International Journal of Analysis and Applications

New Modified Estimators for the Spatial Lag Model with Randomly Missing Data in Dependent Variable: Methods and Simulation Study

Mohamed R. Abonazel[1,*], Ohood A. Shalaby[2], Ahmed H. Youssef[1]

[1]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza 12613, Egypt*

[2]*National Center for Social and Criminological Research, Giza 12513, Egypt*

**Corresponding author:* mabonazel@cu.edu.eg

ABSTRACT. Accurately estimating the spatial lag model (SLM) in the presence of randomly missing data in the dependent variable poses a significant challenge. We introduce some modifications to the two-stage least squares with imputation (I2SLS) estimator previously proposed by Izaguirre [1] and Wang and Lee [2]. Our key contributions include (1) introducing the generalized nonlinear least squares (GNLS) estimator as an alternative imputation method to the previously used nonlinear least squares (NLS) approach in the literature, (2) incorporating additional instrument matrices (IM), and (3) implementing both partial and total imputations for all modified estimators. Through a Monte Carlo simulation (MCS) study, we evaluate the performance of these estimators across various scenarios of sample size, spatial weights matrix densities, and missingness rate. Results are compared in terms of coefficient bias and root mean squares errors (RMSE) for both the parameters and model fit. The findings indicate that all estimators demonstrate relatively strong performance in the context of estimator coefficients bias and RMSE. However, our modified estimators demonstrate slightly better performance compared to those previously documented in the literature in terms of overall RMSE. While both total and partial imputation approaches tend to produce similar results, partial imputation demonstrated superior performance in certain scenarios. Additionally, the estimators proved robust, maintaining their reliability across varying levels of spatial connectivity. However, higher missing data rates led to slightly increased bias and RMSE.

Received Oct. 15, 2024

2020 Mathematics Subject Classification. 62H12, 62F12, 62D10.

Key words and phrases. spatial regression analysis; spatial lag model (SLM); missing data; two-stage least squares estimator (2SLS); GNLS; imputation methods; Monte Carlo simulation (MCS).

https://doi.org/10.28924/2291-8639-23-2025-67

ISSN: 2291-8639

1. Introduction

In recent years, there has been a revival of theoretical and empirical work on the spatial aspects of the economy, giving rise to a new field known as economic geography which has become one of the most exciting fields of contemporary economics. This area of research, pioneered by influential works for Krugman [3] and Fujita et al. [4], has gained prominence due to its exploration of the impact of location on various economic phenomena. As a result, there has been a growing need to employ spatial econometric techniques to analyze data and gain insights into spatial relationships and dependencies among variables. While SLM has become a fundamental tool in spatial data analysis, missing data in the dependent variable poses a significant challenge. With the increasing availability of geo-referenced data over the past few decades, researchers have access to vast amounts of information for spatial analysis. However, this data is often incomplete, with missing values arising from factors such as errors in data collection, non-response, or incomplete records.

Missing data in SLM estimation presents a significant challenge, where leading to biased and inconsistent parameter estimates that can undermine the validity of spatial data analysis [[2],[5]]. Prior research by Yokoi [6] has demonstrated that missing observations in SLM can result in underestimated spatial autocorrelation and incorrect model selection, with negative autocorrelation often going undetected. These findings highlight the critical need for improved SLM estimation techniques in the presence of missing data. Our study addresses these challenges by building upon and modifying the existing alternatives of the I2SLS estimator proposed by Izaguirre [1] and Wang and Lee [2], specifically the best generalized I2SLS (IBG2SLS), series-type efficient I2SLS (ISTE2SLS), and asymptotic IBG2SLS (IABG2SLS).

Our primary modifications focus on three key aspects of the estimation process: the imputation method, the number of imputations, and IM. While previous research by Izaguirre [1] and Wang and Lee [2] employed the NLS consistent estimator for imputation, we introduce the GNLS estimator as an alternative. Furthermore, we incorporate an additional IM used by Lee [7] for the best optimum generalized method of moment (BGMM) estimator and Wang and Lee [2] for GMM in SLM with missing data.

To evaluate the effectiveness of our proposed estimators, we compare their performance to the PIABG2SLS.nls suggested by Izaguirre [1]. This benchmark was chosen because Izaguirre's findings indicated that while all estimators yielded quite similar results, the PIABG2SLS.nls showed slightly better performance in terms of variability for smaller sample sizes.

The following sections are organized as follows: Section 2 outlines the SLM specification and assumptions. Section 3 reviews the estimation methods of the SLM in the case of both complete and incomplete data. Furthermore, we explore the challenges and limitations associated with SLM estimation when dealing with missing data. Section 4 discusses our modification of

Izaguirre's [1] approach and Wang and Lee [2] approach. We discuss the rationale behind this modification and outline the steps involved in implementing the modified estimators. Section 5 provides the MCS study. Sections 6 and 7 highlight the main concluding remarks and outline the practical implications of our research and potential avenues for future exploration in this area.

2. The Spatial Lag Model

The SLM extends the standard linear regression model to incorporate spatial dependence. It captures the idea that the value of a dependent variable at a given location is influenced not only by its characteristics but also by the values of the same variable at neighboring locations.

- Complete Data SLM Specification

Suppose we have n spatial units or locations in our analysis. The SLM can be expressed as follows:

$$\mathrm{y}_n = \lambda_0 \mathrm{W}_n \mathrm{y}_n + \mathrm{X}_n \beta_0 + \varepsilon_n \tag{1}$$

where y_n is an $(n\times1)$ vector containing the values of the dependent variable across all locations, λ_0 is the spatial autoregressive parameter, $\mathrm{W}_n$is an $(n\times n)$ non-negative matrix of known constants capturing the connectivity or proximity among spatial units in our data, $\mathrm{W}_n\mathrm{y}_n$ is a spatially lagged dependent variable, X_n is an $(n\times\mathrm{K})$ matrix representing the observed values of K exogenous explanatory variables for all n spatial units, β_0 is an (K×1) vectors of coefficients that correspond to the explanatory variables in X_n, and ε_n is an $(n\times1)$ vector of $i.i.d.$ disturbances such that $\varepsilon_n \sim (0,\ \sigma^2\mathrm{I}_n)$.

This SLM specification assumes that the dependent variable of unit (i) is directly influenced by the spatially weighted dependent variable of neighboring units [8].

Model (1) can be written in a reduced form as:

$$\mathrm{y}_n = S_n^{-1}(\lambda_0)\mathrm{X}_n\beta_0 + S_n^{-1}(\lambda_0)\varepsilon_n \tag{2}$$

where $S_n(\lambda_0) = (\mathrm{I}_n - \lambda_0\mathrm{W}_n)$.

- Missing Data SLM Specification

Now, considering the scenario where the dependent variable y_n has missing values, we can rewrite the model (1) as shown in (3). In this equation, the vector $\begin{bmatrix}\mathrm{y}_n^o\\ \mathrm{y}_n^m\end{bmatrix}$ represents the full vector of dependent variable values, where the first n^o elements, y_n^o, corresponding to the observed outcomes, and the last n^m elements, y_n^m, corresponding to the missing outcomes, where the total number of elements in the vector is $n = n^o + n^m$.

$$\begin{bmatrix}\mathrm{y}_n^o\\ \mathrm{y}_n^m\end{bmatrix} = \lambda_0\mathrm{W}_n\begin{bmatrix}\mathrm{y}_n^o\\ \mathrm{y}_n^m\end{bmatrix} + \mathrm{X}_n\beta_0 + \varepsilon_n \tag{3}$$

To systematically handle incomplete data structures, we use the selection matrices $J_n^o = [I_{n^o \times n^o}, 0_{n^o \times n^m}]$ and $J_n^m = [0_{n^m \times n^o}, I_{n^m \times n^m}]$. Where J_n^o is an $(n^0 \times n)$ matrix that is used to extract the observed elements from the full vector of y_n, creating a sub-vector of observed outcomes, y_n^o. While, J_n^m is an $(n^m \times n)$ matrix that is used to select the unobserved elements from the full vector of y_n, resulting in a sub-vector of missing outcomes y_n^m. The matrices $I_{n^o \times n^o}$ and $I_{n^m \times n^m}$ are identity matrices of sizes $(n^o \times n^o)$ and $(n^m \times n^m)$ respectively, while $0_{n^o \times n^m}$ and $0_{n^m \times n^o}$ are zero matrices of sizes $(n^o \times n^m)$ and $(n^m \times n^0)$ respectively.

By applying the selection matrix J_n^o on the model (3) to choose the observed elements from y_n. Thus:

$$y_n^o = \lambda_0 J_n^o W_n \begin{bmatrix} y_n^o \\ y_n^m \end{bmatrix} + J_n^o X_n \beta_0 + J_n^o \varepsilon_n \tag{4}$$

- SLM Assumptions

A1. Assumptions of Spatial Weights Matrix:

The spatial weights matrix, i.e., W_n are non-stochastic matrices with zero diagonals.

The spatial transformation matrices, i.e., $S_n(\lambda) = (I_n - \lambda W_n)$ are invertible on the compact parameter spaces of spatial parameters λ.

The admissible parameter space for the true spatial parameters λ_0 is [-1, 1].

Spatial matrices, W_n and $S_n^{-1}(\lambda)$, before W_n is row-standardized, are uniformly bounded in both rows and columns sum in absolute value as n goes to infinity. This assumption originated by Kelejian and Prucha [[9],[10]] to ensure that $S_n(\lambda)$ is not singular. Furthermore, this allows us to express $S_n^{-1}(\lambda) = \sum_{k=0}^{\infty} \lambda^k W_n^k$.

A2. Assumptions of the Error Components: The elements of the disturbance vector ε_n, i.e., $\{\varepsilon_i\}, i = 1, \cdots, n$ are $i.i.d$ with zero mean and finite variance σ_0^2, and $0 < \sigma_0^2 < \infty$.

A3. Assumptions on Covariates: The regressors X_n are non-stochastic and have full column rank. Their elements are uniformly bound constants. Additionally, as n approaches infinity, the limit of $\frac{1}{n} X_n' X_n$ exist and is non-singular.

A4. Assumptions of the Observed Data: Proportion $\frac{n^0}{n}$ of observations tends to c, where c is a finite positive constant, as n goes to infinity. This assumption indicates that the number of observed observations should not be too small relative to n.

A5. Assumptions of the Instrument Matrix: Let Q_n and Q_n^o be $(n \times K)$ and $(n^o \times K)$ instrument matrices, respectively, constructed as functions of X_n and W_n, then:

The elements of $Q_n^o = J_n^o Q_n$ should be uniformly bounded.

$\lim_{n \to \infty} \frac{1}{n} Q_n^{o\prime} Q_n^o$ exists and is a non-singular matrix.

$\lim_{n\to\infty} \frac{1}{n} Q_n^{o\prime}\ [J_n^o G_n X_n \beta_0 \quad J_n^o X_n]$ has full column rank (K + 1), where $G_n = W_n S_n^{-1}(\lambda_0)$.

$\lim_{n\to\infty} \frac{1}{n} Q_n'\ S_n(\hat{\lambda}) J_n^{o\prime} J_n^o S_n^{-1}(\hat{\lambda})\ [G_n X_n \beta_0 \quad X_n]$ has full column rank (K + 1) for any feasible value λ in ʌ. Where $Q_n^{'}$ and $J_n^{o'}$ denote the transposes of the matrices of Q_n and J_n^o, and $\hat{\lambda}$ is the estimated value of λ.

These assumptions are frequently made in spatial econometrics; see [[1],[2],[9],[11]] among others.

3. The Estimation Methods of the SLM

- Review of Estimation Methods for SLM with Complete Data

Because the ordinary least squares estimator is inconsistent in the presence of a spatial weight matrix, much of the early work has focused on a maximum likelihood (ML) approach [12]. the ML estimators for the SLM have been derived and applied by Anselin [13], amongst others. Anselin et al. [14] mentioned that this method faces significant computational complexities when computing the Jacobian term. $|S_n|$, i.e. the determinant of (n × n) matrix. Although Ord [15], Smirnov and Anselin [16], and others have suggested some simplification or approximation techniques to address this issue. The computation process remains challenging, especially for large sample sizes and general spatial weights matrices. To overcome these challenges, Kelejian and Prucha [9] proposed a 2SLS estimator, and Lee [17] further discussed the best one (B2SLS) within the class of instrumental variables. The proposed 2SLS estimators are computationally simpler but inefficient relative to ML estimators. In contrast, Lee [11] demonstrated the consistency and asymptotic normality of the quasi-ML (QML) estimator. Building upon this, Lee [7] proposed a general GMM framework for estimating SLM that combines both the advantages of computational simplicity and efficiency. In general, GMM estimation for SLM can be computationally simpler than the ML or QML methods. Additionally, it may be asymptotically more efficient than the 2SLS estimator and may be asymptotically efficient as the ML estimator.

- Review of Estimation Methods for SLM with Missing Data

To address the issue of missing data, several methods have been developed. Some of these methods include listwise deletion, Pairwise Deletion, Expectation-Maximization (EM) algorithm, and multiple imputation. Detailed information about these methods can be found in works by Dempster et al. [18], Little and Rubin [19], and Yaseen [20]. However, when we deal with spatial models, certain methods, such as listwise deletion, aren't suitable due to the interdependence among the components of the dependent variable vector. Simple deletion of unobserved data from samples may lead to inconsistent estimates, as highlighted by Wang and Lee [2]. Table 1 presents a summary of the proposed methods for handling missing data for SLM in the literature:

Table 1. The Proposed Methods for Handling Missing Data in SLM in Literature

No.	Authors	Missing Data Mechanism*	Variable	The Used Methods
1	LeSage and Pace **[21]**	MAR & MNAR	Response	MLE with EM Algorithm
2	Wolfgang et al. **[22]**	-----	Response	ML and Bayesian Chow-Lin Procedure
3	Wang and Lee **[2]**	MAR	Response	GMM & NLS & TIBG2SLS.nls
4	Boehmke et al. [5]	MCAR & MAR	Response & Covariates	EM Algorithm for Imputation and Estimation
5	Suesse and Zammit-Mangion **[23]**	MAR	Response	Modified EM Algorithm
6	Amitha et al. **[24]**	MAR	Response	Stochastic Regression Imputation
7	Izaguirre **[1]**	MAR	Response	PIBG2SLS.nls & PISTE2SLS.nls & PIABG2SLS.nls
8	Seya et al. **[25]**	MNAR	Response	Modifying BMCMC1
9	Teng et al. **[26]**	MAR	Covariates	MCMCINLA & EM & FIML

Note: MAR: Missing Completely at Random, MAR: Missing at Random, MNAR: Missing not at Random, TIBG2SLS.nls: Best-Generalized 2SLS with NLS Total Imputation, PIBG2SLS.nls: Best-Generalized 2SLS with NLS Partial Imputation, PISTE2SLS.nls: Series-Type Efficient 2SLS with NLS Partial Imputation, BMCMC: Bayesian Markova Chain Monte Carlo, MCMCINLA: Markova Chain Monte Carlo Integrated Nested Laplace Approximation, FIML: Full Information Maximum Likelihood

4. Proposed Estimators

In practice, many empirical researchers "fill in" missing observations, and then estimate the filled equation by conventional methods. Therefore, in line with the approach suggested by Izaguirre [1] and Wang and Lee [2], we replace y_n^m in (3) by its expectation. This expectation can

be expressed as $E(\mathrm{y}_n^m) = \mathrm{E}(\mathrm{J}_n^m \mathrm{y}_n) = \mathrm{J}_n^m \mathrm{E}(\mathrm{y}_n) = \mathrm{J}_n^m S_n^{-1}(\lambda_0)\mathrm{X}_n\beta_0 = F_n(\theta_0)$, where $F_n(\theta_0)$ depends on unknown parameters, $\theta_0 = [\lambda_0 \quad \beta_0']'$. Consequently, it is necessary to estimate θ_0 using a consistent estimator. In this context, both Izaguirre [1] and Wang and Lee [2] used NLS as an initial consistent estimate for imputing missing observations and then estimate the imputed equation by the 2SLS method.

- Modification of Izaguirre's (2021) Approach

Izaguirre's [1] approach revolves around replacing the unobserved observations, y_n^m, in only the spatial lag term of dependent variables in (4) by its expected value, $F_n(\theta_0)$. Through some algebraic manipulations, this replacement leads to the derivation of the following equation:

$$\begin{aligned} \mathrm{y}_n^o &= \lambda_0 \mathrm{J}_n^o \mathrm{W}_n \begin{bmatrix} \mathrm{y}_n^o \\ F_n(\theta_0) \end{bmatrix} + \mathrm{J}_n^o \mathrm{X}_n \beta_0 + \mathrm{J}_n^o u_n, \\ u_n &= \left[\lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m S_n^{-1}(\lambda_0) + \mathrm{I}_n\right]\varepsilon_n \end{aligned} \tag{5}$$

where I_n is an $(n \times n)$ identity matrix. By replacing θ_0 in (5) with an initial consistent estimate, $\hat{\theta}$, and performing further algebraic manipulations. The equation (6) is obtained:

$$\begin{aligned} \mathrm{y}_n^o &= \lambda_0 \mathrm{J}_n^o \mathrm{W}_n \begin{bmatrix} y_n^o \\ F_n(\hat{\theta}) \end{bmatrix} + \mathrm{J}_n^o \mathrm{X}_n \beta_0 + \mathrm{J}_n^o \tilde{u}_n, \\ \tilde{u}_n &= u_n - \lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m S_n^{-1}(\lambda_0) \mathrm{C}_n(\hat{\theta} - \theta_0) \\ &\quad + \lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m \mathrm{R}_n(\hat{\lambda} - \lambda_0) \end{aligned} \tag{6}$$

where

$$\begin{aligned} (\hat{\theta} - \theta_0) &= \begin{bmatrix} (\hat{\lambda} - \lambda_0) \\ (\hat{\beta} - \beta_0) \end{bmatrix}, \\ C_n &= [G_n \mathrm{X}_n \beta_0 \quad \mathrm{X}_n], \\ \mathrm{R}_n &= S_n^{-1}(\hat{\lambda}) G_n \mathrm{X}_n \hat{\beta} - S_n^{-1}(\lambda_0) G_n \mathrm{X}_n \beta_0, \\ G_n &= \mathrm{W}_n S_n^{-1}(\lambda_0) \end{aligned} \tag{7}$$

Unlike the approach of Izaguirre [1] and Wang and Lee [2], we utilized GNLS instead of NLS consistent estimator. The asymptotic distribution of the GNLS estimator is provided in (8).

$$\begin{aligned} \sqrt{n}(\hat{\theta}_{gnls} - \theta_0) &= \left[\frac{1}{n} C_n' B_n' \Omega_{v,n}^{-1} B_n C_n\right]^{-1} \\ &\quad \left[\frac{1}{\sqrt{n}} C_n' B_n' \Omega_{v,n}^{-1} B_n \varepsilon_n\right] + o_p(1) \end{aligned} \tag{8}$$

where $\Omega_{v,n}$ represents the variance-covariance matrix of the error terms for the reduced model in equation (4). This matrix can be rewritten as:

$$\begin{aligned} \Omega_{v,n} &= var(v_n) = var(B_n \varepsilon_n) = \sigma_0^2 B_n B_n'; \\ B_n &= \mathrm{J}_n^o S_n^{-1}(\lambda_0) \end{aligned} \tag{9}$$

Replacing from (8) into error term of (6) and using some algebra, we obtain:

$$\mathrm{y}_n^o = \lambda_0 \mathrm{J}_n^o \mathrm{W}_n \begin{bmatrix} y_n^o \\ F_n(\hat{\theta}_{gnls}) \end{bmatrix} + \mathrm{J}_n^o \mathrm{X}_n \beta_0 + \mathrm{J}_n^o \tilde{u}_n, \qquad \tilde{u}_n = \mathrm{H}_n \varepsilon_n + \mathrm{R}^* \tag{10}$$

where

$$\mathrm{H}_n = \lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m S_n^{-1}(\lambda_0) + \mathrm{I}_n - \lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m$$
$$S_n^{-1}(\lambda_0) \mathrm{C}_n \left[C_n' B_n' \Omega_{v,n}^{-1} B_n C_n \right]^{-1} C_n' B_n' \Omega_{v,n}^{-1} B_n,$$
$$R^* = \lambda_0 \mathrm{W}_n \mathrm{J}_n^{m\prime} \mathrm{J}_n^m \left[\mathrm{R}_n(\hat{\lambda} - \lambda_0) - S_n^{-1}(\lambda_0) \mathrm{C}_n l_{k+1} o_p\left(\frac{1}{\sqrt{n}}\right) \right] \tag{11}$$

and l_{k+1} is an $(k+1 \times 1)$ vector of ones.

Given the above, after imputing θ_0 by $\hat{\theta}_{gnls}$ in (10), we need to estimate the following model by 2SLS:

$$\mathrm{y}_n^o = \mathrm{J}_n^o \tilde{Z}_n \, \theta_0 + \mathrm{J}_n^o \tilde{u}_n \tag{12}$$

where

$$\tilde{Z}_n = [\mathrm{W}_n \tilde{y}_n \quad \mathrm{X}_n], \quad \tilde{y}_n = \begin{bmatrix} \mathrm{y}_n^o \\ F_n(\hat{\theta}_{gnls}) \end{bmatrix}, \qquad \theta_0 = \begin{pmatrix} \lambda_0 \\ \beta_0' \end{pmatrix} \tag{13}$$

- Modification of Wang and Lee's (2013) Approach

Wang and Lee's [2] approach revolve around replacing the unobserved observations, y_n^m, in both the right and left-hand sides of (3) by its expected value, $F_n(\theta_0)$. Through some algebraic manipulations, this replacement leads to the derivation of the following equation:

$$\begin{bmatrix} y_n^o \\ F_n(\hat{\theta}) \end{bmatrix} = \lambda_0 \mathrm{W}_n \begin{bmatrix} y_n^o \\ F_n(\hat{\theta}) \end{bmatrix} + \mathrm{X}_n \beta_0 + \breve{u}_n,$$
$$\breve{u}_n = T_n \varepsilon_n + \mathrm{S}_n(\lambda_0) \mathrm{J}_n^{m\prime} \mathrm{J}_n^m S_n^{-1}(\lambda_0) \mathrm{C}_n(\hat{\theta} - \theta_0) + \mathrm{S}_n(\lambda_0) \mathrm{J}_n^{m\prime} \mathrm{J}_n^m S_n^{-1} \mathrm{R}_n(\hat{\lambda} - \lambda_0),$$
$$T_n = \mathrm{S}_n(\lambda_0) \mathrm{J}_n^{o\prime} B_n \tag{14}$$

By replacing $(\hat{\theta} - \theta_0)$ in (14) with the asymptotic distribution of the GNLS estimator in (8) and performing further algebraic manipulations. The equation (15) is obtained:

$$\breve{u}_n = \mathrm{H}_n^* \varepsilon_n + \mathrm{R}_n^{**} \tag{15}$$

where

$$H_n^* = T_n + S_n(\lambda_0) J_n^{m\prime} J_n^m S_n^{-1}(\lambda_0) C_n \left[\frac{1}{n} C_n' B_n' \Omega_{v,n}^{-1} B_n C_n \right]^{-1} C_n' B_n' \Omega_{v,n}^{-1} B_n,$$
$$R_n^{**} = S_n(\lambda_0) J_n^{m\prime} J_n^m \left[R_n(\hat{\lambda} - \lambda_0) + S_n^{-1}(\lambda_0) C_n l_{k+1} o_p\left(\frac{1}{\sqrt{n}}\right) \right] \tag{16}$$

Following Wang and Lee [2], the Moore-Penrose generalized inverse of $\mathrm{H}_n \mathrm{H}_n' = H_n^{+\prime} H_n^+$ can be used, because $\mathrm{H}_n \mathrm{H}_n'$ is non-invertible, where;

$$H_n^+ = \mathrm{H}'_{n,c}\left(\mathrm{H}_{n,c}\mathrm{H}'_{n,c}\right)^{-1}\left(\mathrm{H}_{n,b}\mathrm{H}'_{n,b}\right)^{-1}\mathrm{H}'_{n,b},$$

$$\mathrm{H}_{n,b} = \mathrm{S}_n(\lambda_0)\left[\mathrm{J}_n^{o\prime} + \mathrm{J}_n^{m\prime}\mathrm{J}_n^m S_n^{-1}(\lambda_0)\mathrm{C}_n\left[\frac{1}{n}C_n'B_n'\Omega_{v,n}^{-1}\,B_nC_n\right]^{-1}C_n'B_n'\Omega_{v,n}^{-1}\right],$$

$$\mathrm{H}_{n,c} = B_n \tag{17}$$

Table 2 provides comprehensive information regarding the notable distinctions between our proposed estimator and the estimators introduced by Izaguirre [1] and Wang and Lee [2]. Table 3 and Table 4 present the final formula of I2SLS estimators with our modification.

5. Monte Carlo Simulation Study

In this section, we compare the performance of the modified estimators with other estimators provided in the literature. The MCS study considers different scenarios that encompass various factors affecting the estimation process. These factors include sample size, spatial weights matrix density level, and the percentage of missing data.

Table 2. Comparison of Our Proposal with Related Estimators in Literature

Estimator	Instrument Matrix		Our Proposals		Izaguirre [1]	Wang and Lee [2]
			Type of Imputation			
IBG2SLS1	Lee [17]	$C_n = [W_nS_n^{-1}(\lambda_0)X_n\beta_0 \quad X_n]$	Partial/ Total	----	Partial	Total
ISTE2SLS	Kelejian et al. [27]	$\hat{Q}_n^{kp} = \left[\sum_{k=0}^{r_n}\hat{\lambda}_0^k\,W_n^{k+1}X_n\hat{\beta}_0 \quad X_n\right]$; r_n is a sequence of natural numbers, such that $r_n \uparrow \infty$.	Partial/ Total	Total	Partial	----
IABG2SLS	Lee [17]	$C_n = [W_nS_n^{-1}(\lambda_0)X_n\beta_0 \quad X_n]$	Partial/ Total	Total	Partial	----
IBG2SLS2	Lee [7]	$Q_{n.1} = \left[W_nS_n^{-1}(\lambda_0) - \frac{1}{n}tr\left(W_nS_n^{-1}(\lambda_0)\right)I_n \quad C_n\right]$	Partial/ Total	Partial/ Total	----	----
IBG2SLS3	Wang and Lee [2]	$Q_{n.2} = T_n'^{+}C_n$; $T_n^+ = B_n'[\,B_n\,B_n'\,]^{-1}[\mathrm{J}_n^oS_n(\lambda_0)'S_n(\lambda_0)\mathrm{J}_n^{o\prime}]^{-1}\,\mathrm{J}_n^oS_n(\lambda_0)'$	Partial/ Total	Partial/ Total	----	----
Imputation Method			GNLS	NLS	NLS	NLS

Table 3. Our Proposal with Partial Imputation

Estimator	Instrument Matrix	Estimator based on Partial Imputation	Notations
PIBG2SLS1	Lee [17]	$\hat{\theta}_{pibg2sls1,gnls} = \left[\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}C_n^o\left(C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}C_n^o\right)^{-1}C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\tilde{Z}_n^o\right]^{-1}$ $\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}C_n^o\left(C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}C_n^o\right)^{-1}C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}y_n^o$	$\Omega_{\tilde{u}_{n,\mathrm{n}}^0} = H_n^o H_n^{o\prime}$, $H_n^o = \mathrm{J}_n^o \mathrm{H}_n$, $C_n^o = \mathrm{J}_n^o C_n$, $\tilde{Z}_n^o = \mathrm{J}_n^o \tilde{Z}_n$
PISTE2SLS	Kelejian et al. [27]	$\hat{\theta}_{pist2sls,gnls} = \left[\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\hat{Q}_n^{okp}\left(\hat{Q}_n^{okp\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\hat{Q}_n^{okp}\right)^{-1}\hat{Q}_n^{okp\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\tilde{Z}_n^o\right]^{-1}$ $\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\hat{Q}_n^{okp}\left(\hat{Q}_n^{okp\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\hat{Q}_n^{okp}\right)^{-1}\hat{Q}_n^{okp\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}y_n^o$	$\hat{Q}_n^{okp} = \mathrm{J}_n^o \hat{Q}_n^{kp}$
PIABG2SLS	Lee [17]	$\hat{\theta}_{piabg2sls,gnls} = \left(C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}C_n^o\right)^{-1}C_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}y_n^o$	-----
PIBG2SLS2	Lee [7]	$\hat{\theta}_{pibg2sls2,gnls} = \left[\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.1}^0\left(Q_{n.1}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.1}^0\right)^{-1}Q_{n.1}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\tilde{Z}_n^o\right]^{-1}$ $\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.1}^0\left(Q_{n.1}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.1}^0\right)^{-1}Q_{n.1}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}y_n^o$	$Q_{n.1}^0 = \mathrm{J}_n^o Q_{n.1}$
PIBG2SLS3	Wang and Lee [2]	$\hat{\theta}_{pibg2sls3,gnls} = \left[\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.2}^0\left(Q_{n.2}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.2}^0\right)^{-1}Q_{n.2}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}\tilde{Z}_n^o\right]^{-1}$ $\tilde{Z}_n^{o\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.2}^0\left(Q_{n.2}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}Q_{n.2}^0\right)^{-1}Q_{n.2}^{0\prime}\Omega_{\tilde{u}_{n,n}^0}^{-1}y_n^o$	$Q_{n.2}^0 = \mathrm{J}_n^o Q_{n.2}$

Table 4. Our Proposal with Total Imputation

Estimator	Instrument Matrix	Estimator based on Total Imputation
TIBG2SLS1	Lee [17]	$\hat{\theta}_{tibg2sls1,gnls} = \left[\tilde{Z}_n' H_n^{+\prime} H_n^+ C_n \left(C_n{}' H_n^{+\prime} H_n^+ C_n\right)^{-1} C_n{}' H_n^{+\prime} H_n^+ \tilde{Z}_n\right]^{-1}$ $\tilde{Z}_n' H_n^{+\prime} H_n^+ C_n \left(C_n{}' H_n^{+\prime} H_n^+ C_n\right)^{-1} C_n{}' H_n^{+\prime} H_n^+ \tilde{y}_n$
TISTE2SLS	Kelejian et al. [27]	$\hat{\theta}_{tist2sls,gnls} = \left[\tilde{Z}_n' H_n^{+\prime} H_n^+ \hat{Q}_n^{kp}\left(\hat{Q}_n^{kp\prime} H_n^{+\prime} H_n^+ \hat{Q}_n^{kp}\right)^{-1}\hat{Q}_n^{kp\prime} H_n^{+\prime} H_n^+ \tilde{Z}_n\right]^{-1}$ $\tilde{Z}_n' H_n^{+\prime} H_n^+ \hat{Q}_n^{kp}\left(\hat{Q}_n^{kp\prime} H_n^{+\prime} H_n^+ \hat{Q}_n^{kp}\right)^{-1}\hat{Q}_n^{kp\prime} H_n^{+\prime} H_n^+ \tilde{y}_n$
TIABG2SLS	Lee [17]	$\hat{\theta}_{tiabg2sls,gnls} = \left(C_n{}' H_n^{+\prime} H_n^+ C_n\right)^{-1} C_n^{o\prime} H_n^{+\prime} H_n^+ \tilde{y}_n$
TIBG2SLS2	Lee [7]	$\hat{\theta}_{tibg2sls2,gnls} = \left[\tilde{Z}_n' H_n^{+\prime} H_n^+ Q_{n.1}\left(Q_{n.1}{}' H_n^{+\prime} H_n^+ Q_{n.1}\right)^{-1} Q_{n.1}{}' H_n^{+\prime} H_n^+ \tilde{Z}_n\right]^{-1}$ $\tilde{Z}_n' H_n^{+\prime} H_n^+ Q_{n.1}\left(Q_{n.1}{}' H_n^{+\prime} H_n^+ Q_{n.1}\right)^{-1} Q_{n.1}{}' H_n^{+\prime} H_n^+ \tilde{y}_n$
TIBG2SLS3	Wang and Lee [2]	$\hat{\theta}_{tibg2sls3,gnls} = \left[\tilde{Z}_n' H_n^{+\prime} H_n^+ Q_{n.2}\left(Q_{n.2}{}' H_n^{+\prime} H_n^+ Q_{n.2}\right)^{-1} Q_{n.2}{}' H_n^{+\prime} H_n^+ \tilde{Z}_n\right]^{-1}$ $\tilde{Z}_n' H_n^{+\prime} H_n^+ Q_{n.2}\left(Q_{n.2}{}' H_n^{+\prime} H_n^+ Q_{n.2}\right)^{-1} Q_{n.2}{}' H_n^{+\prime} H_n^+ \tilde{y}_n$

- Simulation Algorithm

The simulation algorithm can be detailed in the following steps:

1. Determine the factors of our MCS study: We generate two explanatory variables, drawing their observations from a normal distribution (0, 10), while the error terms are defined as $\varepsilon_i \sim i.i.d\ \mathrm{N}(0, \sigma_\varepsilon^2 = 1)$. To assess the impact of sample size, we employ two different values: $n = (50, 100)$, which are deemed sufficient as larger samples yield only marginal improvements in results. The coefficient of spatial dependence, λ, is set to a positive medium value of 0.4. Additionally, we assign a value of 1 to β's, the parameters associated with the explanatory variables.
2. Generate spatial weights matrices by using the following steps:
   - We first draw the coordinates (x_i, y_i) of n spatial units from a uniform distribution (10, 30).
   - Using these coordinates, we calculate the Euclidian distance between each pair of spatial units, creating a full $n \times n$ distance matrix.
   - We then construct the spatial weights matrix by using the K-nearest neighbor structure as follows

$$\mathrm{w_{ij}} = \begin{cases} 1 & \text{if } d_{ij} \leq \ddot{\mathrm{d}}_{i(n)} \forall\ i \neq \mathrm{j} \text{ and } i, \mathrm{j} = 1, \dots, n \\ 0 & \text{otherwise} \end{cases} \tag{18}$$

   where $\mathrm{w_{ij}}$ is an element in the matrix W_n that is predefined to represent the interaction strength between region (i) and region (j) within a set of geographical units, d_{ij} denotes the distance between the two regions (i) and (j), and $\ddot{\mathrm{d}}_{i(n)}$ refers to a critical distance threshold (distance to the K[th] nearest neighbor) [[28]-[30]]. For our analysis, we use $\mathrm{K} = \frac{n}{5}$ for W_1 and $K = \frac{n}{25}$ for W_2.
3. For each unique simulation, we compute the dependent variable (y) using a given weighting matrix, $\mathrm{W}_m, m = 1,2$, and a reduced form of the SLM equation. We then generate a random sample of missing data in the dependent variable at two distinct levels: 10% and 25%. This allows us to examine how the severity of missing data affects model performance and estimation accuracy.
4. Our estimation process begins with a crucial preparatory step: we use the observed data to obtain initial consistent estimates of θ (the parameter vector) through both NLS and GNLS methods. These initial estimates serve as the foundation for inducing the missing observations in our dataset.
5. Once we have these initial estimates and have inputted the missing data, we proceed with our main estimation process:
   - For each set of imputed data, we apply a suite of estimation methods. These include: IBG2SLS1, IST2SLS, IABG2SLS, IBG2SLS2, and IBG2SLS3. We implement these methods using both total and partial imputation approaches.

- To ensure the robustness of our results, we repeat this entire process 1000 times for each unique design described in our study. This repetition helps to mitigate the impact of random variations and provides a more reliable basis for comparing the different estimation methods.

6. After completing all simulations, we calculate two key performance metrics for each model and estimator across the 1000 simulations, as mentioned in Saguatti [31]:
   - The average of coefficient bias, which indicates how far the estimated coefficients tend to deviate from their true values, is calculated as follows:

$$\text{Bias}\left(\bar{\hat{\theta}}\right) = \bar{\hat{\theta}} - \theta \tag{19}$$

   where $\bar{\hat{\theta}}$ is the average of estimates for the coefficient θ over the No. of replications (R), it is calculated as:

$$\bar{\hat{\theta}} = \frac{1}{R}\sum_{r=1}^{R} \hat{\theta}_r \tag{20}$$

   - The RMSE, or standard error, which provides a measure of the overall accuracy of the estimates, taking into account both bias and variance, is defined as:

$$\text{RMSE}\left(\bar{\hat{\theta}}\right) = \left[\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_r - \theta\right)^2\right]^{1/2} \tag{21}$$

Based on the above algorithm, our study encompasses 136 unique simulations for each model and estimator under investigation. Fig. 1 outlines the main algorithm of our MCS study.

- Simulation Results

The finite sample properties of all estimators are presented in Tables 5 through 8, with each table corresponding to a specific spatial weights matrix and sample size (n). These tables showcase bias and RMSE for the coefficient estimates across different values of missing data percentage (∝). Additionally, the tables also include overall RMSE values for each model. To visually represent the interplay of key factors affecting estimator performance, Fig. 2 to Fig. 5 illustrate the effect of spatial weights matrix density, sample size, and missing data rates on the model RMSE for each estimator.

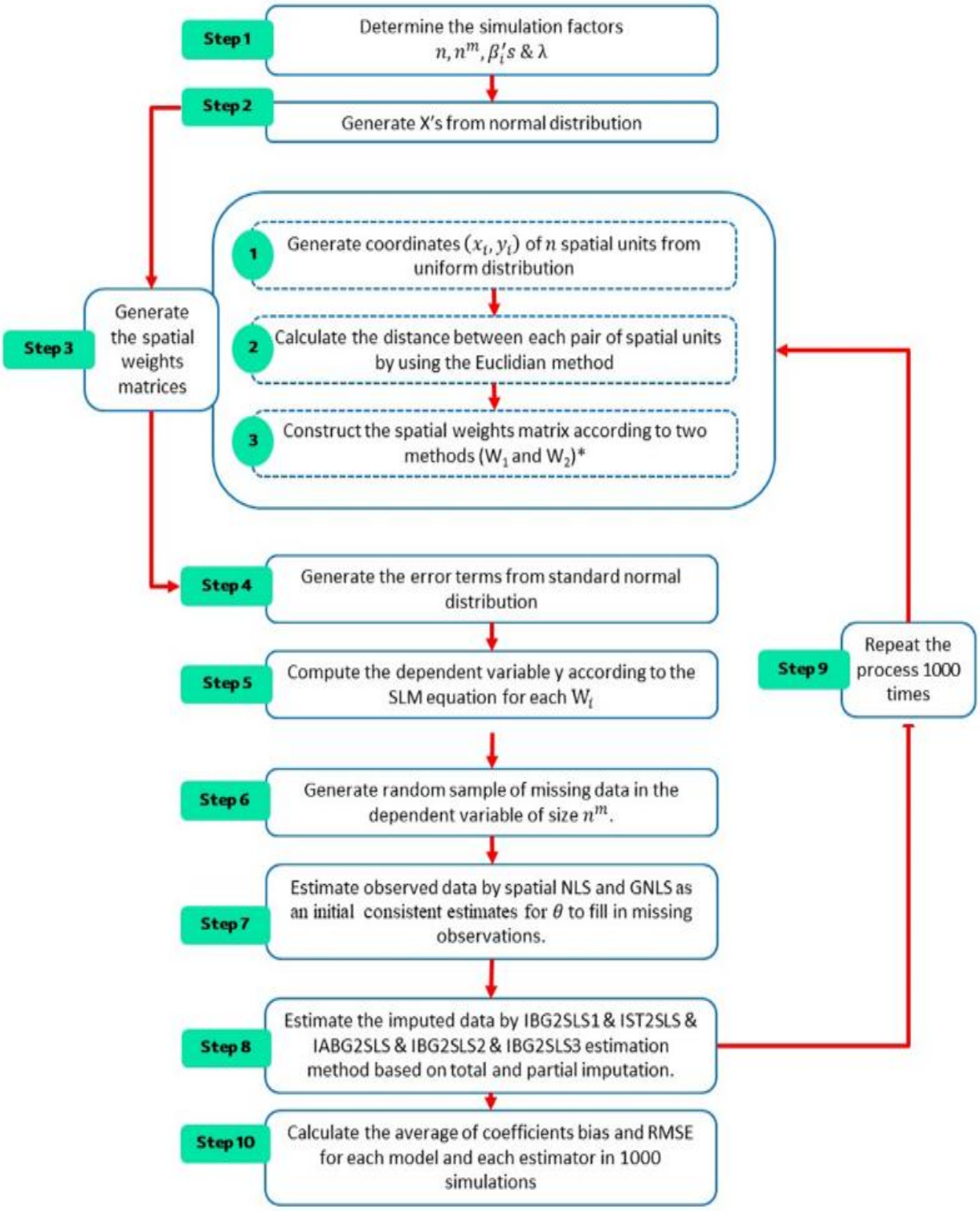

Figure 1. Algorithm of Our MCS

Table 5. MCS Results; $n = 50, and\ \mathrm{W}_1$

Estimator	$\propto = 10\%, n^o = 45$							$\propto = 25\%, n^o = 38$						
	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE
PIABG2SLS.nls	0.001	0.029	0.000	0.016	0.000	0.016	1.205	0.001	0.031	0.001	0.017	0.001	0.017	1.409
TIABG2SLS.nls	0.001	0.029	0.000	0.016	0.000	0.016	1.205	0.001	0.031	0.001	0.017	0.001	0.017	1.409
PIABG2SLS.gnls	0.008	0.037	0.001	0.016	0.000	0.016	1.018	0.019	0.055	0.000	0.017	0.002	0.017	1.078
TIABG2SLS.gnls	0.001	0.031	0.000	0.016	0.001	0.017	1.002	0.005	0.043	0.003	0.019	0.001	0.019	1.052
PIBG2SLS1.gnls	0.002	0.029	0.000	0.016	0.000	0.016	0.993	0.007	0.035	0.000	0.017	0.001	0.017	1.007
TIBG2SLS1.gnls	0.005	0.035	0.001	0.016	0.001	0.017	1.022	0.020	0.064	0.004	0.019	0.001	0.019	1.190
TISTE2SLS.nls	0.000	0.029	0.000	0.016	0.000	0.016	1.200	0.001	0.031	0.001	0.017	0.001	0.017	1.399
PISTE2SLS.gnls	0.002	0.029	0.000	0.016	0.000	0.016	0.994	0.007	0.035	0.000	0.017	0.001	0.017	1.007
TISTE2SLS.gnls	0.005	0.035	0.001	0.016	0.001	0.017	1.022	0.020	0.064	0.004	0.019	0.001	0.019	1.190
PIBG2SLS2.nls	0.000	0.029	0.000	0.016	0.000	0.016	0.993	0.000	0.031	0.001	0.017	0.001	0.017	0.993
TIBG2SLS2.nls	0.000	0.029	0.000	0.016	0.000	0.016	0.992	0.001	0.031	0.001	0.017	0.001	0.017	0.993
PIBG2SLS2.gnls	0.001	0.029	0.000	0.016	0.000	0.016	0.994	0.006	0.035	0.000	0.017	0.001	0.017	1.007
TIBG2SLS2.gnls	0.005	0.035	0.001	0.016	0.001	0.017	1.021	0.018	0.061	0.004	0.019	0.001	0.019	1.173
PIBG2SLS3.nls	0.000	0.029	0.000	0.016	0.000	0.016	0.992	0.001	0.031	0.001	0.017	0.001	0.017	0.993
TIBG2SLS3.nls	0.001	0.029	0.000	0.016	0.000	0.016	0.993	0.000	0.032	0.001	0.017	0.001	0.017	0.995
PIBG2SLS3.gnls	0.002	0.029	0.000	0.016	0.000	0.016	0.994	0.007	0.035	0.000	0.017	0.001	0.017	1.007
TIBG2SLS3.gnls	0.000	0.039	0.000	0.017	0.001	0.017	1.031	0.015	0.120	0.002	0.031	0.004	0.030	1.243

Table 6. MCS Results; $n = 100, and\ W_1$

Estimator	$\propto = 10\%, n^o = 90$							$\propto = 25\%, n^o = 75$						
	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE
PIABG2SLS.nls	0.002	0.028	0.000	0.011	0.000	0.011	1.191	0.000	0.032	0.000	0.012	0.000	0.012	1.308
TIABG2SLS.nls	0.002	0.028	0.000	0.011	0.000	0.011	1.191	0.000	0.032	0.000	0.012	0.000	0.012	1.308
PIABG2SLS.gnls	0.010	0.036	0.000	0.011	0.000	0.011	1.007	0.021	0.057	0.001	0.012	0.001	0.012	1.054
TIABG2SLS.gnls	0.003	0.030	0.000	0.011	0.000	0.011	0.996	0.011	0.047	0.001	0.013	0.001	0.013	1.041
PIBG2SLS1.gnls	0.003	0.028	0.000	0.011	0.000	0.011	0.994	0.007	0.034	0.001	0.012	0.000	0.012	1.007
TIBG2SLS1.gnls	0.004	0.034	0.001	0.011	0.001	0.011	1.006	0.028	0.074	0.001	0.013	0.002	0.013	1.161
TISTE2SLS.nls	0.002	0.028	0.000	0.011	0.000	0.011	1.190	0.000	0.031	0.000	0.012	0.000	0.012	1.300
PISTE2SLS.gnls	0.003	0.028	0.000	0.011	0.000	0.011	0.994	0.007	0.034	0.001	0.012	0.000	0.012	1.007
TISTE2SLS.gnls	0.004	0.034	0.001	0.011	0.001	0.011	1.006	0.028	0.074	0.001	0.013	0.002	0.013	1.161
PIBG2SLS2.nls	0.001	0.028	0.000	0.011	0.000	0.011	0.993	0.000	0.031	0.000	0.012	0.000	0.012	0.999
TIBG2SLS2.nls	0.002	0.028	0.000	0.011	0.000	0.011	0.993	0.000	0.031	0.000	0.012	0.000	0.012	0.999
PIBG2SLS2.gnls	0.002	0.028	0.000	0.011	0.000	0.011	0.994	0.007	0.034	0.001	0.012	0.000	0.012	1.007
TIBG2SLS2.gnls	0.005	0.034	0.001	0.011	0.001	0.011	1.006	0.026	0.070	0.001	0.013	0.002	0.013	1.141
PIBG2SLS3.nls	0.002	0.028	0.000	0.011	0.000	0.011	0.993	0.000	0.031	0.000	0.012	0.000	0.012	0.999
TIBG2SLS3.nls	0.002	0.028	0.000	0.011	0.000	0.011	0.994	0.000	0.032	0.000	0.012	0.000	0.012	1.000
PIBG2SLS3.gnls	0.003	0.028	0.000	0.011	0.000	0.011	0.994	0.007	0.034	0.001	0.012	0.000	0.012	1.007
TIBG2SLS3.gnls	0.001	0.034	0.000	0.011	0.000	0.011	1.006	0.022	0.185	0.003	0.021	0.002	0.022	1.190

Table 7. MCS Results; $n = 50, and\ \mathrm{W}_2$

Estimator	$\propto = 10\%, n^o = 45$							$\propto = 25\%, n^o = 38$						
	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE
PIABG2SLS.nls	0.000	0.013	0.000	0.016	0.000	0.017	1.258	0.000	0.014	0.001	0.018	0.000	0.018	1.422
TIABG2SLS.nls	0.000	0.013	0.000	0.016	0.000	0.017	1.258	0.000	0.014	0.001	0.018	0.000	0.017	1.422
PIABG2SLS.gnls	0.004	0.016	0.001	0.017	0.000	0.017	1.109	0.010	0.023	0.001	0.019	0.001	0.018	1.154
TIABG2SLS.gnls	0.002	0.015	0.001	0.018	0.001	0.018	1.105	0.004	0.018	0.007	0.024	0.005	0.023	1.132
PIBG2SLS1.gnls	0.001	0.013	0.000	0.016	0.000	0.017	1.094	0.003	0.015	0.000	0.017	0.001	0.017	1.102
TIBG2SLS1.gnls	0.001	0.014	0.001	0.017	0.001	0.017	1.108	0.004	0.021	0.006	0.023	0.005	0.022	1.174
TISTE2SLS.nls	0.000	0.013	0.000	0.016	0.000	0.017	1.255	0.000	0.013	0.001	0.017	0.001	0.017	1.420
PISTE2SLS.gnls	0.001	0.013	0.000	0.016	0.000	0.017	1.094	0.003	0.015	0.000	0.017	0.001	0.017	1.102
TISTE2SLS.gnls	0.001	0.014	0.001	0.017	0.001	0.017	1.108	0.004	0.021	0.006	0.023	0.005	0.022	1.174
PIBG2SLS2.nls	0.001	0.013	0.000	0.016	0.000	0.017	1.092	0.001	0.013	0.001	0.017	0.001	0.017	1.089
TIBG2SLS2.nls	0.001	0.013	0.000	0.016	0.000	0.017	1.092	0.000	0.013	0.001	0.017	0.001	0.017	1.089
PIBG2SLS2.gnls	0.000	0.013	0.001	0.016	0.000	0.017	1.094	0.002	0.015	0.000	0.017	0.001	0.017	1.102
TIBG2SLS2.gnls	0.002	0.014	0.001	0.017	0.001	0.017	1.109	0.004	0.020	0.006	0.023	0.005	0.022	1.170
PIBG2SLS3.nls	0.000	0.013	0.000	0.016	0.000	0.017	1.092	0.000	0.014	0.001	0.017	0.000	0.017	1.090
TIBG2SLS3.nls	0.000	0.013	0.000	0.016	0.000	0.017	1.094	0.000	0.014	0.001	0.018	0.000	0.018	1.096
PIBG2SLS3.gnls	0.001	0.013	0.000	0.016	0.000	0.017	1.095	0.003	0.015	0.000	0.018	0.001	0.018	1.104
TIBG2SLS3.gnls	0.000	0.015	0.001	0.017	0.001	0.017	1.117	0.001	0.024	0.004	0.023	0.002	0.022	1.189

Table 8. MCS Results; $n = 100, and\ W_2$

Estimator	$\propto = 10\%, n^o = 90$							$\propto = 25\%, n^o = 75$						
	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE	Bias $\hat{\lambda}$	RMSE $\hat{\lambda}$	Bias $\hat{\beta}_1$	RMSE $\hat{\beta}_1$	Bias $\hat{\beta}_2$	RMSE $\hat{\beta}_2$	Model RMSE
PIABG2SLS.nls	0.001	0.012	0.000	0.011	0.000	0.011	1.199	0.000	0.013	0.000	0.012	0.000	0.013	1.356
TIABG2SLS.nls	0.001	0.012	0.000	0.011	0.000	0.011	1.198	0.000	0.013	0.000	0.012	0.000	0.012	1.357
PIABG2SLS.gnls	0.006	0.015	0.001	0.011	0.001	0.011	1.061	0.012	0.024	0.002	0.013	0.001	0.013	1.099
TIABG2SLS.gnls	0.002	0.014	0.000	0.011	0.000	0.011	1.055	0.002	0.016	0.003	0.014	0.003	0.015	1.073
PIBG2SLS1.gnls	0.001	0.012	0.000	0.011	0.000	0.011	1.051	0.004	0.014	0.001	0.012	0.000	0.012	1.059
TIBG2SLS1.gnls	0.002	0.013	0.001	0.011	0.001	0.011	1.057	0.010	0.024	0.004	0.014	0.004	0.015	1.119
TISTE2SLS.nls	0.000	0.012	0.000	0.011	0.000	0.011	1.195	0.000	0.013	0.000	0.012	0.000	0.012	1.357
PISTE2SLS.gnls	0.001	0.012	0.000	0.011	0.000	0.011	1.051	0.004	0.014	0.001	0.012	0.000	0.012	1.059
TISTE2SLS.gnls	0.002	0.013	0.001	0.011	0.001	0.011	1.057	0.010	0.024	0.004	0.014	0.004	0.015	1.119
PIBG2SLS2.nls	0.001	0.012	0.000	0.011	0.000	0.011	1.050	0.001	0.013	0.000	0.012	0.000	0.012	1.053
TIBG2SLS2.nls	0.000	0.012	0.000	0.011	0.000	0.011	1.050	0.000	0.013	0.000	0.012	0.000	0.012	1.053
PIBG2SLS2.gnls	0.000	0.012	0.000	0.011	0.000	0.011	1.051	0.003	0.014	0.001	0.012	0.000	0.012	1.059
TIBG2SLS2.gnls	0.003	0.013	0.001	0.011	0.001	0.011	1.058	0.010	0.024	0.004	0.014	0.004	0.015	1.116
PIBG2SLS3.nls	0.000	0.012	0.000	0.011	0.000	0.011	1.050	0.000	0.013	0.000	0.012	0.000	0.012	1.054
TIBG2SLS3.nls	0.000	0.012	0.000	0.011	0.000	0.011	1.051	0.000	0.013	0.000	0.012	0.000	0.012	1.056
PIBG2SLS3.gnls	0.001	0.012	0.000	0.011	0.000	0.011	1.051	0.004	0.014	0.001	0.012	0.000	0.012	1.060
TIBG2SLS3.gnls	0.001	0.014	0.001	0.011	0.001	0.011	1.060	0.004	0.022	0.002	0.014	0.002	0.015	1.103

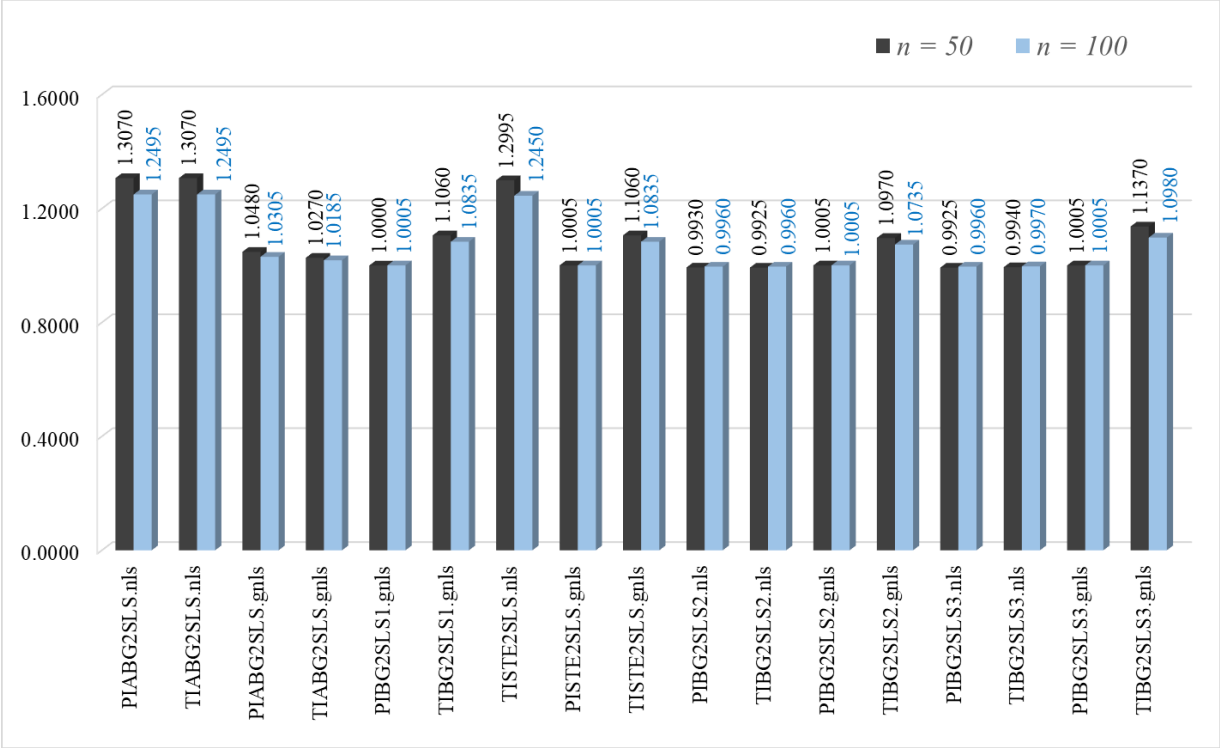

Figure 2. Average RMSE of W_1- based Models Across Various Rates of Missing Data by Sample Sizes

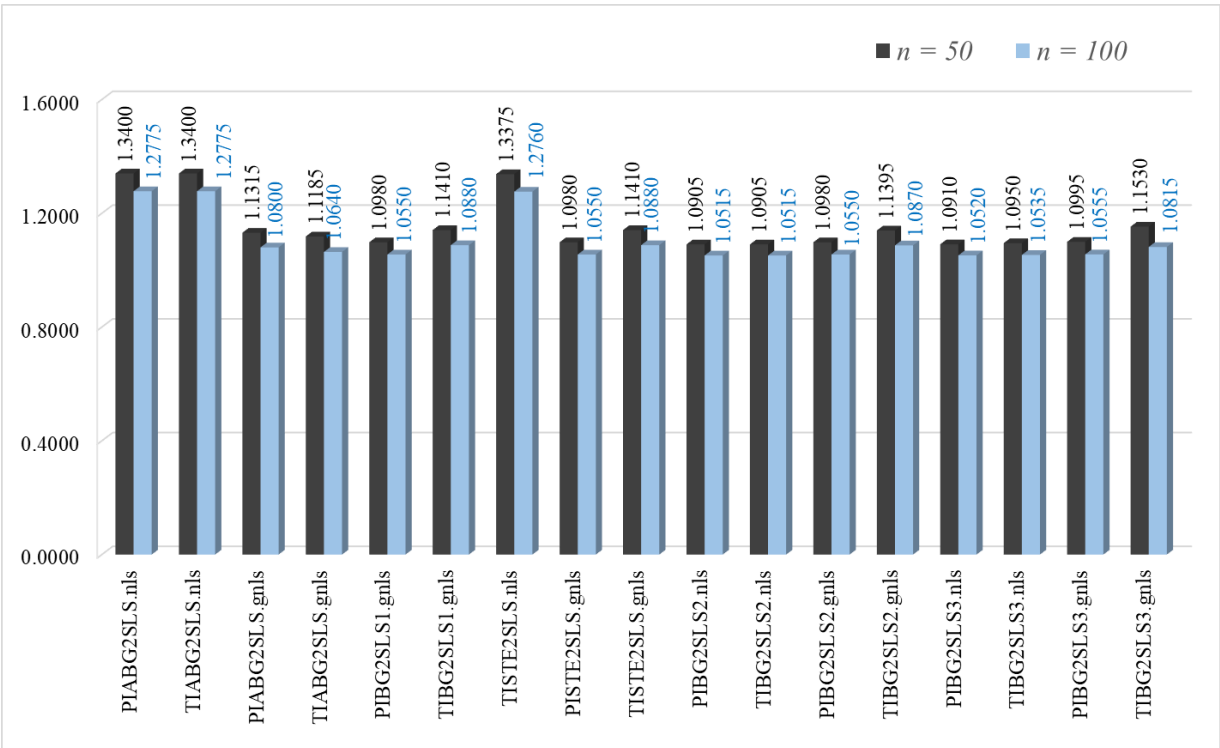

Figure 3. Average RMSE of W_2- based Models Across Various Rates of Missing Data by Sample Sizes

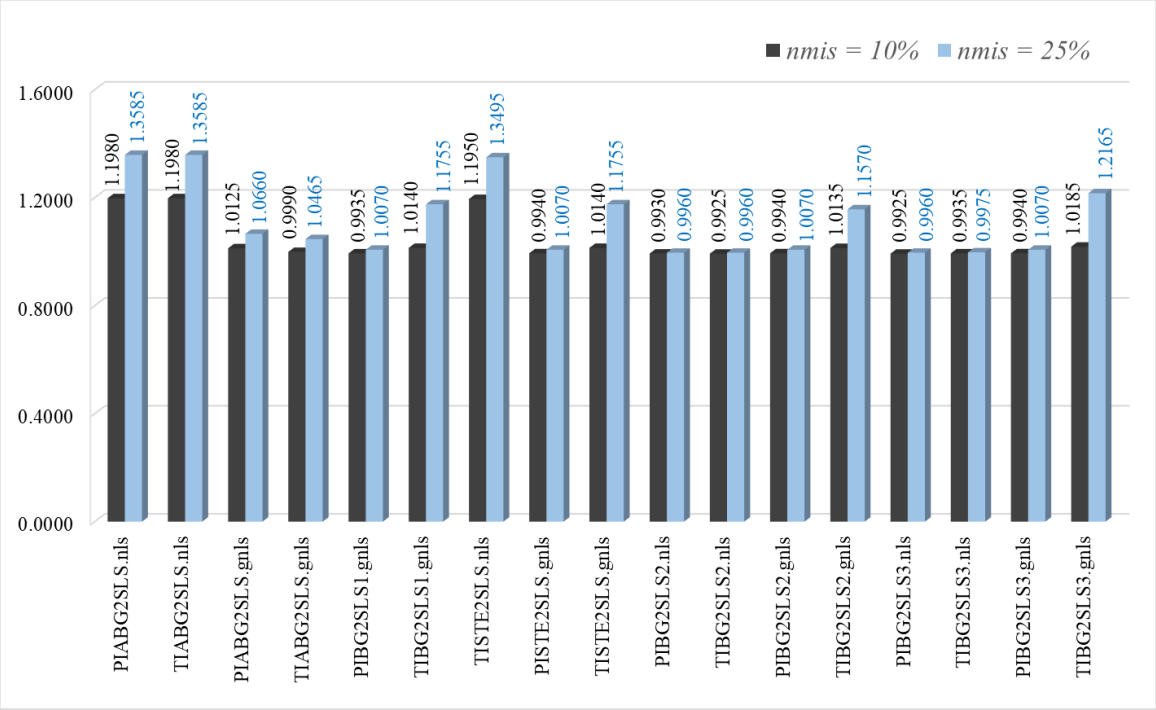

Figure 4. Average RMSE of W_1- based Models Across Various Sample Sizes by Missing Data Rates

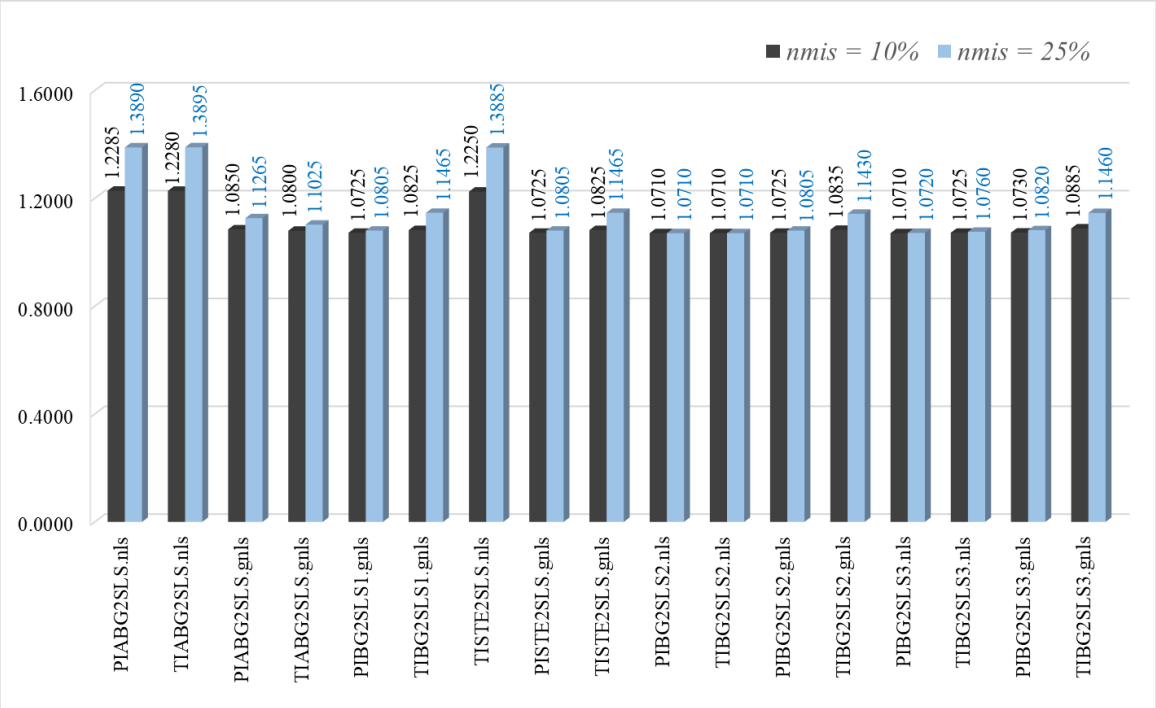

Figure 5. Average RMSE of W_2- based Models Across Various Sample Sizes by Missing Data Rates

6. Discussion

Regarding the sample size effect, there isn't a clear trend showing that smaller sample sizes consistently yield higher bias and RMSE compared to larger sample sizes. The differences between $n = 50$ and $n = 100$ is generally small, with some cases showing slightly better performance for $n = 100$, especially in terms of RMSE and for W_2- based Models.

As for the missing rate impact, the results show that higher missing data rates (25% vs. 10%) lead to increased bias and RMSE across almost all estimators. This aligns with the expected challenges of dealing with more missing data in spatial lag models.

In the context of estimator coefficients bias and RMSE, all estimators demonstrate relatively strong performance.

Moreover, each of our proposed estimators exhibits superior model efficiency (Overall RMSE) compared to Izaguirre's estimator (PIABG2SLS.nls). For a more comprehensive understanding of these results, please refer to Fig. 2 through Fig. 5.

Consistent with Smith [32] and Farber et al. [33], our results indicate that the density of the spatial weights matrix has a negative impact on inference. Specifically, we observe that most of the bias and RMSE results of spatial dependence parameter for the denser spatial weights matrix (W_1) are greater than those for the less dense matrix (W_2). This finding aligns with the notion that increased spatial connectivity can potentially lead to more pronounced estimation challenges in spatial econometric models. Despite this fact, the overall results produced by both matrices are remarkably similar. This finding suggests that, in our specific study context, the density of the spatial weights matrix does not substantially alter the outcomes of the estimation process. Such consistency across different matrix densities indicates robustness in the presented estimators that may be valuable for practitioners dealing with varying levels of spatial connectivity in their data.

Finaly, both total and partial imputation approaches tend to produce similar results, partial imputation demonstrated superior performance in certain scenarios. This superior performance of partial imputation was particularly pronounced for the IBG2SLS1.gnls estimator. The advantage of partial imputation likely stems from it gives rise to the possibility of working only with complete data. Estimators based on partial imputation only require knowing the spatial lag for the observed dependent variable. This characteristic allows for more efficient and potentially more accurate estimation, as it leverages the available observed data more effectively.

For future work, we can develop a robust M-estimator for the spatial lag model when the dataset contains outliers as presented in several regression modes such as [34], [35], [36], [37], and [38].

7. Conclusions

This work has addressed a critical challenge in spatial econometrics: the accurate estimation of SLM in the presence of missing data in the dependent variable. Building upon the works of Izaguirre [1] and Wang and Lee [2] in this field, we have introduced and evaluated several modifications to imputation-based 2SLS estimators. Our key contributions include: (1) proposing the GNLS estimator as an alternative imputation method to the previously used NLS approach in the literature; (2) incorporating the additional IMs, including those used by Lee [7] for BGMM estimator in SLM with complete data, and those employed by Wang and Lee [2] for GMM estimator in SLM with missing data; (3) performing all estimators using both partial and total imputations strategies to balance computational efficiency with estimation accuracy; and (4) conducting a comprehensive MCS study to compare the performance of our modified estimators (see Table 2) across various spatial configurations, sample sizes, and missing data proportions.

Our findings reveal several important insights: (1) partial imputation consistently outperforms total imputation in most cases; (2) all estimators show good performance with low bias and RMSE; (3) all of our proposed estimators outperform Izaguirre's estimator (PIABG2SLS.nls) in terms of model RMSE; (4) although, denser matrix generally yields higher bias and RMSE for spatial dependence parameter than less dense matrix, the results from both matrices are remarkably similar, indicating the robustness of the estimators across different spatial connectivity levels; (5) higher missing data rates (25% *vs*. 10%) lead to slightly increased bias and RMSE for most estimators; and (6) no clear trend that smaller sample sizes yields higher bias or RMSE.

In conclusion, we can say that the proposed estimators demonstrate improved efficiency compared to existing methods in the literature, making them valuable alternatives when dealing with missing data in SLM. Notably, these estimators exhibit robustness across varying spatial weight matrix densities, ensuring their reliability in diverse spatial contexts - from sparsely connected regions to densely interconnected areas. While increasing rates of missing data inevitably affect estimation efficiency, our methods effectively mitigate these impacts, providing dependable results even in scenarios with up to 25% missing data. These findings collectively enhance the toolkit available to researchers and practitioners, enabling more accurate and reliable spatial data analysis across a broad spectrum of real-world applications.

While our study has advanced the understanding of SLM estimation with missing data, several avenues for future research remain, including: (1) investigation of non-random missing data patterns and their impact on estimator performance, and (2) application of these methods to real-world datasets across various disciplines to further validate their practical utility.

Conflicts of Interest: The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] A. Izaguirre, Estimation of Spatial Lag Model Under Random Missing Data in the Dependent Variable. Two Stage Estimator with Imputation, Economia 44 (2021), 1–19. https://doi.org/10.18800/economia.202101.001.

[2] W. Wang, L. Lee, Estimation of Spatial Autoregressive Models with Randomly Missing Data in the Dependent Variable, Econom. J. 16 (2013), 73–102. https://doi.org/10.1111/j.1368-423X.2012.00388.x.

[3] P. Krugman, Geography and Trade, MIT Press, Cambridge, 1991.

[4] M. Fujita, P. Krugman, A. Venables, The Spatial Economics: Cities, Regions and International Trade, MIT Press, Cambridge, 1999.

[5] F.J. Boehmke, E.U. Schilling, J.C. Hays, Missing Data in Spatial Regression, in: Midwest Political Science Association Conference, 16-19, April 2015.

[6] T. Yokoi, Spatial Lag Dependence in the Presence of Missing Observations, Ann. Reg. Sci. 60 (2018), 25–40. https://doi.org/10.1007/s00168-015-0737-2.

[7] L. Lee, GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models, J. Econom. 137 (2007), 489–514. https://doi.org/10.1016/j.jeconom.2005.10.004.

[8] T. Rüttenauer, Spatial Regression Models: A Systematic Comparison of Different Model Specifications Using Monte Carlo Experiments, Sociol. Methods Res. 51 (2022), 728–759. https://doi.org/10.1177/0049124119882467.

[9] H.H. Kelejian, I.R. Prucha, A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances, J. Real Estate Finance Econ. 17 (1998), 99–121. https://doi.org/10.1023/A:1007707430416.

[10] H.H. Kelejian, I.R. Prucha, A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model, Int. Econ. Rev. 40 (1999), 509–533. https://doi.org/10.1111/1468-2354.00027.

[11] L.-F. Lee, Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models, Econometrica 72 (2004), 1899–1925. https://doi.org/10.1111/j.1468-0262.2004.00558.x.

[12] I. Mattsson, J. Lyhagen, Modeling Spatial Regimes With Smooth Transitions, Int. Reg. Sci. Rev. 48 (2025), 38–61. https://doi.org/10.1177/01600176241237180.

[13] L. Anselin, Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity, Geogr. Anal. 20 (1988), 1–17. https://doi.org/10.1111/j.1538-4632.1988.tb00159.x.

[14] L. Anselin, J.L. Gallo, H. Jayet, Spatial Panel Econometrics, in: L. Mátyás, P. Sevestre (Eds.), The Econometrics of Panel Data, Springer, Berlin, Heidelberg, 2008: pp. 625–660. https://doi.org/10.1007/978-3-540-75892-1_19.

[15] K. Ord, Estimation Methods for Models of Spatial Interaction, J. Am. Stat. Assoc. 70 (1975), 120–126. https://doi.org/10.1080/01621459.1975.10480272.

[16] O. Smirnov, L. Anselin, Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach, Comput. Stat. Data Anal. 35 (2001), 301–319. https://doi.org/10.1016/S0167-9473(00)00018-9.

[17] L. Lee, Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with

Autoregressive Disturbances, Econom. Rev. 22 (2003), 307–335. https://doi.org/10.1081/ETC-120025891.

[18] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data Via the EM Algorithm, J. R. Stat. Soc. Ser. B Stat. Methodol. 39 (1977), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

[19] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, Wiley, New York, 2002.

[20] A.S.A. Yaseen, Fractional Imputation Methods for Longitudinal Data Analysis, Thesis, Faculty of Economics and Political Science, Cairo University, Egypt, 2014.

[21] J.P. LeSage, R.K. Pace, Models for Spatially Dependent Missing Data, J. Real Estate Finance Econ. 29 (2004), 233–254. https://doi.org/10.1023/B:REAL.0000035312.82241.e4.

[22] P. Wolfgang, C. Llano, R. Sellner, Bayesian Methods for Completing Data in Spatial Models, Rev. Econ. Anal. 2 (2010), 194–214. https://doi.org/10.15353/rea.v2i2.1472.

[23] T. Suesse, A. Zammit-Mangion, Computational Aspects of the EM Algorithm for Spatial Econometric Models with Missing Data, J. Stat. Comput. Simul. 87 (2017), 1767–1786. https://doi.org/10.1080/00949655.2017.1286495.

[24] P. Amitha, V.S. Binu, B. Seena, Estimation of Missing Values in Aggregate Level Spatial Data, Clin. Epidemiol. Glob. Health 9 (2021), 304–309. https://doi.org/10.1016/j.cegh.2020.10.003.

[25] H. Seya, M. Tomari, S. Uno, Parameter Estimation in Spatial Econometric Models with Non-Random Missing Data, Appl. Econ. Lett. 28 (2021), 440–446. https://doi.org/10.1080/13504851.2020.1758618.

[26] J. Teng, S. Ding, X. Shi, H. Zhang, X. Hu, MCMCINLA Estimation of Missing Data and Its Application to Public Health Development in China in the Post-Epidemic Era, Entropy 24 (2022), 916. https://doi.org/10.3390/e24070916.

[27] H.H. Kelejian, I.R. Prucha, Y. Yuzefovich, Instrumental Variable Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances: Large and Small Sample Results, in: Advances in Econometrics, Emerald, Bingley, 2004: pp. 163–198. https://doi.org/10.1016/S0731-9053(04)18005-5.

[28] M.M. Abdelwahab, O.A. Shalaby, H.E. Semary, M.R. Abonazel, Driving Factors of NOx Emissions in China: Insights from Spatial Regression Analysis, Atmosphere 15 (2024), 793. https://doi.org/10.3390/atmos15070793.

[29] Y. Song, A. Cibin, Optimizing Spatial Weight Matrices in Spatial Econometrics: A Graph-Theoretic Approach Based on Shortest Path Algorithms: A New York City Application of Crime and Economic Indicators, Int. Rev. Spat. Plan. Sustain. Dev. 12 (2024), 181–200. https://doi.org/10.14246/irspsd.12.2_181.

[30] J. Dubé, D. Legros, Spatial Econometrics Using Micro-Data, Wiley, 2014.

[31] A. Saguatti, Modeling the Spatial Dynamics of Economic Models, Thesis, Mater Studiorum Università di Bologna, 2013. https://doi.org/10.6092/UNIBO/AMSDOTTORATO/5978.

[32] T.E. Smith, Estimation Bias in Spatial Models with Strongly Connected Weight Matrices, Geogr. Anal. 41 (2009), 307–332. https://doi.org/10.1111/j.1538-4632.2009.00758.x.

[33] S. Farber, A. Páez, E. Volz, Topology, Dependency Tests and Estimation Bias in Network Autoregressive Models, in: A. Páez, J. Gallo, R.N. Buliung, S. Dall'erba (Eds.), Progress in Spatial

Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010: pp. 29–57. https://doi.org/10.1007/978-3-642-03326-1_3.

[34] M.R. Abonazel, A.H. Youssef, E.G. Ahmed, On Robust M, S and MM Estimations for the Poisson Fixed Effects Panel Model with Outliers: Simulation and Applications, J. Stat. Comput. Simul. (2025). https://doi.org/10.1080/00949655.2024.2449102.

[35] A.A. El-Sheikh, M.C. Ali, M.R. Abonazel, Development of Two Methods for Estimating High-Dimensional Data in the Case of Multicollinearity and Outliers, Int. J. Anal. Appl. 22 (2024), 187. https://doi.org/10.28924/2291-8639-22-2024-187.

[36] A.R. Azazy, M.R. Abonazel, A.M. Shafik, et al. A Proposed Robust Regression Model to Study Carbon Dioxide Emissions in Egypt, Commun. Math. Biol. Neurosci. 2024 (2024), 86. https://doi.org/10.28919/cmbn/8673.

[37] A.H. Youssef, M.R. Abonazel, E.G. Ahmed, Robust M Estimation for Poisson Panel Data Model with Fixed Effects: Method, Algorithm, Simulation, and Application, Stat. Optim. Inf. Comput. 12 (2024), 1292–1305. https://doi.org/10.19139/soic-2310-5070-1996.

[38] E.G. Ahmed, M.R. Abonazel, M.N. Al-Ghamdi, et al. Proposed Robust Estimators for the Poisson Panel Regression Model: Application to COVID-19 Deaths in Europe, Commun. Math. Biol. Neurosci. 2024 (2024), 121. https://doi.org/10.28919/cmbn/8795.