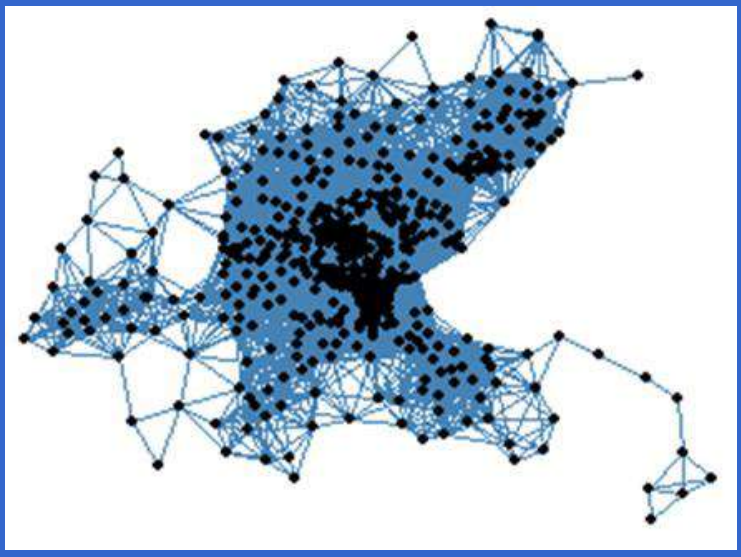
SPATIAL REGRESSION MODELS



Dr. Turgut ÖZALTINDİŞ

EDITOR

Assist. Prof. Dr. Elif Özge ÖZDAMAR

ISBN: 978-625-5753-18-2

Ankara -2025

SPATIAL REGRESSION MODELS

EDITOR

Assist. Prof. Dr. Elif Özge ÖZDAMAR
ORCID ID: 0000-0001-5652-1858

AUTHOR

Dr. Turgut ÖZALTINDİŞ

Mimar Sinan Fine Arts University, Statistics Department, Applied Statistics Department, İstanbul, Türkiye
turgut.ozaltindis@msgsu.edu.tr
ORCID ID: 0000-0002-7811-5428

DOI: https://doi.org/10.5281/zenodo.17698950

E mail: ubakyayinevi@gmail.com
www.ubakyayinevi.org
It is responsibility of the author to abide by the publishing ethics rules.
UBAK Publishing House – 2025©

ISBN: 978-625-5753-18-2
November / 2025
Ankara / Turkey

PREFACE

Spatial regression analysis has emerged as a fundamental subfield within spatial statistics and econometrics, addressing the limitations of classical regression models in the presence of spatial dependence. The recognition that spatially proximate observations frequently exhibit statistical interdependence has necessitated the development of models and methods capable of capturing such relationships. This monograph presents a systematic treatment of the theoretical foundations, methodological frameworks, and applied implementations of spatial regression techniques.

The work is organized to provide a comprehensive exposition, beginning with the conceptualization and formalization of spatial neighborhood structures and spatial weight matrices, which constitute the basis for any spatial analysis. Subsequently, measures of spatial autocorrelation, both global and local, are examined in depth, including their statistical properties, interpretation, and practical considerations. The core chapters focus on the formulation, estimation, and evaluation of major spatial regression models, including the Spatial Autoregressive Model (SAR), the Spatial Error Model (SEM), the Spatial Durbin Model (SDM), the Spatial Autocorrelation Model (SAC), the Spatially Lagged X Model (SLX), and the General Nesting Spatial Model (GNS). Each model is presented with its mathematical derivation, underlying assumptions, estimation procedures, and potential advantages and limitations.

The volume also addresses model specification and diagnostic testing, recognizing that rigorous model selection is essential for valid inference in spatial econometric applications. The final sections demonstrate the practical implementation of spatial regression models using an empirical dataset, thereby bridging theoretical discussion with empirical practice and providing a reproducible analytical framework.

This book is intended as a reference for researchers, graduate students, and practitioners across disciplines such as economics, geography, urban and regional planning, environmental sciences, epidemiology, and political science, fields in which spatial dependence is an intrinsic characteristic of the data. It aims to equip the reader with both the theoretical insights and the applied skills required to model spatial processes accurately and to interpret the resulting analyses with methodological rigor.

By integrating theoretical constructs with empirical applications, this work aspires to contribute to the methodological advancement of spatial analysis and to facilitate the adoption of spatial regression techniques in addressing complex, spatially structured phenomena.

24.11.2025

Assist. Prof. Dr. Elif Özge ÖZDAMAR

TABLE OF CONTENTS

SPATIAL REGRESSION MODELS

INTRODUCTION

Spatial statistics is becoming increasingly central in modern data analysis and plays a fundamental role in critical decision-making across numerous disciplines. The methodology examines spatiality through topological and geometric characteristics within a dataset alongside additional variables, is employed across various domains including urban planning, agriculture, environmental management, logistics, and public health. Spatial statistics analyses integrate locational data, cartographic layers, and analytical techniques to identify patterns, correlations, and trends within the examined areas. One of the most powerful and comprehensive tools in this field, spatial regression analysis not only considers spatial structures but also represents an approach that questions and redefines the fundamental assumptions of classical regression models. The emergence of spatial regression analysis began with the recognition of spatial dependence. In 1948, Patrick A. Moran pioneered the field by developing Moran's I index, which defined the concept of spatial autocorrelation. Subsequently, Charles Geary (1954) proposed the Geary's C ratio as an alternative to Moran's approach. These metrics enabled the mathematical expression of similarity among spatial units and filled a significant gap in literature.

In the 1970s, Cliff and Ord made significant contributions regarding spatial weight matrices and spatial relationship structures, thus leading the institutionalization of spatial statistics. Luc Anselin's work, *Spatial*

Econometrics, published in 1988, systematized spatial regression models and established spatial econometrics as a distinct sub-discipline. With the development of local statistics by Getis and Ord in the 1990s, it became possible to analyze spatial clustering not only at the global but also at the local level.

This evolving literature has led to diversification in spatial regression models. The SAR (Spatial Autoregressive) model, which incorporates interactions between the dependent variable and neighboring units into the model; the SEM (Spatial Error Model), which addresses spatial dependence in the error structure; the SDM and SLX models, which include spatial effects of both dependent and independent variables; and the more comprehensive structures of SAC and GNS models allow for the representation of spatial analyses with varying structures.

The use of spatial regression analysis is not confined to theoretical academic discussions but also finds extensive practical application. In economics, analyses of regional development, income inequality, and housing prices; in health sciences, analyses of disease clustering, access to healthcare, and environmental health risks; in urban planning, studies on population density, land use, and distribution of urban services; in environmental sciences, research on air and water pollution, land degradation, and other spatially sensitive indicators; in agriculture and natural resource management, assessments of crop productivity, soil quality, and the effects of climate change; and in political analyses, evaluations of vote distributions, political trend maps, and regional preference patterns are just a few of the many applications.

This book addresses spatial regression analyses in both theoretical and applied contexts across these diverse domains. It is structured around four main themes. The first chapter introduces the concept of spatial neighborhood, the foundational element of spatial analysis, and explains the construction of weight matrices with different neighborhood structures and weighting methods. The second chapter elaborates on spatial autocorrelation techniques, identifying spatial patterns through global metrics such as Moran's I, Geary's C, and Getis-Ord G, and local metrics such as LISA, Local Geary, and Gi* statistics. The third chapter presents the core of the book spatial regression models. Models such as the Spatial Autoregressive Model (SAR), which directly incorporates the interaction of the dependent variable with neighboring units, the Spatial Error Model (SEM), which considers spatial dependence in the error structure, and advanced models such as SDM, SLX, SAC, and GNS are discussed in detail, including their theoretical background, assumptions, advantages, disadvantages, and estimation techniques. Each of these models offers different advantages depending on the spatial structure and characteristics of the dataset. The fourth chapter explains how to select the most appropriate spatial model for a given dataset using model specification tests, including Lagrange Multiplier tests and other diagnostic tools. The fifth and sixth chapters present the analysis of theoretical information using applied datasets and provide interpretations of the findings. In conclusion, the book evaluates the opportunities offered by spatial regression analysis and provides guidance for future research.

1. SPATIAL NEIGHBORHOOD AND SPATIAL WEIGHT MATRIX

The foundation of spatial data analysis lies in the neighborhood structure of observations within the dataset and the weight matrix that reflects the degree of such neighborhood. Neighborhood refers to the concept that defines the spatial proximity between observations. In spatial data analysis (SDA), neighborhood refers to associating observations that are positioned near or adjacent to a specific area (point, region, polygon, etc.). It can be defined using criteria such as two observations being adjacent in space (contiguous) or located within a specific distance threshold.

A spatial weight matrix for a dataset is represented by a square matrix of size $N \times N$, denoted by W, where each element expresses whether a unit (observation or region) has a neighborhood relationship with another, and to what degree.

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ w_{N1} & w_{N1} & \cdots & w_{NN} \end{bmatrix}$$

Given two observations *i* and *j*, W_{ij} expresses the neighborhood between observation *i* and observation *j*. If this value is 1, the observations are neighbors; if 0, they are not. For polygon-based spatial data, neighborhood is defined through contiguity; for point data, it is defined through a distance threshold. Researchers may also employ other types of neighborhood definitions. In the literature, commonly used

boundary-based neighborhood structures are inspired by chess piece movements and are named as rook, bishop, and queen contiguities. Figure 1.1 illustrates these neighborhood structures. On a map grid, areas that share an edge with the above, below, left, or right neighbors represent rook contiguity (edge-based); areas that touch diagonally at the corners represent bishop contiguity (vertex-based); and areas adjacent in both edge and corner directions form queen contiguity (combined). Distance-based neighborhood, on the other hand, is determined by a radius defined by the researcher or through a *k-nearest neighbors* algorithm.

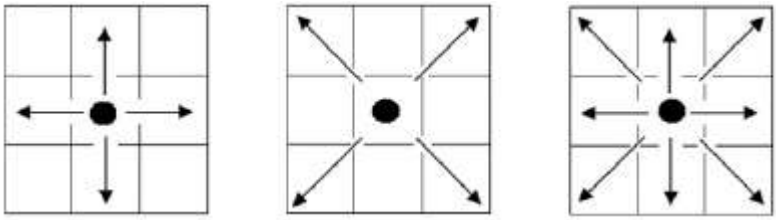


Figure 1.1 Rook, Bishop, and Queen Contiguities

After defining spatial neighborhood, the spatial weight matrix must be constructed. The weight matrix (W) is used to grade the defined neighborhood. Determining the weight matrix is a critical step in spatial analysis, as it can significantly impact the statistical test results derived from the analysis (Tiefelsdorf et al., 1999).

Since the pioneering studies of Moran (1948), Geary (1954), Cliff and Ord (1973 and 1981), determining the spatial weight matrix has been considered a complex and debated issue (A. C. Cliff & Ord, 1973; A. D. Cliff & Ord, 1981; Geary, 1954; P. A. Moran, 1948). Nevertheless,

a shared understanding from these works is that spatial weights should reflect accessibility between observations. Furthermore, spatial weights are expected to decrease as the distance between observations increases and to increase proportionally as the shared boundary lengths increase for adjacent units. Although there is consensus on these points, the literature has not converged on a standardized approach. It is recommended that researchers construct a spatial weight matrix that best reflects the spatial characteristics of the dataset being used (A. C. Cliff & Ord, 1973).

Getis, one of the leading figures in the field, proposed three perspectives for constructing spatial weight matrices: theoretical, topological, and empirical (Getis, 2009):

- **Theoretical Perspective:** Uses one of the general approaches established in literature. Approaches based on spatial distance are usually preferred. An example is a function where weight decreases with increasing distance. The challenge here is that such weightings may not always be suitable for real-world conditions.
- **Topological Perspective:** Arises from the need to realistically define the physical properties of spatial units within a study area. In standard weight matrices, all adjacent observations are weighted equally without regard to topological differences. Thus, observations with different spatial structures are represented in the same way. In this approach, topological features are reflected in the weight matrix. For instance, the ratio

of the shared edge length of adjacent regions to their area can be used.

- **Empirical Perspective:** According to Cliff and Ord, this is the most consistent approach. They noted that "the researcher can highlight the spatial features they consider important using a flexible weighting system" (A. D. Cliff & Ord, 1969). Here, the researcher builds a weight matrix that best represents the spatial relationship structure.

Based on a review of the literature, the principal methods for constructing spatial weight matrices include:

- Contiguity-based weighting
- Distance-based weighting
- Inverse-distance weighting
- K-nearest neighbors weighting
- Shared boundary length weighting

The spatial weight matrix proposed by Getis and Aldstadt, based on the local Getis-Ord Gi statistics, differs from other approaches. In this method, not only the neighborhood status but also whether neighbors exhibit similar characteristics is considered during matrix construction (Getis & Aldstadt, 2004).

Despite the diversity in the literature, general practice tends toward using distance-based neighborhoods for point data, boundary contiguity for polygon data, and rook or queen contiguity for raster (grid) data.

1.1 Contiguity-Based Weighting

The weight matrix constructed based on contiguity is fundamentally binary in structure, as it solely accounts for the condition of adjacency. If two observations share a common border, the matrix assigns a value of 1; otherwise, it assigns a value of 0. The weight matrix based on contiguity can be expressed as follows:

$$w_{ij} = \begin{cases} 1, & \text{if unit } i \text{ and } j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases} \quad (1.1)$$

The W_{ij} matrix is symmetric, and the diagonal elements are zero. When necessary, it can be row-standardized such that the row sums equal 1, using the following expression:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_{j \in C} w_{ij}} \sum_{j \in C} w_{ij}^{*} = 1 \quad (1.2)$$

In Equation (1.2), C represents the set of observations that are contiguous (i.e., neighbors) to observation *i*.

1.2 Distance-Based Weighting

Distance-based weighting is determined according to the distance between observations. Observations that fall within a distance threshold specified by the researcher are considered neighbors. Observations identified as neighbors are assigned a weight of 1 in the matrix, while non-neighbors receive a weight of 0. The matrix is defined as follows:

$$w_{ij} = \begin{cases} 1, & d_{ij} \leq \delta \\ 0, & \text{otherwise} \end{cases} \quad (1.3)$$

Here, d_{ij} is the distance between observations *i* and *j*, and δ is the threshold distance defined to determine neighborhood. The choice of distance metric may vary. This approach is commonly used with point data. For polygon data, weighting can be applied based on the distance between centroid points instead of using contiguity. As with the previous approach, standardization using Equation (1.2) may be applied when necessary.

1.3 Inverse Distance-Based Weighting

Inverse distance weighting is based on Tobler's First Law of Geography, which states that "everything is related to everything else, but near things are more related than distant things." When the W matrix is constructed using this method, the weights decrease as the distance between observations increases. The weights are calculated using the following equation:

$$w_{ij} = \frac{1}{d_{ij}^{p}} \tag{1.4}$$

Here, d_{ij} is the distance between observations i and j, and p is a power parameter determined by the researcher. Increasing the power parameter p increases the standardized weights for closer observations while decreasing those for distant ones.

1.4 K-Nearest Neighbors-Based Weighting

Another binary structure is the k-nearest neighbors weighting matrix. In this approach, each observation defines as neighbors the *k* closest observations to itself. The weighting matrix is obtained as follows:

$$w_{ij} = \begin{cases} 1, & \text{if } j \text{ is among the } k \text{ nearest neighbors of } i \\ 0, & \text{otherwise} \end{cases} \quad (1.5)$$

The resulting matrix may not be symmetric. The number of neighbors k is the sole parameter and must be specified by the researcher. An advantage of this method is that every observation is guaranteed to have at least one neighbor, regardless of the dataset structure.

1.5 Shared Boundary Length-Based Weighting

This approach considers the topographical characteristics of the regions and is a refined version of boundary-based contiguity weighting matrices. Various methods exist for constructing this type of matrix, with one of the most used being the generalized weight matrix, which accounts for both the shared boundary length and the distance between the centroids of the observations. The matrix is calculated as:

$$w_{ij} = \frac{1}{d_{ij} \sum_{j \in C} l_{ij}} \quad (1.6)$$

Here, C represents the set of observations neighboring observation i, l_{ij} denotes the length of the shared boundary between observations i and j, and d_{ij} is the distance between their centroids.

2. MEASURES OF SPATIAL AUTOCORRELATION

In natural systems, the random distribution of observations across space is rarely observed. As Tobler stated, all things are related, but near things are more related than distant things. Therefore, observations that are geographically close to each other are expected to exhibit higher

similarity compared to those that are far apart. At this point, spatial autocorrelation emerges as the statistical measure that reveals such relationship structures among observations. In its simplest form, spatial autocorrelation refers to the investigation of whether spatially adjacent (neighboring) observations on a map exhibit similarity in terms of a given variable. Similar values of nearby observations indicate positive spatial autocorrelation, whereas different values indicate negative spatial autocorrelation. In practice, negative spatial autocorrelation is rarely observed. Figure 2.1 illustrates structures of positive and negative autocorrelations respectively.

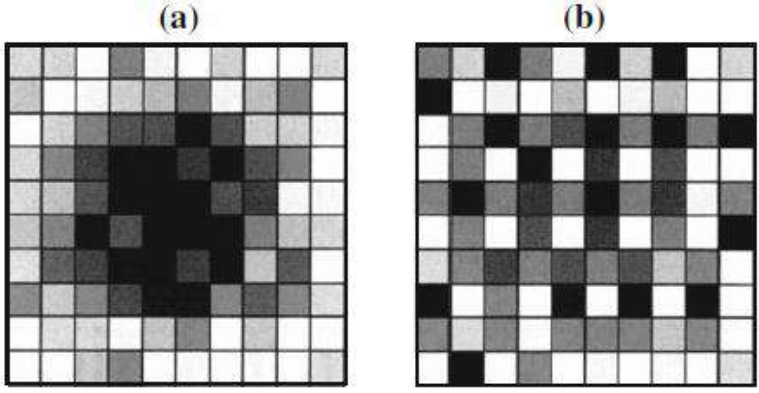


Figure 2.1: Positive spatial autocorrelation (a), Negative spatial autocorrelation (b)

The concept of spatial autocorrelation was first introduced in the late 1950s by geographer Michael F. Dacey. His efforts to develop thisstatistics and received significant support from W. L. Garrison and Edward Ullman. German economic geographer Walter Christaller's works was known to have influenced all three of these geographers

(Çubukçu, 2015). *Although the term “autocorrelation” did not yet exist in the literature prior to Dacey studies had already acknowledged that nearby spatial units tend to be similar and interact strongly, while distant units exhibit relatively weak interactions (Ravenstein, 1885; von Thünen, 1826; Zipf, 1949). The development of the concept was later best summarized by Tobler’s First Law of Geography. Although the term spatial autocorrelation was first formally introduced by Garrison in the late 1960s, its theoretical foundation had already been established in 1948 when Patrick Alfred Pierce Moran published his calculation. In 1954, Robert Charles Geary proposed an alternative approach for calculating spatial autocorrelation (Geary, 1954). Since the term autocorrelation had not yet been coined, both researchers referred to their measures as contiguity ratios. These two approaches are now widely accepted and frequently used in the literature. Up until 1964, the concept of spatial autocorrelation was referred to in the literature by various other names such as spatial dependence, spatial association, or spatial interaction. More recently, the studies by Getis and Ord have further popularized the concept (L. Anselin & Getis, 1992).

Spatial autocorrelation is generally analyzed in two ways: global and local. Global spatial autocorrelation evaluates the structure of spatial relationships across the entire study area using a single summary statistic. It is used to provide an overall view and to identify spatial patterns. Local spatial autocorrelation, on the other hand, calculates a separate statistic for each observation, revealing localized patterns and clusters.

While global measures determine whether spatial clustering exists, local methods also identify where such clustering occurs. Several techniques exist in the literature to measure both global and local spatial autocorrelation. These measures are detailed in the following sections.

2.1 Global Measures of Spatial Autocorrelation

Global spatial autocorrelation measures determine whether neighboring observations exhibit similar characteristics. The most well-known global measures for spatial autocorrelation, are Moran's I Index and Geary's C Ratio, introduced by Austrian statistician Patrick Alfred Pierce Moran and Irish statistician Robert Charles Geary respectively. Even though developed in the 1950s, these statistics continue to be widely accepted by practitioners today. Apart from the statistics mentioned, there is another global spatial autocorrelation measure that is named as Getis-Ord General G. Compared to the other two measures, Getis-Ord General G is used less frequently.

2.1.1 Moran's I Index

Moran's I Index was introduced by Moran in 1950, and it is the most widely used measure of global spatial autocorrelation in the literature. Moran's I Index simultaneously considers both the spatial proximity of observations and the values of the variable under study (P. A. P. Moran, 1950). It is calculated as in Equation (2.1):

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad (2.1)$$

In this equation, *n* represents the number of observations in the sample;

x_i and x_j, the values of the i-th and j-th observations respectively; $\bar{x}$, the overall-mean; w_{ij}, the spatial weight between observations i and j, representing the degree of spatial proximity, and S_0 ($S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$) denotes the sum of all spatial weights in the weight matrix.

The I Index takes values in the range of -1 to 1. A positive value indicates positive spatial autocorrelation, meaning spatially adjacent observations tend to have similar values and generating spatial clusters. Conversely a negative value indicates negative spatial autocorrelation, meaning that spatially close observations have dissimilar values, so that no spatial clustering is present.

The closer the index approaches to 1, the stronger the spatial relationship becomes. If the I Index is close or equal to 0, it indicates that the observations are randomly distributed, with a conclusion there is no spatial clustering. To test whether the spatial autocorrelation is statistically significant, the normal distribution is used. The z-values of the calculated I indices are obtained using Equation (2.2).

$$z_I = \frac{I - E(I)}{\sqrt{V(I)}} \qquad (2.2)$$

The expected value and variance of the I Index are calculated as follows:

$$E(I) = -1/n - 1 \ and \qquad (2.3)$$

$$V(I) = E(I^2) - E(I)^2 = \frac{nP_1 - P_1P_2}{(n-1)(n-2)(n-3)W} - E(I)^2 \qquad (2.4)$$

In the literature, Equation (2.4) is expressed in various forms. Here, it has been reformulated and presented in a simplified manner by decomposing it into the terms P_1, P_2 and P_3, where they are denoted as follows respectively:

$$P_1 = (n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2 \tag{2.5}$$

$$P_2 = \frac{m_4}{m_2^2} \tag{2.6}$$

$$P_3 = (n^2 - n)S_1 - 2nS_2 + 6S_0^2 \tag{2.7}$$

Due to simplification, the term S emerges in the equations, and it is calculated according to whether the weight matrix is symmetric or asymmetric. In practice, weight matrix w_{ij} is usually taken as symmetric ($w_{ij} = w_{ji}$), in order to simplify computation.

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij} \tag{2.8}$$

S_0 represents the total sum of all spatial weights in the spatial weight matrix W, providing a measure of the overall spatial connectivity in the dataset.

$$S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(w_{ij} + w_{ji}\right)^2 \tag{2.9}$$

S_1 shows the symmetry and magnitude of spatial relationships between observations, indicating bidirectional spatial interactions. Taking the weight matrix as symmetric, the computation of S_1

simplifies. It reflects the second-order moment of the spatial weight matrix

$$S_2 = \sum_{i=1}^{n}\left(\sum_{j=1}^{n} w_{ij} + \sum_{j=1}^{n} w_{ji}\right)^2 \qquad (2.10)$$

S2 captures the variability in the row and column sums of the weight matrix, reflecting heterogeneity in spatial connectivity per location. Under the assumption that the data follow a normal distribution, the variance of the I Index can also be calculated as below:

$$V(I) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{S_0^2(n^2-1)} - E(I)^2 \qquad (2.11)$$

However, in practice, the variance formula (2.4) is preferred for more realistic results.

2.1.2 Geary's C Ratio

Geary's C Ratio is considered an alternative to Moran's I Index and is one of the most frequently used measures in calculating global spatial autocorrelation. It was introduced by Charles Geary in 1954 (Geary, 1954). Analogous to Moran's I, the Geary's C statistics incorporates both spatial relationships and attribute values and is formally computed as shown in Equation (2.12).

$$C = \frac{n-1}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - x_j\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (2.12)$$

In this equation, n represents the number of observations in the sample; x_i and x_j, the values of the i-th and j-th observations respectively; $\bar{x}$, the overall mean; and w_{ij} denotes the spatial weight between the two

observations. The C Ratio is always positive and takes a value between 0 and 2. Its interpretation differs from that of Moran's I Index. While a higher Moran's I value indicates positive spatial autocorrelation, this interpretation does not apply for Geary's C Ratio. When the calculated C value is less than 1, it indicates the presence of positive spatial autocorrelation, whereas a value greater than 1 indicates negative spatial autocorrelation. In cases where the C value is equal to or approximately 1, the spatial distribution of observations is random, implying no spatial clustering. Like as Moran's I Index, the statistical significance of the calculated spatial autocorrelation is tested under the assumption of normal distribution. The expected value and variance of Geary's C Ratio are calculated as follows:

$$E(C) = 1 \; and \tag{2.13}$$

$$V(C) = \frac{T_1+T_2+T_3}{4n(n-2)(n-3)S_4} \tag{2.14}$$

In Equation (2.14), the formula is simplified by breaking it down into parts named as T_1, T_2, and T_3. These parts are calculated with the usage of the terms P_1, P_2 and P_3, with referring to the Equations (2.5), (2.6), and (2.7) respectively, which used in the calculation of Moran's I Index.

$$T_1 = 4(n-1)S_1[n^2 - 3n + 3 - (n-1)P_2] \tag{2.15}$$

$$T_2 = -(n-1)S_2[n^2 - 3n - 6 - (n^2 - n + 1)P_2] \tag{2.16}$$

$$T_3 = 4S_0^2[n^2 - 3 - (n-1)^2P_2] \tag{2.17}$$

Under the assumption that the data follow a normal distribution, the variance of the C Ratio can also be calculated as follows:

$$V(C) = \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_0^2} \quad (2.18)$$

Here, the terms S_0, S_1 and S_2 used here correspond to those defined in Equations (2.8), (2.9), and (2.10) in the section on Moran's I Index. Comparing their usage in the literature, Geary's C Ratio is employed less frequently than Moran's I Index.

Among the global spatial autocorrelation measures, the Getis-Ord General G statistic offers a distinct analytical perspective. Unlike Moran's I and Geary's C, the Getis-Ord G specifically targets the spatial clustering of particularly high or low values. Although it is less widely utilized in comparison to Moran's and Geary's measures, it is conceptually regarded as an extension of the local Getis-Ord Gi statistics. By omitting the terms $d()$ and $w_{ij}(d)$ from the local formulation, the global version is derived as follows (Bivand & Wong, 2018):

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \forall j \neq i \quad (2.19)$$

The expected value and variance of the Getis-Ord General G statistic are calculated as follows:

$$E(G) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)} = \frac{S_0}{n(n-1)} \forall j \neq i \quad (2.20)$$

$$V(G) = E(G^2) - E(G)^2 \quad (2.21)$$

$$E(G^2) = \frac{A+B}{C} \quad (2.22)$$

The terms used in these calculations are defined below:

$$A = D_0\left(\sum_{i=1}^{n} x_i^2\right)^2 + D_1\sum_{i=1}^{n} x_i^4 + D_2\left(\sum_{i=1}^{n} x_i\right)^2\sum_{i=1}^{n} x_i^2 \qquad (2.23)$$

$$B = D_3\left(\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i^3 + D_4\left(\sum_{i=1}^{n} x_i\right)^4 \qquad (2.24)$$

$$C = n(n-1)(n-2)(n-3)\left[\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2\right]^2 \qquad (2.25)$$

The terms used in the calculations of A, B, and C are as follows:

$$D_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2 \qquad (2.26)$$

$$D_1 = -[n(n-1)S_1 - 2nS_2 + 6S_0^2] \qquad (2.27)$$

$$D_2 = -[2nS_1 - (n+3)S_2 + 6S_0^2] \qquad (2.28)$$

$$D_3 = 4(n-1)S_1 - (n+1)S_2 + 8S_0^2 \qquad (2.29)$$

$$D_4 = S_1 - S_2 + S_0^2 \qquad (2.30)$$

The values S_0, S_1, and S_2 used above were defined in previous measures. To eliminate the condition $\forall j \neq i$, the Getis-Ord General G statistic can be redefined as follows:

$$G = \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j\right) - \left(\sum_{i=1}^{n} w_{ii}x_i^2\right)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right) - \sum_{i=1}^{n} x_i^2} \qquad (2.31)$$

Unlike Moran's I Index and Geary's C Ratio, the Getis-Ord General G statistic is not bounded within a predefined numerical range. Therefore, to enable meaningful interpretation, a standardization process is

required. In general, larger values of the statistic suggest a stronger degree of spatial association.

2.1.3 Getis-Ord General G

Among the global spatial autocorrelation measures, the Getis-Ord General G statistic offers a distinct analytical perspective. Unlike Moran's I and Geary's C, the Getis-Ord G specifically targets the spatial clustering of particularly high or low values. Although it is less widely utilized in comparison to Moran's and Geary's measures, it is conceptually regarded as an extension of the local Getis-Ord Gi statistic. By omitting the terms $d()$ and $w_{ij}(d)$ from the local formulation, the global version is derived as follows (Bivand & Wong, 2018):

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \forall j \neq i \qquad (2.32)$$

The expected value and variance of the Getis-Ord General G statistic are calculated as follows:

$$E(G) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)} = \frac{S_0}{n(n-1)} \forall j \neq i \qquad (2.33)$$

$$V(G) = E(G^2) - E(G)^2 \qquad (2.34)$$

$$E(G^2) = \frac{A+B}{C} \qquad (2.35)$$

The terms used in these calculations are defined below:

$$A = D_0\left(\sum_{i=1}^{n} x_i^2\right)^2 + D_1\sum_{i=1}^{n} x_i^4 + D_2\left(\sum_{i=1}^{n} x_i\right)^2\sum_{i=1}^{n} x_i^2 \qquad (2.36)$$

$$B = D_3\left(\sum_{i=1}^{n} x_i\right)\sum_{i=1}^{n} x_i^3 + D_4\left(\sum_{i=1}^{n} x_i\right)^4 \qquad (2.37)$$

$$C = n(n-1)(n-2)(n-3)\left[\left(\sum_{i=1}^{n} x_i\right)^2 - \sum_{i=1}^{n} x_i^2\right]^2 \qquad (2.38)$$

The terms used in the calculations of A, B, and C are as follows:

$$D_0 = (n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2 \qquad (2.39)$$

$$D_1 = -[n(n-1)S_1 - 2nS_2 + 6S_0^2] \qquad (2.40)$$

$$D_2 = -[2nS_1 - (n+3)S_2 + 6S_0^2] \qquad (2.41)$$

$$D_3 = 4(n-1)S_1 - (n+1)S_2 + 8S_0^2 \qquad (2.42)$$

$$D_4 = S_1 - S_2 + S_0^2 \qquad (2.43)$$

The values S_0, S_1, and S_2 used above were defined in previous measures. To eliminate the condition $\forall j \neq i$, the Getis-Ord General G statistic can be redefined as follows:

$$G = \frac{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j\right) - \left(\sum_{i=1}^{n} w_{ii}x_i^2\right)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j\right) - \sum_{i=1}^{n} x_i^2} \qquad (2.44)$$

Unlike Moran's I Index and Geary's C Ratio, the Getis-Ord General G statistic is not bounded within a predefined numerical range. Therefore, to enable meaningful interpretation, a standardization process is

required. In general, larger values of the statistic suggest a stronger degree of spatial association.

2.2 Local Spatial Autocorrelation Measures

Global spatial autocorrelation measures whether there is an overall spatial relation within a dataset. Specifically, they determine whether geographically proximate observations tend to exhibit similar values. These measures produce a single summary statistic, which allows for general inferences about the presence of spatial clustering. However, such global indicators do not provide information about the specific locations or spatial units involved in the clustering. To address this limitation, local spatial autocorrelation measures have been developed to identify the spatial extent and intensity of clustering more precisely. These measures assign an individual association value to each spatial unit, thereby enabling the identification of localized clusters. Among the various techniques introduced in the literature, the most widely used are Local Moran's I, Local Geary's C, and Local Getis-Ord G.

The moments of Local Moran's I and Local Geary's C statistics often violate the assumption of normality. Therefore, standard hypothesis testing based on p-values, as used in global spatial autocorrelation analysis, may not be appropriate in this context. As a result, a spatial unit may appear significantly different from its neighbors purely by chance, which can lead to statistically unreliable test outcomes. To overcome this issue, a permutation-based measure is employed to generate a pseudo p-value. This approach involves repeatedly permuting the values of neighboring observations for a selected spatial

unit, where each permutation yields a new test statistic and its associated p-value. The pseudo p-value is then calculated by comparing the distribution of the permuted values with the observed value.

$$p = \frac{L + 1}{M + 1} \tag{2.45}$$

Here, L denotes the number of permuted statistics that are less than or equal to the observed statistic, and M represents the total number of permutations. The pseudo p-value serves as a robust alternative when the assumption of normality is questionable (Emrehan, 2022).

2.2.1 Local Moran's I

Local Moran's I is one of the most extensively utilized statistics for assessing local spatial autocorrelation. It represents the localized extension of the global Moran's I index and was introduced into the literature by Belgian statistician and economist Luc Anselin in 1995 (Luc Anselin, 1995). It is also commonly referred to as LISA (Local Indicators of Spatial Association). This statistic not only detects the spatial locations of clusters in datasets exhibiting spatial dependence, but it also identifies observations that significantly deviate from their spatial neighbors. Consequently, it serves as a valuable tool for detecting spatial outliers (Çubukçu, 2015). It is computed for each observation in the sample using the following expression:

$$I_i = z_i \sum_{j=1}^{n} w_{ij} z_j \tag{2.46}$$

In this formulation, n denotes the number of observations in the sample, w_{ij} represents the spatial weight between observations i and j, and z_i and z_j are the standardized values associated with these observations. Since the I_i statistic does not fall within a fixed numerical range, it is interpreted in a relative manner, rather than absolute. High positive values of I_i indicate that both the observation and its neighbors possess similarly high values. Conversely, negative values of I_i, deviating significantly from zero, indicate that the observation's value substantially differs from those of its neighbors, suggesting spatial dissimilarity or outlier status.

To evaluate the statistical significance of the detected spatial association, the expected value of I_i is calculated as follows:

$$E(I_i) = \frac{-w_i}{(n-1)} \tag{2.47}$$

Where w_i represents the total spatial weight of the neighbors of observation i, and is given by:

$$w_i = \sum_{j=1}^{n} w_{ij}\, j \neq i \tag{2.48}$$

The variance of the Local Moran's I statistic is computed using the following expression:

$$V(I_i) = \frac{w_{i(2)}(n-b_2)}{(n-1)} - \frac{2w_{i(kh)}(2b_2-n)}{(n-1)(n-2)} - \frac{w_i^2}{(n-1))^2} \tag{2.49}$$

Where the components are defined as:

$$w_{i(2)} = \sum_{j=1}^{n} \left(w_{ij}^2\right) j \neq i \tag{2.50}$$

$$b_2 = \frac{n \sum_{i=1}^{n} z_i^4}{(\sum_{i=1}^{n} z_i^2)^2} \tag{2.51}$$

$$2w_{i(kh)} = \sum_{k=1}^{n} \sum_{h=1}^{n} (w_{ik} w_{ih})\, k, h \neq i \tag{2.52}$$

Based on the resulting standardized z-score, the statistical significance of spatial association is tested at the predefined confidence level α.

2.2.2 Local Geary's C

The Local Geary's C statistic constitutes another widely used measure for evaluating local spatial autocorrelation. Like other approaches, it is derived from its global counterpart and is employed to determine the spatial location of clusters (Luc Anselin, 1995). The statistic is calculated using the following expression:

$$C_i = \frac{1}{m_2} \sum_{j=1}^{n} w_{ij} \left(z_i - z_j\right)^2 \tag{2.53}$$

In this formulation, n denotes the total number of observations in the sample, w_{ij} represents the spatial weight between observations i and j, z_i and z_j are their standardized values, and m_2 corresponds to the second-order moment. The expected value and variance of the C_i statistic are defined as:

$$E(C_i) = \frac{2nw_i}{n-1} \tag{2.54}$$

$$V(C_i) = \left(\frac{n}{n-1}\right)\left(w_i^2 + w_{i(2)}\right)(3 + b_2) - E(C_i)^2 \qquad (2.55)$$

The terms w_i, $w_{i(2)}$, and b_2 are identical to those introduced in the section on the Local Moran's I statistic. Owing to the lack of a clearly defined distribution, the Local Geary's C statistic is considerably less prevalent in the spatial autocorrelation literature compared to alternative local measures.

2.2.3 Local Getis-Ord G_i

The Local Getis-Ord G_i statistic is the most widely used technique for local spatial autocorrelation. It was developed by geographer Arthur Getis and statistician J.K. Ord (Getis & Ord, 1992). Similar to other measures, it assesses whether observations exhibit similar values based on a given variable, thereby identifying potential clustering. This statistic is commonly employed in the mapping of hot and cold spots. Hot spots indicate areas where observations and their neighbors exhibit similarly high values, while cold spots indicate clusters of similarly low values.

In their initial work, Getis and Ord calculated local statistics using a binary spatial weight matrix based on distance. Accordingly, local statistics for each observation are calculated using the following formula:

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j}{\sum_{j=1}^{n} x_j} i \neq j \qquad (2.56)$$

Here, n denotes the number of observations, $w_{ij}(d)$ represents the spatial weight between observations i and j based on a threshold distance d, and x_j indicates the observed value for unit j. The parameter d defines the neighborhood structure within the spatial weight matrix. When an inverse-distance weighting scheme is used, which is common in applied studies, the binary neighborhood structure is no longer preserved. In such cases, the weight of an observation with itself becomes equal to one.

To allow for analysis with non-binary weight structures, the G_i statistic is reformulated as follows:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{x}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{\left(n\sum_{j=1}^{n} w_{ij}^2\right) - \left(\sum_{j=1}^{n} w_{ij}\right)^2}{n-1}}} \qquad (2.57)$$

In this equation, x_j is the observed value of unit j, w_{ij} is the spatial weight between units i and j, $\bar{x}$ denotes the global mean, and S represents the standard deviation, which is calculated as:

$$S = \sqrt{\frac{\left(\sum_{j=1}^{n} x_j^2\right)}{n} - \bar{x}^2} \qquad (2.58)$$

The G_i^* statistic follows the standard normal distribution, and the resulting values are interpreted as z-scores. High positive values indicate that an observation and its neighbors exhibit high attribute values. Conversely, low negative values suggest clustering of low

attribute values. The statistical significance of the computed G_i^* values is evaluated at a predefined confidence level α.

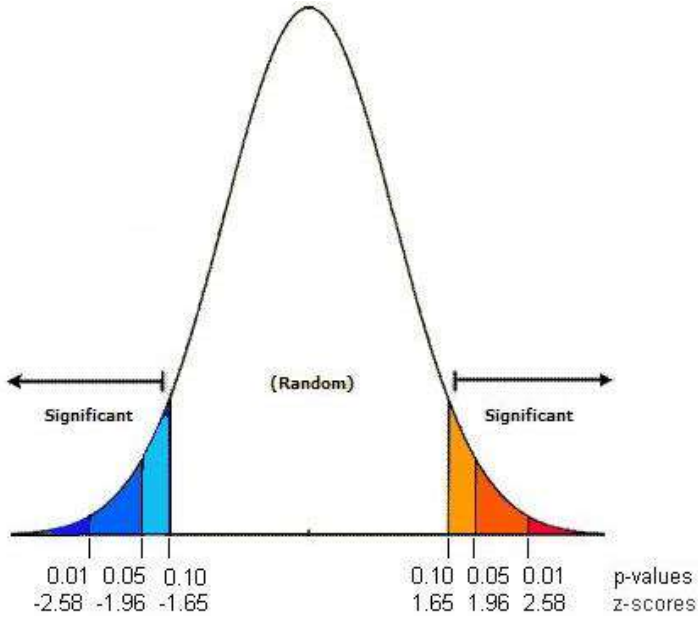


Figure 2.2: Identification of statistically significant hot and cold spots

3. SPATIAL REGRESSION MODELS

Spatial regression analysis has increasingly become one of the most prominent branches of statistics in recent years. In studies where space plays a critical role, it has been recognized that classical statistical methods are inadequate for explaining statistical variation and for making sound inferences. As a result, spatial statistical methods have been adopted in place of traditional approaches. These methods include spatial models that incorporate spatial information and consider the influence of location on observations.

Data that exhibit spatial relationships do not satisfy one of the

fundamental assumptions of classical statistics, namely the independence of observations. In spatial datasets, the presence of spatial dependence or spatial autocorrelation causes classical statistical methods to produce biased or inconsistent results. Therefore, specialized methods and techniques have been developed for analyzing spatially dependent data. Spatial regression models are among the most important of these methods, as they provide more reliable results by accounting for spatial relationships among observations.

Spatial regression models are generally based on the classical linear regression model. The classical linear regression model is a statistical framework that describes the linear relationship between a dependent variable and one or more independent variables, and it is expressed as follows:

$$y_i = \sum_{q=1}^{Q} X_{iq}\beta_q + \varepsilon_i \quad , \qquad i = 1,2,\dots,n \tag{3.1}$$

In this equation, y_i represents the value of the dependent variable for the ith observation, X_{iq} represents the value of the qth explanatory variable for the ith observation, β_q is the regression coefficient for the qth variable, and ε_i denotes the error term for the ith observation. In classical regression, error terms are assumed to have a mean of zero $E[\varepsilon_i] = 0$, constant variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$, and to be mutually uncorrelated $E[\varepsilon_i, \varepsilon_j] = E[\varepsilon_i]E[\varepsilon_j] = 0$. The matrix form of the model is:

$$y = X\beta + \varepsilon \tag{3.2}$$

Here, y denotes an $n \times 1$ vector of the dependent variable, X is an $n \times Q$ matrix of independent variables, β is a $Q \times 1$ vector of regression coefficients, and ε is an $n \times 1$ vector of error terms. The assumption of independent observations simplifies the model significantly. However, this assumption is generally not valid in spatial data analysis. If the explanatory variables, the residuals, or the dependent variable exhibit spatial dependence, the model becomes misspecified and the resulting estimators may be biased or inconsistent (Fischer & Wang, 2011).

Spatial dependence refers to the situation in which observations located near each other in space are interconnected. Based on this assumption, three main approaches have been proposed for incorporating spatial dependence into regression models. The first approach is known as endogenous spatial interaction models. These models examine and include in the model the relationship between the dependent variable and the values of the dependent variable in neighboring areas. This type of model is commonly referred to in the literature as the Spatial Autoregressive Model (SAR).

The second approach is exogenous spatial interaction models. These models investigate how independent variables in neighboring areas influence the dependent variable of a given unit. This structure is referred to as the Spatial Lag of X Model (SLX) or the Cross-Regressive Model in the literature.

The third approach is spatial error interaction models. In these models, spatial dependence is not introduced through the dependent variable but rather through unobserved influences captured in the error terms. In

other words, spatial relationships arise when error terms display similar patterns across neighboring units. This approach is known as the Spatial Error Model (SEM) in the literature.

3.1 Spatial Autoregressive Models (SAR)

The SAR model is used to capture situations in which the dependent variable is directly influenced by the values of the dependent variable in neighboring regions. This type of interaction represents an endogenous form of spatial autocorrelation. By incorporating the spatial relationship directly through the dependent variable, the model aims to produce more accurate and consistent results in contexts where classical regression methods are inadequate. In the literature, it is also referred to as the Spatial Lag Model. The SAR model is expressed as follows:

$$y_i = \rho \sum_{j=1}^{n} W_{ij} y_j + \sum_{q=1}^{Q} X_{iq} \beta_q + \varepsilon_i \qquad (3.3)$$

In this model, y_i denotes the dependent variable for the ith unit; ρ is the spatial autoregressive coefficient; w_{ij} represents the spatial weight between units i and j; y_j is the dependent variable for unit j; x_{iq} is the qth independent variable for unit i; β_q is the regression coefficient corresponding to the qth variable; and ε_i is the error term for unit i. The weight matrix is row-standardized such that each row sums to one. The matrix form of the model is given as:

$$y = \rho W y + X\beta + \varepsilon \qquad (3.4)$$

In this formulation, y is the $n \times 1$ vector of the dependent variable, X is the $n \times Q$ matrix of independent variables, β is the $Q \times 1$ vector of regression coefficients, W is the $n \times n$ spatial weights matrix, ρ is the spatial autoregressive coefficient, and ε is the $n \times 1$ vector of error terms. Solving the SAR model for y yields the following expression:

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon \tag{3.5}$$

The expected value and variance of y are then computed as follows:

$$E[y] = (I - \rho W)^{-1} X\beta \tag{3.6}$$

$$Var[y] = \sigma^2 (I - \rho W)^{-1}[(I - \rho W)^{-1}]^T \tag{3.7}$$

The matrix $(I - \rho W)^{-1}$ is referred to as the spatial multiplier, emphasizing that the expected value of each observation y_j depends on a linear combination of X values from neighboring observations (Fischer & Wang, 2011). The spatial autoregressive coefficient ρ is one of the most critical parameters in the SAR model, indicating the extent to which the dependent variable of a given observation is influenced by the dependent variables of its neighbors. This coefficient typically ranges between -1 and 1. A positive value of ρ suggests that neighboring observations tend to have similar values of the dependent variable. In other words, a high value in one observation exerts an upward influence on surrounding observations. Conversely, a negative value of ρ indicates a negative relationship among neighboring observations, implying that spatially proximate observations exhibit dissimilar values. A value of ρ close to zero implies that the spatial relationship is negligible, in which case classical regression methods

may be appropriate.

In SAR models, coefficient estimation cannot be performed using the Ordinary Least Squares (OLS) method, as in classical regression models. This is because the spatially lagged dependent variable Wy is correlated with the dependent variable itself, and hence with the error term. This correlation leads to biased and inconsistent estimates. Therefore, Maximum Likelihood (ML) estimation is typically used in SAR models. ML estimation possesses desirable asymptotic properties such as consistency, efficiency, and asymptotic normality. It is based on the assumption that the error terms follow a normal distribution. Accordingly, the reduced form of the SAR model and the corresponding log-likelihood function are given as follows:

$$y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon \tag{3.8}$$

$$\begin{aligned} \mathcal{L}(\rho, \beta, \sigma^2) =) = &-\frac{n}{2} ln(2\pi) + ln|I - \rho W| - \frac{1}{2\sigma^2}(y - \rho W y - X\beta)'(y \\ &- \rho W y - X\beta) - \frac{n}{2} ln(\sigma^2) \end{aligned} \tag{3.9}$$

Since the transformed model has the structure of a classical linear regression, a new dependent variable is defined as $y^* = (I - \rho W)y$. In this way, the β parameter can be estimated using OLS as follows:

$$\hat{\beta} = (X'X)^{-1} X' y^* = (X'X)^{-1} X'(y - \rho W y) \tag{3.10}$$

This estimator is conditionally unbiased and consistent for a given value of ρ. The estimator for σ^2 is obtained using the residual term:

$$\hat{\varepsilon} = y - \rho W y - X\hat{\beta} \tag{3.11}$$

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n}(y - \rho Wy - X\hat{\beta})'(y - \rho Wy - X\hat{\beta}) \qquad (3.12)$$

The final variance estimate is obtained by substituting the optimized value of ρ. This procedure is consistent with the concentrated log-likelihood approach and constitutes an integral part of SAR model estimation. By substituting the above expressions into the log-likelihood function, the number of unknowns is reduced, resulting in a simplified concentrated log-likelihood function that depends only on the parameter ρ:

$$\ell_C(\rho) = ln|I - \rho W| - \frac{n}{2} ln\left[(y - \rho Wy - X\hat{\beta}(\rho))'(y - \rho Wy - X\hat{\beta}(\rho))\right] \quad (3.13)$$

Because the spatial autoregressive coefficient ρ cannot be solved analytically, numerical optimization methods are used to find the value that maximizes the log-likelihood function. One of the simplest such methods is grid search. In this approach, a plausible interval for ρ is selected, typically between -1 and 1, although a narrower interval may be chosen based on the eigenvalues of the W matrix to ensure model stability and invertibility. In particular, the matrix $(I - \rho W)$ must be invertible, which requires that $\rho\lambda_i < 1$ for all eigenvalues λ_i. Therefore, ρ is often constrained to be less than $1/\lambda_{\max}$. Within the chosen interval, a set of equally spaced grid points is defined and $\ell_C(\rho)$ is calculated for each. The ρ value that yields the highest function value is selected. Although this method is straightforward, it typically has lower accuracy.

Another method for estimating ρ is the Newton-Raphson method. This technique involves taking the first derivative (score function) and the

second derivative (Hessian) of the concentrated log-likelihood function, which are expressed as follows (LeSage & Pace, 2009):

$$\frac{\partial \ell_C(\rho)}{\partial \rho} = -\text{tr}[(I - \rho W)^{-1} W] + \frac{e_{0'} e_L - \rho e_{L'} e_L}{(e_0 - \rho e_L)'(e_0 - \rho e_L)} \quad (3.14)$$

$$e_0 = y - X\beta_0 \quad \beta_0 = (X'X)^{-1} X'y \quad (3.15)$$

$$e_L = Wy - X\beta_L \quad \beta_L = (X'X)^{-1} X'Wy \quad (3.16)$$

The value of ρ that maximizes this function is the maximum likelihood estimate $\hat{\rho}$. However, since an analytical solution does not exist, the second derivative is also required:

$$H(\rho) = \sum_{i=1}^{n} \frac{\lambda_i^2}{(1 - \rho\lambda_i)^2} + \frac{e_{L'} e_L}{(e_0 - \rho e_L)'(e_0 - \rho e_L)} - \left[\frac{e_{0'} e_L - \rho e_{L'} e_L}{(e_0 - \rho e_L)'(e_0 - \rho e_L)}\right]^2 \quad (3.17)$$

The first term here represents the second derivative of the log-determinant (expressed in terms of eigenvalues), while the other two terms arise from the derivative of the sum of squared errors. The iterative update formula for ρ used in the Newton-Raphson method is given as:

$$\rho^{(t+1)} = \rho^{(t)} - \left[\frac{\partial^2 \ell_C(\rho)}{\partial \rho^2}\right]^{-1} \cdot \left[\frac{\partial \ell_C(\rho)}{\partial \rho}\right] \quad (3.18)$$

Determining when to terminate the iterative process is critical for the accuracy and efficiency of the algorithm. The literature generally follows three criteria. The first is the parameter change criterion, which is the most commonly used:

$$\left|\rho^{(t+1)} - \rho^{(t)}\right| < \varepsilon_1 \quad (3.19)$$

If the difference between consecutive estimates is sufficiently small, the algorithm is considered to have converged. Another criterion is based on the score function:

$$\left|\frac{\partial \ell_C(\rho)}{\partial \rho}\right| < \varepsilon_2 \qquad (3.20)$$

A score function value close to zero suggests that the estimate is near the maximum, at which point the iterations are halted. In the literature, the threshold is typically chosen as $\varepsilon_2 = 10^{-5}$ or smaller. A third commonly used criterion is based on changes in the log-likelihood function. If the function no longer increases during iterations, further updates are deemed unnecessary and the procedure is terminated (Kazar & Celik, 2012).

3.2 Spatial Error Models (SEM)

Spatial Error Models represent a class of spatial regression models that account for spatial dependence in the error terms. When the independent variables included in the regression model fail to fully capture the spatial variation, spatial autocorrelation may manifest in the residuals. This form of dependence typically arises from spatially structured but unobserved factors, the influence of latent variables, or systematic measurement errors that follow a spatial pattern.

SEM captures such dependence as an exogenous form of spatial autocorrelation. It is particularly useful when classical regression models yield biased or inefficient results due to unmodeled spatial effects. The model is commonly referred to in the literature as the

Spatial Error Model. The core structure of SEM is defined as follows (Fischer & Wang, 2011):

$$\varepsilon_i = \lambda \sum_{j=1}^{n} w_{ij}\, \varepsilon_j + u_i \qquad (3.21)$$

In this formulation, ε_i denotes the total error for unit i, λ is the spatial error dependence coefficient, w_{ij} represents the spatial weight between units i and j, ε_j is the error term for unit j, and u_i is an independently distributed disturbance term. This equation forms the foundation of the spatial error model and indicates that the error terms are not independent from one another but are influenced by the errors of neighboring units. In other words, the spatial structure is explicitly modeled within the residual terms. The matrix representation of the model is expressed as follows:

$$\varepsilon = \lambda W \varepsilon + u \qquad (3.22)$$

$$u \sim \mathcal{N}(0, \sigma^2 I) \qquad (3.23)$$

By substituting this spatial error structure into the regression model, the SEM formulation becomes:

$$\varepsilon = (1 - \lambda W)^{-1} + u \qquad (3.24)$$

$$y = X\beta + (I - \lambda W)^{-1}\varepsilon \qquad (3.25)$$

This model structure clearly demonstrates that the errors do not follow a random distribution but instead propagate according to spatial adjacency relationships. It illustrates why classical regression models can produce misleading results in the presence of spatial error

dependence. In this context, $E[\varepsilon\varepsilon'] = \sigma^2 I$, and the covariance matrix is given as follows:

$$\text{Var}[\varepsilon\varepsilon'] = \sigma^2(I - \lambda W)^{-1}[(I - \lambda W)^{-1}]^T \qquad (3.26)$$

The variance-covariance matrix further indicates that the error terms are not independent and that the errors of neighboring units are interdependent. The spatial error dependence coefficient λ in the model is generally not interpreted as a direct structural effect coefficient. Rather, it is regarded as a technical component aimed at improving the model's estimation accuracy. In the literature, this coefficient is frequently referred to as a nuisance parameter. In other words, it is considered an element that falls outside the primary interest of the analysis but must be accounted for to ensure the structural consistency of the model. A positive value of the spatial error dependence coefficient λ implies that the unobserved effects in neighboring units move in similar directions, whereas a negative value indicates that these effects move in opposite directions. If λ is close to zero, it suggests that spatial error dependence is weak or negligible and that classical regression models may be sufficient.

In the SEM model, parameter estimation cannot be performed using the classical Ordinary Least Squares method due to the presence of spatial dependence among the error terms. This dependence violates the assumptions required for OLS to yield efficient and reliable estimates. Therefore, parameter estimation in SEM is typically carried out using the Maximum Likelihood (ML) method. The ML approach simultaneously estimates the regression coefficients β, the spatial error

dependence coefficient λ, and the error variance σ² by explicitly accounting for spatial dependence in the error structure. The log-likelihood function for the SEM is given as follows:

$$\mathcal{L}(\lambda, \beta, \sigma^2) = -\frac{n}{2} ln(2\pi) + ln|I - \lambda W| - \frac{1}{2\sigma^2} e'e - \frac{n}{2} ln\sigma^2 \quad (3.27)$$

$$e = (I - \lambda W)(y - X\beta) \quad (3.28)$$

For a given value of λ, the conditional least squares estimate of β is:

$$\hat{\beta}(\lambda) = (X'X)^{-1}X'(y - \lambda W y) \quad (3.29)$$

Alternatively, the generalized least squares (GLS) estimator may be employed using the error covariance matrix $\Omega = [(I - \lambda W)'(I - \lambda W)]^{-1}$. The variance of the error term is estimated as:

$$\hat{\sigma}^2(\lambda) = \frac{1}{n} e'e = \frac{1}{n}(y - X\hat{\beta})'(I - \lambda W)'(I - \lambda W)(y - X\hat{\beta}) \quad (3.30)$$

$$e = (I - \lambda W)\left(y - X\hat{\beta}\right) \quad (3.31)$$

As with the SAR model, the spatial error coefficient λ is typically estimated by maximizing the concentrated log-likelihood function with respect to λ:

$$\ell_C(\lambda) = ln|I - \lambda W| - \frac{n}{2} ln\left[(y - X\hat{\beta})'(I - \lambda W)'(I - \lambda W)(y - X\hat{\beta})\right] \quad (3.32)$$

The maximum of this function is found using numerical techniques such as grid search or the Newton-Raphson method. The first derivative of the concentrated log-likelihood with respect to λ is:

$$\frac{\partial \ell_C(\lambda)}{\partial \lambda} = \text{tr}[W(I - \lambda W)^{-1}] - \frac{1}{\sigma^2} e'W\left(y - X\hat{\beta}\right) \quad (3.33)$$

Since this derivative cannot be solved analytically, the Newton-Raphson algorithm is applied. The second derivative of the concentrated log-likelihood is:

$$\frac{\partial^2 \ell_C(\lambda)}{\partial \lambda^2} = \text{tr}[W^2(I - \lambda W)^{-2}] + \frac{2}{\sigma^2} e' W^2 e \qquad (3.34)$$

Using these expressions, the Newton-Raphson iterations are conducted until convergence is achieved. While the ML approach is widely adopted in the literature, alternative methods such as the Generalized Method of Moments (GMM) can be employed for large samples (Luc Anselin, 1988).

3.3 Spatial Durbin Model (SDM)

The Spatial Durbin Model (SDM) is an extended spatial regression framework that incorporates both the spatial lag of the dependent variable and the spatial lags of the independent variables. It enables the assessment of both direct effects (on the observation itself) and indirect effects (from neighboring observations), thereby allowing a more comprehensive understanding of spatial interactions. The model accounts for both endogenous and exogenous spatial autocorrelation simultaneously. Its primary objective is to capture the influence not only of neighboring regions' dependent variables but also of their explanatory variables. The SDM is specified as:

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + \sum_{q=1}^{Q} X_{iq}\, \beta_q + \sum_{q=1}^{Q} \sum_{j=1}^{n} W_{ij}\, X_{jq} \theta_q + \varepsilon_i \qquad (3.35)$$

In this model, y_i is the dependent variable for unit i, ρ is the spatial autoregressive coefficient, W_{ij} represents the spatial weight between units i and j, X_{iq} is the qth explanatory variable for unit i, X_{jq} is the same variable for neighboring unit j, β_q denotes the standard regression coefficient, θ_q captures the effect of neighboring units' covariates, and ε_i is the error term.

The matrix notation of the model is:

$$y = \rho W y + X\beta + WX\theta + \varepsilon \qquad (3.36)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$$

Here, y denotes the $n \times 1$ vector of the dependent variable, X is the $n \times Q$ matrix of independent variables, β represents the $Q \times 1$ vector of regression coefficients, W is the $n \times n$ spatial weight matrix, θ is the $Q \times 1$ vector of coefficients for the spatially lagged independent variables, and ε is the $n \times 1$ vector of error terms. The model can be algebraically solved for y and expressed in the following reduced form.

$$y = (I - \rho W)^{-1}(X\beta + WX\theta) + (I - \rho W)^{-1}\varepsilon \qquad (3.37)$$

While the SAR model incorporates only the dependent variables of neighboring regions, the SDM model extends this framework by also including spatially lagged versions of the independent variables. This extension endows the SDM model with greater flexibility and allows it to be interpreted as a general specification that nests both the SAR and SEM models as special cases. In this model, the coefficient ρ captures the influence of neighboring regions' dependent variable values on the

dependent variable of the focal region, whereas the coefficient θ_q reflects the impact of the independent variables in neighboring regions on the dependent variable of the focal unit. A positive θ coefficient indicates that an increase in a variable in neighboring areas leads to an increase in the dependent variable in the current region. Conversely, a negative θ implies that an increase in neighboring regions' explanatory variables has a negative effect on the dependent variable in the current unit. Notably, when $\theta = 0$, the SDM simplifies to the SAR model, and when $\rho = 0$ with $\theta \neq 0$, it becomes equivalent to the Spatially Lagged X Model (SLX).

Like the SAR and SEM models, the SDM cannot be estimated using the conventional ordinary least squares method, as the presence of lagged dependent and independent variables introduces correlation with the error terms. Consequently, parameter estimation is typically conducted via the maximum likelihood method, which provides consistent and efficient estimates of the parameters ρ, β, and θ. The corresponding log-likelihood function for estimation via maximum likelihood is formulated as follows (Fischer & Wang, 2011).

$$\mathcal{L}(\rho, \beta, \theta, \sigma^2) = -\frac{n}{2} ln(2\pi) + ln|I - \rho W| - \frac{1}{2\sigma^2} \varepsilon'\varepsilon - \frac{n}{2} ln\sigma^2 \quad (3.38)$$

$$\varepsilon = y - \rho W y - X\beta - WX\theta \quad (3.39)$$

Since both X and WX are included in the SDM model, their corresponding coefficients are estimated simultaneously. Given a fixed value of ρ, the coefficients can be obtained using a method analogous to classical least squares. The estimator is defined as follows:

$$\begin{bmatrix}\hat{\beta}\\ \hat{\theta}\end{bmatrix} = (Z'Z)^{-1}Z'(y - \rho Wy) \qquad Z = [X \quad WX] \tag{3.40}$$

These estimates represent the direct and indirect effects of the covariates in the deterministic component of the SDM. Based on this, the estimated residuals and the error variance can be computed as follows:

$$\hat{\varepsilon} = y - \rho Wy - X\hat{\beta} - WX\hat{\theta} \tag{3.41}$$

$$\hat{\sigma}^2 = \frac{1}{n}(y - \rho Wy - X\hat{\beta} - WX\hat{\theta})'(y - \rho Wy - X\hat{\beta} - WX\hat{\theta}) \tag{3.42}$$

The spatial autoregressive coefficient ρ, which reflects the feedback structure of spatial dependence in the model, cannot be estimated directly. Therefore, a concentrated log-likelihood function is used and is given by:

$$\ell_C(\rho) = ln|I - \rho W| - \frac{n}{2} ln[\hat{\varepsilon}'\hat{\varepsilon}] \tag{3.43}$$

This function is numerically maximized with respect to ρ. To achieve this, iterative methods such as grid search or the Newton-Raphson algorithm are commonly employed. For estimation using the Newton-Raphson method, the first and second derivatives of the concentrated log-likelihood function with respect to ρ are derived as follows:

$$\frac{\partial \ell_C(\rho)}{\partial \rho} = -\text{tr}[(I - \rho W)^{-1}W] + \frac{1}{\sigma^2}\hat{\varepsilon}'Wy \tag{3.44}$$

$$\frac{\partial^2 \ell_C(\rho)}{\partial \rho^2} = \text{tr}[W(I - \rho W)^{-1}W(I - \rho W)^{-1}] + \frac{2}{\sigma^2}(Wy)'(Wy) \tag{3.45}$$

With these expressions, the Newton-Raphson steps can be implemented to obtain the estimate of the spatial autoregressive coefficient ρ.

3.4 Spatial Autocorrelation Model (SAC)

The Spatial Autocorrelation Model (SAC) is an integrated spatial regression specification that simultaneously accounts for spatial dependence in both the dependent variable and the error terms. In this respect, it can be viewed as a combination of the Spatial Lag Model (SAR) and the Spatial Error Model (SEM). Accordingly, the SAC model incorporates both endogenous and exogenous forms of spatial autocorrelation. In the literature, it is also referred to as the SARAR model, short for Spatial Autoregressive Model with Autoregressive Residuals.

The primary motivation for this model is to capture more comprehensive and realistic spatial interactions by modeling spatial dependence not only through the spatially lagged dependent variable but also through spatially autocorrelated errors. This approach is particularly useful in situations where the influence of neighboring observations and the impact of unobserved spatially patterned factors must be simultaneously addressed. The SAC model is formally expressed as follows (LeSage & Pace, 2009):

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + \sum_{q=1}^{Q} X_{iq}\, \beta_q + \varepsilon_i \qquad (3.46)$$

$$\varepsilon_i = \lambda \sum_{j=1}^{n} W_{ij}\, \varepsilon_j + u_i \qquad (3.47)$$

Here, y_i denotes the dependent variable for unit i; ρ is the spatial autoregressive coefficient of the dependent variable; W_{ij} is the spatial weight between units i and j; X_{iq} represents the qth independent variable for unit i; β_q is the corresponding regression coefficient; λ is the spatial error dependence coefficient; ε_i is the composite error term; and u_i denotes the random error term assumed to be independently and identically distributed with constant variance.

The matrix form of the model is given by:

$$y = \rho W y + X\beta + \varepsilon \tag{3.48}$$

$$\varepsilon = \lambda W \varepsilon + u, \quad u \sim N(0, \sigma^2 I) \tag{3.49}$$

Combining equations (4.48) and (4.49), the SAC model can be rewritten in its reduced form as:

$$y = \rho W y + X\beta + (I - \lambda W)^{-1} u \tag{3.50}$$

In this formulation, the parameter ρ captures the effect of spatial dependence in the dependent variable and corresponds to the autoregressive coefficient in the SAR model. Conversely, λ quantifies the spatial dependence in the error structure, reflecting the spatial error coefficient from the SEM specification. Joint estimation of both parameters enables simultaneous modeling of direct (through the dependent variable) and indirect (through the error term) spatial effects. The expectation and variance of y in this model are expressed as follows:

$$E[y] = (I - \rho W)^{-1} X\beta \quad (3.51)$$

$$\text{Var}[y] = \sigma^2 (I - \rho W)^{-1} (I - \lambda W)^{-1} (I - \lambda W)^{-T} (I - \rho W)^{-T} \quad (3.52)$$

As with other spatial models, ordinary least squares estimation is not appropriate for the SAC model, given the endogeneity introduced by both the spatially lagged dependent variable and the autocorrelated error structure. Maximum likelihood estimation is typically employed, allowing for the joint estimation of ρ, λ, β, and σ^2. Due to the model's complexity, iterative procedures or simulation-based methods such as Bayesian MCMC may also be applied. The log-likelihood function for the SAC model is defined as:

$$\mathcal{L}(\rho, \lambda, \beta, \sigma^2) = -\frac{n}{2} ln(2\pi) + ln|I - \rho W| + ln|I - \lambda W| - \frac{1}{2\sigma^2} \varepsilon' \varepsilon - \frac{n}{2} ln\sigma^2 \quad (3.53)$$

$$\varepsilon = (I - \lambda W)(y - \rho W y - X\beta) \quad (3.54)$$

The regression coefficients β represent the direct influence of the explanatory variables. Given fixed values for the spatial dependence parameters, the coefficients may be estimated using generalized least squares as follows:

$$\hat{\beta} = (X' \Omega^{-1} X)^{-1} X' \Omega^{-1} (y - \rho W y) \quad (3.55)$$

$$\Omega = [(I - \lambda W)'(I - \lambda W)]^{-1} \quad (3.56)$$

Based on these expressions, the estimated residuals and the corresponding error variance are calculated as:

$$\hat{\varepsilon} = (I - \lambda W)\left[y - \rho W y - X\hat{\beta}\right] \quad (3.57)$$

$$\hat{\sigma}^2 = \frac{1}{n} \hat{\varepsilon}' \hat{\varepsilon} \quad (3.58)$$

As in previous spatial models, the spatial parameters ρ and λ are estimated jointly using the concentrated log-likelihood function:

$$\ell_C(\rho, \lambda) = ln|I - \rho W| + ln|I - \lambda W| - \frac{n}{2} ln(\hat{\varepsilon}'\hat{\varepsilon}) \qquad (3.59)$$

Since closed-form solutions for ρ and λ are not available, numerical optimization techniques such as Newton-Raphson, grid search, or BFGS (Broyden–Fletcher–Goldfarb–Shanno Algorithm) are employed to maximize the likelihood function. In the Newton-Raphson framework, partial derivatives of the log-likelihood function with respect to ρ and λ must be computed individually. These first derivatives are given by:

$$\frac{\partial \ell_C}{\partial \rho} = -\text{tr}[(I - \rho W)^{-1} W] + \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}'(I - \lambda W) W y \qquad (3.60)$$

$$\frac{\partial \ell_C}{\partial \lambda} = -\text{tr}[(I - \lambda W)^{-1} W] + \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}' W (y - \rho W y - X\hat{\beta}) \qquad (3.61)$$

Computing second-order derivatives of the log-likelihood function in the SAC model is highly nontrivial due to the complex matrix operations involved, particularly the presence of trace operators and the fact that the residuals $\hat{\varepsilon}$ depend on both ρ and λ. As a result, second derivatives are often approximated numerically, or substitute structures such as the observed Fisher information matrix are used. The Newton-Raphson update step is given as:

$$\theta^{(t+1)} = \theta^{(t)} - [H(\theta^{(t)})]^{-1} \nabla \ell_C(\theta^{(t)}) \quad \theta = \begin{bmatrix} \rho \\ \lambda \end{bmatrix} \qquad (3.62)$$

In this formulation, $\nabla \ell(\theta)$ denotes the gradient (score function) and $H(\theta)$ is the Hessian matrix. The iterative procedure continues until the change between successive parameter estimates falls below a specified tolerance level. Upon convergence, the estimates are considered maximum likelihood solutions.

In conclusion, the SAC model provides a comprehensive framework that simultaneously accounts for spatial dependence in both the dependent variable and the error structure. While the SDM also captures complex spatial interactions through lagged explanatory variables, the SAC model emphasizes error propagation alongside spatial lag. Consequently, model selection should be guided by the data structure and the type of spatial effects expected.

3.5 Spatially Lagged X Model (SLX)

The Spatially Lagged X (SLX) model is a spatial regression framework that incorporates the spatially lagged values of independent variables into the regression equation. In this model, the dependent variable is influenced not only by the independent variables of the corresponding unit, but also by the independent variables of neighboring units. In other words, the effects of explanatory variables from neighboring regions are explicitly included in the model, while spatial dependence in the dependent variable (as in the ρWy term of the SAR model) or in the error terms (as in the λWε term of the SEM model) is not considered. The main objective of this model is to analyze the impact of explanatory variables from neighboring units on the target unit. This structure

allows for the separate estimation of both direct and indirect effects. The SLX model is expressed as follows (LeSage & Pace, 2009):

$$y_i = \sum_{q=1}^{Q} X_{iq}\,\beta_q + \sum_{q=1}^{Q}\sum_{j=1}^{n} W_{ij}\,X_{jq}\theta_q + \varepsilon_i \qquad (3.63)$$

In this model, y_i denotes the dependent variable for unit i; X_{iq} represents the qth independent variable for unit i; β_q is the standard regression coefficient (direct effect); W_{ij} denotes the spatial weight between units i and j; X_{jq} is the qth independent variable for neighboring unit j; θ_q is the coefficient associated with the spatial lag of the independent variable (indirect effect); and ε_i denotes the error term that satisfies the classical regression assumptions. The matrix notation of the model is given as:

$$y = X\beta + WX\theta + \varepsilon \quad \text{ve} \quad \varepsilon \sim N(0, \sigma^2 I) \qquad (3.64)$$

In this formulation, y is an $n \times 1$ vector of the dependent variable, X is an $n \times Q$ matrix of independent variables, WX is the matrix of spatially lagged independent variables, β and θ are $Q \times 1$ coefficient vectors, and ε is an $n \times 1$ vector of error terms. The SLX model serves as a simple yet powerful tool in spatial analysis. Since it includes only the neighborhood effects of explanatory variables, it avoids the problem of structural endogeneity commonly encountered in models such as SAR or SEM. As a result, ordinary least squares (OLS) can be used for parameter estimation.

By combining the X and WX matrices, the extended design matrix is defined as follows:

$$Z = [X \quad WX] \tag{3.65}$$

Accordingly, the parameter vector to be estimated and the OLS estimator are expressed as:

$$\delta = \begin{bmatrix} \beta \\ \theta \end{bmatrix} \tag{3.66}$$

$$\hat{\delta} = (Z'Z)^{-1}Z'y \quad \Rightarrow \quad \hat{\beta}, \hat{\theta} \tag{3.67}$$

Based on this, the residuals and the error variance are estimated as follows:

$$\hat{\varepsilon} = y - Z\hat{\delta} = y - X\hat{\beta} - WX\hat{\theta} \tag{3.68}$$

$$\hat{\sigma}^2 = \frac{1}{n - 2Q}\hat{\varepsilon}'\hat{\varepsilon} \tag{3.69}$$

In this context, $2Q$ represents the number of parameters estimated for both β and θ. The variance-covariance matrix of the estimated coefficients is given by:

$$\text{Var}(\hat{\delta}) = \hat{\sigma}^2(Z'Z)^{-1} \tag{3.70}$$

This matrix facilitates the construction of confidence intervals and the execution of statistical significance tests for the estimated parameters.

3.6 General Nesting Spatial Model (GNS)

The General Nesting Spatial Model (GNS) is one of the most comprehensive models in spatial regression analysis. This model aims to represent spatial dependence in a multidimensional manner by

simultaneously accounting for the spatial lag of the dependent variable, the spatial lag of the independent variables, and the spatial dependence in the error terms. In this respect, it can be regarded as a combination of the SAR, SEM, and SLX models. The GNS model is expressed as follows (Elhorst, 2014).

$$y_i = \rho \sum_{j=1}^{n} W_{ij}\, y_j + \sum_{q=1}^{Q} X_{iq}\, \beta_q + \sum_{q=1}^{Q} \sum_{j=1}^{n} W_{ij}\, X_{jq} \theta_q + \varepsilon_i \qquad (3.71)$$

$$\varepsilon_i = \lambda \sum_{j=1}^{n} W_{ij}\, \varepsilon_j + u_i \qquad (3.72)$$

In this context, y_i denotes the dependent variable for unit i, ρ is the spatial autoregressive coefficient of the dependent variable (reflecting the SAR effect), W_{ij} represents the spatial weight between units i and j, X_{iq} is the q^{th} explanatory variable for unit i, β_q denotes the regression coefficient indicating the direct effect of the q^{th} explanatory variable, X_{jq} refers to the q^{th} explanatory variable for neighboring unit j, θ_q represents the coefficient of the spatially lagged independent variables (reflecting the SLX effect), λ is the spatial dependence coefficient among the error terms (reflecting the SEM effect), ε_i denotes the total error term for unit i, and u_i is the independent error term that satisfies the classical regression assumptions. Additionally, the model can be expressed in matrix notation as follows.

$$y = \rho W y + X\beta + WX\theta + \varepsilon \qquad (3.73)$$

$$\varepsilon = \lambda W \varepsilon + u, \quad u \sim N(0, \sigma^2 I) \qquad (3.74)$$

Equations (4.73) and (4.74) can be combined to yield the following simplified form of the model:

$$y = \rho W y + X\beta + WX\theta + (I - \lambda W)^{-1}u \qquad (3.75)$$

The coefficient ρ captures the effect of the dependent variable values in neighboring regions on the dependent variable of the current unit, as in the SAR model. A positive value of ρ indicates spatial clustering of similar values, whereas a negative value suggests spatial adjacency of dissimilar values. The coefficient θ, analogous to that in the SLX model, measures the influence of explanatory variables in neighboring units on the dependent variable of the focal unit. A positive θ suggests that increases in explanatory variables in adjacent units elevate the dependent variable in the target unit. The parameter λ reflects the degree of spatial dependence in the error terms and captures the propagation of unobserved factors through spatial proximity. The expected value and variance of the dependent variable are given as follows:

$$E[y] = (I - \rho W)^{-1}(X\beta + WX\theta) \qquad (3.76)$$

$$Var[y] = \sigma^2(I - \rho W)^{-1}(I - \lambda W)^{-1}(I - \lambda W)^{-T}(I - \rho W)^{-T} \qquad (3.77)$$

This structure allows both direct and indirect spatial effects as well as neighborhood interactions to be reflected in the model through both the dependent variable and the error structure. Therefore, the General Nesting Spatial (GNS) model is considered the most comprehensive umbrella model in spatial regression analysis. In the GNS model parameter estimation cannot be conducted using the classical Ordinary

Least Squares method because spatial lag in the dependent variable and spatial dependence in the error terms violate its assumptions. Consequently, the most widely used estimation approach is the Maximum Likelihood (ML) method. In addition, the Generalized Method of Moments (GMM) and Instrumental Variables (IV) methods are also employed especially in large samples and complex spatial structures. Furthermore, due to the large number of parameters and the complexity of the error structure Bayesian estimation approaches and Markov Chain Monte Carlo (MCMC) techniques are sometimes preferred.

Assuming the error terms are normally distributed the log-likelihood function is expressed as follows:

$$\mathcal{L}(\rho, \lambda, \beta, \theta, \sigma^2) = \ln|I - \rho W| + \ln|I - \lambda W| - \frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\varepsilon'\varepsilon \qquad (3.78)$$

$$\varepsilon = (I - \lambda W)(y - \rho W y - X\beta - WX\theta) \qquad (3.79)$$

The GNS model, due to its structure involving many parameters, gives rise to a complex likelihood function. Consequently, estimating all parameters simultaneously is mathematically demanding both in terms of computational intensity and complexity. To address this problem more efficiently, some of the parameters that can be directly estimated are calculated beforehand. For the remaining key parameters, a simplified log likelihood function, known as the concentrated likelihood method, is used as in other methods. The parameters β, θ, and σ^2 in the model can be estimated under the assumption that the spatial dependence coefficients ρ and λ are held constant. Since these

coefficients appear in a linear structure within the model, they can be directly computed within the classical regression framework. However, because the error term includes spatial dependence through the λ parameter, the classical ordinary least squares method is not appropriate. Instead, the generalized least squares method is applied. This method takes into account the covariance structure of the error term and yields the best linear unbiased estimators. The estimators for these parameters are expressed as follows.

$$Z = [X \quad WX], \quad \delta = \begin{bmatrix} \beta \\ \theta \end{bmatrix} \tag{3.80}$$

$$\hat{\delta} = (Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}(y - \rho Wy) \tag{3.81}$$

The matrix $\Omega = [(I - \lambda W)' (I - \lambda W)]^{-1}$ represents the covariance structure of the error terms. Based on this formulation, the residuals and the corresponding estimate of the error variance are computed as follows.

$$\hat{\varepsilon} = (I - \lambda W)\left(y - \rho Wy - X\hat{\beta} - WX\hat{\theta}\right) \tag{3.82}$$

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon} \tag{3.83}$$

Once the estimates of $\hat{\beta}$, $\hat{\theta}$, and $\hat{\sigma}^2$ are obtained, the concentrated log-likelihood function for the remaining spatial dependence parameters ρ and λ is constructed as follows, upon which numerical optimization techniques are applied to obtain their estimates.

$$\ell_C(\rho, \lambda) = ln|I - \rho W| + ln|I - \lambda W| - \frac{n}{2} ln(\hat{\varepsilon}'\hat{\varepsilon}) \tag{3.84}$$

As in other models, the widely used Newton-Raphson algorithm will be employed as the numerical method. Accordingly, the first derivatives of the concentrated log-likelihood function are obtained as follows:

$$\frac{\partial \ell_C}{\partial \rho} = -\text{tr}[(I-\rho W)^{-1}W] + \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}'(I-\lambda W)Wy \qquad (3.85)$$

$$\frac{\partial \ell_C}{\partial \lambda} = -\text{tr}[(I-\lambda W)^{-1}W] + \frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}'W(y-\rho Wy - X\hat{\beta} - WX\hat{\theta}) \qquad (3.86)$$

The second derivatives, in turn, are computed as follows.

$$\frac{\partial^2 \ell_C}{\partial \rho^2} = \text{tr}[(I-\rho W)^{-1}W(I-\rho W)^{-1}W] - \frac{\partial}{\partial \rho}\left(\frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}'(I-\lambda W)Wy\right) \qquad (3.87)$$

$$\frac{\partial^2 \ell_C}{\partial \lambda^2} = \text{tr}[(I-\lambda W)^{-1}W(I-\lambda W)^{-1}W] - \frac{\partial}{\partial \lambda}\left(\frac{1}{\hat{\varepsilon}'\hat{\varepsilon}} \cdot \hat{\varepsilon}'W(y-\rho Wy - X\hat{\beta} - WX\hat{\theta})\right) \qquad (3.88)$$

$$\frac{\partial^2 \ell_C}{\partial \rho \partial \lambda} = \frac{\partial}{\partial \lambda}\left(\frac{\partial \ell_C}{\partial \rho}\right) \qquad (3.89)$$

Using this information, new parameter estimates are obtained at each iteration as specified below. This iterative process continues until the difference between successive estimates falls below a predetermined tolerance level, yielding a solution that converges to the maximum likelihood estimates. The iterative update equation for the Newton-Raphson algorithm is given below, where $\nabla \ell(\theta)$denotes the first derivative of the log-likelihood function (the score function), and $H(\theta)$ represents the matrix of second derivatives (the Hessian).

$$\theta^{(t+1)} = \theta^{(t)} - [H(\theta^{(t)})]^{-1}\nabla \ell_C(\theta^{(t)}), \quad \theta = \begin{bmatrix}\rho \\ \lambda\end{bmatrix} \qquad (3.90)$$

4. MODEL SPECIFICATION TESTS

Prior to conducting spatial regression analysis, it is critical to determine whether spatial dependence exists in the dataset and, if so, through which components this dependence manifests. This step not only ensures the appropriate model selection but also reveals whether the classical regression assumptions are violated. Therefore, in spatial analysis, it is recommended to begin the modeling process by performing various diagnostic tests. Identifying spatial autocorrelation highlights the necessity of establishing a spatial model, but it also raises the question of which type of spatial structure should be integrated. Specification tests have been developed to address this question.

4.1 Testing for Spatial Dependence in Residuals using Moran's I Test

Moran's I test constitutes an essential component in the evaluation of model specification in spatial regression models. It is primarily used to test whether spatial autocorrelation exists in the residual terms. This test examines whether the residuals obtained from a classical regression model exhibit spatial dependence. If the Moran's I statistic is found to be statistically significant, it implies that the model fails to adequately account for spatial dependence and should be re-estimated using a more appropriate spatial specification. Moran's I can be applied to residuals post-regression or during the initial data exploration phase to detect the presence of spatial dependence. Consequently, it helps assess the necessity of employing more advanced spatial models such as SAR,

SEM, SDM, or SLX. Detailed explanation of Moran's I is provided in Section 3.1.1.

4.2 Lagrange Multiplier (LM) Test

Lagrange Multiplier (LM) tests are employed to examine whether spatial dependence exists in the residuals of the classical regression model, thus serving as diagnostic tools that justify the use of spatial models. These tests aim to determine whether the dependent variable or the error term is influenced by neighboring observations. The LM test for the SAR model evaluates whether the spatial lag of the dependent variable is significant, while the LM test for the SEM model examines spatial autocorrelation in the residuals. The test statistics are assessed under the null hypothesis using the chi-square (χ^2) distribution. If the test result is statistically significant, it suggests that the classical regression model is inadequate and a spatial model (e.g., SAR or SEM) should be employed instead. Additionally, in cases where both SAR and SEM structures may be present, Robust LM tests are used for clearer model selection (Luc Anselin, 1988).

LM (lag): Tests whether the spatially lagged dependent variable, Wy, should be included in the model.

$$LM_{lag} = \frac{(e'Wy)^2}{\sigma^2 \cdot tr(W'W + W^2)} \qquad (4.1)$$

e: vector of OLS residuals

W: spatial weights matrix

y: dependent variable

σ^2: OLS-based error variance

Robust LM (lag): Tests the significance of the SAR component controlling for the SEM effect.

$$RLM_{lag} = LM_{lag} - \frac{\text{Cov}(LM_{lag}, LM_{error})}{\text{Var}(LM_{error})} \quad (4.2)$$

LM (error): Tests for spatial dependence in the residuals.

$$LM_{error} = \frac{(e'We)^2}{\sigma^2 \cdot tr(WW + W^2)} \quad (4.3)$$

Robust LM (error): Tests for residual dependence controlling for the SAR effect.

$$RLM_{error} = LM_{error} - \frac{\text{Cov}(LM_{lag}, LM_{error})}{\text{Var}(LM_{lag})} \quad (4.4)$$

LM SARMA: Tests for simultaneous spatial dependence in both the dependent variable and the residuals.

$$M_{SARMA} = LM_{lag} + LM_{error} \quad (4.5)$$

If this test is statistically significant, it indicates the presence of spatial dependence in both components. A significant result from these tests reveals misspecification in the classical model, warranting the use of a more suitable spatial model. For instance, if both LM (lag) and LM (error) are significant, it would be more appropriate to consider mixed models such as SAC or SDM.

4.3 Other Specification Tests

Anselin–Kelejian Test: Developed by Anselin and Kelejian (1997), this test allows for the detection of spatial dependence in the residuals under conditions of non-constant variance; heteroskedasticity It is considered as a robust alternative in situations where the traditional Moran's I test loses power (L. Anselin & Kelejian, 1997).

Ramsey Regression Specification Error Test (RESET): The RESET test was developed by Ramsey in 1969 with the purpose of evaluating the correctness of a model's functional form and testing for the presence of omitted variables or potential nonlinear relationships (Ramsey, 1969). It helps to identify whether the regression equation suffers from specification errors due to missing variables or neglected nonlinear effects. In the context of spatial analysis, it contributes to the decision-making process regarding whether certain transformations of the dependent variable should be incorporated into the model.

Heteroskedasticity Tests: In spatial models, the assumption that error terms possess constant variance often does not hold. To test this assumption, the Breusch–Pagan and Koenker–Bassett tests are commonly employed. A statistically significant result from these tests indicates the necessity of adopting estimation methods that are robust to heteroskedasticity. The Breusch–Pagan test was developed to assess whether the error terms in a regression model exhibit constant variance (Breusch & Pagan, 1979). In contrast, the Koenker–Bassett test is a more flexible alternative that does not rely on the assumption of normality, making it suitable for detecting heteroskedasticity in a

broader range of cases. It is regarded as a more robust version of the Breusch–Pagan test (Koenker & Bassett, 1982).

In conclusion, selecting the appropriate model in spatial data analysis depends not only on theoretical understanding, but also on the careful interpretation of diagnostic tests. Once the existence of spatial dependence is established through Moran's I, the structure of this dependence can be further explored using LM and other specification tests. An extensive evaluation of these tests not only enhances the accuracy of the selected model but also strengthens the reliability and interpretability of the results.

5. APPLICATION

The primary objective of this section is to clarify the concept of spatial regression models explained previous chapters, by an empirical study with demonstrating how spatial statistics can be employed through a practical application. Specifically, it will provide a step-by-step analysis of how spatial dependence affects classical regression models, how this dependence can be addressed using spatial models, and how model outcomes change accordingly. Spatial regression models enable the incorporation of similarities observed among geographically proximate observations into statistical modeling. Neglecting such dependencies may lead to biased and unreliable estimates. In order to establish a clear analytical flow in readers minds, a widely known data set including socioeconomic and environmental variables is selected.

Boston Housing Data, related to housing in various neighborhoods of Boston, Massachusetts contains geographic coordinates (latitude and

longitude) for each observation. The dependent variable determined for the analysis is the median value of owner-occupied homes in a neighborhood (CMEDV), and the explanatory variables are crime rate (CRIM), proportion of residential land zoned for large lots (ZN), proportion of non-retail business acres (INDUS), a binary variable indicating proximity to the Charles River (CHAS), nitrogen oxide concentration as a measure of air pollution (NOX), average number of rooms per dwelling (RM), and proportion of old buildings (AGE).

Firstly, a classical Ordinary Least Squares (OLS) regression model will be estimated, and the residuals will be tested for spatial dependence using Moran's I test. Subsequently, Lagrange Multiplier tests (both lag and error types, as well as their robust versions) will be conducted to determine whether spatial interaction operates through the dependent variable or the error terms. These preliminary tests will guide for the decision of selecting the appropriate spatial model. Following this, spatial regression models such as the Spatial Lag Model (SAR), Spatial Error Model (SEM), Spatial Durbin Model (SDM), and, if necessary, the Spatial Durbin Error Model (SDEM) will be estimated.

The results of each model will be compared in terms of the significance of explanatory variables and spatial parameters, as well as model fit criteria such as the Log-Likelihood and Akaike Information Criterion (AIC). This process will facilitate understanding of the nature of spatial dependence and provide a methodological framework for selecting appropriate spatial models. Finally, by comparing model performances,

the contribution of accounting for spatial interaction to the accuracy and robustness of the analysis will be clearly demonstrated.

Installing Required Packages: To implement spatial regression models in R, several specialized packages are required. These can be installed and loaded into the working environment using the following commands:

```
library(spdep)
library(sf)
library(spatialreg)
library(ggplot2)
library(dplyr)
```

The `spdep` package provides essential tools for the estimation of spatial regression models, the construction of spatial weight matrices, and the implementation of spatial dependence tests such as Moran's I. The *sf* package has been employed for reading, writing, transforming, and performing geometric operations on spatial data in accordance with the Simple Features standard, thereby ensuring that the data are converted into a spatial format suitable for the analysis process. The `spatialreg` package enables the parametric estimation of spatial econometric models such as SAR, SEM, SDM, SAC, SLX, and GNS, facilitating the modeling of various forms of spatial dependence structures. `ggplot2` has been chosen for the high-quality visual presentation of spatial and statistical results, particularly for creating maps, scatter plots, and other visualizations supporting model outputs. Finally, the `dplyr` package streamlines data manipulation tasks, including filtering, transforming, and summarizing data frames, thereby contributing to the efficient execution of data preparation steps within the analysis process.

Loading and Preparing the Dataset: the Boston Housing dataset utilized for the application is included in the `spdep` package. However, to perform spatial analysis, the coordinate information and variables within the dataset must be appropriately structured as follows:

```
data(boston)
# Dependent variable (median housing value)
y <- boston.c[, "CMEDV" ]
# Independent variables
x <- boston.c[, c("CRIM", "ZN", "INDUS", "CHAS", "NOX",
"RM", "AGE")]

# Coordinates
coords <- boston.c[, c("LON", "LAT")]

# Creation of the sf object
boston_sf <- st_as_sf(data.frame(x, y, coords), coords =
c("LON", "LAT"), crs = 4326)
```

Defining the Neighborhood Structure and Creating the Spatial Weights Matrix: One of the core components of spatial regression models is the neighborhood structure, which describes spatial relationships between observations, and the corresponding spatial weights matrix. In the Boston Housing dataset, observations are defined by point coordinates. Therefore, a distance-based neighborhood approach is employed. The goal is to ensure that each observation has at least one neighbor. To achieve this, the distance to each observation's nearest neighbor is calculated, and this maximum distance is then used as the threshold to define the neighborhood. The following R code block implements these steps:

```
# Coordinates
coords <- st_coordinates(boston_sf)
```

```r
# Determination of the minimum distance that ensures eac
h observation has at least one neighbor
dmax <- max(unlist(nbdists(knn2nb(knearneigh(coords, k =
1)), coords)))

# Distance-based neighborhood structure
nb_dist <- dnearneigh(coords, d1 = 0, d2 = dmax, longlat
= FALSE)

# Weight matrix: Row-standardized
listw_dist <- nb2listw(nb_dist, style = "W")

# Inspect the neighborhood structure
summary(nb_dist)
```

```
## Neighbour list object:
## Number of regions: 506
## Number of nonzero links: 45746
## Percentage nonzero weights: 17.86702
## Average number of links: 90.40711
## Link number distribution:
```

The `st_coordinates()` function extracts the coordinates from the object in simple feature (`sf`) format. Using the `knearneigh()` and `nbdists()` functions, a distance threshold (`dmax`) is calculated to ensure that each observation has at least one neighbor. The `dnearneigh()` function defines distance-based neighborhood relationships based on this threshold. The *`nb2listw()`* function constructs a row standardized spatial weights matrix based on the defined neighborhood structure. In this matrix, the influence of each observation on its neighbors is normalized to sum to one.

```r
# Visualization of neighborhoods
plot(nb_dist, coords = coords, pch = 20, col = "steelblu
e", cex = 0.8)
```

Figure 5.1 Visualization of the neighborhood structure

In the application, a distance-based neighborhood structure is employed due to the point-based spatial format of the dataset. Since each observation in point data is represented by geographic coordinates, neighborhood relationships are typically defined based on Euclidean distance or within a specified radius. In this case, ensuring that each point has at least one neighbor, a distance-based neighborhood matrix is constructed using the `dnearneigh()` function. On the other hand, when the data structure is polygon-based, such as in the case of spatial units with defined boundaries like neighborhoods, districts, or regions, neighborhood relationships are established based on shared borders (rook contiguity) or corners (queen contiguity). In such cases, methods like queen or rook contiguity become appropriate. Therefore, the selection of the neighborhood structure to be used in spatial analysis should be directly based on the type of data and the geometric nature of the spatial entities.

Examining Spatial Autocorrelation in Independent Variables: Before conducting spatial regression analysis, it is crucial to assess how the variables in the dataset are distributed across the spatial structure, as this plays a significant role in model selection. In this context, it is necessary to test for spatial autocorrelation not only in the dependent variable but also in the independent variables. The presence of spatial dependence in some independent variables may lead to biased and inconsistent estimates when using classical regression models. This issue becomes particularly important when deciding whether to include spatially lagged independent variables, as in models like SLX or SDM. For this purpose, Moran's I test is applied to the continuous variables used in the model. Moran's I evaluates whether a given variable exhibits spatial autocorrelation. A statistically significant Moran's I value indicates that the corresponding variable displays similar values in neighboring regions (positive autocorrelation) or dissimilar values (negative autocorrelation).

The Moran's I statistics and corresponding p-values calculated for the independent variables are reported below. The results of the analysis indicate that most of these variables exhibit statistically significant spatial patterns. Therefore, it is deemed appropriate to incorporate the spatially lagged versions of the independent variables into the spatial regression models. The R code used to conduct these analyses is provided below.

```
# Moran's I test
vars <- c("CRIM", "ZN", "INDUS", "NOX", "RM", "AGE")
moran_results <- lapply(vars, function(var) {moran.test(
```

```
boston.c[[var]], listw = listw_dist) })

# Converting Moran's I results into a table
moran_table <- lapply(vars, function(var) {
  test <- moran.test(boston.c[[var]], listw = listw_dist
)
  data.frame(
    Variable = var,
    Morans_I = test$estimate[["Moran I statistic"]],
    p_value = test$p.value
  )
}) %>% bind_rows()

# Presentation of results
print(moran_table)
```

```
##   Variable  Morans_I       p_value
## 1     CRIM 0.2309408 1.724062e-100
## 2       ZN 0.5248370  0.000000e+00
## 3    INDUS 0.5430339  0.000000e+00
## 4      NOX 0.6680633  0.000000e+00
## 5       RM 0.1663515  6.308489e-50
## 6      AGE 0.6172407  0.000000e+00
```

Moran's I Test for Residuals and Lagrange Multiplier Tests: In this section, the presence of spatial autocorrelation in the residuals of the OLS model will be tested to assess the adequacy of the classical model. Furthermore, in order to determine which spatial regression model (SAR, SEM, or SARMA) is most appropriate, Lagrange Multiplier (LM) and Robust LM tests will be conducted. The corresponding R code used for the implementation is provided below.

```
# OLS model
ols_model <- lm(y ~ CRIM + ZN + INDUS + CHAS + NOX + RM
+ AGE, data = boston_sf)
summary(ols_model)
```

```
##
## Call:
## lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NOX + RM
+ AGE, data = boston_sf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.447  -3.209  -0.701   2.089  39.882
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.12992    3.20092  -5.664 2.50e-08 ***
## CRIM         -0.17301    0.03449  -5.016 7.34e-07 ***
## ZN            0.01365    0.01445   0.945   0.3453
## INDUS        -0.12929    0.06406  -2.018   0.0441 *
## CHAS1         4.84977    1.05352   4.603 5.28e-06 ***
## NOX          -4.60617    4.07140  -1.131   0.2585
## RM            7.38341    0.41571  17.761  < 2e-16 ***
## AGE          -0.02353    0.01469  -1.602   0.1099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
##
## Residual standard error: 5.909 on 498 degrees of free
dom
## Multiple R-squared:  0.5916, Adjusted R-squared:  0.5
859
## F-statistic: 103.1 on 7 and 498 DF,  p-value: < 2.2e-
16
```

In order to establish a benchmark for comparison with spatial models, the analysis begins with the estimation of a classical linear regression model using Ordinary Least Squares (OLS). In the Boston Housing dataset, the dependent variable is defined as "CMEDV" (median value of owner-occupied homes), while the explanatory variables include crime rate (CRIM), proportion of residential land zoned for large lots (ZN), proportion of non-retail business acres per town (INDUS),

proximity to the Charles River (CHAS), nitrogen oxide concentration as an indicator of air pollution (NOX), average number of rooms per dwelling (RM), and proportion of older housing units (AGE). This model estimates the linear relationships without incorporating spatial effects. However, if spatial dependence is present in the data, the resulting estimates may be biased. Therefore, in the following section, specification tests are conducted to determine whether spatial dependence exists in the model residuals.

```
# Moran's I test for residuals
ols_resid <- residuals(ols_model)
moran.test(ols_resid, listw_dist)

##
##  Moran I test under randomisation
##
## data:  ols_resid
## weights: listw_dist
##
## Moran I statistic standard deviate = 6.249, p-value =
2.065e-10
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      0.0685786476     -0.0019801980      0.0001274917
```

The residuals obtained from the OLS model represent the estimation errors for each observation. In the presence of spatial structure, residuals from neighboring observations may exhibit similarity or correlation. To assess this possibility, the Moran's I statistic is employed. This test evaluates whether spatial autocorrelation exists among the residuals. A significantly positive Moran's I value indicates that neighboring residuals are similar, suggesting the presence of spatial

dependence. Conversely, a negative value implies an inverse relationship among neighboring residuals. A statistically significant Moran's I statistic implies that the classical OLS model fails to account for underlying spatial structure, and that a more appropriate spatial model should be considered. In the current analysis, the Moran's I value was found to be positive and statistically significant ($p < 0.01$). This result indicates a meaningful degree of positive spatial autocorrelation in the model residuals, implying that the error terms tend to behave similarly across neighboring units. Consequently, it can be concluded that the OLS model is insufficient and that spatial dependence must be explicitly incorporated into the model specification.

```
# Lagrange Multiplier tests
lm.LMtests(ols_model, listw_dist, test = "all")

## Please update scripts to use lm.RStests in place of l
m.LMtests

##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics
for spatial
##  dependence
##
## data:
## model: lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NO
X + RM + AGE, data
## = boston_sf)
## test weights: listw
##
## RSerr = 34.167, df = 1, p-value = 5.058e-09
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics
for spatial
##  dependence
##
```

```
## data:
## model: lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NO
X + RM + AGE, data
## = boston_sf)
## test weights: listw
##
## RSlag = 21.277, df = 1, p-value = 3.974e-06
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics
for spatial
##  dependence
##
## data:
## model: lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NO
X + RM + AGE, data
## = boston_sf)
## test weights: listw
##
## adjRSerr = 15.315, df = 1, p-value = 9.099e-05
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics
for spatial
##  dependence
##
## data:
## model: lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NO
X + RM + AGE, data
## = boston_sf)
## test weights: listw
##
## adjRSlag = 2.4255, df = 1, p-value = 0.1194
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics
for spatial
##  dependence
##
## data:
## model: lm(formula = y ~ CRIM + ZN + INDUS + CHAS + NO
```

```
X + RM + AGE, data
## = boston_sf)
## test weights: listw
##
## SARMA = 36.593, df = 2, p-value = 1.133e-08
```

Moran's I test serves to detect the presence of spatial autocorrelation, while Lagrange Multiplier (LM) tests offer guidance in selecting the most appropriate spatial regression model. These diagnostic tests are applied to the residuals obtained from the Ordinary Least Squares (OLS) model and help determine which type of spatial structure is more suitable for the data. The LM lag test examines whether the inclusion of the spatially lagged dependent variable supports the adoption of the Spatial Autoregressive (SAR) model. In contrast, the LM error test evaluates the presence of spatial dependence within the error terms, thus indicating the appropriateness of the Spatial Error Model (SEM). The robust versions of these tests control alternative forms of spatial dependence to provide more reliable conclusions regarding model selection.

If both LM lag and LM error tests produce statistically significant results, but only one of the robust tests is significant, the corresponding spatial model is typically preferred. However, if both robust tests are statistically significant, this suggests that a more complex model structure should be considered. Models such as the Spatial Autoregressive Combined (SAC) or the Spatial Durbin Model (SDM), which incorporate multiple sources of spatial dependence, may offer a more accurate specification. Utilizing these tests ensures that spatial

model selection is based on rigorous statistical evidence and enhances the robustness of the subsequent analysis.

Based on the diagnostic results obtained in this study, both the LM error and the Robust LM error tests were found to be statistically significant. This outcome provides strong evidence of spatial dependence within the error structure and indicates that the SEM model may be an appropriate choice. Although the LM lag test also yielded a significant result, the Robust LM lag test was not significant. This implies that when the error-based spatial dependence is taken into account, the contribution of the spatial lag of the dependent variable may be relatively limited. Furthermore, the LM SARMA test was statistically significant, indicating the presence of simultaneous spatial dependence in both the dependent variable and the error terms.

In summary, the SEM model is supported due to the strong spatial autocorrelation observed in the residuals. Nonetheless, the significance of the SARMA test points to the potential advantages of employing more comprehensive spatial specifications, such as SAC or SDM, which allow for a more general representation of spatial dependence in the data structure.

Estimation and Comparative Analysis of Spatial Regression Models: In this section, multiple spatial regression models—including the Spatial Autoregressive Model (SAR), Spatial Error Model (SEM), Spatial Durbin Model (SDM), Spatial Lag of X Model (SLX), Spatial Autoregressive Combined Model (SAC), and the General Nesting Spatial Model (GNS)—are estimated alongside the conventional

Ordinary Least Squares (OLS) model using the Boston housing dataset. The primary objective is to determine the most appropriate model by statistically comparing their performance in capturing the underlying structure of spatial dependence. The analysis aims to identify which model best reflects the nature of spatial relationships present in the data. The corresponding R code implementations and output results are presented below.

```
sar_model <- lagsarlm(y ~ CRIM + ZN + INDUS + CHAS + NOX
+ RM + AGE, data = boston_sf, listw = listw_dist)
summary(sar_model)

##
## Call:lagsarlm(formula = y ~ CRIM + ZN + INDUS + CHAS
+ NOX + RM +
##     AGE, data = boston_sf, listw = listw_dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -19.21804  -3.11530  -0.63815   2.19951  39.82708
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -25.0338728   3.8882435 -6.4384 1.208e-10
## CRIM         -0.1642686   0.0337173 -4.8719 1.105e-06
## ZN            0.0071581   0.0143851  0.4976  0.618765
## INDUS        -0.1308520   0.0628036 -2.0835  0.037205
## CHAS1         3.4360398   1.0724773  3.2038  0.001356
## NOX          -0.4669262   4.1435493 -0.1127  0.910278
## RM            7.0357636   0.4097395 17.1713 < 2.2e-16
## AGE          -0.0174085   0.0146149 -1.1911  0.233596
##
## Rho: 0.29325, LR test value: 14.377, p-value: 0.00014
959
## Asymptotic standard error: 0.082058
##     z-value: 3.5737, p-value: 0.00035198
## Wald statistic: 12.771, p-value: 0.00035198
```

```
##
## Log likelihood: -1605.636 for lag model
## ML residual variance (sigma squared): 33.299, (sigma:
5.7705)
## Number of observations: 506
## Number of parameters estimated: 10
## AIC: 3231.3, (AIC for lm: 3243.6)
## LM test for residual autocorrelation
## test value: 5.9465, p-value: 0.014746
```

The Spatial Autoregressive (SAR) model is one of the regression structures that directly accounts for spatial dependence in the dependent variable. In this model, the value of the dependent variable for a given unit is influenced not only by the explanatory variables associated with that unit but also by the values of the dependent variable in neighboring units. Hence, spatial dependence is intrinsically incorporated into the model structure. This feature allows for more reliable estimates in spatial data contexts where the independence assumption of classical regression models is violated.

According to the estimation results, the spatial autoregressive coefficient is $\rho = 0.293$, and this value is statistically significant ($p < 0.001$). This positive and significant coefficient indicates the presence of spatial spillover effects. In other words, housing values in one region are influenced by those in adjacent regions. This confirms a tendency toward spatial clustering and aligns with Tobler's first law of geography, which posits that "everything is related to everything else, but near things are more related than distant things."

In terms of model performance, the SAR model demonstrates superior fit compared to the classical linear regression model. The Akaike

Information Criterion (AIC) value for the SAR model is 3231.3, whereas it is 3243.6 for the classical model. Additionally, the log-likelihood value of -1605.64 suggests a better model fit. These differences reveal that incorporating spatial dependence significantly enhances the explanatory power of the model and underscores the importance of accounting for spatial structure in the analysis.

Furthermore, the likelihood ratio (LR) test confirms that the SAR model provides a statistically significant improvement over the classical model ($p < 0.001$). However, the results of the Lagrange Multiplier (LM) test indicate that spatial autocorrelation remains present in the residuals at a statistically significant level ($p = 0.0147 < 0.05$). This suggests that the SAR model may not fully capture all aspects of spatial dependence, and a more comprehensive model such as the Spatial Durbin Model (SDM) might be more appropriate.

In conclusion, the SAR model successfully identifies spatial dependencies in the data and provides a notable improvement over the classical model. Based on the model assumptions and output, it can be inferred that housing values are determined not only by local characteristics but also by the broader spatial environment.

```
sem_model <- errorsarlm(y ~ CRIM + ZN + INDUS + CHAS + N
OX + RM + AGE, data = boston_sf, listw = listw_dist)
summary(sem_model)

##
## Call:errorsarlm(formula = y ~ CRIM + ZN + INDUS + CHA
S + NOX + RM +
##     AGE, data = boston_sf, listw = listw_dist)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -17.1010  -3.1115  -0.7162   1.7524  39.2296
##
## Type: error
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept) -7.677553   3.721690 -2.0629 0.0391201
## CRIM        -0.169448   0.033376 -5.0769 3.837e-07
## ZN           0.032935   0.016828  1.9571 0.0503338
## INDUS       -0.228422   0.064560 -3.5382 0.0004029
## CHAS1        3.156952   1.052567  2.9993 0.0027061
## NOX         -8.580410   4.304592 -1.9933 0.0462269
## RM           6.388227   0.415361 15.3799 < 2.2e-16
## AGE         -0.055712   0.015608 -3.5694 0.0003578
##
## Lambda: 0.78728, LR test value: 37.971, p-value: 7.18
07e-10
## Asymptotic standard error: 0.064145
##     z-value: 12.273, p-value: < 2.22e-16
## Wald statistic: 150.64, p-value: < 2.22e-16
##
## Log likelihood: -1593.84 for error model
## ML residual variance (sigma squared): 30.918, (sigma:
5.5604)
## Number of observations: 506
## Number of parameters estimated: 10
## AIC: 3207.7, (AIC for lm: 3243.6)
```

The estimation results of the Spatial Error Model (SEM) indicate that spatial dependence is not transmitted directly through the dependent variable but rather through spatially structured unobserved effects, which are incorporated into the model via the error term. In this context, the SEM model aims to account for the influence of residuals that are systematically structured in space but are not captured by the included explanatory variables. This approach effectively addresses spatial autocorrelation arising from omitted or unmeasurable spatial factors.

According to the estimation results, the spatial error dependence coefficient is $\lambda = 0.787$, which is highly positive and statistically significant ($z = 12.27$, $p < 0.001$). This outcome reveals the significant influence of unobserved factors that follow a spatial pattern. In other words, although certain common factors affecting housing prices in neighboring regions are not explicitly included in the model, their effects are indirectly captured through the error structure. This underscores both the rationale for employing the SEM model and its analytical advantage.

The overall model fit indicators further support the robustness of the SEM model. The AIC value is 3207.7, which indicates a substantial improvement compared to the classical model (AIC = 3243.6). The log-likelihood value of -1593.84 also signals a better fit than that of the SAR model. Furthermore, the likelihood ratio (LR) test result ($LR = 37.97$, $p < 0.001$) strongly confirms that the SEM model significantly outperforms the classical linear model.

The regression coefficients obtained from the SEM estimation are both statistically significant and interpretable. Notably, variables such as CRIM, INDUS, and AGE exhibit a significant negative influence on housing values, while RM and CHAS1 demonstrate a positive association. These findings suggest that by incorporating the spatial error structure, the model successfully disentangles both structural and spatially unaccounted effects.

In conclusion, the SEM model provides more accurate and valid results in the presence of spatially structured but unobserved influences. The

model demonstrates high predictive power and effectively captures spatial error dependence. These results affirm that the SEM model constitutes a strong alternative in analyses where residual spatial autocorrelation is present.

```
sdm_model <- lagsarlm(y ~ CRIM + ZN + INDUS + CHAS + NOX
+ RM + AGE, data = boston_sf, listw = listw_dist, type =
"mixed")
summary(sdm_model)

##
## Call:lagsarlm(formula = y ~ CRIM + ZN + INDUS + CHAS
+ NOX + RM +
##     AGE, data = boston_sf, listw = listw_dist, type =
"mixed")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17.86324  -2.83045  -0.29627   1.93063  38.01095
##
## Type: mixed
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -49.3127399  12.8428330 -3.8397 0.0001232
## CRIM         -0.1750376   0.0328972 -5.3207 1.033e-07
## ZN            0.0560098   0.0175457  3.1922 0.0014118
## INDUS        -0.1577912   0.0661566 -2.3851 0.0170738
## CHAS1         3.0793660   1.0519052  2.9274 0.0034179
## NOX          -8.6539197   4.5038577 -1.9214 0.0546755
## RM            6.3343950   0.4102214 15.4414 < 2.2e-16
## AGE          -0.0685472   0.0159957 -4.2854 1.824e-05
## lag.CRIM     -0.6145716   0.2396966 -2.5640 0.0103487
## lag.ZN        0.0057274   0.0344982  0.1660 0.8681412
## lag.INDUS     0.5523612   0.2009769  2.7484 0.0059890
## lag.CHAS1     5.8981766   4.4997219  1.3108 0.1899298
## lag.NOX       7.3084275  16.4617475  0.4440 0.6570684
## lag.RM        3.9935452   2.2990382  1.7371 0.0823783
## lag.AGE       0.1132758   0.0455573  2.4864 0.0129026
##
## Rho: 0.066013, LR test value: 0.13591, p-value: 0.712
```

```
38
## Asymptotic standard error: 0.15251
##     z-value: 0.43284, p-value: 0.66513
## Wald statistic: 0.18735, p-value: 0.66513
##
## Log likelihood: -1574.238 for mixed model
## ML residual variance (sigma squared): 29.496, (sigma:
5.4311)
## Number of observations: 506
## Number of parameters estimated: 17
## AIC: 3182.5, (AIC for lm: 3180.6)
## LM test for residual autocorrelation
## test value: 0.14231, p-value: 0.70599
```

The Spatial Durbin Model (SDM) extends conventional spatial regression frameworks by incorporating spatially lagged values of both the dependent and explanatory variables. This structure allows for the simultaneous estimation of direct and indirect spatial effects, making it possible to analyze both the direction and the source of spatial interactions in greater detail. As such, the SDM can be viewed as a generalization that encompasses both the SAR and SLX models as special cases.

According to the estimation results, the spatial autoregressive coefficient is estimated as $\rho = 0.066$, which is relatively low and statistically insignificant ($z = 0.4328$, $p > 0.05$). This finding suggests that, within the context of this model, the spatially lagged values of the dependent variable do not have a significant influence on the outcome variable. However, when considering the effects arising from the spatial lags of the explanatory variables, it becomes evident that certain variables in neighboring regions exert significant impacts on the dependent variable. In particular, the variables lag.CRIM, lag.INDUS,

and lag.AGE are found to have statistically significant positive effects ($p < 0.01$), indicating that some socioeconomic characteristics influence housing prices not only within a given region but also through their presence in adjacent areas.

The model fit indicators warrant cautious interpretation. The AIC value for the SDM is 3182.5, which is marginally higher than that of the classical model (AIC = 3180.6). Despite the broader structure and larger number of estimated parameters in the SDM, this result indicates a limited gain in model fit. The lack of statistical significance for the spatial autoregressive coefficient and the non-significant likelihood ratio test suggest that the SDM performs less favorably compared to the SAR model in this dataset. This could be attributed to the fact that spatial dependence in this context is primarily mediated through the lagged explanatory variables rather than through the dependent variable itself.

Nonetheless, the residual spatial autocorrelation test returns an insignificant result ($p = 0.706 > 0.05$), indicating that the SDM sufficiently captures spatial dependence in the error structure. In other words, while direct spatial dependence is weak, the indirect effects are effectively modeled within this framework. Thus, the SDM provides meaningful insights in cases where spatial spillovers from explanatory variables play a dominant role.

Although the model demonstrates flexibility by allowing the decomposition of spatial effects associated with neighboring covariates, the lack of support for the spatial lag of the dependent variable,

combined with the modest improvement in model performance, suggests that the inclusion of ρ may not be essential in this setting. As a result, a simpler alternative such as the SLX model, which focuses exclusively on the spatial lags of explanatory variables, may offer a more parsimonious and interpretable solution. By retaining statistically significant spatial spillover variables while avoiding unnecessary parametric complexity, the SLX model may deliver advantages in both explanatory clarity and estimation efficiency in this context.

```
slx_model <- lmSLX(y ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE, data = boston_sf, listw = listw_dist)
summary(slx_model)
```

```
##
## Call:
## lm(formula = formula(paste("y ~ ", paste(colnames(x)[
-1], collapse = "+"))),
##     data = as.data.frame(x), weights = weights)
##
## Coefficients:
##              Estimate    Std. Error  t value     Pr(>
|t|)
## (Intercept)  -5.198e+01   1.180e+01  -4.406e+00   1.2
96e-05
## CRIM         -1.755e-01   3.338e-02  -5.257e+00   2.1
87e-07
## ZN            5.630e-02   1.781e-02   3.161e+00   1.6
71e-03
## INDUS        -1.533e-01   6.697e-02  -2.289e+00   2.2
50e-02
## CHAS1         3.088e+00   1.067e+00   2.893e+00   3.9
85e-03
## NOX          -8.578e+00   4.573e+00  -1.876e+00   6.1
26e-02
## RM            6.344e+00   4.161e-01   1.525e+01   2.9
73e-43
## AGE          -6.842e-02   1.624e-02  -4.213e+00   3.0
00e-05
```

```
## lag.CRIM     -6.531e-01   2.364e-01  -2.763e+00   5.9
49e-03
## lag.ZN        7.850e-03   3.500e-02   2.243e-01   8.2
26e-01
## lag.INDUS     5.685e-01   1.985e-01   2.865e+00   4.3
55e-03
## lag.CHAS1     6.650e+00   4.287e+00   1.551e+00   1.2
15e-01
## lag.NOX       7.123e+00   1.649e+01   4.319e-01   6.6
60e-01
## lag.RM        4.637e+00   1.722e+00   2.692e+00   7.3
48e-03
## lag.AGE       1.125e-01   4.625e-02   2.431e+00   1.5
40e-02
```

The Spatial Lag of X (SLX) model operates under the assumption that the dependent variable is influenced not only by the explanatory variables in the local unit but also by the values of these variables in neighboring units. In contrast to models that incorporate a spatial lag of the dependent variable, the SLX framework does not include an endogenous spatial autoregressive component. Instead, it captures spatial spillover effects through the inclusion of lagged explanatory variables. This structure offers a parsimonious yet powerful approach, particularly in empirical contexts where both direct and indirect spatial effects are of analytical interest.

The estimation results obtained from the SLX model indicate that a substantial proportion of the spatially lagged explanatory variables exert statistically significant effects. Among the local explanatory variables, CRIM, ZN, INDUS, CHAS1, RM, and AGE are found to be statistically significant. Notably, the variables CRIM and AGE exhibit significant negative impacts on housing values, whereas RM,

representing the average number of rooms per dwelling, shows a strong positive effect. These findings suggest that the SLX model successfully reproduces expected and meaningful estimates within a classical regression framework.

More importantly, the spatially lagged explanatory variables, including lag.CRIM, lag.INDUS, lag.RM, and lag.AGE, are also statistically significant ($p < 0.05$). This result confirms that housing prices are shaped not only by conditions within a given neighborhood but also by socioeconomic characteristics of adjacent areas. For instance, a high crime rate in neighboring locations (lag.CRIM) may negatively influence housing values in the focal area, while a higher average number of rooms in nearby dwellings (lag.RM) may exert a positive externality. These patterns highlight that spatial diffusion effects are driven not only by geographic proximity but also by shared structural and social characteristics across regions.

An important advantage of the SLX model lies in its structural simplicity. By excluding the spatial autoregressive coefficient ρ, the model avoids unnecessary parametric complexity while still capturing spatial dependence through lagged covariates. Considering the SDM model's results, in which the spatial autoregressive coefficient was found to be statistically insignificant and the likelihood ratio test failed to demonstrate superiority over the classical model, the SLX model emerges as a more appropriate alternative. It retains the significant spillover effects identified in the SDM model but omits the

uninformative spatial autoregressive term, thereby yielding a more consistent and interpretable structure.

In conclusion, the SLX model offers a transparent and analytically effective representation of both direct and indirect spatial effects. Within the scope of this analysis, it produces statistically significant and substantively meaningful results while maintaining a streamlined model structure. Compared to more complex alternatives, the SLX model stands out for its balance between explanatory power and interpretability.

```
sac_model <- sacsarlm(y ~ CRIM + ZN + INDUS + CHAS + NOX
+ RM + AGE, data = boston_sf, listw = listw_dist)
summary(sac_model)

##
## Call:sacsarlm(formula = y ~ CRIM + ZN + INDUS + CHAS
+ NOX + RM +
##     AGE, data = boston_sf, listw = listw_dist)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17.35491  -2.96176  -0.75006   1.80923  38.87902
##
## Type: sac
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)  -2.610789   5.692804 -0.4586 0.6465128
## CRIM         -0.173075   0.033350 -5.1897 2.106e-07
## ZN            0.033726   0.016865  1.9997 0.0455280
## INDUS        -0.234234   0.064603 -3.6258 0.0002881
## CHAS1         3.333625   1.058078  3.1506 0.0016291
## NOX         -10.162825   4.448201 -2.2847 0.0223302
## RM            6.353781   0.414818 15.3170 < 2.2e-16
## AGE          -0.059099   0.015845 -3.7299 0.0001916
##
## Rho: -0.17678
```

```
## Asymptotic standard error: 0.14892
##     z-value: -1.187, p-value: 0.23522
## Lambda: 0.82136
## Asymptotic standard error: 0.067849
##     z-value: 12.106, p-value: < 2.22e-16
##
## LR test value: 39.758, p-value: 2.3259e-09
##
## Log likelihood: -1592.946 for sac model
## ML residual variance (sigma squared): 30.652, (sigma:
5.5365)
## Number of observations: 506
## Number of parameters estimated: 11
## AIC: 3207.9, (AIC for lm: 3243.6)
```

The Spatial Autoregressive Combined (SAC) model provides a comprehensive framework for spatial regression analysis by simultaneously incorporating the spatial lag of the dependent variable, as in the Spatial Autoregressive (SAR) model, and spatial dependence in the error terms, as in the Spatial Error Model (SEM). This dual structure makes the SAC model particularly suitable for contexts in which spatial dependence arises both through observed neighborhood interactions and through unobserved or omitted spatially structured factors captured in the error term. The SAC model can be viewed as a generalization that nests both SAR and SEM models, making it an appropriate choice when both types of spatial dependence are jointly significant.

According to the estimation results, the spatial autoregressive coefficient ρ is estimated at −0.177 and is not statistically significant ($p = 0.2352 > 0.05$). This indicates that the housing prices in neighboring regions do not exert a direct influence on those in the target area. In

contrast, the spatial autocorrelation in the error terms, represented by λ, is estimated at 0.821 and is strongly statistically significant ($p < 0.001$). This finding suggests that unobserved but spatially correlated factors play a substantial role in shaping housing prices, and that the spatial dependence is primarily channeled through the error structure rather than through the dependent variable itself.

An evaluation of model fit metrics further supports the SAC model's effectiveness. The AIC value of 3207.9 is notably lower than that of the classical linear regression model (AIC = 3243.6), indicating superior model performance. Similarly, the log-likelihood value of −1592.946 is higher than in the benchmark model, reflecting improved overall model fit. Moreover, the likelihood ratio (LR) test result (LR = 39.76, $p < 0.001$) provides strong statistical evidence that the SAC model offers a significantly better fit compared to the standard OLS model. This supports the necessity of explicitly incorporating spatial structure into the modeling process.

The estimated coefficients of the explanatory variables are generally statistically significant and align with theoretical expectations. Variables such as CRIM, INDUS, AGE, and NOX have statistically significant negative effects on housing prices, while RM and CHAS1 display strong and positive influences. These findings confirm the model's ability to accurately capture the fundamental structural relationships while integrating spatial considerations.

In conclusion, for the dataset under investigation, although the spatial autoregressive component does not appear to be statistically significant,

the presence of strong spatial autocorrelation in the error terms underscores the importance of accounting for unobserved spatial factors. While the insignificance of the ρ coefficient weakens the case for employing SAR or SDM models alone, it strengthens the argument for using SEM or SAC models as more appropriate alternatives. Given that the source of spatial dependence appears to be concentrated in the error structure, the SEM model may suffice. However, the SAC model, due to its broader scope, offers a more robust and comprehensive solution in such contexts.

```
gns_model <- sacsarlm(y ~ CRIM + ZN + INDUS + CHAS + NOX
+ RM + AGE, data = boston_sf, listw =listw_dist, type =
"sacmixed")
summary(gns_model)
```

```
##
## Call:sacsarlm(formula = y ~ CRIM + ZN + INDUS + CHAS
+ NOX + RM +
##     AGE, data = boston_sf, listw = listw_dist, type =
"sacmixed")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -17.89727  -2.83364  -0.28754   1.93271  37.99537
##
## Type: sacmixed
## Coefficients: (asymptotic standard errors)
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -50.1260123  19.4748139 -2.5739  0.010056
## CRIM         -0.1754590   0.0330326 -5.3117 1.086e-07
## ZN            0.0559717   0.0175268  3.1935  0.001406
## INDUS        -0.1582480   0.0675349 -2.3432  0.019119
## CHAS1         3.0969681   1.0558328  2.9332  0.003355
## NOX          -8.7317872   4.5071415 -1.9373  0.052706
## RM            6.3382068   0.4132672 15.3368 < 2.2e-16
## AGE          -0.0685829   0.0159871 -4.2899 1.788e-05
## lag.CRIM     -0.6201600   0.2844534 -2.1802  0.029244
```

```
## lag.ZN         0.0056669   0.0354579  0.1598  0.873022
## lag.INDUS      0.5673074   0.2396629  2.3671  0.017928
## lag.CHAS1      6.2388870   6.1820800  1.0092  0.312884
## lag.NOX        6.3317315  18.3139255  0.3457  0.729543
## lag.RM         4.2995528   4.7281277  0.9094  0.363162
## lag.AGE        0.1133200   0.0465859  2.4325  0.014995
##
## Rho: 0.034186
## Asymptotic standard error: 0.43372
##     z-value: 0.078822, p-value: 0.93717
## Lambda: 0.052012
## Asymptotic standard error: 0.46047
##     z-value: 0.11295, p-value: 0.91007
##
## LR test value: 77.215, p-value: 5.7643e-13
##
## Log likelihood: -1574.217 for sacmixed model
## ML residual variance (sigma squared): 29.494, (sigma:
5.4309)
## Number of observations: 506
## Number of parameters estimated: 18
## AIC: 3184.4, (AIC for lm: 3243.6)
```

The General Nesting Spatial (GNS) model represents the most comprehensive and flexible structure within the framework of spatial regression analysis. In this model, the spatial lag of the dependent variable, the spatial autocorrelation in the error terms, and the spatial lags of the explanatory variables are all incorporated simultaneously. As such, the GNS model encompasses all fundamental spatial models including SAR, SEM, SDM, SAC, and SLX, offering the most parameter-rich specification. This structure enables the simultaneous modeling of both direct and indirect effects as well as unobserved spatial interactions.

According to the model estimation results, the spatial lag coefficient of the dependent variable ρ is estimated at 0.034 and the spatial autocorrelation coefficient of the error terms λ is estimated at 0.052. However, both coefficients are statistically insignificant ($p > 0.05$). These findings indicate that, for this specific dataset, direct forms of spatial dependence, whether through the dependent variable or through the error structure, do not contribute significantly to the model. In other words, spatial dependence appears to be transmitted primarily through the spatially lagged explanatory variables.

Indeed, among the spatially lagged covariates, the variables lag.CRIM, lag.INDUS, and lag.AGE are found to be statistically significant ($p < 0.05$). This supports the view that the relevant spatial dynamics are more appropriately captured through the exogenous covariates and their spatial spillover effects.

Regarding overall model performance, the GNS model demonstrates a strong fit. Its AIC value is 3184.4, which is substantially lower than that of the classical linear regression model (AIC = 3243.6). Additionally, the log-likelihood value of −1574.217 indicates a high degree of model fit. The likelihood ratio (LR) test further supports this, with an LR statistic of 77.215 ($p < 0.001$), signifying a statistically significant improvement over the classical model. However, this improvement appears to stem primarily from the SLX component of the model.

In summary, while the GNS model offers the most extensive parametric structure by incorporating all relevant spatial processes, the empirical results for this dataset reveal that neither spatial lag dependence in the

dependent variable nor spatial autocorrelation in the residuals are statistically meaningful. Therefore, although the GNS model includes all possible components, it yields a predictive performance comparable to that of more parsimonious models. In this specific empirical context, a simpler specification such as the SLX model provides similar explanatory power and may offer a more efficient and interpretable alternative.

```
# Pseudo R²
pseudo_r2_general <- function(y, y_hat) {
  1 - (sum((y - y_hat)^2) / sum((y - mean(y))^2))}

# Computation of a model-specific R² for the SLX model
pseudo_r2_slx <- function(model) {
  y <- model.response(model.frame(model))
  y_hat <- fitted(model)
  pseudo_r2_general(y, y_hat)}

# R² calculation for other models
pseudo_r2_default <- function(model) {
  y <- model$y
  y_hat <- model$fitted.values
  pseudo_r2_general(y, y_hat)}

# Comprehensive compilation of metrics
model_metrics <- data.frame(
  Model = c("SAR", "SEM", "SDM", "SLX", "SAC", "GNS"),
  AIC = c(AIC(sar_model), AIC(sem_model), AIC(sdm_model)
, AIC(slx_model), AIC(sac_model), AIC(gns_model)),
  BIC = c(BIC(sar_model), BIC(sem_model), BIC(sdm_model)
, BIC(slx_model), BIC(sac_model), BIC(gns_model)),
  LogLikelihood = c(logLik(sar_model), logLik(sem_model)
, logLik(sdm_model), logLik(slx_model), logLik(sac_model
), logLik(gns_model)),

# Pseudo R²
  Pseudo_R2 = c(pseudo_r2_default(sar_model), pseudo_r2_
default(sem_model), pseudo_r2_default(sdm_model), pseudo
```

```
_r2_slx(slx_model), pseudo_r2_default(sac_model), pseudo
_r2_default(gns_model)))


# Presentation of results in tabular form
print(model_metrics)

##   Model      AIC      BIC LogLikelihood Pseudo_R2
## 1   SAR 3231.273 3273.538     -1605.636 0.6042684
## 2   SEM 3207.679 3249.944     -1593.840 0.6325613
## 3   SDM 3182.477 3254.328     -1574.238 0.6494597
## 4   SLX 3180.613 3248.237     -1574.306 0.6493170
## 5   SAC 3207.892 3254.384     -1592.946 0.6357223
## 6   GNS 3184.435 3260.512     -1574.217 0.6494833
```

The R codes provided above generate model evaluation metrics including information criteria such as AIC and BIC, log-likelihood values, and pseudo R^2 statistics. According to the findings, the models with the lowest AIC values are SLX (3180.61), SDM (3182.48), and GNS (3184.43). These values indicate that these three models exhibit better fit with the data and offer greater parametric efficiency compared to alternative specifications. In terms of log-likelihood, the SLX, SDM, and GNS models perform similarly, confirming their comparable levels of model fit. Furthermore, pseudo R^2 values are highest in these three models, reinforcing their strong explanatory power.

However, model selection should not be based solely on information criteria. Factors such as parametric parsimony, interpretability, and the structural characteristics of spatial dependence must also be considered. In this regard, both SDM and GNS models involve a relatively large number of parameters (SDM: 17, GNS: 18). Yet in both models, the spatial lag coefficient rho is statistically insignificant. In particular, the simultaneous insignificance of both rho and lambda in the GNS model

suggests that direct spatial dependence through the dependent variable and the error structure is limited for this dataset.

Therefore, the model fit achieved by SDM and GNS appears to result primarily from the inclusion of spatially lagged explanatory variables. This finding implies that the more parsimonious SLX model may represent a theoretically and practically preferable alternative. The SLX model captures the significant components of SDM and GNS while excluding the insignificant spatial autoregressive structure, thereby enhancing interpretability without compromising model performance.

On the other hand, the SEM and SAC models are suitable when spatial autocorrelation is primarily present in the error terms. Although the SEM model performs strongly in this analysis, its inability to account for spillover effects from explanatory variables in neighboring regions constitutes a key limitation when compared to SLX and SDM. The SAC model incorporates both rho and lambda parameters, yet the insignificance of rho indicates that the spatial autoregressive effect is not supported in this context.

In summary, when evaluating model selection based on parametric efficiency, interpretability, and model fit collectively, the SLX model emerges as the most appropriate specification for the context of this analysis.

6. CONCLUSION

This book offers a comprehensive examination of the fundamental concepts, theoretical foundations, and applied dimensions of spatial

statistics with a particular focus on spatial regression models. The main objective is to identify issues of spatial dependence that violate the assumptions of classical regression analysis, introduce model types specifically developed to address these issues, and demonstrate how these models can be evaluated through empirical applications.

The initial chapters cover essential topics such as types of spatial data, the structure of spatial autocorrelation, and the construction of spatial weight matrices. These are followed by detailed explanations of statistical techniques used to measure spatial dependence, including Moran's I, Geary's C, and LISA. The book also illustrates how spatial dependence can be analyzed at both global and local levels, supported by visualization techniques.

Subsequent chapters explain the limitations of classical regression analysis when applied to spatial data and introduce spatial regression models developed to address these limitations, including SAR, SEM, SDM, SLX, SAC, and GNS. Each model is presented systematically, covering underlying assumptions, mathematical structure, estimation techniques, and methods of interpretation, supported by illustrative examples.

In the applied section of the book, the effects of spatial dependence on economic indicators are analyzed using the Boston housing dataset. Different spatial regression models are compared in terms of model performance, based on information criteria such as AIC, log-likelihood, and pseudo R^2. Moreover, each model is assessed comparatively in terms of its ability to capture spatial structure, the significance of

estimated parameters, and interpretability. The findings reveal that spatial effects are predominantly transmitted through the values of explanatory variables in neighboring regions, highlighting the practical effectiveness of models such as SLX and SDM.

In conclusion, this book aims to serve as a comprehensive reference for researchers seeking to engage with spatial regression modeling, offering both a robust theoretical foundation and practical application insights. By demonstrating how spatial relationships can be integrated into the statistical modeling process, this work provides valuable contributions across a range of disciplines, including the social sciences, urban planning, economics, and environmental studies.

REFERENCES

Anselin, L., & Getis, A. (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, *26*(1), 19–33.

Anselin, L., & Kelejian, H. H. (1997). Testing for Spatial Error Autocorrelation in the Presence of Endogenous Regressors. *International Regional Science Review*, *20*(1–2), 153–182.

Anselin, Luc. (1988). Spatial Econometrics: Methods and Models. In *Econometrica*. https://doi.org/https://doi.org/10.1007/978-94-015-7799-1

Anselin, Luc. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis*, *27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Bivand, R. S., & Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *Test*, *27*(3), 716–748. https://doi.org/10.1007/s11749-018-0599-x

Breusch, T. S., & Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, *47*(5), 1287–1294.

Cliff, A. C., & Ord, J. K. (1973). Spatial autocorrelation. London: Pion. *Progress in Human Geography*, *19*(2), 245–249. https://doi.org/https://doi.org/10.1177/030913259501900205

Cliff, A. D., & Ord, J. . (1969). The Problem of Spatial Autocorrelation.

In London Papers in Regional Science, *1*(Studies in Regional Science), 22–55.

Cliff, A. D., & Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion.

Çubukçu, K. M. (2015). *Planlamada ve coğrafyada temel istatistik ve mekansal istatistik*. Nobel.

Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer Berlin, Heidelberg.

Emrehan, A. F. (2022). *Quantile Approach To Contiguity Based Spatial Autocorrelation : Spatial Theta Lag And Conditional Weighting*. Yıldız Technical University.

Fischer, M. M., & Wang, J. (2011). *Spatial Data Analysis Models, Methods and Techniques*. Springer Berlin, Heidelberg.

Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician*, *5*(3), 115–146.

Getis, A. (2009). Spatial Weights Matrices. *Geographical Analysis*, *41*(4), 404–410.

Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, *36*(2), 90–104. https://doi.org/10.1111/j.1538-4632.2004.tb01127.x

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, *24*(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Kazar, B. M., & Celik, M. (2012). *Spatial AutoRegression (SAR) Model: Parameter Estimation Techniques*. Springer.

Koenker, R., & Bassett, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, *50*(1), 43–61.

LeSage, J. P., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society*, *10*(2), 243–251.

Moran, P. A. P. (1950). *Notes On Continuous Stochastic Phenomena*. *37*((1/2)), 17–23. http://biomet.oxfordjournals.org/

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*, *48*, 167–227.

Tiefelsdorf, M., Griffith, D. A., & Boots, B. (1999). A variance-stabilizing coding scheme for spatial link matrices. *Environment and Planning A*, *31*(1), 165–180.

von Thünen, J. H. (1826). *Der isolierte Staat in Beziehung auf Landwirtschaft und Nationalökonomie*. Perthes.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley.

SPATIAL REGRESSION MODELS

ISBN: 978-6-25575-318-2