance and travel time, thereby improving the model's fit (Lu *et al*. 2014). Another study proposed the use of Minkowski distance, arguing it to be the most appropriate distance metric for the GWR model (Lu *et al*. 2015). Despite these substantial efforts to adopt more flexible distance metrics within the GWR framework, the essence of these distance measures largely aligns with the foundational concept of the classical GWR model, which primarily accounts for the physical distance between spatial units. To the best of our knowledge, all preceding research exclusively focused on geographical proximity as the determinant for creating a weighting matrix. While geographical proximity is a critical factor in GWR models, it does not necessarily imply congruity across other pertinent attributes. Indeed, there exist instances where regions in close spatial contiguity may exhibit substantial contrasts (Zhu and Turner 2022). Such observations challenge the fundamental postulate of spatial homogeneity embedded within the traditional GWR models, invoking a need for more comprehensive spatial analysis methods.

As the third principle of geography essentially emphasizes that if two locations share a similar set of geographic conditions, they are likely to exhibit comparable outcomes for a specific geographic variable (Zhu and Turner 2022). It underlines the concept that similarity is not confined to geographic proximity but extends to the arrangement and hierarchy of geographic conditions and other attributes in the data. The concept of attribute similarity has been largely considered in environmental science, particularly in predicting soil properties (Jiaogen *et al*. 2017, Zhu *et al*. 2018). On the other hand, the ongoing process of globalization and the decline in transportation expenses in recent years could imply a diminished influence of geographical distance (Pedersen *et al*. 2008). As such, the phenomenon of distant yet closely linked relationships is becoming more evident, and geographical distance may not accurately represent actual closeness. There have been considerable studies highlighting the various ways in which distant locations can significantly influence one another. For instance, economies around the globe are becoming more integrated due to globalization, leading to economic shocks or policy changes in one country having widespread effects (Rey 2016). With the increase of globalization, supply chains have become more complex and geographically dispersed. An issue in one part of the chain can lead to significant impacts in distant locations (Park *et al*. 2013). Similarly, with the advent of

digital communication technologies, information can be transmitted almost instantaneously across large distances, leading to impacts on public sentiment, financial markets, and political landscapes worldwide (Karim *et al*. 2020). A recent study proposed a new model of geographically weighted regression based on network weight, which is derived according to population mobility data (He *et al*. 2023). From the perspective of public health, diseases can spread quickly across large distances due to modern travel networks (Lessani *et al*. 2023). The COVID-19 pandemic is a recent example of this, where a local outbreak in one city (Wuhan, China) led to a global pandemic, which highlights the remote close association (Nicola *et al*. 2020). Moreover, despite lacking geographical proximity, certain regions in the United States exhibited similar attitudes of hesitancy or opposition towards the COVID-19 vaccine. These trends indicate that shared sociocultural and political attributes can influence behaviors. This highlights how similarity in attributes can significantly shape behavior outcomes and phenomena, regardless of physical distance (Yasmin *et al*. 2021).

Recognizing this, the inclusion of attribute similarity in spatial analysis represents a novel and promising research trajectory within the field of geography. We posit that geographical proximity and attribute similarity are not mutually exclusive but rather complementary in determining the influence of an observation on another, acknowledging the complex interplay between these two aspects. This introduces a new way of conceptualizing and operationalizing 'proximity', and it expands our understanding of the concept from mere geographical closeness to include similarity in other important attributes. This expanded perspective enables a more holistic and thorough analysis of spatial data, particularly in cases where geographic proximity does not align with similarity in other important characteristics. In this regard, this paper introduces a novel approach that combines geographical proximity and attribute similarity in a local regression model, called Similarity and Geographically Weighted Regression (SGWR). This approach seeks to harness the strengths of both concepts, capturing the complex interplay between spatial proximity and attribute similarity for a more robust understanding of spatial patterns.

2. Methodology

2.1. Geographically weighted regression

Global regression models operate on the assumption that the relationships being analyzed through the parameters of the model are spatially invariant. GWR, however, provides a more elaborated approach by relaxing the spatial constancy assumption inherent in traditional regression models. GWR recognizes that the parameters being examined may, in fact, exhibit spatial variation and the relationships between independent and dependent variables can vary across different geographical spaces. The mathematical formula (Equations 1 and 2) is established in detail in Fotheringham *et al*. (1997). It is important to note that the choice of the kernel function and the bandwidth parameter, which determines the extent of the geographical proximity considered 'local' can significantly influence the weight matrix and, consequently, the GWR model results.

$$y_i = \beta_0(u_i, v_i) + \sum_j^q \beta_j(u_i, v_i)X_{ij} + \varepsilon_i, i = 1, 2, 3, \ldots, n \quad (1)$$

$$\beta_j(u_i, v_i) = \left[X^T W(u_i, v_i)X\right]^{-1} X^T W(u_i, v_i)y \quad (2)$$

In the given equations, the terms (u_i, v_i) denote the geographical coordinates corresponding to the observation point (i), while ε_i represents the random error associated with the same observation point. The term $\beta_j(u_i, v_i)$ refers to the *j*-th regression parameter at observation (i), emphasizing that the value of the regression coefficient can fluctuate based on spatial location. $W(u_i, v_i)$ symbolizes the weight matrix specific to the sample (i).

2.2. Similarity and geographically weighted regression

2.2.1. Problem statement

The GWR primarily emphasizes the 'near' concept, rooted in geographical proximity (geographical similarity), while neglecting the 'related' notion based on attribute similarity (data space) (Anselin and Li 2020). As presented in Figure 1, we assume locations $(t\&p)$ are in equitable distance from regression point (i); thus, they receive equal weights in the weight matrix based on GWR. Location (j) will be weighted lowest among these labeled locations since it is situated farther from the regression point. However, in real-world scenarios, it is unlikely that always the closest locations will be most related to the target location (Zhu and Turner 2022). For instance, though locations $(t\&p)$ are in the same geographical proximity to the target location, it is unlikely that their data attributes are also in the same proximity as their geographical distance to the regression point (i). If two locations exhibit dissimilar characteristics, it suggests less interaction between them. When interaction is minimal, their influence on each other is likewise diminished. The GWR approach has often shown limitations in accounting for such cases where spatial proximity does not necessarily correspond to the similarity in various relevant attributes (Griffith 2019). Therefore, an alternative approach that accounts for attribute similarity – not just geographical proximity – enables the model to capture more accurately the spatial relationships and dependencies across space. This study aims to address this gap in GWR by developing and applying a similarity weight matrix that considers attribute similarity alongside geographical proximity. The analysis will be founded based on the hypothesis that regions with

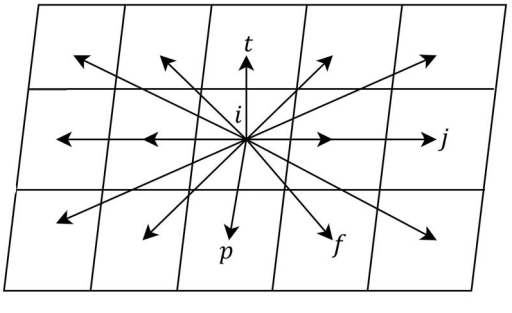

Figure 1. Illustration of weighting matrix in GWR and the proposed model, where weights are assigned based on both geographical proximity and data attribute similarity of locations.

similar attributes are likely to exhibit similar trends or patterns. The model enhances its ability to accurately identify spatial relationships by assigning higher weights to neighbors with greater similarities and lower weights to those with less similarity to the regression point. Assuming location (t) is more similar (related) to regression point (i) in terms of data attribute compared to location (p), in this case, region (t) will be receiving higher weight compared to location (p) though they are located in the same geographical distance from the target location. This suggests that these locations have a stronger interaction, leading to a greater degree of influence on one another.

2.2.2. Similarity weight matrix

The principle underlying the similarity matrix is the notion that regions sharing similar attributes tend to exhibit similar trends or patterns, suggesting a stronger interaction. Several well-established methods exist for conducting similarity analyses across a wide range of fields. In statistics and data science, techniques such as Pearson Correlation Coefficient (Pearson 1985) and Cosine Similarity (Salton 1983) have been extensively used for determining the degree of similarity between two entities. In the field of machine learning, clustering techniques like K-Means (MacQueen 1967), Hierarchical Clustering (Johnson 1967), and a density-based algorithm for clustering (DBSCAN) (Ester *et al*. 1996) rely on similarity measures to group similar data points together. In bioinformatics, the Smith-Waterman algorithm (Smith and Waterman 1981) and Basic Local Alignment Search Tool (Altschul *et al*. 1990) are employed for sequence alignment, revealing similarities between biological sequences. Each of these techniques brings its unique perspective to the notion of similarity, contributing to a richer, more detailed understanding of the patterns in data.

In this study, the pairwise distance method is used to calculate attribute similarity between the regression location and the other observations (Cai *et al*. 2020) based on their independent variables. The equations can be expressed as follows:

$$d_k(i,j) = |x_{k(i)} - x_{k(j)}| \tag{3}$$

$$d(i,j) = \frac{1}{m}\sum_{k=1}^{m} d_k(i,j) \tag{4}$$

$$W_S(u_i, v_i) = \exp(-d(i,j)^2) \tag{5}$$

Equation 3 calculates the pairwise distance for each variable (k) between the regression location (i) and the observation (j). Then the pairwise distance is averaged using Equation 4, where m represents the number of independent variables. Based on pairwise distance, the lower value indicates more similarities and vice versa. However, the weight in GWR is designed such that the value close to 1 hint that the observation and the regression location is much closer while close to zero highlights the farther distance between (i) and (j). Equation 5 is used to align the similarity weight with the GWR weight, where values close to one indicate higher similarity and values close to zero signify dissimilarity between the attributes of two locations. Note that the data is standardized prior to calculating the pairwise distances for similarity assessment, ensuring each variable has a mean of zero and a standard deviation of one. This standardization process neutralizes the impact of varying magnitudes among the

variables, which allows a more accurate and scale-independent similarity measurement between locations.

2.2.3. Incorporation of geographically weighted and similarity weighted matrices

After the construction of geographical and similarity weight matrices, these weights should be combined such that one weight matrix is generated as the final weight matrix. To elucidate this integration, Equations 6–9 detail how a similarity weight matrix can be synergistically combined with a geographically weighted matrix within the framework of GWR.

$$\beta_j(u_i, v_i) = \left[X^T W(u_i, v_i) X\right]^{-1} X^T W(u_i, v_i) y \tag{6}$$

According to Equation 6, which utilizes the original weight matrix derived from geographical weights, we now modify it in alignment with the proposed model by incorporating $W_S(u_i, v_i)$ from Equation 5. In this stage, another parameter (α) is added, which controls the contribution of the geographically weighted matrix ($W_G(u_i, v_i)$) and the similarity weighted matrix ($W_S(u_i, v_i)$) in the final weight matrix (Equation 7). For instance, an optimal α value of 0.5 implies that the geographic and similarity weight matrices have an equal influence on the composition of the final weight matrix. Therefore, the final weight matrix and regression formula are expressed in Equations 8 and 9. Figure 2 visualizes geographical weight and its incorporation with the similarity weight matrix.

$$\gamma = 1 - \alpha, \alpha \ \epsilon(0, 1] \tag{7}$$

$$W_{GS}(u_i, v_i) = \alpha * \ W_G(u_i, v_i) + \gamma * W_S(u_i, v_i) \tag{8}$$

$$\beta_j(u_i, v_i) = \left[X^T W_{GS}(u_i, v_i) X\right]^{-1} X^T W_{GS}(u_i, v_i) y \tag{9}$$

2.2.4. Determining the optimal α value

While combining the similarity weight matrix with the geographically weighted matrix, understanding the proportional influence of each matrix is essential. The parameter α

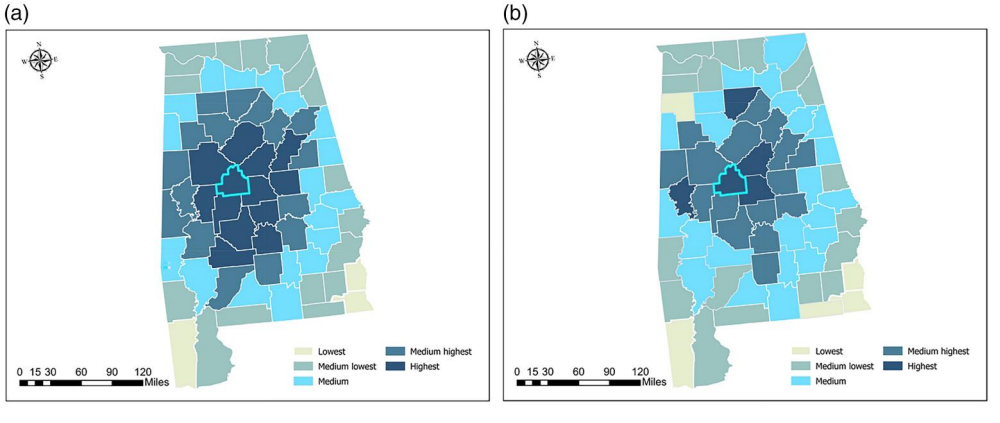

Figure 2. (a) The geographically weighted values associated with the target location. (b) The weight value for the same location derived from a combination of geographically and similarly weight matrices, with an α parameter set to 0.5.

in SGWR functions in a similar way to the bandwidth in GWR, wherein the model seeks a bandwidth that yields the lowest AICc. For SGWR, the α value is optimized when it leads to minimizing the AICc value.

The process of identifying the optimal α value follows the concept of divide and conquer (Bentley 1980), albeit with a minor modification. The iterative process initiates by partitioning the prospective range value (0, 1] into a series of segments at five-unit intervals (e.g., 1, 0.5, 0.1). With every iteration, the model selects a new α value and assesses the corresponding AICc value. This process gradually narrows down the range to identify an α value that not only reduces the AICc but also enhances other evaluation metrics, as detailed in Section 2.3.2. The equations can be expressed as follows:

$$S(\alpha) = \{[1, 0.5], [0.5, 0.1], [0.1, 0.05], \ldots\} \tag{10}$$

$$\alpha_i^{(new)} = \frac{a_i + b_i}{2} \tag{11}$$

$$\alpha_{opt} = argmin_{\alpha \in (0,1]} \ AICc(\alpha) \tag{12}$$

Equation 10 defines an initial segmentation function $S(\alpha)$ that maps α to a series of subranges. For each subrange $[a_i, b_i] \in S(\alpha)$, the divide and conquer steps are applied, by extracting the midpoint of the subrange (Equation 11). Finally, for each derived α from Equation 11, the corresponding AICc value is calculated, and the model aims to select the optimal values across all subranges that yield the lowest AICc value.

Consider, for example, the subrange $[0.5,\ 0.1]$. In this scenario, two outcomes are possible: either the AICc value at 0.1 is greater than at 0.5 or it is less. If the AICc value at 0.1 is lower, indicating a better fit, the process progresses to the subsequent subrange $[0.1,\ 0.05]$ without evaluating other values within $[0.5,\ 0.1]$. Conversely, if the AICc at 0.1 is higher than at 0.5, then the process seeks a better fit by evaluating the midpoint, in this case, 0.3. Again, the model encounters two possibilities: if the AICc at 0.3 is lower than at 0.1, the subrange narrows to [0.3, 0.1], and the midpoint of this new subrange is selected for further evaluation. If not, the midpoint of the range [0.5, 0.3] will be considered as a new subrange. This divide-and-conquer approach continues iteratively, refining the range until identifying an optimal α that minimizes the AICc value.

The proposed model consists of three stages (Figure 3): (1) bandwidth selection and calculating weight matrices; (2) alpha (α) optimization, which aims to minimize

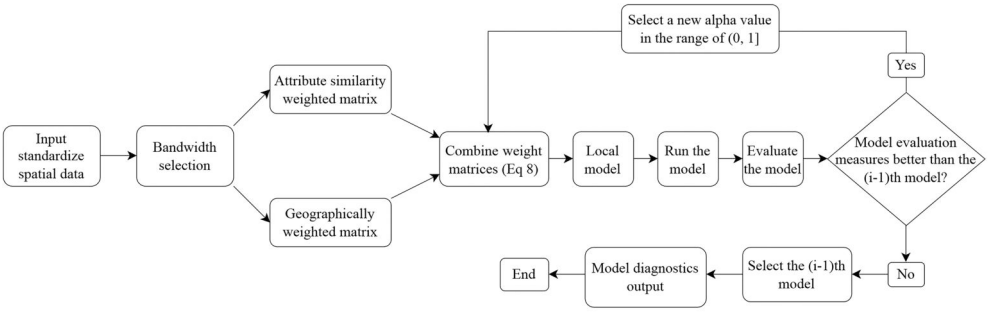

Figure 3. Flowchart of the SGWR model. $(i-1)$ corresponds to varying values of α within the range of (0, 1].

the AICc value while improving other evaluation metrics; and (3) model diagnostics output. The careful selection of α is critical in ensuring model robustness and accuracy similar to bandwidth selection in GWR.

2.3. Model evaluation

2.3.1. Experimental datasets

Five distinct datasets are used to evaluate the proposed model: housing prices, crime rates, and three health outcomes – focusing on mental health, depression prevalence, and HIV. The housing dataset pertains to King County, Washington, US, and it consists of six explanatory variables that are used in the regression analysis, with the housing price acting as the response variable. The crime dataset is comprised of 13 predictor variables that reflect demographic characteristics, socioeconomic status, health conditions, and environmental factors. Lastly, the health outcomes dataset encompasses 12 predictor variables from a broad range of categories, including demographic information, socioeconomic indicators, health-related metrics, beliefs, and environmental factors, while HIV includes 7 predictors. The Variance Inflation Factor (VIF) values for all predictor variables in the model were ensured to be less than 10 to mitigate the effects of multicollinearity in the analysis. It is worth noting that the datasets used in this study exhibit varying levels of spatial resolutions and sample sizes, as shown in Table 1.

2.3.2. Evaluation metrics

The effectiveness of SGWR is assessed using key evaluation metrics including the adjusted coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Akaike Information Criterion Corrected (AICc), and Residual Sum of Squares (RSS). These metrics are widely recognized for gauging model performance, as highlighted in the literature (Du *et al.* 2020, Fotheringham *et al.* 2017, He *et al.* 2023, Brewer *et al.* 2016). The formulas for these evaluation measures are detailed below:

$$Adj_{R^2} = 1 - \left(\frac{(1-R^2)(n-1)}{n-p-1}\right) \tag{13}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{14}$$

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \tag{15}$$

Table 1. Experimental datasets.

Data	Observations	Predictors	Dependent Variable	Geographical Unit
Housing	21,613	6	Price	Neighborhood
Crime	2,841	13	Crime rate	County
Mental health	68,356	12	Mental health prevalence	Census tract (Contagious US)
Depression	1,072	12	Depression prevalence	Census tract (South Carolina)
HIV	2,526	7	Rate per 10^5	County

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{16}$$

$$AIC_C = n\log_e(\hat{\sigma}^2) + n\log_e(2\pi) + n\left(\frac{n + tr(S)}{n - 2 - tr(S)}\right) \tag{17}$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{18}$$

where R^2 is the coefficient of determination, n is the number of observations, p presents the number of explanatory variables, $\hat{y}_i$ represents the predicted value of the dependent variable for the ith observation, $\overline{y}$ represents the mean value of the dependent variables across all observations, and $tr(S)$ refers to the trace of hat matrix S.

Adjusted R^2 quantifies the alignment between observed and predicted values, indicating how well the model fits the data. On the other hand, RMSE, MAE, and MAPE gauge the precision of the predictions, where a reduced value denotes a superior model performance in terms of accuracy. The AICc is a measure that assesses the goodness of fit of a statistical model, while also accounting for the model's complexity (Brewer *et al*. 2016).

3. Results

3.1. The role of similarity weight matrix

Tables 2–6 delineate the impact of varying α values on the performance of the model across five different datasets – housing, crime, mental health, depression, and HIV dataset, respectively. An α value lower than 1 signifies an improvement, indicating the proportion that the geographically weighted matrix contributes to the generation of the final weight matrix; however, it varies across various datasets. As α decreases at a specific level, the model's goodness of fit, as indicated by adjusted R^2 values, improves significantly. This is coupled with a marked reduction in the AICs and RSS, further affirming enhanced model performance. While the adjusted R^2 continues to improve, AIC values start to rise again at a certain α value, for example, $\alpha < 0.1$ represents the turning point for the housing dataset (Table 2). In the current study, the α value that results in the lowest AICc in the model is deemed the optimal value for α as it implies that the model has a better balance of goodness of fit and simplicity. In other words, the model fits the data well without being overly complicated, which can lead to overfitting (Akaike 1974). These observations underscore the significance of the α parameter in the model.

Table 2. Various values of α for the housing dataset.

α	Adjusted R^2	AICs	RSS
1	0.841	22,465.894	3,302.826
0.5	0.855	21,108.933	2,899.692
0.1	0.903	19,402.566	1,555.813
0.05	0.926	21,921.454	973.565
0.03	0.941	28,209.858	628.115

Table 3. Various values of α for the crime dataset.

α	Adjusted R^2	AICs	RSS
1	0.580	5,691.625	1,154.919
0.5	0.661	5,138.987	914.047
0.1	0.827	3,552.855	427.288
0.05	0.888	2,615.204	254.388
0.02	0.948	1,267.923	98.328
0.001	0.997	23,298.167	1.007

Table 4. Various values of α for the health outcome dataset related to mental health prevalence.

α	Adjusted R^2	AICs	RSS
1	0.853	64,920.647	9,647.566
0.5	0.862	63,319.380	8,844.214
0.2	0.893	61,595.212	7,241.106
0.1	0.900	62,893.579	5,672.961
0.09	0.906	63,557.646	5,419.847
0.05	0.910	71,506.745	4,014.088

Table 5. Various values of α for health outcome dataset related to depression.

α	Adjusted R^2	AICs	RSS
1	0.570	2,229.396	427.007
0.5	0.607	2,212.537	366.139
0.3	0.622	2,213.339	343.114
0.1	0.728	2,463.986	179.784
0.05	0.789	3,012.919	105.233

Table 6. Various values of α for the HIV dataset.

α	Adjusted R^2	AICs	RSS
1	0.679	4,412.398	777.050
0.5	0.694	4,359.397	720.925
0.3	0.705	4,352.323	635.424
0.1	0.735	4,439.021	545.862
0.05	0.766	4,630.091	426.196

Figure 4 depicts that the optimal α values can be identified as the lowest point in the α-AICc plots using the 'elbow' method. While the optimal α values differ among datasets, a commonality is observed where an α below one consistently yields superior performance. This indicates that the dual consideration of geographic and similarity weight matrices enhances the efficacy of the regression model beyond what is achieved by the GWR alone. Specifically, factoring in the similarity between neighbors and the point of regression more precisely captures spatial variations than relying solely on geographical closeness. For instance, setting α to 0.1 optimizes the AICc value for the housing price dataset. Additionally, the figure demonstrates a general trend where the AICc value tends to decrease as α values diminish. As can be also seen in Figure 4, as the α value passes the optimal point, for the housing and crime datasets, the AICc values exhibit a gradual increase. In contrast, a significant rise in AICc values can be observed as α decreases beyond the optimal value for health outcome datasets.

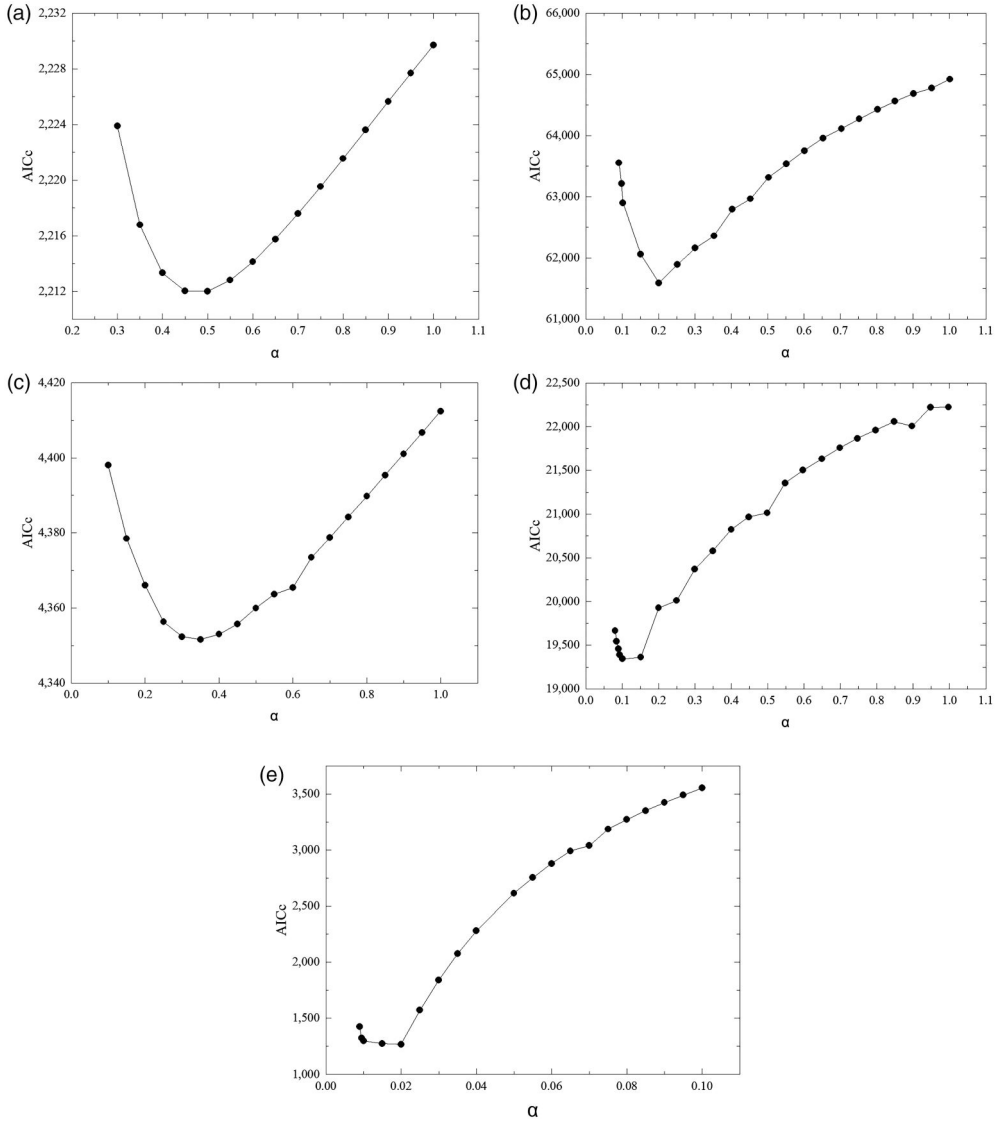

Figure 4. Identifying the optimal α value using the α-AICc plot for the five datasets: (a) depression, (b) mental health, (c) HIV, (d) housing, and (e) crime.

3.2. Performance of the models

The performance of the three models, namely GWR, ordinary least squares (OLS), and SGWR, were evaluated based on several statistical measures. Additionally, we briefly discussed the results of SGWR with MGWR model. The OLS model, serving as our baseline model, provided us with a general idea about the relationships between the predictor variables and the response variable. The models are evaluated and compared using several key metrics, including the adjusted R^2 the AIC, and the RSS. In addition, MAE, RMSE, and MAPE metrics are also considered at this stage of performance analyses; these metrics are measures of error, so smaller values indicate that the predictions from the model are closer to the actual values.

Table 7. Results of different models based on the housing dataset.

Model	Adjusted R^2	MAE	RMSE	AICs	RSS	MAPE (%)
OLS	0.477	0.466	0.722	47,324.407	11,295.417	786.36
GWR	0.841	0.219	0.390	22,465.894	3,302.826	260.00
SGWR	0.903	0.138	0.225	19,402.566	1,555.813	203.09

Table 8. Results of different models based on crime dataset at the county level.

Model	Adjusted R^2	MAE	RMSE	AICs	RSS	MAPE (%)
OLS	0.302	0.228	0.833	7,053.911	1,972.537	289.40
GWR	0.580	0.166	0.637	5,691.625	1,154.919	150.33
SGWR	0.948	0.065	0.186	1,267.923	98.328	74.45

Table 9. Results of different models based on health outcome datasets related to mental health prevalence at census tract level in contagious US states.

Model	Adjusted R^2	MAE	RMSE	AICs	RSS	MAPE (%)
OLS	0.611	0.472	0.623	129,352.481	26,543.317	190.87
GWR	0.853	0.273	0.375	64,920.647	9,647.566	127.71
SGWR	0.893	0.238	0.315	61,595.212	7,241.106	113.53

Tables 7 and 8 respectively present the performance characteristics of various models applied to the housing and crime datasets, with α set to 0.1 and 0.02 respectively for the proposed model, as these optimal values were obtained based on Section 2.2.4. Regarding the housing dataset, as depicted in Table 7, the proposed model was superior, evidenced by an adjusted R^2 of 0.903, an AIC score of 19,402.566, and the lowest RSS at 1,555.813. The outcomes derived from the crime dataset, as indicated in Table 8, followed a similar trajectory. In this scenario, the SGWR model once again outperformed its counterparts, yielding an adjusted R^2 of 0.948 and the most favorable AICs and RSS scores. In terms of MAE and RMSE, the SGWR model shows the lowest error rates (MAE of 0.138, RMSE of 0.225), outperforming both the OLS and GWR models as shown in Table 7. Similarly, as presented in Table 8, the SGWR model exhibits the lowest errors, with MAE of 0.065 and RMSE of 0.186, demonstrating improved performance compared to the OLS and GWR models.

Tables 9–11 present the performance attributes of assorted models when applied to health outcomes (mental health, and depression prevalence) and HIV datasets, with α set to 0.2, 0.5, and 0.3 respectively for the SGWR model; similar patterns can be observed. For instance, as presented in Table 9, the SGWR model delivered the highest performance. As indicated by an adjusted R-squared of 0.893, an AICs score of 61,595.212, and the smallest RSS of 7,241.106, the SGWR model provides the most robust fit for the data, able to account for around 89.3% of the variability in mental health outcomes. While GWR model exhibits a lower value of adjusted R-squared and higher AICc and RSS values compared to the proposed model. In terms of MAE and RMSE, analogous to previous datasets, the SGWR model consistently yields the lowest error rates across all datasets (Tables 9–11).

In addition to OLS and GWR models, the results of SGWR were also compared with those from the MGWR model, although the basis of the proposed model is GWR and OLS. For the datasets on housing, depression, and HIV, the adjusted R-squared values are 0.91, 0.77, and 0.73 for the MGWR, respectively. For SGWR, these values are 0.90,

Table 10. Results of different models based on health outcome dataset related to depression prevalence at census tract level in South Carolina US.

Model	Adjusted R^2	MAE	RMSE	AICs	RSS	MAPE (%)
OLS	0.157	0.730	0.919	2,876.561	892.793	183.11
GWR	0.570	0.499	0.631	2,229.396	427.007	174.68
SGWR	0.607	0.464	0.584	2,212.537	366.139	158.40

Table 11. Results of different models based on the HIV dataset at the county level.

Model	Adjusted R^2	MAE	RMSE	AICs	RSS	MAPE (%)
OLS	0.322	0.561	0.822	6,198.549	1,708.310	270.52
GWR	0.679	0.337	0.554	4,412.398	777.050	241.25
SGWR	0.705	0.312	0.517	4,352.323	635.424	231.02

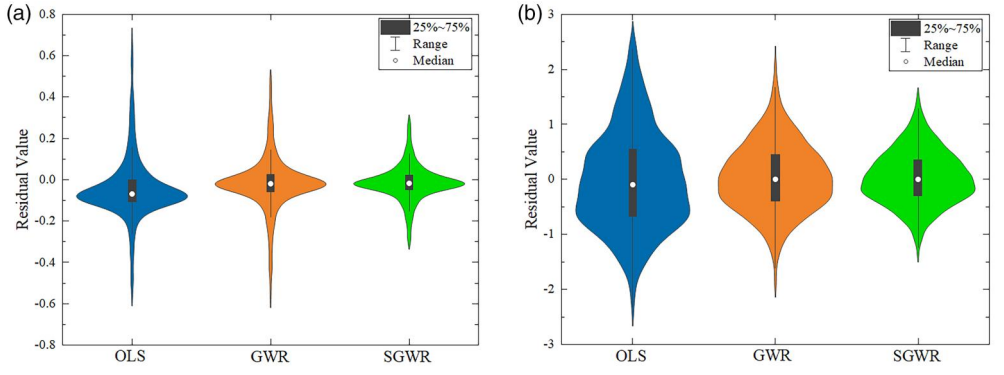

Figure 5. Residual values of different models based on two datasets: (a) the crime rate dataset, and (b) the depression prevalence dataset.

0.61, and 0.71 respectively. For the crime dataset, the SGWR outperforms the MGWR, with adjusted R-squared values of 0.95 and 0.65, respectively. For the mental health dataset, the MGWR model was not executed due to its demanding computational requirements. To illustrate, the housing dataset contains 21,613 observations and 6 predictors, and the execution time was recorded at 23.76 hours. In comparison, the mental health dataset comprises 68,356 observations with 12 predictors, given these parameters, the computational time for MGWR could be considerably lengthy, potentially extending to days. According to other four datasets, it is evident that for the housing, depression, and HIV datasets, the MGWR model marginally outperforms the SGWR in terms of adjusted R-squared values while its computation time is extensively longer than the SGWR, as detailed in Section 3.5. One possible explanation for obtaining a higher R-square could be that in the MGWR model, a unique bandwidth is calculated for each predictor variable, whereas in the SGWR model, a single bandwidth is applied across all variables, like the approach in the GWR model.

3.3. Residual values comparison

As visualized in Figure 5(a), the OLS and GWR models exhibit a tall and narrow distribution, indicating a broad range of residuals. In contrast, the residual distribution is

more tightly centered around zero with shorter tails in the SGWR model. A similar pattern is observed for the depression prevalence dataset, as shown in Figure 5(b). Note, the extreme outliers (>0.75) and (< -0.75) are removed from the results of the models since they led to an extremely skewed plot in Figure 5(a). In essence, both figures reveal that the OLS and GWR model has a wider spread of residuals, indicating greater variance in prediction errors, compared to the SGWR model.

The residual values generated by various models based on the crime dataset and depression prevalence vary. For example, in GWR (Figure 6(b)), the residual value spans from −7.11 to 23.88, while in SGWR, it has a closer range between −1.18 and 13.64 (Figure 6(c)). However, in the OLS model, the residual values vary between −5.96 and 26.18 (Figure 6(a)). Nearly all values are within the range of (−1, 1) with a few exceptions in SGWR. In the OLS (Figure 6(a)) and the GWR results (Figure 6(b)), 29 and 20 locations out of 2,841 have values surpassing 1.99, respectively. This number decreases to only one in the proposed model, as shown in Figure 6(c). Similarly, below −1, the count stands at 43, 34, and 3 for OLS (Figure 6(a)), GWR (Figure 6(b)), and SGWR (Figure 6(c)), respectively. This demonstrates that 98.09% and 99.85% of the data maintain residual values within the −1 to 1 range for GWR and SGWR model, respectively. The results derived from the depression prevalence data exhibit similar patterns. The range of residual values for OLS (Figure 6(d)), GWR (Figure 6(e)), and SWGR (Figure 6(f)) are −2.51–4.60, −2.04–2.44, and 1.79–1.94, respectively. Additionally, a notable proportion of regions display residual values between −1 and 1, with 831 out of 1103 for OLS, 987 for GWR, and 1,054 for SWGR. These observations affirm the robustness of the SGWR model in managing outliers and enhancing model fit.

The spatial residual distribution shows distinct patterns depending on the model used. In the OLS and GWR models, higher residual measurements are observed in densely populated areas like Washington, California, and the East Coast, as illustrated in Figures 6(a,b). Conversely, the Midwest, northern regions, and areas to the west

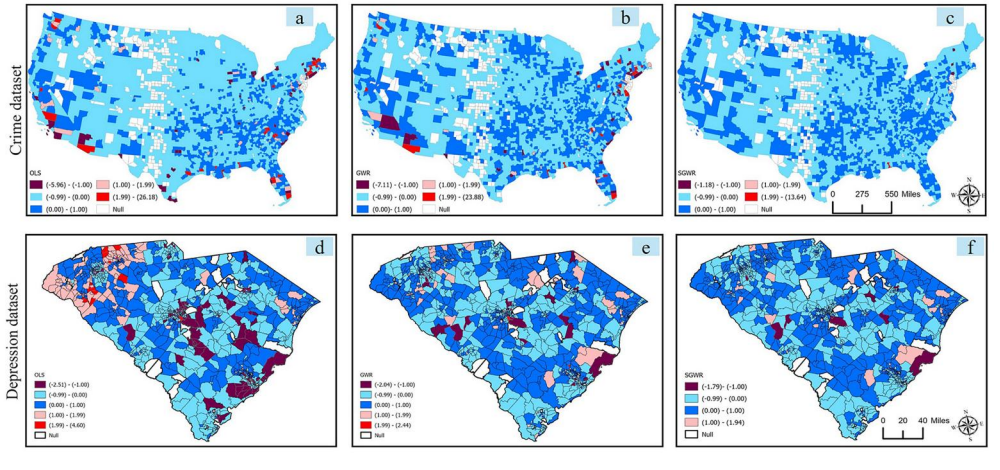

Figure 6. Spatial distribution of residuals from various models using crime and depression prevalence datasets is noted. Due to data limitations, some counties and neighborhoods have null values.

display lower residuals. In contrast, the SGWR model primarily reveals elevated residuals in the Northeast and Florida, with the distribution becoming notably more concentrated and less, as seen in Figure 6(c). The spatial distribution of residuals for the depression prevalence dataset displays consistent patterns across all three models. Specifically, elevated residual values are observed in the Northwest, Central, and Southeast regions, which correspond to areas with higher populations.

3.4. Model interpretation

We particularly discuss the relationship between housing price and its relevant independent variables using two models among five datasets since it has been often used in the literature (Mathur 2013, Gunasilan 2021). The independent variables are bedrooms, bathrooms, sqft_lot, sqft_living15, sqft_lot15, and grade. The models' parameter estimates are visualized by analyzing the relationship individually, as shown in Figure 7. The descriptive statistic is presented in Table 12 based on GWR and SGWR models. As can be observed, for both the GWR and SGWR models, the estimated coefficients for all variables are statistically significant.

Figure 7 depicts spatial variation in the estimated coefficient parameters for housing prices, using both models across six independent variables. There is a noticeable spatial heterogeneity in the coefficients for all the variables across the geographical extent. Based on the GWR model, the bedrooms variable, for instance, exhibits higher coefficients in the northern regions while tapering towards the south. In contrast, Grade, Bathrooms, and sqft_living15 variables display higher values more towards central regions and diminish towards south and east. The variables sqft_lot and sqft_lot15 both display a broad range of values. In the central region, they demonstrate a positive correlation, but this relationship diminishes as one moves further from the center. The SGWR model, on the other hand, also illustrates significant spatial variation, but with some distinctions from GWR results. The bedroom variable, for example, has a more dispersed pattern of high coefficients. Similarly, the sqft_lot15 shows a more pronounced variation towards the eastern regions compared to its GWR counterpart. In both the GWR and SGWR models, positive coefficients for the Bedroom suggest that additional bedrooms drive up housing prices. Each model presents varied trends for the Bathrooms, with SGWR particularly noting the influence of local factors. The sqft_lot indicates that larger lots generally decrease prices, a trend consistently shown in the SGWR model. In both models, the coefficient given to 'Grade' emphasizes the importance of housing grades in determining prices. Furthermore, the sqft_living15 and sqft_lot15 in each model illuminate the intricate connection between living spaces. Note, that the interpretation of the estimated coefficient in SGWR is analogous to that in the GWR model. Taking the bedroom variable as an example, a coefficient of 0.094 suggests that for every additional bedroom in a house, the housing price is anticipated to rise by 0.094 units, while keeping all other variables unchanged.

3.5. Computational time

Computational time is a critical metric when evaluating the efficiency of regression models, particularly in cases involving large datasets or complex calculations. Efficient

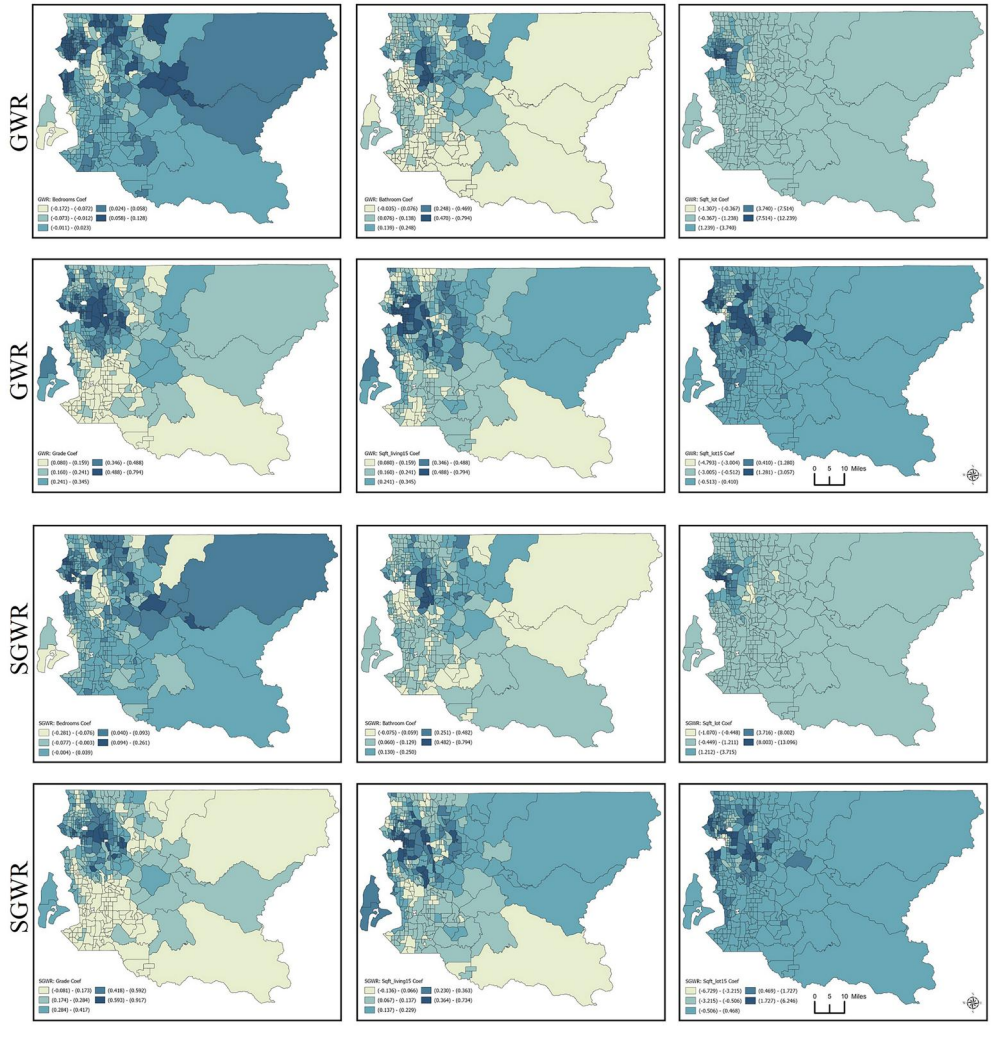

Figure 7. Illustration of estimated parameters for six independent variables based on GWR and SGWR.

Table 12. Descriptive statistics of the coefficient parameters based on GWR and SGWR models for housing dataset for the estimated parameters.

Model	Variable	Mean	STD	Min	Median	Max
GWR	Intercept***	0.260224	0.897400	−0.705972	0.052866	4.174773
	Bedrooms***	0.017701	0.050186	−0.203596	0.019518	0.161760
	Bathrooms***	0.128881	0.112164	−0.053478	0.095359	0.905304
	Sqft_lot***	1.161779	1.988914	−1.408633	0.234691	13.063038
	Grade***	0.265000	0.137281	0.047866	0.237882	0.883812
	Sqft_living15***	0.168281	0.101141	−0.074468	0.144643	0.522961
	Sqft_lot15***	0.329933	0.783678	−5.713785	0.119962	3.523045
SGWR	Intercept***	0.258479	0.900861	−0.767433	−0.053535	5.891888
	Bedrooms***	0.018134	0.057546	−0.573229	0.020478	0.397994
	Bathrooms***	0.128298	0.118016	−0.475722	0.095744	1.499524
	Sqft_lot***	1.162855	2.021162	−3.076638	0.234224	21.470264
	Grade***	0.263886	0.143554	−0.333431	0.234966	1.750847
	Sqft_living15***	0.166977	0.109067	−0.604066	0.143332	0.849110
	Sqft_lot15***	0.326242	0.855641	−11.402725	0.117810	12.651284

Note: ***Indicates highly significant.

Table 13. The computational time of different models.

Model	Depression	Crime	HIV	Housing	Mental health
GWR	8.01s	25.90s	15.43s	19.45m	7.28h
SGWR	8.09s	33.54s	19.61s	21.05m	7.44h
MGWR	3.02m	21.54m	18.76m	23.76h	//

Note: s, m, and h represent seconds, minutes, and hours, respectively. To maintain consistency in reporting computation times, each model was executed five times and the average of these five runs is what is presented in the table.

models that provide accurate results in less time are highly valuable in the data science community, as they allow for quicker insights and decisions, saving both resources and time. In this study, a similarity weight matrix is computed with $O(k\ n^2)$ time complexity for n observations and k predictors. However, as GWR's total time complexity is higher at $O(k^3n^2logn)$ when using the golden search for bandwidth selection (Li *et al*. 2019a), this dominates, and the overall time complexity remains $O(k^3n^2logn)$ and its influence is insignificant. As presented in Table 13, GWR demonstrated the shortest computational time, with SGWR trailing closely. It is noted that α optimization is conducted in parallel to expedite the process.

Additionally, the computational time for the MGWR model is shown for all datasets except the mental health outcome dataset, which comprises 68,356 observations and requires an excessive amount of time. The results reveal that the computational time for MGWR is significantly longer than that of SGWR (Table 13), although the differences in their R^2 are very marginal, as discussed in Section 3.2. While a parallel version of MGWR exists, which employs parallel processing techniques, it does not fundamentally alter the original computational complexity of the model; thus, this does not mean less computational resources are required. Instead, as the size of the data partition assigned to each processor increases, the computational time also increases, but this increase is still governed by the polynomial nature of the algorithm's complexity (Li and Fotheringham 2020). According to the experimental datasets, SGWR is able to achieve results approximately close to MGWR with significantly less computation time, and even outperforms in crime dataset.

4. Discussion

4.1. Beyond geographical distance

Waldo Tobler's First Law of Geography underscores the significance of spatial proximity in shaping relationships and interactions, thereby serving as a cornerstone in spatial analysis and geographical studies. Consequently, prior research primarily employed geographical distance as the foundation for the weighting matrix in traditional GWR and its extended models. Nevertheless, it has become increasingly apparent that geographical distance does not always accurately represent the patterns present in real-world datasets. Indeed, studies have unveiled that factors other than physical proximity, such as social media-based networks or place connectivity beyond geographical neighborhoods (Jing *et al*. 2022, Li *et al*. 2021, Jing *et al*. 2024), can influence human behaviors, which leads to the generation of data for each geographical region (Yasmin *et al*. 2021, Karim *et al*. 2020). In response to these limitations, SGWR

introduces a weight matrix based on the similarity of data attributes into GWR, enabling more accurate identification of patterns within the data.

In the formulation of the similarity matrix in SGWR, additional parameters are introduced to the traditional GWR framework. These include the parameter α and a chosen method for conducting similarity analysis, which could involve techniques such as cosine similarity, k-means clustering, and pairwise distance approach, among others, depending on the nature of the data. The parameter α modulates the balance between the geographically weighted matrix and the attribute-based similarity weighted matrix in the final weight matrix. Selecting an optimal value for this parameter is critical, as it governs the trade-off between geographical and attribute-based similarity, with the appropriate balance varying based on the specifics of the input data and the research question being addressed.

Our findings carry significant implications, with broad impacts on spatial analysis. This is particularly notable given that our experiments span a diverse range of datasets, encompassing housing, health outcomes, and crime dataset, and extend across various geographical units, including neighborhoods, census tracts, and county levels. SGWR demonstrates superior performance compared to traditional models like OLS and GWR, which establishes a promising new approach to understanding and predicting complex spatial phenomena. This advancement in spatial regression models could potentially enhance how we analyze, interpret, and predict patterns within geographic datasets, beyond geographical proximity.

This study provides various directions for future research. The proposed (SGWR) model showed promising results across tested datasets. However, its application across a broader range of domains and datasets warrants investigation. Future studies examining the impact of a differential weighting of explanatory variables on model performance and their individual influence on the model's output could also be explored. In addition, this study individually calculates similarities between regression and observation points' variable values, but averages these for the final similarity matrix, overlooking unique interactions. Future research could use a more flexible approach in creating similarity weight matrices to account for these interactions by assigning varied weights to different variables.

4.2. Attribute similarity

The two key concepts of Tobler's first law are 'near' and 'related'. A recent study discussed in detail regarding the first law of geography (Anselin and Li 2020), indicates that geographical similarity is related to the concept of 'near' while attribute similarity corresponds to 'related'. According to this law, neighboring locations in close physical proximity have a greater influence on each other than those further apart, as they often share similar characteristics. However, in the GWR model, only the geographical similarity (near) is considered. We posit that if two neighboring locations are geographically proximate, the data generated from these sites should similarly be closely aligned.

As stated, the foundational principle of the GWR model rests on the idea that things closer in proximity are more closely related than those further apart. Yet, in

GWR and its extended models, geographical closeness is often the sole criterion to determine the influence of surrounding observations on a regression point, overlooking the actual data originating from these locations. In today's world, vast amounts of data are being generated across various domains, such as health, social behavior, and the environment. These data can offer a genuine reflection of reality in terms of connectedness and the influence of things on each other. For instance, despite the physical proximity of two neighboring states, one being Republican and the other Democrat, they often adopt different policies and may exert limited influence upon each other (Motta 2021, Lang *et al*. 2021), such as the stances on mask-wearing or vaccination. In the GWR approach, higher weights may be assigned to these neighboring states due to their geographical closeness, even if their actual influence on each other is minimal, which can lead to misinterpretation and unable to capture the true spatial variation. Conversely, in SGWR, not only physical distance is considered, but also the nature of interactions between the states based on data attributes, offering a more accurate reflection of their real-world relationship. Thus, shaping current spatial patterns can be more precise when data attribute is used, rather than solely depending on geographical proximity. Therefore, this study extends beyond mere geographical distance by also factoring in the attributes of the data to evaluate how locations influence each other.

Concerns regarding overfitting the model may arise if we give more weight to locations with similar patterns. This reflects the core principle of the first law of geography and the GWR model, which suggests that locations with a higher proportion of similarities have a stronger influence on each other than those that are dissimilar or further apart; and higher weights need to be assigned to these locations. However, when developing the model, a balance must be struck between optimizing model's performance and preventing overfitting regardless of considering geographical proximity or data attribute similarity. To determine the right balance, various evaluation metrics are available, as discussed in Section 2.3.2 and corroborated by previous studies (Du *et al*. 2020, Fotheringham *et al*. 2017, He *et al*. 2023, Brewer *et al*. 2016).

4.3. Evaluating SGWR model against the enhanced GWR models

Our analysis reveals that while there are several extensions to the GWR model, such as incorporating non-Euclidean distances (Lu *et al*. 2014) and using geographically neural network weighted regression (Du *et al*. 2020), these are often tailored to specific scenarios. For instance, the integration of mobility data for location weighting (He *et al*. 2023) presents a challenge due to the complexity of procuring such data. Additionally, approaches like the neural network weighted regression lean more towards predictive rather than explanatory analysis. Different from previous approaches, the SGWR offers straightforward implementation, requiring only location coordinates, dependent, and independent variables, akin to the GWR model's requirements.

The performance of SGWR is compared with MGWR in addition to OLS and GWR. The results showed that SGWR surpasses MGWR in performance for specific dataset, such as crime. However, for datasets concerning depression, housing, and HIV, MGWR exhibits a marginal lead as discussed in Section 3.2. A critical aspect, however, to

highlight is the computational efficiency of the proposed model. While MGWR shows competitive performance for some datasets, it demands substantially longer computation times as shown in Section 3.5. The significant reduction in computational time in SGWR, without a notable compromise in performance, underscores the practical advantages of the model, particularly in scenarios where the number of observations is large, or time and computational resources are limiting factors. However, integrating a similarity-weighted matrix into the MGWR model could be a promising avenue for future research. This integration has the potential to further improve the SGWR performance in terms of capturing spatial variation.

5. Conclusion

GWR uses a localized approach to regression, where each observation is assessed in relation to its neighboring data points based on geographical proximity. In the proposed SGWR, we augment this approach by considering the similarity of data attributes in addition to geographical proximity when constructing the weight matrix, to more accurately capture underlying data patterns. The core idea behind the similarity-weighted matrix is that greater similarity between two locations suggests a higher likelihood of interaction. The performance of SGWR was evaluated using five real-world diverse datasets with varying geographical units. The results reveal that SGWR consistently outperforms the global regression model and the traditional GWR based on several statistical measures across all experimental datasets. The improved model performance reinforces our view that geographical proximity and attribute similarity should not be viewed as mutually exclusive factors. Instead, they are complementary in assessing the influence of one observation on another, acknowledging the intricate relationship between these two dimensions.

Despite the promising results, the study acknowledges certain limitations. Primarily, the efficacy of SGWR has been tested on only five distinct datasets. Its applicability across a broader range of domains remains to be explored. In addition, the choice of equal importance scores for all predictor variables when constructing the final weight matrix might also overlook the differential impacts of these variables. Finally, future research could explore the integration of this similarity concept into the MGWR model to potentially further enhance its performance.

Acknowledgements

The authors extend their sincere gratitude to the three anonymous reviewers and the editor for their insightful comments, which significantly improves the manuscript.

Disclosure statement

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

Notes on contributors

M. Naser Lessani is currently a PhD student in the Department of Geography at The Pennsylvania State University. His primary research interests are geospatial big data analytics,

human mobility, parallel spatial computing, Machine Learning, and GIS. He contributed to conceptualization of the research idea, data analysis, code development, and writing.

Zhenlong Li is an Associate Professor in the Department of Geography and Director of the Geoinformation and Big Data Research Lab at The Pennsylvania State University. His primary research field is GIScience with a focus on geospatial big data analytics, spatial computing, and geospatial AI with applications to disaster management, human mobility, and public health. He contributed to conceptualization of the research idea, data analysis, and writing.

Data and codes availability statement

The datasets and codes used for this study are publicly available at https://github.com/Lessani252/SGWR.

References

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716–723.

Altschul, S.F., *et al.*, 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215 (3), 403–410.

Anselin, L., and Li, X., 2020. Tobler's Law in a Multivariate World. *Geographical Analysis*, 52 (4), 494–510.

Bentley, J.L., 1980. Multidimensional divide-and-conquer. *Communications of the ACM*, 23 (4), 214–229.

Brewer, M.J., *et al.*, 2016. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7 (6), 679–692.

Brunsdon, C., Stewart Fotheringham, A., and Charlton, M.E., 2010. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28 (4), 281–298.

Cai, Z., *et al.*, 2020. A new similarity combining reconstruction coefficient with pairwise distance for agglomerative clustering. *Information Sciences*, 508, 173–182.

Cui, X., *et al.*, 2021. Spatial-temporal responses of ecosystem services to land use transformation driven by rapid urbanization: a case study of Hubei Province, China. *International Journal of Environmental Research and Public Health*, 19 (1), 178.

Du, Z., *et al.*, 2020. Geographically neural network weighted regression for the accurate estimation of spatial non-stationarity. *International Journal of Geographical Information Science*, 34 (7), 1353–1377.

Ester, M., *et al.*, 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 96 (34), 226–231.

Fotheringham, A.S., Crespo, R., and Yao, J., 2015. Geographical and temporal weighted regression (GTWR). *Geographical Analysis*, 47 (4), 431–452.

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., Martin, C., and Christopher, B., 1997. Two techniques for exploring non-stationarity in geographical data. *Geographical Systems*, 4 (1), 58–82.

Fotheringham, A.S., Chris, B., and Martin, C., 2003. *Geographically weighted regression: The analysis of spatially varying relationships*. Hoboken, NJ: John Wiley & Sons.

Griffith, D.A., 2019. Negative Spatial Autocorrelation: One of the Most Neglected Concepts in Spatial Statistics. *Stats*, 2 (3), 388–415.

Gunasilan, U., 2021. Managing and mitigating housing - price predictions using a genetic algorithm: A case of King County in Washington, U.S.A. *International Journal of Management*, 12 (9), 9–18.

He, J., Wei, Y., and Yu, B., 2023. Geographically weighted regression based on a network weight matrix: a case study using urbanization driving force data in China. *International Journal of Geographical Information Science*, 37 (6), 1209–1235.

Hu, X., and Xu, H., 2019. Spatial variability of urban climate in response to quantitative trait of land cover based on GWR model. *Environmental Monitoring and Assessment*, 191 (3), 194.

Huang, B., Wu, B., and Barry, M., 2010. Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. *International Journal of Geographical Information Science*, 24 (3), 383–401.

Jiao, J., Chen, Y., and Azimian, A., 2021. Exploring temporal varying demographic and economic disparities in COVID-19 infections in four U.S. areas: based on OLS, GWR, and random forest models. *Computers, Environment and Urban Systems*, 1 (1), 27.

Jiaogen, Z., Daming, D., and Yuyuan, L. 2017. Local attribute-similarity weighting regression algorithm for interpolating soil property values. *International Journal of Agricultural and Biological Engineering*, 10 (5), 95–103.

Jing, F., *et al.*, 2022. Investigating the relationships between concentrated disadvantage, place connectivity, and COVID-19 fatality in the United States over time. *BMC Public Health*, 22 (1), 2346.

Jing, F., *et al.*, 2024. From neighborhood contexts to human behaviors: cellphone-based place visitation data contribute to estimating neighborhood-level depression prevalence in the United States. *Cities*, 148, 104905.

Johnson, S.C., 1967. Hierarchical clustering schemes. *Psychometrika*, 32 (3), 241–254.

Karim, F., *et al.*, 2020. Social media use and its connection to mental health: a systematic review. *Cureus*, 12 (6), e8627.

Lang, J., Erickson, W.W., and Jing-Schmidt, Z., 2021. #MaskOn! #MaskOff! Digital polarization of mask-wearing in the United States during COVID-19. *PLoS One*, 16 (4), e0250817.

Lessani, M.N., *et al.*, 2023. Human mobility and the infectious disease transmission: a systematic review. *Geo-Spatial Information Science*, 1–28. https://doi.org/10.1080/10095020.2023.2275619

Li, Z., *et al.*, 2019a. 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, S., *et al.*, 2019b. Spatial heterogeneity in the determinants of urban form: an analysis of chinese cities with a GWR approach. *Sustainability*, 11 (2), 479.

Li, Z., *et al.*, 2021. Measuring global multi-scale place connectivity using geotagged social media data. *Scientific Reports*, 11 (1), 14694.

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

Lu, B., *et al.*, 2015. The Minkowski approach for choosing the distance metric in geographically weighted regression. *International Journal of Geographical Information Science*, 30 (2), 351–368.

Lu, B., *et al.*, 2014. Geographically weighted regression with a non-Euclidean distance metric: a case study using hedonic house price data. *International Journal of Geographical Information Science*, 28 (4), 660–681.

MacQueen, J., 1967. Classification and analysis of multivariate observations. *In*: Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Mathur, S., 2013. Impact of urban growth boundary on housing and land prices: evidence from King County, Washington. *Housing Studies*, 29 (1), 128–148.

Motta, M., 2021. Republicans, not democrats, are more likely to endorse anti-vaccine misinformation. *American Politics Research*, 49 (5), 428–438.

Nicola, M., *et al.*, 2020. The socio-economic implications of the coronavirus pandemic (COVID-19): a review. *International Journal of Surgery*, 78, 185–193.

Park, Y.W., Hong, P., and Roh, J.J., 2013. Supply chain lessons from the catastrophic natural disaster in Japan. *Business Horizons*, 56 (1), 75–85.

Pearson, K., 1985. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58 (347–352), 240–242.
Pedersen, P.J., Pytlikova, M., and Smith, N., 2008. Selection and network effects—migration flows into OECD countries 1990–2000. *European Economic Review*, 52 (7), 1160–1186.
Rey, H., 2016. International channels of transmission of monetary policy and the Mundellian trilemma. *IMF Economic Review*, 64 (1), 6–35.
Salton, G., 1983. *Introduction to modern information retrieval*. London: McGraw-Hill.
Shi, B., *et al.*, 2023. Spatial effects of public health laboratory emergency testing institutions under COVID-19 in China. *International Journal for Equity in Health*, 22 (1), 88.
Smith, T.F., and Waterman, M.S., 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147 (1), 195–197.
Stapleton, J.H., 2009. *Linear statistical models*. Hoboken, NJ: John Wiley & Sons, 719.
Tasyurek, M., and Celik, M., 2022. 4D-GWR: geographically, altitudinal, and temporally weighted regression. *Neural Computing and Applications*, 34 (17), 14777–14791.
Tobler, W.R., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234–240.
Tomal, M., 2020. Modelling housing rents using spatial autoregressive geographically weighted regression: a case study in Cracow, Poland. *ISPRS International Journal of Geo-Information*, 9 (6), 346.
Wang, Z., Zhao, Y., and Zhang, F., 2022. Simulating the spatial heterogeneity of housing prices in Wuhan, China, by regionally geographically weighted regression. *ISPRS International Journal of Geo-Information*, 11 (2), 129.
Wende, D., 2019. Spatial risk adjustment between health insurances: using GWR in risk adjustment models to conserve incentives for service optimisation and reduce MAUP. *The European Journal of Health Economics*, 20 (7), 1079–1091.
Yasmin, F., *et al.*, 2021. COVID-19 vaccine hesitancy in the United States: a systematic review. *Frontiers in Public Health*, 9, 770985.
Zhang, S., Wang, L., and Lu, F., 2019. Exploring housing rent by mixed geographically weighted regression: a case study in Nanjing. *ISPRS International Journal of Geo-Information*, 8 (10), 431.
Zhu, A.-X., *et al.*, 2018. Spatial prediction based on Third Law of Geography. *Annals of GIS*, 24 (4), 225–240.
Zhu, A.X., and Turner, M., 2022. How is the Third Law of Geography different? *Annals of GIS*, 28 (1), 57–67.
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.
Zhu, H., *et al.*, 2022. Spatiotemporal dynamics and driving forces of land urbanization in the Yangtze river delta urban agglomeration. *Land*, 11 (8), 1365.