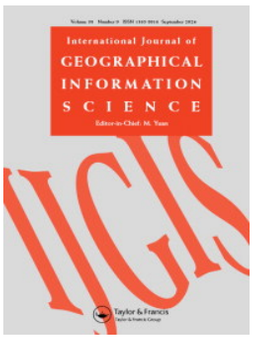

International Journal of Geographical Information Science

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

On ignoring the heterogeneity in spatial autocorrelation: consequences and solutions

Zehua Zhang, Ziqi Li & Yongze Song

To cite this article: Zehua Zhang, Ziqi Li & Yongze Song (20 Aug 2024): On ignoring the heterogeneity in spatial autocorrelation: consequences and solutions, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2024.2391981

To link to this article: https://doi.org/10.1080/13658816.2024.2391981

Published online: 20 Aug 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2024.2391981

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

On ignoring the heterogeneity in spatial autocorrelation: consequences and solutions

Zehua Zhang[a], Ziqi Li[b] and Yongze Song[a]

[a]School of Design and the Built Environment, Curtin University, Bentley, Australia; [b]Department of Geography, Florida State University, Tallahassee, Florida, USA

ABSTRACT

Spatial autoregressive (SAR) models are often used to explicitly account for the spatial dependence underlying geographic phenomena. However, traditional SAR models are specified using a single SAR coefficient, assuming constant spatial dependence over space. This assumption oversimplifies the situation where the true spatial autoregressive process varies in strength; the consequences of ignoring heterogeneous autocorrelation remain to be discussed. This study proposes a heterogeneous spatial autocorrelation model by extending the spatial lag model (SLM). The new model includes change point detection for identifying patterns of spatially varying autocorrelation strengths, a SAR coefficient matrix for representing heterogeneous spatial autocorrelation, and maximum likelihood estimation for determining multiple SAR coefficients. Monte Carlo simulations demonstrate that the proposed method is effective in modeling SAR processes with heterogeneous autocorrelation patterns, while traditional SLM inflates uncertainties in the regression coefficients when a heterogeneous autocorrelation structure is not accounted for. We further applied the new method to an empirical analysis of traffic crashes in the Greater Perth Area, Australia. The heterogeneous spatial autocorrelation model reduces model RMSE by 42% (compared with traditional SLM). Results from both simulation and empirical studies indicate that spatially varying autocorrelation strengths should be considered for SAR processes and relevant applications.

ARTICLE HISTORY
Received 10 April 2024
Accepted 9 August 2024

KEYWORDS
Heterogeneous spatial autocorrelation assumption; spatial lag model; spatial autoregressive coefficient matrix; transport geography

1. Introduction

Spatial dependence refers to the phenomenon in which observations across space are interdependent, and their degree is often measured by spatial autocorrelation (Anselin 1988, Anselin 2010). Spatial autoregressive (SAR) models are often used to explicitly account for spatial dependence, and the model's spatial impacts underlie geographic

CONTACT Yongze Song yongze.song@curtin.edu.au

phenomena (Fischer and Wang 2011). Models within the SAR class include a series of revised model specifications such as the spatial lag model (SLM), spatial error model (SEM) and spatial Durbin model, with additional spatial lagged effects from geographical proximity (Fotheringham 2009, Anselin *et al.* 2010). Spatially lagged effects in SAR models are represented by the matrix product of a SAR coefficient indicating spatial autocorrelation strength and direction, a spatial weights matrix defining spatial connectivity among locations, and the values of spatial variables (Anselin and Griffith 1988, Anselin and Rey 2010). The development of SAR models is pivotal in spatial econometrics (Baltagi *et al.* 2007, Arbia and Baltagi 2009), and their applications extend to various research domains that require the interpretation of geographical information. These fields encompass, but are not limited to, transport planning (Rhee *et al.* 2016), urban analysis (Gao *et al.* 2020), social science (Lambert *et al.* 2010) and environmental modeling (Yin *et al.* 2018).

Traditional SAR models assume that the strength and direction of spatial autocorrelation are homogeneous within a geographical space because of the estimation of a single SAR coefficient value (Harris 2019). However, this assumption ignores the variation in spatial autocorrelation strength, in which case multiple autoregressive coefficients should be estimated to reflect this complexity and avoid potential model misspecification. There have been several developments in this direction, and research efforts have been made to re-estimate the spatial autocorrelation strength with its spatial variation using geographically weighted regression (GWR) (Brunsdon *et al.* 1998, Geniaux and Martinetti 2018).

The nonstationarity of spatial autocorrelation can be modeled by two categories of spatial processes in general, including second-order variance-based models and SAR models. Previous discussions on nonstationary spatial autocorrelation effects for spatial modeling have been extensively explored, but primarily through second-order variance-based methods (Fouedjio 2016). Within these Kriging models, the spatial dependence structure or understanding of spatial autocorrelation, is typically represented in semi-variograms or spatial covariances (Goovaerts 1997). The nonstationarity of spatial autocorrelation, which reflects a feature of second-order variance effects, requires careful consideration to avoid misidentifying first-order trend effects (Schabenberger and Gotway 2005). In large or complex study domains, the spatial dependence structure may remain stationary only within local regions, while exhibiting nonstationarity from a global perspective (Sampson *et al.* 2001). To more accurately model nonstationary spatial dependence, a range of techniques have been proposed, including partitioning (Stein *et al.* 1988), moving windows (Haas 1990), kernel-based models (Fuentes 2001, Harris *et al.* 2010), basis functions (Holland *et al.* 1999) and convolution methods (Higdon 1998, Higdon *et al.* 1999, Paciorek and Schervish 2006), among others (Lindgren *et al.* 2011).

However, the effects of nonstationary spatial dependence in SAR models or spatial-weight-matrix-based indicators remain underexplored. The variation in spatial autocorrelation, as reflected by the Local Indicator of Spatial Association (LISA) or the spatial lag term in SAR models, is heavily influenced by the values of geographic neighbors (Anselin 1988, Anselin 1995). The strength of spatial autocorrelation for each individual

geographic unit is not fully accounted for. Further investigation is needed to comprehensively understand and model these effects within SAR frameworks.

In the SAR process, the spatial nonstationarity of spatial autocorrelation was initially explored and quantified through spatially varying autoregressive models, where SAR coefficients were re-estimated using a geographically weighted approach (Brunsdon *et al.* 1998). Despite the lack of discussion on model assumptions and the necessity of analyzing variations in spatial autocorrelation strength, spatially varying autoregressive models proved the feasibility of geographically weighted approaches for quantifying the variability of SAR coefficients. With further exploration of the SAR processes, a new spatial data generation process involving nonstationary spatial autocorrelation strength, known as Mixed GWR-SAR, was proposed (Geniaux and Martinetti 2018). However, an ultimate conclusion on the consequences of ignoring heterogeneous spatial autocorrelation, which could be solid evidence proving the necessity of considering the nonstationary spatial autocorrelation strength for SAR processes, was not clearly presented in this research. Furthermore, how MGWR-SAR could assist in informative decision advice from the variation in spatial autocorrelation strength, or a discussion of its association with geographic proximity or feature interactions, was not shown. In the latest investigation of SAR models using geographically weighted approaches (Mei and Chen 2022), knowledge gaps on the essentiality of considering heterogeneous spatial autocorrelation and its spatial decision-making potential are still skipped.

Table 1 summarizes the representation of spatial dependence for two categories of spatial processes, together with corresponding techniques to reflect nonstationary spatial dependence. Research progress on nonstationary spatial dependence within second-order-variance-based spatial models is comparatively mature, while SAR models mainly relies on geographically weighted approaches to demonstrate nonstationary

Table 1. A summary of nonstationary spatial dependence for spatial processes.

	Second-order-variance-based (Kriging) spatial process	Spatial autoregressive process
The representation of spatial dependence	Semi-variogram, Spatial covariance (Fouedjio 2016)	Spatial autocorrelation coefficient, Spatial weight matrix (Anselin 1988)
Fundamental models with stationary spatial dependence	Ordinary Kriging, Simple Kriging, etc. (Goovaerts 1997)	Spatial lag model, Spatial error model, etc. (Anselin 1988)
Further techniques to reflect nonstationary spatial dependence	• Category or strata based partitioning model (Stein *et al.* 1988) • Moving window based nonstationary model (Haas 1990) • Smoothing and kernel-based methods (Fuentes 2001, Harris *et al.* 2010) • Basis functions from Gaussian random function (Holland *et al.* 1999) • Spatial deformation models (Sampson and Guttorp 1992) • Convolution (Higdon 1998, Higdon *et al.* 1999, Paciorek and Schervish 2006) • Stochastic partial differential equations (Lindgren *et al.* 2011)	• Geographically weighted approach (Brunsdon *et al.* 1998, Geniaux and Martinetti 2018) • Heterogeneous spatial autocorrelation model (category or strata based) – this study

spatial autocorrelation strength at current stage. Previous geographically weighted-based SAR processes assume continuous variability in spatial autocorrelation. Alternatively, we addressed the issue of heterogeneous spatial autocorrelation through residual analysis. Our heterogeneous spatial autocorrelation model is the extension of traditional SAR models proposed by Anselin (1988) and assumes that the variation of spatial autocorrelation strength can be stratified or categorized.

This study aims to explore the impact of heterogeneous spatial autocorrelation strength and develop methods to capture this feature of spatial nonstationarity through simulation studies and empirical spatial analysis of transport geography. In this study, we designed a series of Monte Carlo simulations to demonstrate (1) the consequences of ignoring heterogeneous spatial autocorrelation in the traditional SLM and (2) the capability of our proposed method to capture heterogeneous autocorrelation patterns. We further applied the new method to an empirical analysis of traffic crashes in the Greater Perth Area of Australia. The remainder of the article is organized as follows: Section 2 presents our developed method to handle the existence of heterogeneous spatial autocorrelation patterns. Section 3 demonstrates the consequences of ignoring heterogeneous spatial autocorrelation using the traditional SLM and our model's capability through a series of Monte Carlo simulations. Section 4 presents the results of applying our adjusted SLM to a traffic crash study, followed by a discussion and conclusion in Sections 5 and 6, respectively.

2. Modeling the heterogeneity of spatial autocorrelation

Figure 1 shows the framework illustrating our proposed strategy for extracting patterns of heterogeneous autocorrelation and re-estimating the varying spatial

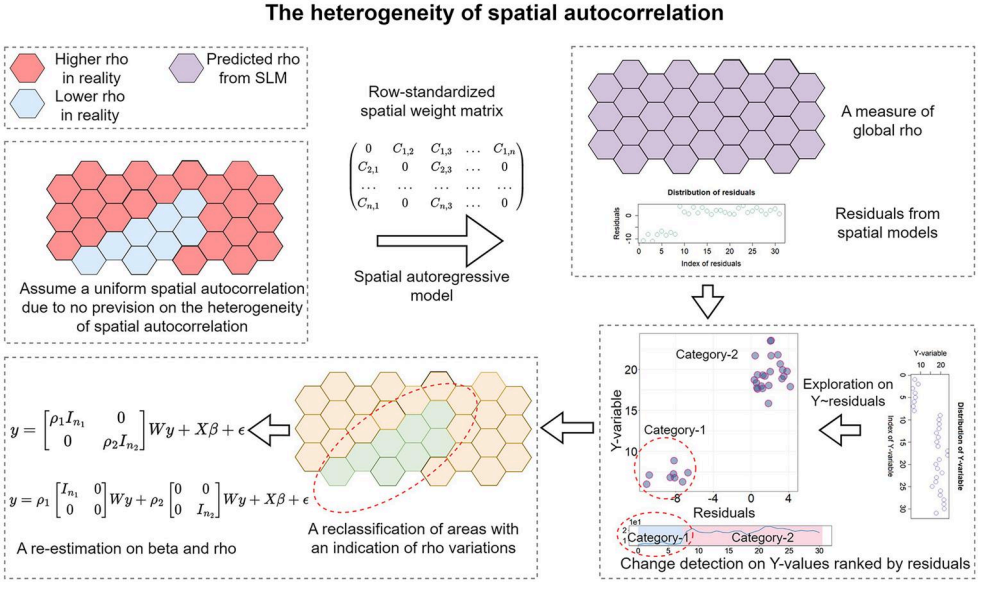

Figure 1. Framework of extracting and re-estimating heterogeneous spatial autocorrelation from residuals.

autocorrelation strength. Traditional SAR models assume that the strength of spatial dependence is homogeneous across space by estimating a single value of the SAR coefficient as a global indicator. Uncaptured heterogeneity in the autocorrelation will be left over as residuals. Thus, the proposed analytical approach was designed to capture heterogeneous autocorrelation patterns through additional residual analysis and to re-estimate the SAR processes using an adjusted generalized SAR coefficient matrix. The following two subsections present the details of the framework, including pattern extraction and model re-estimation using SLM.

2.1. Extracting patterns of heterogeneous spatial autocorrelation from residuals

Residuals from models that ignore heterogeneous autocorrelation may have remaining statistical associations with the dependent variables. Regions with different spatial autocorrelation strengths may have different degrees of association between the residuals and the dependent variable. Analysis of this statistical association can help to identify areas where different SAR processes occur. To this end, we propose the use of a change point detection algorithm from a Robust Geographical Detector (RGD) as a candidate method for extracting heterogeneous autocorrelation patterns from traditional SLM residuals when heterogeneous autocorrelation is overlooked.

The geographical detector is effective in exploring the statistical associations between variables according to the least-squared costs of the dependent variable categorized by statistical groups, which are determined through the discretization of the rank of the independent variable (Song *et al.* 2020, Guo *et al.* 2022). A higher sum of the least-squared costs from all statistical groups indicates a lower statistical association between the variables. The RGD has been proven to be a more effective method for indicating the association between variables than other geographical-detector-based approaches because of the introduction of an optimization algorithm for detecting changing points (Zhang *et al.* 2022). This optimization algorithm returns the specified number of breaking points that guarantee the highest level of association between variables after searching and comparing the least-squared costs for all possible combinations of subseries from the dependent variable values sorted by the independent variable values (Zhang *et al.* 2024). We keep the general algorithm structure identical to previous RGD research because the change point detection algorithm fits the need to identify statistical associations between the dependent variable and the residual (Zhang *et al.* 2022).

Algorithm 1: Change point detection from RGD (Zhang *et al.* 2022) for the statistical association between the dependent variable and the residuals from the SLM.

1 **function** Change Point Detection (minimal group size: r, number of changing points: k, input data: y-values and corresponding residual values from SLM)

Note: The number of different groups of spatial autocorrelation strength is $k+1$.

 # Preparation: compute the least cost for all pairs of sub-series

2 Reorder y-values according to the corresponding rank of residuals from SLM

3 Store the reordered y-values as a 1-dimensional series, and note as $y_{SLM-resi}$

4 **for all** sub-series with acceptable length (length $> r$) **do**

5 store all sub-series and corresponding least squared costs
6 **end for**
7 **for all** possible sub-series pairs with acceptable length (length $>$ *r*) **do**
8 record these sub-series pairs and their least costs
9 **end for**
 # Using Dynamic Programming to find changing points starting with the head or the tail of the $y_{SLM-resi}$
10 **while** not all changing points are found **do**
11 find two sub-series pairs with the least costs for the remaining *y*-series
12 store this changing point to the list *L*
13 set this last found changing point as the next starting point
14 **end while**
15 categorize $y_{SLM-resi}$ according to the list *L*, and save it as a new attribute to the file *CPD_by_k*
16 **return** *CPD_by_k*

The change point detection algorithm operates on a one-dimensional series of *y*-values, initially sorted by the SLM residual, along with two user-defined parameters: number of intervals and minimum size of the statistical group. The Algorithm 1, spanning Lines 2 to 7, starts by computing and storing the least-squared cost values for all possible subseries. Subsequently, from Lines 8 to 16, the algorithm uses dynamic programming to identify potential change points. Starting from the ascending or descending peaks of the residual, it iteratively seeks the change point that partitions the current subseries of the *y*-values with the least cost. This process continues until all change points are found, with each newly identified change point serving as the starting point for the next iteration. Typically, the minimum size of the statistical group is set to one, enabling change point detection on all subseries when the number of observations is not substantial. The number of intervals initially begins at two, indicating the requirement of one change point to delineate a binary pattern of variation. The capability of change point detection in identifying spatial autocorrelation patterns hinges on the statistical association between residuals and the *y*-value within the context of the SAR data generation process. Observations exhibiting higher spatial autocorrelation strength may receive additional values from their geographical neighbors, resulting in a higher *y*-value compared to others. Traditional SLM, which assumes constant spatial autocorrelation, tends to leave larger residuals for such observations. In other words, observations with higher spatial dependence values will also exhibit greater estimation errors when modelled by stationary spatial dependence models. This statistical association between spatial dependence and estimation error can be captured by change point detection.

In this study, we applied the change point detection algorithm to the *y*-variable and residuals from the SLM, as an example of an SAR model. The SLM is expressed by Eq. (1) (Anselin 1988).

$$y = \rho Wy + X\beta + \varepsilon \tag{1}$$

where ρWy is the spatial term with ρ as the SAR coefficient, W as a spatial weights matrix and y as the dependent variable; X is a matrix of independent variables; β are coefficients of independent variables; and ε is the error term.

2.2. Re-estimation on the heterogeneity of spatial autocorrelation

Change point detection has the potential to identify heterogeneous autocorrelation patterns, but the strength of spatial autocorrelation with variations from different statistical groups still needs to be re-estimated. In a traditional SLM, the strength of the spatial autocorrelation can be represented by the SAR coefficient (rho) as a constant value. As demonstrated in the matrix representation, the constant value of rho can also be equivalently transformed into a diagonal matrix, where all nonzero elements are equivalent to the value of rho, as shown in Eq. (2).

$$y = \begin{bmatrix} \rho & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho \end{bmatrix} Wy + X\beta + \varepsilon \tag{2}$$

If the strength of spatial autocorrelation varies from place to place, the diagonal matrix of the SAR coefficient can be maximally extended to the form $\begin{bmatrix} \rho_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \rho_n \end{bmatrix}$, where SAR coefficients from ρ_1 to ρ_k $(k < n)$ are different from each other and the number of different SAR coefficient 'k' is no greater than the number of total observations minus the number of parameters to be estimated, as restricted by the degrees of freedom.

To simplify the problem, we start with the impact of heterogeneous autocorrelation, shown as a binary difference, in which regions with variations in spatial autocorrelation strength can be roughly categorized into two groups. A binary difference in spatial autocorrelation strength is the simplest representation of variation, and this research conducts a pilot investigation on the consequences of ignoring the variation in spatial autocorrelation strength resulting from this simple representation.

$$y = \begin{bmatrix} \rho_1 I_{n_1} & 0 \\ 0 & \rho_2 I_{n_2} \end{bmatrix} Wy + X\beta + \varepsilon \tag{3}$$

The SAR coefficient matrix in Eq. (3) is shown as $\begin{bmatrix} \rho_1 I_{n_1} & 0 \\ 0 & \rho_2 I_{n_2} \end{bmatrix}$, where $\rho_1 I_{n_1}$ and $\rho_2 I_{n_2}$ represent two sets of areas with different strengths of spatial autocorrelation (with $n_1 + n_1 = n$, and n is the number of observations). These two sets of areas were disjointed and the union was the entire study area. Thus, in Eq. (3), $\begin{bmatrix} \rho_1 I_{n_1} & 0 \\ 0 & \rho_2 I_{n_2} \end{bmatrix} Wy$ is the spatial lag term with patterns of heterogeneous autocorrelation; X is a matrix of independent variables; β are coefficients, and ε is the error term. Equation (3) can be converted into Eq. (4) when re-estimating the values of the beta and SAR coefficients. For a heterogeneous autocorrelation with binary differences, an extra degree of freedom must be allocated to estimate the additional SAR coefficient. The terms $\rho_1 \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} Wy$ and $\rho_2 \begin{bmatrix} 0 & 0 \\ 0 & I_{n_2} \end{bmatrix} Wy$ in Eq. (4) represent the spatial impacts from the first and the second groups of areas respectively. Eqs. (3) and (4) are extensions of the SLM, and the SLM with the heterogeneous autocorrelation assumption remains a

nonlinear model. Thus, we re-estimate the values of the beta and SAR coefficients using maximum likelihood estimation (MLE) rather than ordinary least squares.

$$y = \rho_1 \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} Wy + \rho_2 \begin{bmatrix} 0 & 0 \\ 0 & I_{n_2} \end{bmatrix} Wy + X\beta + \varepsilon \tag{4}$$

Algorithm 2 illustrates the computational process for estimating regression coefficients within the heterogeneous spatial autocorrelation model. The 'Heterogeneous Spatial Autocorrelation Model' function estimates the coefficients for categories of spatial autocorrelation and independent variables using the MLE estimation and PORT routines method for optimal parameter search. In the MLE process, the objective of maximizing the log-likelihood values derived from residuals is transformed into minimizing the negative log-likelihood values. The steps from Lines 1 to 7 detail the computation of the log-likelihood values of residuals, which serve as the cost function for identifying optimal regression coefficients. The 'Heterogeneous Spatial Autocorrelation Model' function is initiated at Line 8, and the spatial lag values for different groups of spatial autocorrelation are computed between Lines 9 and 14. Subsequently, regression coefficients are determined using the PORT routines via the 'nlminb' function in R. The algorithm concludes by returning the estimated coefficients along with the mean square error (MSE) of the estimator.

Algorithm 2: Computation of Log-Likelihood and estimation of spatial regression coefficients.

Input: data[*Y*_value, *X*_variables, Category of spatial autocorrelation], spatial weight matrix (*SWM*)

Note: The column of variable 'Category of spatial autocorrelation' can be generated through Change Point Detection or other method that can identify the heterogeneity of spatial autocorrelation.

1 **function** Log-Likelihood (parameters[1:k], data)

Note: there are *k* parameters to be estimated with *p* ($p < k$) parameters for coefficients of *X*_variables, *(k-1-p)* parameters for coefficients of spatial autocorrelation, and the remaining one for σ; the number of observation is *n*.

2 *coef* ← *parameters[1:k-1]*

3 *sigma* ← *parameters[k]*

4 *predicted_Y* ← $\sum_{i=1}^{p} X_i$*coef[i] + $\sum_{j=1}^{k-1-p} Lag\_y\_Category_j$*coef[j]

5 *residuals* ← *Y_value - predicted_Y*

6 *log-likelihood* ← $\sum_{m=1}^{n} \log(f(residual_m|0,\sigma))$
(where $\log(f(z|0,\sigma)) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{z^2}{2\sigma^2}$)

7 **return** *-log-likelihood*

8 **function** Heterogeneous Spatial Autocorrelation Model (data, *SWM*)
 # Compute spatial lag values for each category

9 *full_lag_y* ← *SWM* %*% *Y_value*

10 append *full_lag_y* to the data

11 **for all** categories of spatial autocorrelation **do**

12 *lag_category_j* ← *full_lag_y*

```
13        lag_category_j[category != "category_j"] ← 0
14      end for
    # Find optimal parameters and get coefficients using PORT routines
15    initial_parameters ← (1, 1, 1, … … , 1)
16    para_estimated ← nlminb(initial_parameters, Log-Likelihood, data)
17    coef ← para_estimated[1:k-1]
18    predicted_Y ← ∑_{i=1}^{p} X_i*coef[i] + ∑_{j=1}^{k-1-p} Lag_y_Category_j*coef[j]
19    MSE ← mean((Y_value − predicted_Y)^2)
20    return (coef, MSE)
```

3. Monte Carlo simulation: extracting and re-estimating binary heterogeneous spatial autocorrelation

In this research, we conducted a series of Monte Carlo simulations to demonstrate (1) the consequences of ignoring heterogeneous spatial autocorrelation in the traditional SLM and (2) the capability of our proposed method to capture heterogeneous autocorrelation patterns under different heterogeneous autocorrelation patterns. Each pattern under each scenario was tested using 1,000 random sequences of independent variables and errors. Estimations of beta values for independent variable coefficients and SAR coefficients from both the traditional SLM and our adjusted SLM are summarized and compared. We designed a series of heterogeneous autocorrelation patterns under different ranges of spatial autocorrelation strength, as described in six scenarios:

- Scenario 1: Spatial autocorrelation is generally very weak, with rho = 0.1, whereas in some local areas, spatial autocorrelation is very strong, with rho = 0.7.
- Scenario 2: Spatial autocorrelation is generally very strong, with rho = 0.7, whereas in some local areas, spatial autocorrelation is very weak, with rho = 0.1.
- Scenario 3: Spatial autocorrelation is generally comparatively weak, with rho = 0.2, whereas in some local areas, spatial autocorrelation is comparatively strong, with rho = 0.6.
- Scenario 4: Spatial autocorrelation is generally comparatively strong, with rho = 0.6, whereas in some local areas, spatial autocorrelation is comparatively weak, with rho = 0.2.
- Scenario 5: Spatial autocorrelation is generally slightly weak with rho = 0.3, whereas in some local areas, spatial autocorrelation is slightly strong, with rho = 0.5.
- Scenario 6: Spatial autocorrelation is generally slightly strong, with rho = 0.5, whereas in some local areas, spatial autocorrelation is slightly weak, with rho = 0.3.

To simulate the SAR processes, we designed 10-by-10 grids as the study area, and two independent variables '*x*1' and '*x*2' were randomly drawn from a uniform distribution *U (4, 8)*, where independent variables are all valued positive. The coefficient values for both independent variables were designed and equivalent to 1. The random error term followed a normal distribution of *N* (0, 0.5), with a mean value of 0 and standard deviation of 0.5. The spatial weights matrix followed the geometric contiguity

of the queen on grids and was row-standardized. As explained in Eqs. (2) and (3), the heterogeneous autocorrelation patterns can be controlled by the values of the diagonal elements of the SAR coefficient matrix. Taking Scenario 1 as an example, the majority of the diagonal elements of the SAR coefficient matrix were set to 0.1, and in some local areas, the corresponding diagonal elements were set to 0.7. The dependent variable y can be generated using Eq. (5) once the SAR coefficient matrix with specified heterogeneous autocorrelation patterns, spatial weight matrix, error, independent variables and their coefficients are determined.

$$y = \left(I - \begin{bmatrix} \rho_1 I_{n_1} & 0 \\ 0 & \rho_2 I_{n_2} \end{bmatrix} W\right)^{-1} (\beta_1 X_1 + \beta_2 X_2 + \varepsilon) \tag{5}$$

The residuals as an input of change point detection refer to the residuals after the traditional SLM generation, which contain information on the ignorance of heterogeneous autocorrelation patterns. The following subsections demonstrate the performance of change point detection in identifying heterogeneous autocorrelation patterns and compare SLM with heterogeneous and homogeneous spatial autocorrelation assumptions in estimating the value of beta and rho when facing the occurrence of heterogeneous autocorrelation.

3.1. Extracting patterns of heterogeneous spatial autocorrelation

Figures 2–4 show the results of change point detection for extracting different heterogeneous autocorrelation patterns under different scenarios. Scenario 1 in Figure 2(a), Scenario 3 in Figure 3(a) and Scenario 5 in Figure 4(a) have similar globally lower SAR coefficients; and Scenario 2 in Figure 2(b), Scenario 4 in Figure 3(b) and Scenario 6 in Figure 4(b) have similar higher SAR coefficients in more areas. Scenarios 1 and 2, 3 and 4 and 5 and 6 are three groups of cases with inversely designed global and local spatial dependencies.

In general, change point detection can identify heterogeneous autocorrelation patterns, and the algorithm's performance is influenced by the pattern geometry, level of difference between spatial autocorrelation strengths, and differences in both the overall and local spatial dependence (global-lower autocorrelation or global-higher autocorrelation). Regarding pattern geometry, the change point detection algorithm excels

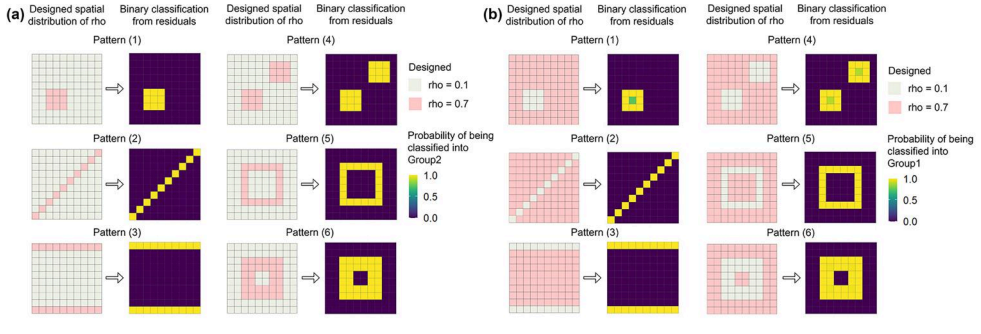

Figure 2. Extracting heterogeneous spatial autocorrelation patterns with rho from 0.1 to 0.7 for (a) Scenario 1 and (b) Scenario 2.

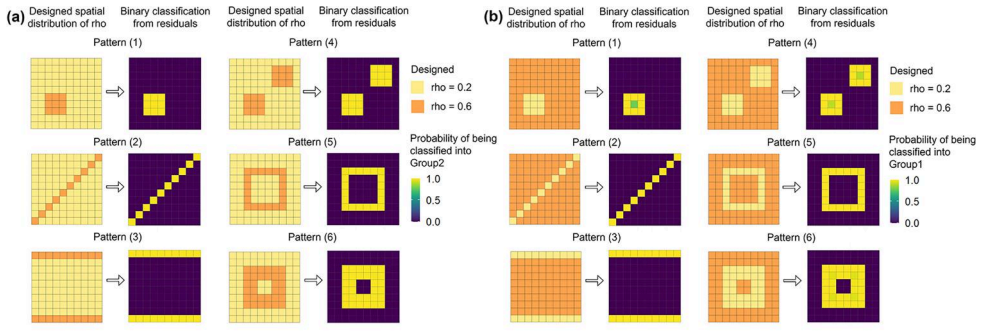

Figure 3. Extracting heterogeneous spatial autocorrelation patterns with rho from 0.2 to 0.6 for (a) Scenario 3 and (b) Scenario 4.

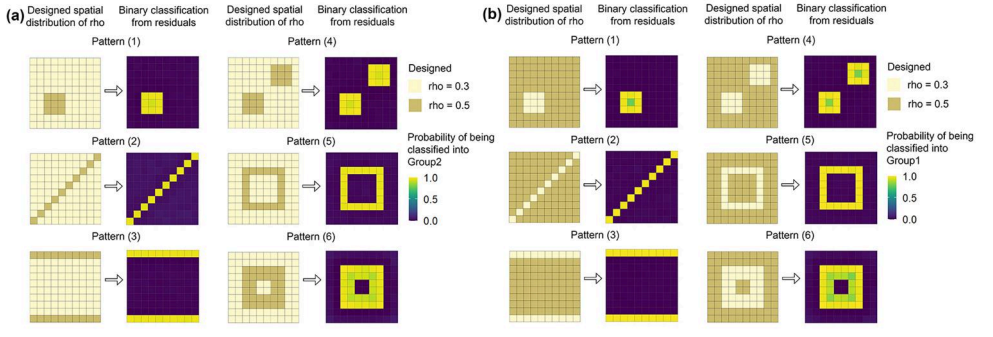

Figure 4. Extracting heterogeneous spatial autocorrelation patterns with rho from 0.3 to 0.5 for (a) Scenario 5 and (b) Scenario 6.

in extracting geographical structures, particularly in identifying regions with variations in transmitting and receiving spatial impacts from neighbors. These geographical structures included line and ring structures with a single unit, as shown in Patterns (2), (3) and (5), and the boundary where the clustered regions with two different SAR coefficients met, as shown in Patterns (1), (4) and (6). For both global-lower and global-higher SAR processes, the level of difference between spatial autocorrelation strengths can affect the performance of change point detection. The greater the differences in the strength of spatial autocorrelation, the easier the change point detection in extracting structures with SAR coefficient variations, which is especially evident in Pattern (6) for all scenarios and Patterns (1) and (4) for the three global-lower autocorrelation scenarios. By comparing each pair of inversely designed spatial dependence scenarios, change point detection can extract geographical structures with variations in transmitting and receiving spatial impacts for all scenarios, and it performs better in global-lower but local-higher cases when identifying a clustering structure of regions with higher spatial autocorrelation.

3.2. Re-estimation on the coefficients of independent variables

Distributions from Monte Carlo simulations of the re-estimated beta based on our adjusted SLM are shown in Figures 5–7, together with comparisons of beta

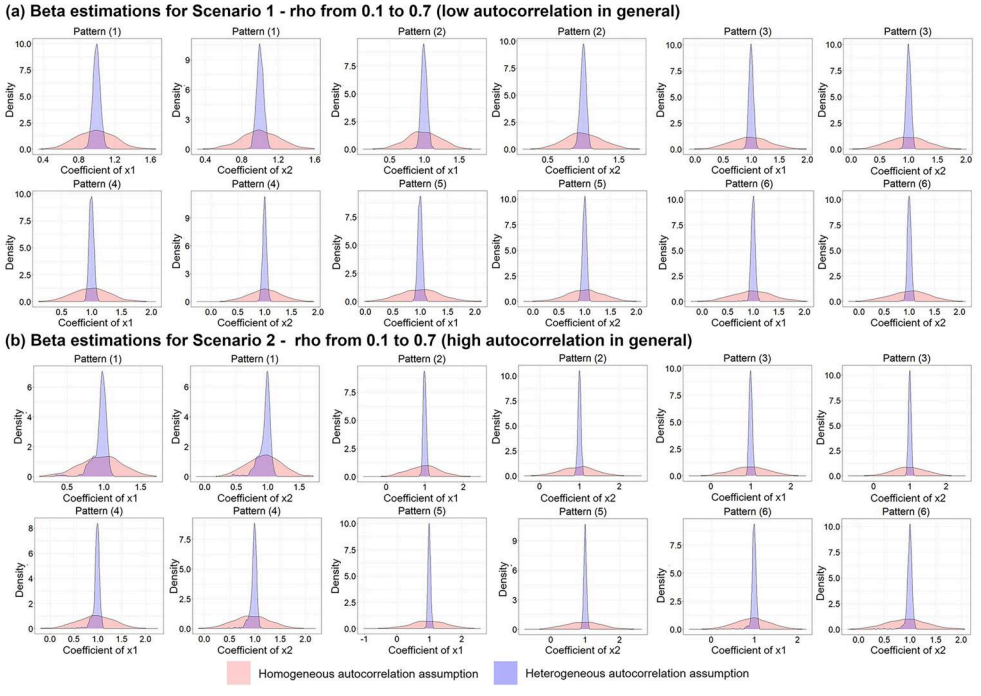

Figure 5. A comparison between beta values estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 1 and 2.

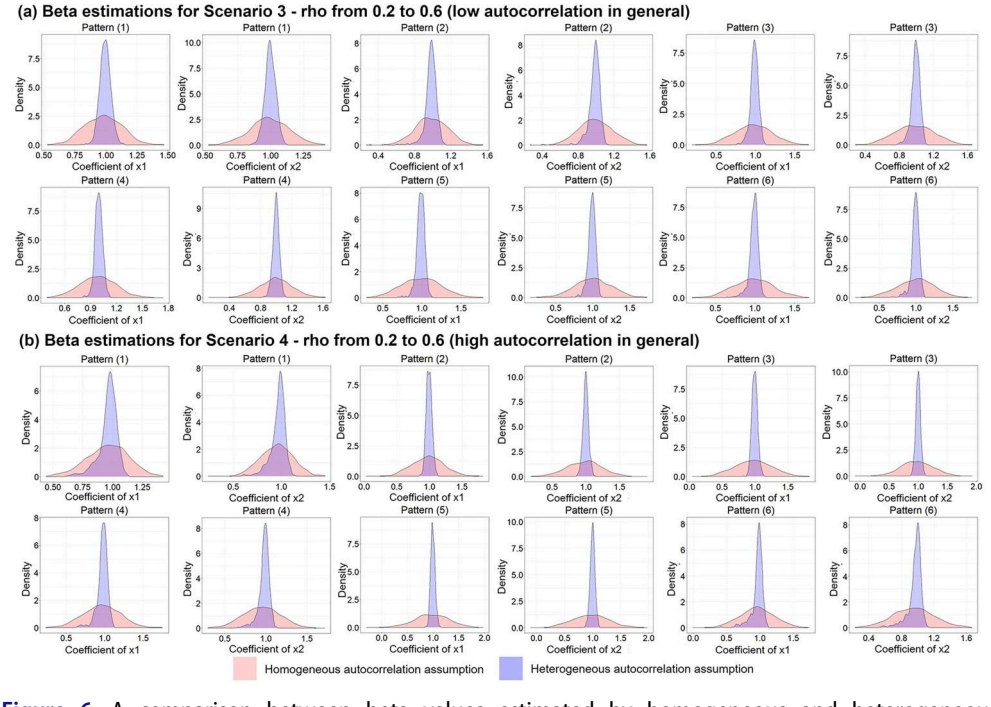

Figure 6. A comparison between beta values estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 3 and 4.

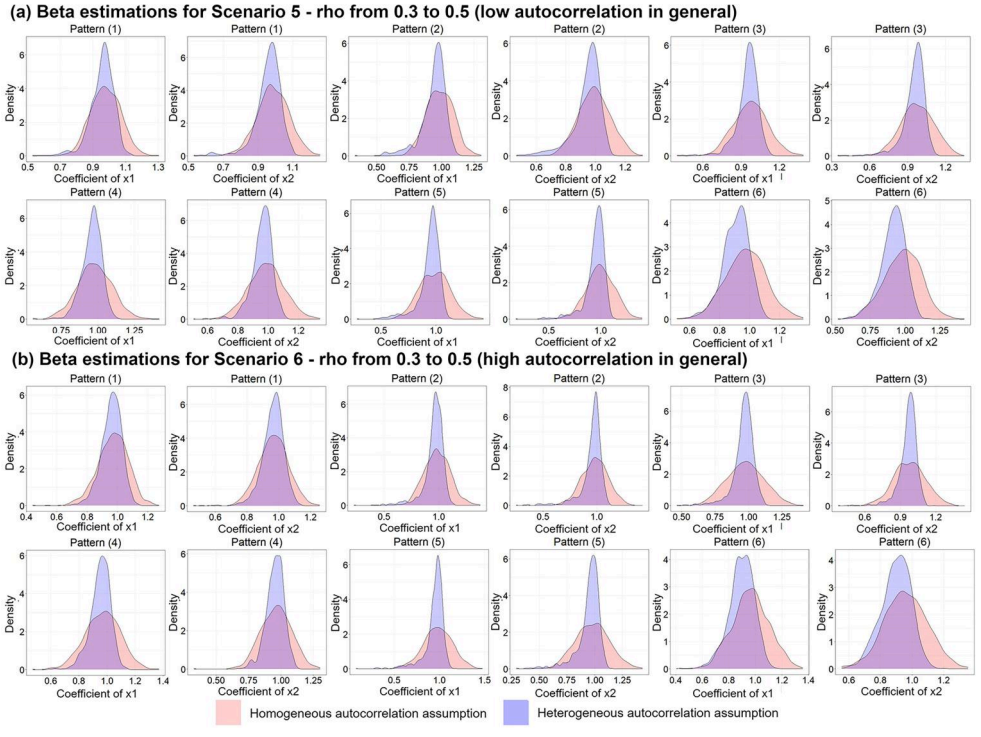

Figure 7. A comparison between beta values estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 5 and 6.

estimations by the traditional SLM with the homogeneous autocorrelation assumption. These results answer (1) the extent to which our adjusted SLM precisely returns the values of beta once the geographical structure showing variations in spatial autocorrelation strength is captured and (2) the model performance of the traditional SLM when facing heterogeneous autocorrelation patterns.

The results in Figures 5 and 6 show that our SLM with a generalized SAR coefficient matrix can easily re-estimate the beta values for both independent variables with high reliability when the difference between spatial autocorrelation strength is no less than 0.4, despite minor disturbances on Pattern (1) under Scenarios 2 and 4 and Pattern (6) under Scenario 4, where the geographical structure of heterogeneous autocorrelation is not fully extracted by change point detection. For patterns under these four scenarios, the vast majority of the estimated beta values for both independent variables fell into the value range of 0.9 to 1.1. However, when there is heterogeneous autocorrelation, the traditional SLM estimates beta values with higher uncertainties. The ranges of the estimated beta values for both independent variables using the traditional SLM were much wider than those from the adjusted SLM. In other words, the beta values estimated by the traditional SLM are not precise when the difference between the two SAR coefficients in the SAR process is greater than 0.4.

As shown in Figure 7, with the decrease in the difference between the two SAR coefficients to 0.2, our adjusted SLM can still accurately re-estimate the beta values for both independent variables for all patterns, excluding Pattern (6) with smaller average

estimated values, where the entire geographical structure of heterogeneous autocorrelation cannot be fully captured. With more minor parts not fully extracted in Scenarios 5 and 6, the distributions of the estimated beta for both independent variables are slightly negatively skewed. On the other hand, the traditional SLM with a homogeneous autocorrelation assumption has less uncertainty in beta estimations when the SAR coefficient difference declines, despite the overall performance paling in comparison to that of SLM with the heterogeneous assumption in a majority of the cases. The ranges of the estimated beta values for both independent variables using the traditional SLM are still somewhat wider than our adjusted SLM in the majority of cases.

3.3. Re-estimation on heterogeneous spatial autocorrelation

Figures 8–10 show the re-estimated values of the SAR coefficients for the six scenarios based on SLM with a generalized SAR coefficient matrix. It is clear that, by costing an extra degree of freedom, our adjusted SLM can re-estimate the SAR coefficients well once a real heterogeneous autocorrelation structure is captured. The estimation of the SAR coefficients may be slightly skewed when a minority of the heterogeneous autocorrelation structure is not extracted, with Patterns (1) and (4) in Scenario 2 as an example. The overall value ranges for the estimated SAR coefficients are small in the majority of cases when the difference in the SAR coefficients is 0.4 or greater. With the difference in SAR coefficients dropping to 0.2, a greater part of the estimated rho values can still be estimated with an approximation to the designed values, even though the overall estimated value range is wider than those under Scenarios 1 to 4. In contrast to the estimations of beta, with minor unidentified heterogeneous

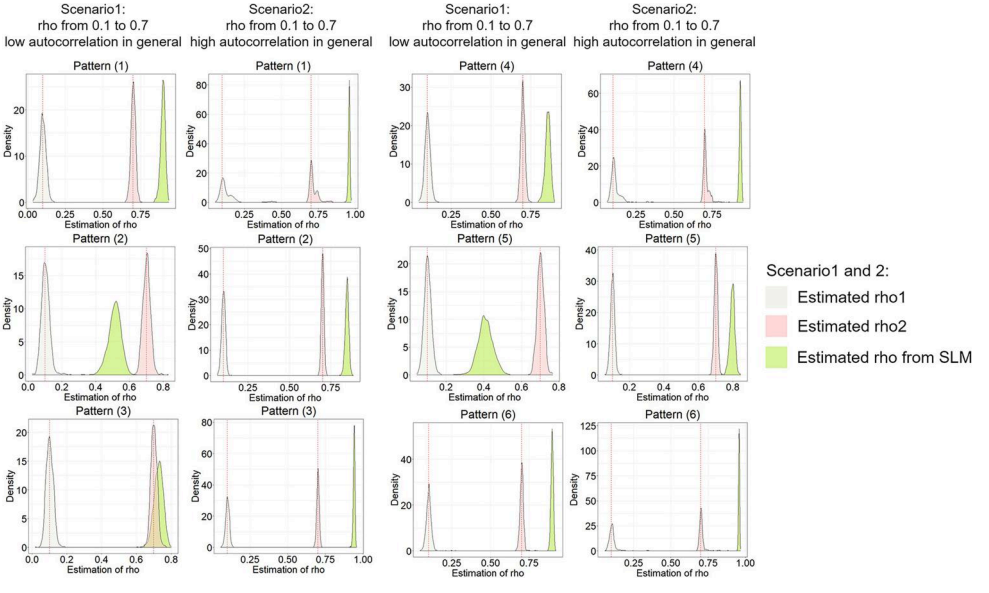

Figure 8. A comparison between rho estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 1 and 2.

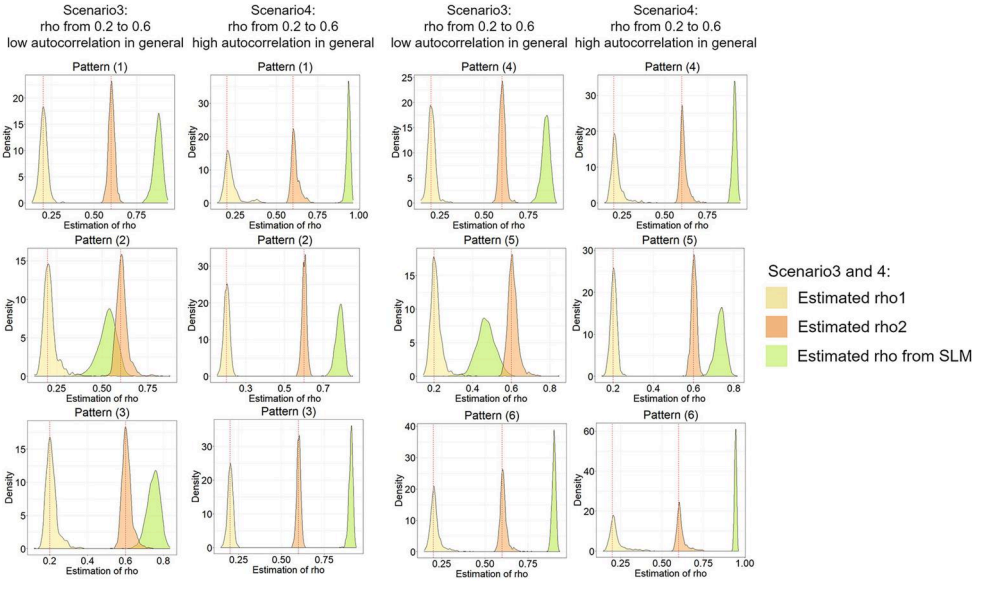

Figure 9. A comparison between rho estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 3 and 4.

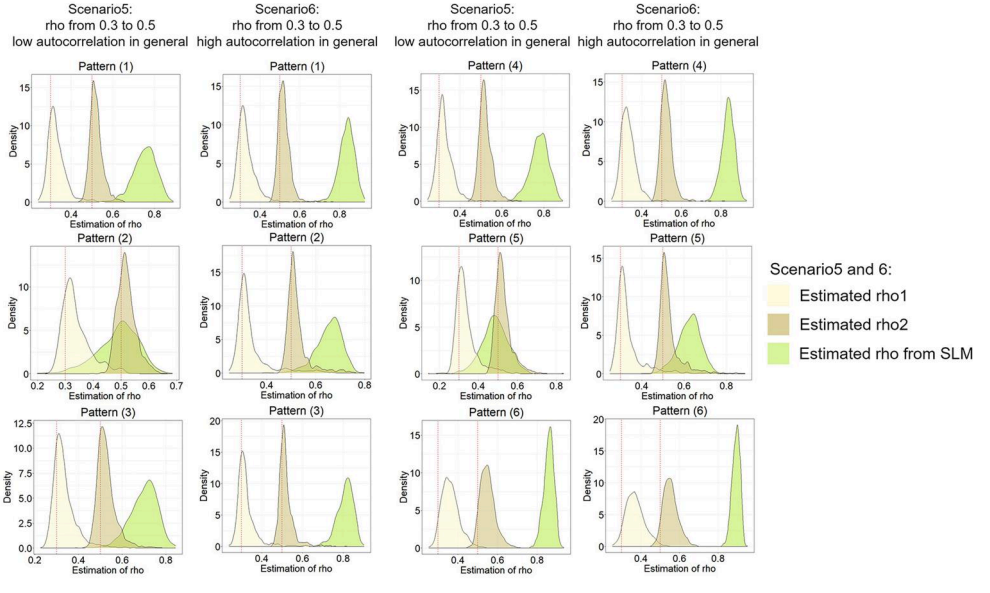

Figure 10. A comparison between rho estimated by homogeneous and heterogeneous autocorrelation assumptions for Scenarios 5 and 6.

autocorrelation structures, the estimated values of rho are slightly positively skewed, and this phenomenon is more evident under Scenarios 5 and 6, as shown in Figure 10. The accuracy of the estimation of the SAR coefficients also relies on the identification of heterogeneous autocorrelation patterns. With more heterogeneous autocorrelation structures not captured in Pattern (6) under Scenarios 5 and 6, the

values of the two designed SAR coefficients were underestimated. The traditional SLM estimates the strength of spatial autocorrelation for various heterogeneous autocorrelation cases, as shown in these figures. Intuitively, the traditional SLM may return an SAR coefficient that indicates the overall averaged strength of the spatial autocorrelation for the entire study area. In some cases, with Patterns (2) and (5) under global-lower SAR strength scenarios, in particular, the traditional SLM would return a value that falls between the two designed SAR coefficients. However, in other cases, the traditional SLM provides an overestimated rho value, representing an extremely high overall autocorrelation that does not indicate the true scenario.

3.4. The impact of heterogeneous spatial autocorrelation on model performance

This section further demonstrates the impact of heterogeneous spatial autocorrelation on spatial autoregressive processes, focusing on the mean squared error (MSE) of estimators and the statistical significance of regression coefficients. Figure 11 provides a summary of the MSE of estimators for the heterogeneous spatial autocorrelation model and the traditional SLM under six scenarios. Although spatial dependence structures may not be fully captured by change point detection in some cases, as discussed in Subsection 3.1, the heterogeneous spatial autocorrelation model demonstrates robust performance and stability despite variations in spatial autocorrelation strength and patterns. The model consistently exhibits a better goodness of fit than traditional SLM across all 36 cases, evidenced by much lower MSE values.

Heterogeneous spatial autocorrelation also has impacts on the statistical significance of regression coefficient, especially for independent variables in specific cases. Both the heterogeneous spatial autocorrelation model and the SLM yield statistically significant spatial autocorrelation coefficients. However, the significance of the regression coefficients for independent variables estimated by the SLM is affected by different spatial autocorrelation patterns. As presented in Table 2, the heterogeneous spatial autocorrelation model consistently estimates statistically significant regression coefficients across all patterns, whereas the SLM cannot guarantee statistical significance, especially in Scenarios 2 and 4.

4. Case study: traffic crash factor analysis in Perth

The heterogeneous autocorrelation assumption was applied to a case study involving traffic crash analysis in Perth. The Greater Perth Area is the functional geographical extent of the capital city of Western Australia. The capital city area has a population of over 80% of all statewide residents (Australian Bureau of Statistics 2023). Transportation plays a key role in modern urban systems (Zou *et al.* 2012), and Perth is no exception. From 2016 to 2021, the number of residents choosing to drive to work in Perth increased from 656,000 to 728,000 and the number of residents using at least one means of transport increased from 784,000 to 850,000 (Australian Bureau of Statistics 2022). Detailed transportation and road asset information was recorded by the Australian Bureau of Statistics (ABS) and the local road authority (Main Roads,

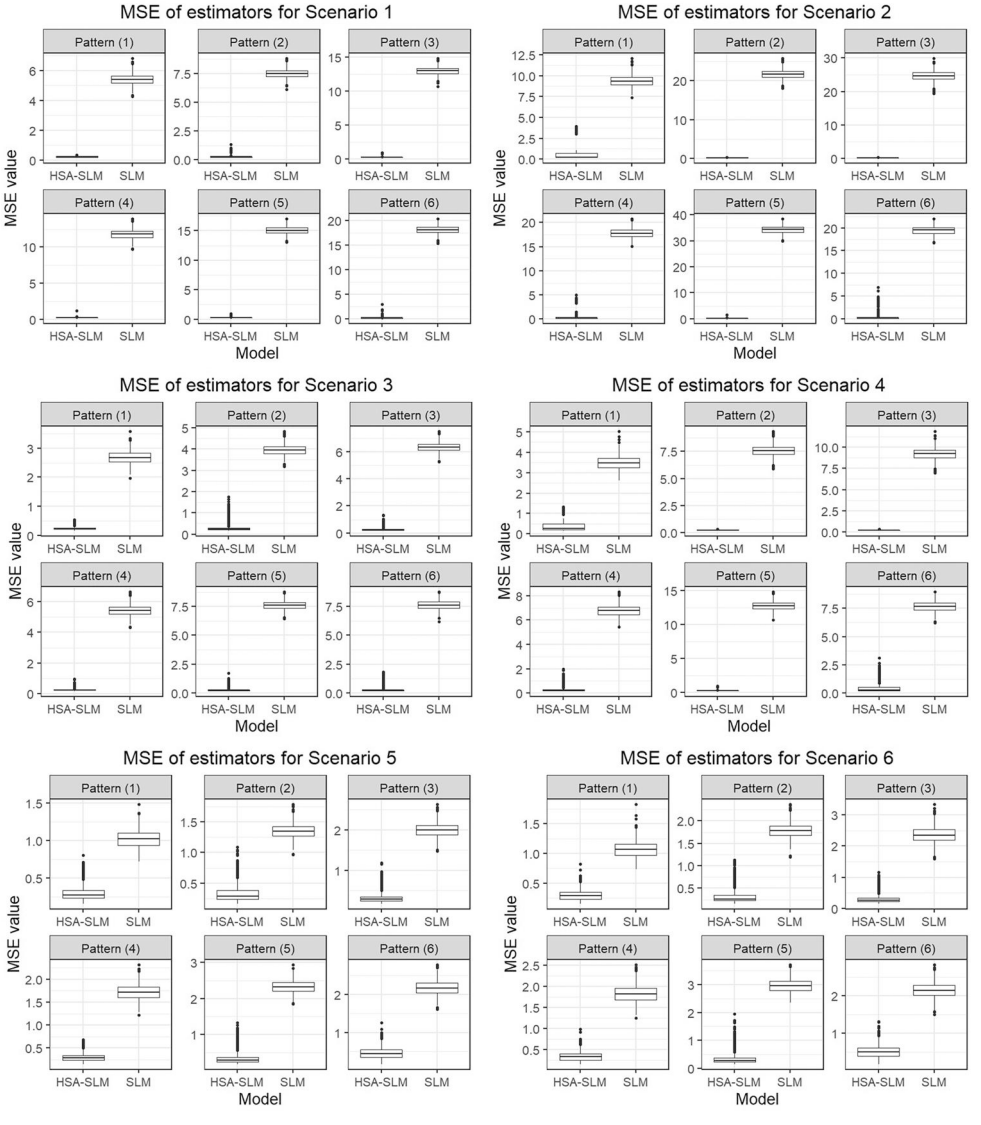

Figure 11. MSE of estimators for the SLM with heterogeneous spatial autocorrelation assumption (HSA-SLM) compared to traditional SLM.

Western Australia). Traffic crash data in Perth are managed by the road information system of Main Roads, Western Australia and are updated at a frequency of at least once a year. There are approximately 22,000 reported traffic crashes in the Greater Perth Area each year, and approximately one-fifth of these are severe traffic crashes with damaged property and injured people (Main Roads Western Australia 2023a).

A detailed spatial analysis of the factors associated with traffic crashes can aid in smart transport planning and data-driven decision-making for road safety. Thus, in this study, we undertook a spatial analysis of traffic crashes in Perth at Statistical Area Level 2 (SA2) for 2021 from the perspective of heterogeneous spatial autocorrelation to assist local transport decision-making.

Table 2. A summary on the percentage of coefficients with statistical significance using *z*-test for all patterns from Monte Carlo simulations.

	Percentage of significant coefficients (p-value <0.05)				
	x1 (HSA-SLM / SLM)	x2 (HSA-SLM / SLM)		x1 (HSA-SLM / SLM)	x2 (HSA-SLM / SLM)
	Scenario 1			Scenario 2	
Pattern (1)	100% / 99.9%	100% / 99.8%	Pattern (1)	99.9% / 92.9%	99.9% / 92.6%
Pattern (2)	100% / 98.7%	100% / 98.2%	Pattern (2)	100% / 67.1%	100% / 65.3%
Pattern (3)	100% / 84.8%	100% / 85.4%	Pattern (3)	100% / 59.9%	100% / 59.5%
Pattern (4)	100% / 90.3%	100% / 92.2%	Pattern (4)	99.9% / 72.8%	100% / 71.4%
Pattern (5)	100% / 81.9%	100% / 84%	Pattern (5)	100% / 47.8%	100% / 46.5%
Pattern (6)	100% / 74.1%	100% / 76.1%	Pattern (6)	99.6% / 68%	99.9% / 66.7%
	Scenario 3			Scenario 4	
Pattern (1)	100% / 100%	100% / 100%	Pattern (1)	100% / 100%	100% / 99.9%
Pattern (2)	100% / 100%	100% / 100%	Pattern (2)	100% / 98.5%	100% / 97.1%
Pattern (3)	100% / 99.2%	100% / 99.2%	Pattern (3)	100% / 94.8%	100% / 95.7%
Pattern (4)	100% / 99.9%	100% / 99.9%	Pattern (4)	100% / 98.3%	100% / 98.4%
Pattern (5)	100% / 98.4%	100% / 97.9%	Pattern (5)	100% / 86.4%	100% / 85.8%
Pattern (6)	100% / 98%	100% / 98.1%	Pattern (6)	100% / 96.5%	100% / 96.6%
	Scenario 5			Scenario 6	
Pattern (1)	100% / 100%	100% / 100%	Pattern (1)	100% / 100%	100% / 100%
Pattern (2)	100% / 100%	100% / 100%	Pattern (2)	100% / 100%	100% / 100%
Pattern (3)	100% / 100%	100% / 100%	Pattern (3)	100% / 100%	100% / 100%
Pattern (4)	100% / 100%	100% / 100%	Pattern (4)	100% / 100%	100% / 100%
Pattern (5)	100% / 100%	99.9% / 100%	Pattern (5)	99.9% / 100%	100% / 100%
Pattern (6)	100% / 100%	100% / 100%	Pattern (6)	100% / 100%	100% / 100%

Note: HSA-SLM refers to the SLM with heterogeneous spatial autocorrelation assumption. SLM refers to traditional SLM.

4.1. Datasets

Traffic crashes are complex incidents that are influenced by various factors. Commonly associated factors include road infrastructure (Mane and Pulugurtha 2018, Siuhi *et al.* 2021), traffic conditions (Ahmed *et al.* 2021) and regional commuting patterns (Hu and Wang 2020). The occurrence of traffic crashes on roads can be influenced by road speed limits, traffic volume and congestion, traffic signal location, road design and commuting patterns (Mane and Pulugurtha 2018, Di Stasi *et al.* 2022). This data-driven study considered five transport variables associated with the occurrence of traffic crashes, summarized in medium-level statistical areas. These independent variables included road speed at crash sites, estimated traffic volume, distance from the crash site to the nearest traffic signal and the percentage of residents choosing to drive or walk to work.

The Main Roads Western Australia release open-access transportation information including traffic crashes, road speed limits and traffic volume. In this research, we focused on severe traffic crashes that caused property damage and reported injuries that occurred in 2021 (Main Roads Western Australia 2023a). Information on road speed limits at each crash site can be inferred from the spatial layers of traffic crashes and road networks with road speed attributes (Main Roads Western Australia 2023b). Information on the distance from the crash site to the nearest signal and regional average traffic volume can be obtained from government-released spatial data on traffic signals and volumes (Main Roads Western Australia 2018, Main Roads Western Australia 2022).

SA2 is a medium-sized geographic boundary that indicates local communities with social interactions within local government areas or significant urban areas (Australian

Table 3. A summary of traffic crash and independent variables.

Variable name	Unit	Primary raw datasets name	Data type	Data source
Traffic crash (Y-variable)	count	Crash Information	Spatial data: crash site point	Main Roads WA
Average road speed at crash sites	km/h	Legal speed limits	Spatial data: road network with speed limit as an attribute	Main Roads WA
Average distance from crash site to the nearest traffic signal	meter	Traffic signal sites	Spatial data: traffic signal point	Main Roads WA
Estimated averaged regional traffic volume	count/day	Traffic digest	Spatial data: traffic count sites with traffic volume as an attribute	Main Roads WA
Percentage of residents: Work by car	%	Data by region: Family and community	Statistical data: Australian national census	ABS
Percentage of residents: Walk to work	%	Data by region: Family and community	Statistical data: Australian national census	ABS

Bureau of Statistics 2021a, 2021b). In this study, we analyzed the spatial relationship between the traffic crash occurrence count and its influential variables at the granularity of SA2 to support data-driven analysis for decision-making. Traffic crashes causing property damage and injury, as the dependent variable, were counted and summed for the SA2 areas. We used the SA2 areas to obtain a regional summary through the spatial average of independent variables, including road speed at each crash site, regional traffic volume and distance from the crash site to its nearest traffic signal, to estimate the regional level of the crash-associated variables. The percentage of residents driving or walking to work can be estimated from the total number of residents and residents choosing different commuting patterns using ABS census data. The details of the variables and raw datasets are summarized in Table 3. Traffic crashes and independent variables at the SA2 level are spatially visualized in Figure 12.

4.2. Data preprocessing and preparation

In the data preprocessing stage, the spatial data accessed from the Main Roads and ABS were processed and transformed into variables at SA2 for further analysis. Traffic crash points, which happened in the year 2021 and categorized into 'medical', 'hospital' and 'fatal', are extracted from raw datasets and summed up at each SA2 region.

The original regional traffic crash count values followed a long-tailed distribution, and the *y*-variable was logarithmically transformed to fulfill the normality assumption of the regression. The independent variable 'average road speed at crash site' is generated by spatial joining the road network with the speed attribute to the crash site points in GIS, and averaging road speed at crash sites by SA2 regions. The variable 'average distance from crash site to the nearest traffic signal' is generated by calculating the distance between each traffic crash point to the nearest traffic signal location

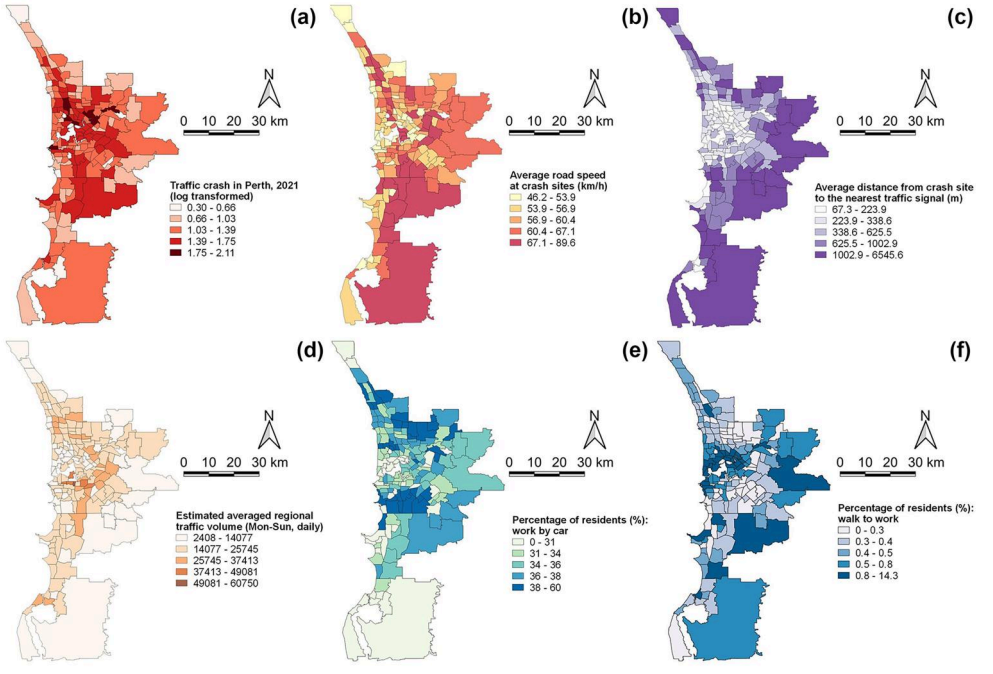

Figure 12. Spatial distribution of traffic crashes and independent variables at Statistical Area Level 2 (SA2). (a) Traffic crash in Perth; (b) Average road speed at crash sites; (c) Average distance from crash site to the nearest traffic signal; (d) Estimated averaged regional traffic volume; (e) Percentage of residents: work by car; and (f) Percentage of residents: walk to work.

using GIS, and taking an average on the attribute of distance by SA2 regions. The traffic volume in each SA2 area was estimated using the mean traffic count at each monitoring site within that area. ABS surveyed the total number of residents and residents driving or walking to work at the SA2 granularity in the 2021 census, and the percentage of residents commuting on foot or by car was calculated as the number of residents following the commuting pattern divided by the total number of residents.

Prior to the regression, a multicollinearity test using the variance inflation factor (VIF) value was applied to validate the basic regression assumption, and the VIF threshold was set to 2.5 (Zhang *et al.* 2023). Explanatory variables with VIF greater than 2.5 will be filtered before the linear model computation. Spatial regression models can be executed once the independent variables pass the multicollinearity test.

4.3. Case study results

The five selected independent variables shown in Table 3 passed the multicollinearity test, with all VIF values less than 2.5. The results of the models with comparisons of goodness of fit are shown in Table 4, and the spatial distribution of the residuals from the SLM is shown in Figure 13(a). No evident improvement was made by the traditional SLM by introducing a spatial lag term with a global SAR coefficient to a linear model, and the root mean squared error value changed from 0.283 using a linear model to 0.273 using the traditional SLM. By assuming the existence of heterogeneous

Table 4. Summary of regression models for traffic crash analysis.

	SLM – Heterogeneous assumption (p-value)	SLM – Homogeneous assumption (p-value)	Linear regression (p-value)
Average road speed at crash sites	6.79 e – 03 ($p < 0.001$)	7.47 e – 03 ($p < 0.01$)	8.19 e – 03 ($p < 0.005$)
Average distance from crash site to the nearest traffic signal	−6.49 e – 05 ($p < 0.001$)	−8.89 e – 05 ($p < 0.001$)	−1.09 e – 04 ($p < 0.001$)
Estimated averaged regional traffic volume	2.30 e – 06 ($p > 0.05$)	2.62 e – 06 ($p > 0.05$)	2.78 e – 06 ($p > 0.05$)
Percentage of residents: Work by car	7.77 e – 01 ($p < 0.001$)	6.18 e – 01 ($p < 0.05$)	6.57 e – 01 ($p < 0.05$)
Percentage of residents: Walk to work	1.03 e + 01 ($p < 0.001$)	7.96 ($p < 0.001$)	8.80 ($p < 0.001$)
Rho value	rho 1 = 1.67 e – 01 ($p < 0.005$)		NA
	rho 2 = 5.28 e – 01 ($p < 0.001$)	2.9 e – 01 ($p < 0.005$)	
RMSE value	0.158	0.273	0.283
AIC value	−130.9	60.2	68.3

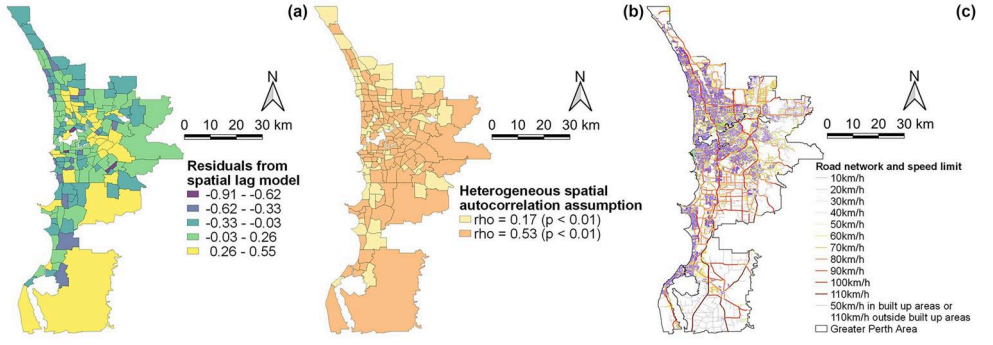

Figure 13. Heterogeneous spatial autocorrelation assumption for traffic crash in Perth. (a) Residuals from spatial lag model and (b) Spatial variation of autocorrelation. (c) Road network in the Greater Perth Area.

spatial autocorrelation strength in this transportation study, we further re-estimated regional traffic crashes using the SLM with a generalized SAR coefficient matrix. By re-ranking a series of regional traffic crash values following the order of SLM residuals, the change point detection algorithm returns potential change points that categorize the entire Greater Perth Area into two regions with different spatial autocorrelation strengths. Our preferred regional divisions, after comparing multiple changing points, for the heterogeneous spatial autocorrelation are shown in Figure 13(b). Our adjusted SLM shows that the majority of the SA2 units in Perth transmit comparatively strong spatial dependence, with rho values over 0.5, whereas spatial impacts from the rest of the region, located at the edge of the city, are weak, with rho values less than 0.2. Two different SAR coefficients, indicating the variation in the spatial dependence strength on traffic crashes in urban areas, were statistically significant at the level of at least 0.01. The entire road network of Perth is illustrated in Figure 13(c). Major roads with high speed limits are indicated by thick red lines. The variation in the spatial autocorrelation strength of traffic crashes was consistent with the distribution of urban

roads. The SA2 regions with high-level autocorrelation contain major roads or dense urban arterial roads. It is expected that car crashes are more likely to spill over along major road networks than across geographical neighbors. The heterogeneous spatial autocorrelation model demonstrates an expected improvement in goodness of fit by providing a deeper understanding of nonstationary spatial dependence effects. The heterogeneous SLM reduced the value of RMSE by 42% to less than 0.158 from both the traditional SLM and linear regression. Furthermore, the coefficients of the independent variables estimated by the adjusted SLM using MLE are more statistically significant than those of the previous models. All three models listed in Table 4 imply that road speed, regional traffic volume and the percentage of residents driving and walking to work were positively associated with regional traffic crashes.

5. Discussion

This study discusses the impact of heterogeneous autocorrelation within SAR processes and provides a strategy for analyzing the variation in spatial autocorrelation strength. To the best of our knowledge, this study is a pilot investigation that incorporates heterogeneity in SAR models through matrix representation. The Monte Carlo simulation indicates that the traditional SLM with a homogeneous autocorrelation assumption introduces additional uncertainties in the estimation of beta if there are variations in the strength of the spatial autocorrelation. Furthermore, the SAR coefficient estimated by the traditional SLM cannot represent the global average of the spatial autocorrelation strength. However, the geographical structure of the heterogeneous autocorrelation can be captured through statistical associations between the residuals and the dependent variable. The variation in the SAR coefficients can be effectively re-estimated using our adjusted SLM with a generalized SAR coefficient matrix. Our strategies for exploring spatial variations in autocorrelation were applied to a transport geography study and have shown significant improvements in the goodness of fit of the SAR models.

Our proposed method is robust for detecting geographical structures with variations in imposing and receiving spatial effects. This type of geographical structure shows the heterogeneity of spatial autocorrelation strength through directional variability, which is similar to the results found in ecosystems and landscapes (Liu *et al.* 2018). Our adjusted SLM assumes that the variation in spatial autocorrelation strength may not be linked to distance. The extension from a single SAR coefficient to a matrix suggests that spillover effects can be attributed to long-range anisotropic environmental and social interactions that do not strictly adhere to geographical proximity. At the cost of additional degrees of freedom, an SLM with an SAR coefficient matrix can also provide global statements for statistical associations.

Despite the effective evidence obtained from both Monte Carlo simulations and real-world transport geography data analyses, the proposed method has several limitations. The limitations of the methodology design are the coordination between the heterogeneous autocorrelation identification process and the heterogeneous autocorrelation re-estimation process, and the performance of change point detection in capturing specific patterns of heterogeneous autocorrelation.

First, our adjusted SLM requires an accurate identification of the heterogeneous autocorrelation pattern or the prior knowledge of spatial dependence structure. As shown by the Monte Carlo simulations, the SLM with a generalized SAR coefficient matrix has uncertainties in re-estimating the coefficients without a clear identification of the heterogeneous autocorrelation structure. Thus, we propose a change point detection algorithm as a candidate method for identifying heterogeneous autocorrelations. The accurate estimations on regression coefficients using the heterogeneous spatial autocorrelation model depend on the performance of change point detection, despite stable and robust model performance measured by MSE.

Second, the change point detection algorithm may demonstrate diminished sensitivity when identifying patterns using closely clustered SAR coefficients. Our algorithm is robust in recognizing geographical structures with variations in imposing and receiving spillover effects. These patterns include structures with a single unit and a spatial boundary where clustered regions with different SAR coefficients meet. However, this algorithm is not sensitive to detecting such a distribution of SAR coefficients in which similar values are tightly clustered. The algorithm can easily identify the boundary where two clusters intersect through a binary division but requires more breaking points and further complicated tests on merging homogeneous areas to capture the full internal structure (eg central points in Patterns (1) and (4), and the internal ring in Pattern (6)).

Third, the change point detection algorithm does not have specific criteria for determining the optimal breaking points, and requires multiple trials by adjusting the parameters to capture accurate heterogeneous spatial autocorrelation patterns. Binary heterogeneous spatial autocorrelation is the simplest form of variation in autocorrelation and can be easily captured by change point detection with the number of intervals starting with the smallest value. However, the real world can be more complex, and more SAR coefficients may be required to model variations in the spatial autocorrelation strength. With no prior knowledge of the spatial autocorrelation strength distribution, we must always experiment with the change point detection algorithm parameters. The optimal number of breaking points was determined by comparing the statistical significance of the estimated coefficients and the overall model performance using a generalized SAR coefficient matrix. It is also recommended that model performance be tested using different combinations of changing points based on different parameters because the real pattern may be disguised by other unknown causes. Further diagnostics should be developed to aid users in parameter and model selection.

6. Conclusion

SAR models with a homogeneous assumption of spatial autocorrelation remain prevalent in the modeling of SAR processes in various research fields. However, traditional SAR models estimate the beta values of independent variables and SAR coefficients with unreliability or uncertainty when there are variations in the strength of spatial autocorrelation. This study conducted a pilot investigation of the impact of heterogeneous spatial autocorrelation on traditional SAR models and proposed strategies to

extract geographical structures representing the variation in spatial autocorrelation strength through residual analysis, together with an adjusted SLM with a generalized SAR coefficient matrix to re-estimate the SAR processes.

A Monte Carlo simulation study showed that the adjusted SLM can precisely re-estimate SAR processes after the identification of heterogeneous spatial autocorrelation patterns. Our methods are particularly robust in recognizing geographical structures with variations in imposing and receiving spillover effects. Our adjusted SLM also has a better goodness of fit than the traditional SLM, given a real-world case study on transport geography. The performance of our adjusted SLM relies largely on the identification of heterogeneous spatial autocorrelation patterns. Thus, future research efforts are required to find better approaches to identify the variation in spatial autocorrelation strength with suitability for more complicated cases where more groups of SAR coefficients are significantly different.

Acknowledgement

We thank the editor and anonymous reviewers for their valuable comments and suggestions for improving the quality of this manuscript.

Disclosure statement

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

Notes on contributors

Zehua Zhang obtained his PhD degree with Chancellor's Commendation from Curtin University in 2023 and was awarded 2023 Humanities Greg Crombie Postgraduate Work of the Year Award. He is currently a researcher of Geospatial Intelligence Lab and working as a Sessional Academic at Curtin University. His research interests include spatial statistics, spatial data science, GIS and transport and population geography. Email: Zehua.Zhang@curtin.edu.au

Ziqi Li is an Assistant Professor in quantitative geography at Florida State University (FSU). His research focuses on the methodological development of spatially explicit and interpretable statistical/machine learning models to investigate human behavior across space and place. He is one of the primary developers of Multi-scale Geographically Weighted Regression (MGWR) and Python Spatial Analysis Library (PySAL). Email: Ziqi.Li@fsu.edu

Yongze Song is a Senior Lecturer at Curtin University. His research interests include spatial statistics, geospatial intelligence, sustainable infrastructure and sustainable development. He servers as an Associate Editor for the journals GIScience & Remote Sensing, and the International Journal of Applied Earth Observation and Geoinformation. Email: Yongze.Song@curtin.edu.au

ORCID

Zehua Zhang http://orcid.org/0000-0003-3462-4025
Ziqi Li http://orcid.org/0000-0002-6345-4347
Yongze Song http://orcid.org/0000-0003-3420-9622

Data and codes availability statement

Data and codes supporting the findings of this study are available at https://doi.org/10.6084/m9.figshare.25557009.v1

References

Ahmed, A., Sadullah, A.F.M., and Yahya, A.S., 2021. Analysis of the effect of directional traffic volume and mix on road traffic crashes at three-legged unsignalized intersections. *Transportation Engineering*, 3, 100052.

Anselin, L., 1988. *Spatial econometrics: methods and models*. Dordrecht, Netherlands: Kluwer Academic.

Anselin, L., 1995. Local indicators of spatial association – LISA. *Geographical Analysis*, 27 (2), 93–115.

Anselin, L., 2010. Thirty years of spatial econometrics. *Papers in Regional Science*, 89 (1), 3–26.

Anselin, L., and Griffith, D.A., 1988. Do spatial effects really matter in regression analysis? *Papers in Regional Science*, 65, 11–34.

Anselin, L., and Rey, S.J., 2010. *Perspectives on spatial data analysis*. Heidelberg, Germany: Springer.

Anselin, L., Syabri, I., and Kho, Y., 2010. GeoDa: an introduction to spatial data analysis. *In*: *Handbook of applied spatial analysis*. Berlin Heidelberg: Springer, 73–89.

Arbia, G., and Baltagi, B.H., 2009. *Spatial econometrics: methods and applications*. Heidelberg, Germany: Physica-Verlag.

Australian Bureau of Statistics, 2021a. Statistical area level 2. Australian Statistical Geography Standard (ASGS) Edition 3. Available from: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/statistical-area-level-2

Australian Bureau of Statistics, 2021b. Digital boundary files. Australian Statistical Geography Standard (ASGS) Edition 3. [Data set]. Available from: https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files

Australian Bureau of Statistics, 2022. Data by region methodology. [Data set]. Available from: https://www.abs.gov.au/methodologies/data-region-methodology/2011-22#data-downloads

Australian Bureau of Statistics, 2023. National, state and territory population. Available from: https://www.abs.gov.au/statistics/people/population/national-state-and-territory-population/mar-2023

Baltagi, B.H., Kelejian, H.H., and Prucha, I.R., 2007. Analysis of spatially dependent data. *Journal of Econometrics*, 140 (1), 1–4.

Brunsdon, C., Fotheringham, A.S., and Charlton, M., 1998. Spatial nonstationarity and autoregressive models. *Environment and Planning A: Economy and Space*, 30 (6), 957–973.

Di Stasi, L.L., *et al.*, 2022. The effect of traffic light spacing and signal congruency on drivers' responses at urban intersections. *Transportation Engineering*, 8, 100113.

Fischer, M.M., and Wang, J., 2011. *Spatial data analysis: models, methods and techniques*. Heidelberg, Germany: Springer.

Fotheringham, A.S., 2009. The problem of spatial autocorrelation and local spatial statistics. *Geographical Analysis*, 41 (4), 398–403.

Fouedjio, F., 2016. Second-order non-stationary modeling approaches for univariate geostatistical data. *Stochastic Environmental Research and Risk Assessment*, 31 (8), 1887–1906.

Fuentes, M., 2001. A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12 (5), 469–483.

Gao, C., *et al.*, 2020. Modeling urban growth using spatially heterogeneous cellular automata models: comparison of spatial lag, spatial error and GWR. *Computers, Environment and Urban Systems*, 81, 101459.

Geniaux, G., and Martinetti, D., 2018. A new method for dealing simultaneously with spatial autocorrelation and spatial heterogeneity in regression models. *Regional Science and Urban Economics*, 72, 74–85.

Goovaerts, P., 1997. *Geostatistics for natural resources evaluation*. Oxford, UK: Oxford University Press.

Guo, J., *et al.*, 2022. Modeling of spatial stratified heterogeneity. *GIScience & Remote Sensing*, 59 (1), 1660–1677.

Haas, T.C., 1990. Kriging and automated variogram modeling within a moving window. *Atmospheric Environment*, 24 (7), 1759–1769.

Harris, P., 2019. A simulation study on specifying a regression model for spatial data: choosing between autocorrelation and heterogeneity effects. *Geographical Analysis*, 51 (2), 151–181.

Harris, P., Charlton, M., and Fotheringham, A.S., 2010. Moving window kriging with geographically weighted variograms. *Stochastic Environmental Research and Risk Assessment*, 24 (8), 1193–1209.

Higdon, D., 1998. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5 (2), 173–190.

Higdon, D., Swall, J., and Kern, J., 1999. Non-stationary spatial modeling. In: Bayesian statistics, vol. 6. New York: Oxford University Press, 761–768.

Holland, D.M., *et al.*, 1999. Spatial prediction of sulfur dioxide in the Eastern United States. *In geoENV II—Geostatistics for environmental applications*. Netherlands: Springer, 65–76.

Hu, Y., and Wang, F., 2020. *GIS-based simulation and analysis of intra-urban commuting*. Boca Raton, FL: CRC Press.

Lambert, D.M., Brown, J.P., and Florax, R.J.G.M., 2010. A two-step estimator for a spatial lag model of counts: theory, small sample performance and an application. *Regional Science and Urban Economics*, 40 (4), 241–252.

Lindgren, F., Rue, H., and Lindström, J., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach: link between Gaussian Fields and Gaussian Markov Random Fields. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73 (4), 423–498.

Liu, Z., *et al.*, 2018. Spatial heterogeneity of leaf area index in a temperate old-growth forest: spatial autocorrelation dominates over biotic and abiotic factors. *The Science of the Total Environment*, 634, 287–295.

Main Roads Western Australia, 2018. Traffic signal sites. [Data set]. Available from: https://portal-mainroads.opendata.arcgis.com/datasets/mainroads::traffic-signal-sites/about

Main Roads Western Australia, 2022. Traffic Digest. [Data set]. Available from: https://portal-mainroads.opendata.arcgis.com/datasets/mainroads::traffic-digest/about

Main Roads Western Australia, 2023a. Crash Information (Last 5 Years). [Data set]. Available from: https://portal-mainroads.opendata.arcgis.com/datasets/mainroads::crash-information-last-5-years/about

Main Roads Western Australia, 2023b. Legal Speed Limits. [Data set]. Available from: https://portal-mainroads.opendata.arcgis.com/datasets/mainroads::legal-speed-limits/about

Mane, A.S., and Pulugurtha, S.S., 2018. Influence of on-network, traffic, signal, demographic, and land use characteristics by area type on red light violation crashes. *Accident; Analysis and Prevention*, 120, 101–113.

Mei, C.-L., and Chen, F., 2022. Detection of spatial heterogeneity based on spatial autoregressive varying coefficient models. *Spatial Statistics*, 51, 100666.

Paciorek, C.J., and Schervish, M.J., 2006. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17 (5), 483–506.

Rhee, K.-A., *et al.*, 2016. Spatial regression analysis of traffic crashes in Seoul. *Accident; Analysis and Prevention*, 91, 190–199.

Sampson, P., Damian, D., and Guttorp, P., 2001. Advances in modeling and inference for environmental processes with nonstationary spatial covariance. NRCSE-TRS No 61, National Research Centre for Statistics and the Environment Technical Report Series, 2001

Sampson, P., and Guttorp, P., 1992. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87 (417), 108–119.

Schabenberger, O., and Gotway, C., 2005. *Statistical methods for spatial data analysis*. London: Chapman & Hall.

Siuhi, S., Mamun, M.M.H., and Mwakalonge, J., 2021. The significance of the posted minimum speed limits along interstate highways in South Carolina on traffic operation and safety. *Journal of Traffic and Transportation Engineering (English Edition)*, 8 (5), 715–724.

Song, Y., *et al.*, 2020. An optimal parameters-based geographical detector model enhances geographic characteristics of explanatory variables for spatial heterogeneity analysis: cases with different types of spatial data. *GIScience & Remote Sensing*, 57 (5), 593–610.

Stein, A., Hoogerwerf, M., and Bouma, J., 1988. Use of soil map delineations to improve (co)kriging of point data on moisture deficits. *Geoderma*, 43 (2–3), 163–177.

Yin, C., *et al.*, 2018. Effects of urban form on the urban heat island effect based on spatial regression model. *The Science of the Total Environment*, 634, 696–704.

Zhang, Z., Song, Y., and Wu, P., 2022. Robust geographical detector. *International Journal of Applied Earth Observation and Geoinformation*, 109, 102782.

Zhang, Z., *et al.*, 2023. Geocomplexity explains spatial errors. *International Journal of Geographical Information Science*, 37 (7), 1449–1469.

Zhang, Z., *et al.*, 2024. Robust interaction detector: a case of road life expectancy analysis. *Spatial Statistics*, 59, 100814.

Zou, H., *et al.*, 2012. An improved distance metric for the interpolation of link-based traffic data using kriging: a case study of a large-scale urban road network. *International Journal of Geographical Information Science*, 26 (4), 667–689.