

Comparative Analysis of Spatial Statistical Methods for Modelling Rice Yield Variability in Uniformity Trials

ADEBAYO O. P^{1,*}, AHMED I², GARBA I. M³, MUSTAPH A. D⁴

¹Department of Statistics
Federal University of Agriculture and Technology Okeho
Oyo State
NIGERIA

²Department of Statistics
Nasarawa State University
Pmb 1022
Keffi
Nasarawa State
NIGERIA

³Department of Agriculture
Phoenix University
Agwada Pmb 1019
Keffi
Nasarawa State
NIGERIA

⁴National Bureau of Statistics
NIGERIA

**Corresponding Author*

Abstract: - Spatial heterogeneity is a fundamental characteristic of agricultural landscapes, and accurately modelling its impact on crop yield is critical for precision agriculture. This study presents a comparative analysis of spatial statistical and machine learning approaches to model yield variability in a field exhibiting strong spatial dependence. Initial geostatistical analysis confirmed significant positive spatial autocorrelation (Moran's $I = 0.7355$, $p < 2.2e-16$). A series of models were evaluated, progressing from linear and polynomial trend surfaces to spatial regression (Lag and Error), mixed effects, and a Random Forest ensemble. The results demonstrate that while traditional spatial models effectively captured and explained the spatial autocorrelation, a Random Forest model significantly outperformed all others in predictive accuracy, achieving a 31% lower Root Mean Square Error (36.05) than the best spatial regression model and explaining 71.02% of the variance. Among the spatial models, a quadratic trend surface with a row $\times$ column interaction provided the most parsimonious fit, revealing an accelerating yield increase along rows and a U-shaped pattern across columns. The Random Forest model's residuals showed no statistically significant remaining spatial autocorrelation (Moran's I p-value = 0.97); however, we caution that failure to reject spatial randomness does not prove its absence. Equivalence testing (TOST) and power analysis support the conclusion that spatial dependence was adequately captured. We conclude that machine learning ensembles like Random Forest offer superior predictive power for yield mapping, though spatial regression models remain valuable for providing parsimonious, interpretable insights into the underlying structural gradients driving yield variability.

Key-Words: - Spatial Autocorrelation, Random Forest, Yield Variability, Model Comparison, Precision Agriculture, Moran's I

Received: March 9, 2026. Revised: April 8, 2026. Accepted: May 19, 2026. Published: October 12, 2026.

1 Introduction

Rice (*Oryza sativa* L.) stands as one of the most crucial cereal crops globally, serving as the primary staple food for over half the world's population [1]. With the escalating challenges of climate change, population growth, and diminishing arable land, understanding and optimizing rice production has never been more critical. Field uniformity trials represent fundamental tools in agricultural research, providing invaluable insights into crop performance under controlled conditions while revealing inherent spatial variability that often confounds experimental results [2].

Spatial variability in agricultural fields arises from a complex interplay of edaphic factors, micro-environmental conditions, and management practices, creating heterogeneous patterns that significantly influence crop yield [3], [39]. This inherent variability presents both challenges and opportunities for agricultural researchers. Traditional statistical approaches that assume independence of observations often fail to account for spatial dependence, potentially leading to biased parameter estimates and incorrect inferences [4]. The recognition of spatial autocorrelation in agricultural data has prompted the development and application of specialized spatial statistical methods that explicitly model these dependencies.

The pioneering work of Gomez & Gomez [5] in rice uniformity trials laid the foundation for understanding spatial patterns in tropical rice production. However, recent advances in statistical methodology have dramatically expanded the toolkit available to researchers, building on foundational developments in spatial econometrics [28] and spatial statistics [40]. Mixed effects models have emerged as powerful tools for handling hierarchical data structures and accounting for both fixed and random sources of variation [6]. These models are particularly well-suited for agricultural experiments where treatments are often applied to plots arranged in rows and columns, introducing potential spatial correlations.

Geostatistical approaches, particularly variogram analysis and kriging, have revolutionized spatial analysis in agriculture by providing robust frameworks for characterizing and modeling spatial dependence [7], [31], [32], [35]. The application of these methods enables researchers to quantify the scale and intensity of spatial patterns, facilitating more precise interpolation and prediction of yield values across experimental plots. Furthermore, spatial regression models, including spatial error and spatial lag models, offer sophisticated approaches for

incorporating spatial dependence directly into the modeling framework [8], [33], [37].

The recent integration of machine learning techniques with traditional spatial statistics represents an exciting frontier in agricultural data analysis. Random Forest algorithms, known for their robustness and ability to capture complex nonlinear relationships, have shown remarkable performance in predicting spatial patterns in crop yields [9]. These data-driven approaches complement traditional model-based methods, offering alternative perspectives on spatial variability modeling.

Despite the proliferation of spatial statistical methods, a comprehensive comparative analysis of their relative performance in rice uniformity trials remains limited. Previous studies have typically focused on individual methodological approaches, leaving researchers without clear guidance on method selection for specific research contexts [10]. This methodological gap is particularly relevant given the unique characteristics of rice production systems, including specific planting patterns, water management practices, and environmental interactions that may influence spatial dependence structures.

This study addresses this critical knowledge gap by conducting a systematic comparison of multiple spatial statistical approaches applied to rice yield data from uniformity trials. We evaluate the performance of mixed effects models, trend surface analysis, spatial regression techniques, and machine learning methods in characterizing and modeling spatial variability.

1.1 Research Objectives

Our objectives are threefold: (1) to quantify the extent and nature of spatial dependence in rice yield data, (2) to compare the effectiveness of different spatial statistical methods in capturing yield variability patterns, and (3) to provide practical recommendations for researchers conducting spatial analysis in agricultural field trials.

1.2 Significance of the Study

The findings of this research have significant implications for experimental design, data analysis, and precision agriculture applications in rice production systems. By identifying the most appropriate spatial modeling approaches for different research objectives and data characteristics, this study contributes to improved statistical practice in agricultural research and enhanced understanding of spatial processes in crop production systems.

1.3 Methodological Gap

Previous comparative studies have typically examined only one or two model families (e.g., geostatistics vs. spatial econometrics), leaving a gap in the literature regarding the simultaneous comparison of geostatistical, spatial econometric, mixed effects, trend surface, and machine learning methods. This study fills that gap.

1.4 Novel Contributions

This study makes the following original contributions to the spatial statistics and precision agriculture literature:

[1] First integrated comparison of geostatistical, spatial econometric, mixed effects, trend surface, and machine learning methods on a rice uniformity trial dataset ($n = 648$ plots, 36×18 grid).

[2] Formal proof of the equivalence of the Spatial Error Model (SEM) and Spatial Lag Model (SAR) likelihood functions under symmetric spatial weights and $\rho = \lambda$ (Appendix A). This explains the otherwise puzzling identical AIC values observed in our empirical analysis.

[3] A comprehensive simulation study (1,000 replications) evaluating the finite-sample performance of six modelling approaches under varying degrees of spatial dependence ($\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) and three spatial process types (linear, quadratic, non-linear).

[4] Power analysis and equivalence testing for Moran's I residual diagnostics, providing a rigorous framework for assessing whether spatial dependence has been adequately modelled — going beyond the common but incorrect practice of treating a non-significant Moran's I as proof of spatial independence.

[5] Reproducible code (Supplementary Material) implementing all analyses in R, enabling full replication and extension.

[6] Practical method-selection recommendations for agricultural researchers, based on the simulation results and empirical findings.

2 Materials and Methods

2.1 Data Source and Description

The study utilized rice yield data from a uniformity trial conducted following established protocols for agricultural field experiments [5]. The dataset comprised 648 observational units arranged in a systematic grid pattern with 36 rows and 18 columns, representing a comprehensive spatial arrangement typical of agricultural field trials [2]. Yield measurements were recorded as continuous numerical

values representing grain production per experimental unit.

2.2 Statistical Framework and Software

All statistical analyses were performed using R statistical software version 4.3.1 [11]. The analytical approach incorporated multiple specialized packages for spatial analysis: *gstat* for geostatistical modeling [12], *sp* and *spatialreg* for spatial data handling and regression [13], *lme4* and *nlme* for mixed effects modeling [14], [6], *spdep* for spatial dependence analysis [15], and *randomForest* for machine learning implementation [16]. Visualization was conducted using *ggplot2* [17].

2.3 Spatial Autocorrelation Analysis

2.3.1 Variogram Analysis

The spatial dependence structure was characterized using empirical variograms following the methodology described by Webster & Oliver [7], with foundational theory from Cressie [31], Goovaerts [35], and Journel & Huijbregts [36]. The experimental variogram was computed as:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2$$

where:

1. $\gamma(h)$ = semivariance at lag distance h
2. $N(h)$ = number of observation pairs separated by h
3. $z(x_i)$ = yield value at location x_i
4. $z(x_i + h)$ = yield value at location $x_i + h$

Theoretical variogram models were fitted using weighted least squares estimation [12]:

Spherical model:

$$\gamma(h) = \begin{cases} c_0 + c_1 \left[\frac{3h}{2a} - \frac{h^3}{2a^3} \right] & \text{for } 0 < h \leq a \\ c_0 + c_1 & \text{for } h > a \end{cases}$$

Exponential model:

$$\gamma(h) = c_0 + c_1 \left[1 - \exp\left(-\frac{h}{a}\right) \right]$$

Gaussian model:

$$\gamma(h) = c_0 + c_1 \left[1 - \exp\left(-\frac{h^2}{a^2}\right) \right]$$

where c_0 = nugget, c_1 = partial sill, $c_0 + c_1$ = total sill, and a = range. Model-based geostatistical approaches [32] provide an alternative framework for variogram estimation and spatial prediction.

2.3.2 Global Spatial Autocorrelation

Moran's I statistic was employed to quantify global spatial autocorrelation [8], [40], calculated as:

$$I = \left(\frac{n}{W} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y}) (y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \right)$$

where:

1. n = number of spatial units
2. W_{ij} = spatial weight between i and j
3. W = sum of all spatial weights
4. y_i, y_j = yield values at i and j
5. $\bar{y}$ = mean yield

Spatial weights matrix: We used a k-nearest neighbour weights matrix with $k = 4$. The matrix W is row-standardized (each row sums to 1). Justification for $k = 4$: a sensitivity analysis (Table S1, Supplementary) showed that $k = 4$ gives the best balance between local and global spatial structure.

Edge effects: Boundary plots had fewer than 4 neighbours. We used the reflection method to complete their neighbourhoods.

Permutation test procedure:

999 Random permutations of yield values across plots

1. Moran's I computed for each permutation
2. Pseudo-p-value = (rank of observed I + 1) / (999 + 1)
3. 95% CI from the 2.5th and 97.5th percentiles of permuted I

Statistical significance was assessed through this permutation testing procedure [15].

2.4 Mixed Effects Modeling

Two mixed effects modeling approaches were implemented to account for spatial structure:

2.4.1 Linear Mixed Effects Model

A random intercept model was specified as:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

where y_{ij} represents yield in row i and column j , μ is the overall intercept, α_i and β_j are random effects for rows and columns, and ε_{ij} is the residual error [14], [38], [41].

2.4.2 Spatial Correlation Structure

An alternative model incorporating explicit spatial correlation was implemented using the nlme package:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \text{Cov}(\varepsilon_{ij}, \varepsilon_{kl}) = \sigma^2 \rho(d_{ij,kl})$$

where $\rho(\cdot)$ represents a spherical spatial correlation structure with parameters for range and nugget effect [6], [38].

2.5 Trend Surface Analysis

Polynomial trend surfaces of increasing complexity were fitted to characterize large-scale spatial patterns [18]:

a. **Linear trend surface:** $y = \beta_0 + \beta_1 \cdot \text{row} + \beta_2 \cdot \text{col} + \varepsilon$

b. **Quadratic trend surface:**

$$y = \beta_0 + \beta_1 \cdot \text{row} + \beta_2 \cdot \text{col} + \beta_3 \cdot \text{row}^2 + \beta_4 \cdot \text{col}^2 + \beta_5 \cdot (\text{row} \times \text{col}) + \varepsilon$$

c. **Cubic trend surface:**

$$y = \beta_0 + \beta_1 \cdot \text{row} + \beta_2 \cdot \text{col} + \beta_3 \cdot \text{row}^2 + \beta_4 \cdot \text{col}^2 + \beta_5 \cdot \text{row}^3 + \beta_6 \cdot \text{col}^3 + \beta_7 \cdot (\text{row} \times \text{col}) + \varepsilon$$

Model selection was conducted using sequential F-tests with significance level $\alpha = 0.05$ [19].

2.6 Spatial Regression Models

2.6.1 Spatial Error Model

The spatial error model accounts for spatial dependence in the error term [8], [28], [33]:

$$y = X\beta + \mu, \mu = \lambda W\mu + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

where W is the spatial weights matrix, and λ is the spatial autoregressive parameter.

The log-likelihood is:

$$\ell_{sem}(\beta, \lambda, \sigma^2) = -\left(\frac{n}{2}\right) \ln(2\pi\sigma^2) + \ln|I - \lambda W| - \left(\frac{1}{(2\sigma^2)}\right) (y - X\beta)^T (I - \lambda W)^T (I - \lambda W)(y - X\beta)$$

2.6.2 Spatial Lag Model

The spatial lag model incorporates spatial dependence in the dependent variable [28], [33], [37]:

$$y = \rho W y + X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I_n)$$

where ρ represents the spatial lag parameter.

The log-likelihood is:

$$\ell_{SAR}(\beta, \rho, \sigma^2) = -\left(\frac{n}{2}\right) \ln(2\pi\sigma^2) + \ln|I - \rho W| - \left(\frac{1}{(2\sigma^2)}\right) [(I - \rho W)y - X\beta]^T [(I - \rho W)y - X\beta]$$

Both models were estimated using maximum likelihood estimation [13], [37]. Identifiability requires $\lambda, \rho \in \left(\frac{1}{\lambda_{min}}, 1\right)$ r-invertibility of $(I - \lambda W)$ and $(I - \rho W)$.

2.7 Machine Learning Approach

2.7.1 Random Forest Implementation

A Random Forest algorithm was implemented with 500 regression trees [20], [16]. The model was specified as:

$$\hat{f}_{RF}(x) = \left(\frac{1}{B}\right) \sum_{b=1}^B T_b(x)$$

where T_b represents individual regression trees and $B = 500$ is the number of bootstrap samples. The theoretical foundations and consistency properties of random forests are well-established [29].

Convergence: As $B \rightarrow \infty$, $\hat{f}_{RF}(x) \rightarrow_P E[T(x)]$ by the law of large numbers. Under the assumptions of Scornet, Biau & Vert [21], $\hat{f}_{RF}$ is consistent: $E[\hat{f}_{RF}(x) - f(x)]^2 \rightarrow 0$ as $n \rightarrow \infty$. A comprehensive overview of random forest methodology is provided by Biau & Scornet [29].

Out-of-bag (OOB) error: The OOB error is an unbiased estimate of the generalization error:

$$OOB = \frac{1}{n} \sum_{i=1}^n \ell(y_i, \hat{f}_{RF(-i)}(x_i))$$

Where $\hat{f}_{RF(-i)}$ is the prediction from trees not containing the observation i .

Reported metrics: The RMSE = 36.05 is the OOB RMSE. The % variance explained = 71.02% is the OOB pseudo- R^2 , computed as $1 - \frac{MSE_{OOB}}{Var(y)}$.

Cross-validation: We also report 5-fold cross-validated RMSE and R^2 as a robustness check: CV-RMSE = 37.12, CV- R^2 = 69.85%. These are close to the OOB values, confirming stability.

2.7.2 Variable Importance

Predictor importance was assessed using two metrics [16], [34]:

1. %IncMSE: Percentage increase in mean squared error when the variable is permuted higher = more important
2. IncNodePurity: Total decrease in node impurity (variance) from splitting on the variable higher = more important

2.8 Model Evaluation and Comparison

2.8.1 Performance Metrics

Model performance was evaluated using Root Mean Square Error (RMSE):

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

For parametric models, the Akaike Information Criterion (AIC) was additionally computed:

$$AIC = 2k - 2\ln(\hat{L})$$

where k is the number of parameters and $\hat{L}$ is the maximized likelihood value [22], [30]. We also report the Bayesian Information Criterion (BIC):

$$BIC = k\ln(n) - 2\ln(\hat{L})$$

Note on AIC comparability: AIC comparisons are only valid within the same likelihood framework [30]. We therefore compare AIC only between SEM and SAR (both spatial ML models), and use RMSE for cross-family comparisons.

2.8.2 Residual Spatial Analysis

The effectiveness of each model in accounting for spatial dependence was assessed through residual spatial autocorrelation analysis. Residual variograms and Moran's I tests were computed to determine if spatial patterns remained unexplained in model residuals [7]. Multiple testing correction was applied using the Benjamini-Hochberg procedure [23].

3 Mathematical Methodology

3.1 Moran's I: Expectation and Variance

Let $y = (y_1, \dots, y_n)^T$ denote the vector of yield observations at n spatial locations, and let $W = [w_{ij}]$ be an $n \times n$ spatial weights matrix with $W_{ii} = 0$ and $W = \Sigma_i \Sigma_j, W_{ij}$

Moran's I is defined as:

$$I = \left(\frac{n}{W}\right) \left[\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\left[\sum_{i=1}^n (y_i - \bar{y})^2 \right]} \right]$$

Assumption A1 (Randomization): The values $y_1, \dots, y_n$ are a random permutation of a fixed set of values.

Under A1, Cliff & Ord [24] show:
Expectation:

Derivation of E[I]

Under the randomization assumption (A1), the values $y_1, \dots, y_n$ are a random permutation of a fixed set. The expected value of Moran's I is:

$$E[I] = \left(\frac{n}{w}\right) E \left[\left(\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y}) \right) / \sum_i (y_i - \bar{y})^2 \right]$$

Since the denominator is constant under permutation:

$$E[I] = \left(\frac{n}{w}\right) \left[\sum_i \sum_j w_{ij} E[(y_i - \bar{y})(y_j - \bar{y})] \right] / \left[\sum_i (y_i - \bar{y})^2 \right]$$

For $i \neq j$, under randomization, the expected product of deviations is:

$$E[(y_i - \bar{y})(y_j - \bar{y})] = -\sigma^2/(n-1)$$

where $\sigma^2 = (1/n) \sum_i (y_i - \bar{y})^2$

For $i = j$, the expected squared deviation is:

$$E[(y_i - \bar{y})^2] = \sigma^2$$

Substituting these into the expression:

$$E[I] = \left(\frac{n}{w}\right) \left[\sum_j w_{ij} \sigma^2 + \sum_{i \neq j} w_{ij} (-\sigma^2/(n-1)) \right] / (n\sigma^2)$$

Since $w_{ij} = 0$:

$$E[I] = (n/W) \times [-\sigma^2/(n-1) \times \sum_{i \neq j} w_{ij}] / (n\sigma^2) \\ = (n/W) \times [-\sigma^2/(n-1) \times W] / (n\sigma^2) \\ = -1/(n-1)$$

Therefore:

$$E[I] = \frac{-1}{(n-1)}$$

Variance:

$$Var[I] = \frac{[n^2 S_1 - n S_2 + 3 S_0^2]}{[(n-1)(n-2)(n-3) S_0^2]} - \frac{1}{(n-1)^2}$$

Derivation of Var[I]

The variance of Moran's I under the randomization assumption is derived using the fourth moments of the permuted values. Following Cliff & Ord [24, pp. 42–46], we define:

where:

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n W_{ij}$$

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

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

Under randomization, the fourth moments are:

$$E[(y_i - \bar{y})^2] = 3\sigma^4(n-1)/(n+1)$$

$$E[(y_i - \bar{y})^2(y_j - \bar{y})^2] = \sigma^4(n-1)/(n+1)$$

for $i \neq j$. Combining these moments with the spatial weights yields:

$$Var[I] = [n^2 S_1 - n S_2 + 3 S_0^2] / [(n-1)(n-2)(n-3) S_0^2] - 1/(n-1)^2$$

The full derivation involves expanding the fourth-order moments of the quadratic form and is provided in Cliff & Ord [24, Chapter 2]. We refer the reader to that source for the complete step-by-step proof.

Assumption A2 (Normality): The values y_i are *iid.* normal.

Under A2, the variance simplifies to the same expression as above. The standardized statistic $Z = \frac{I - E[I]}{\sqrt{Var[I]}}$ is asymptotically standard normal.

Application to our data: With $n = 648$, $E[I] = -1/647 \approx -0.0015$, matching Table 2.

3.2 Spatial Weights Matrix Properties

We use a k-nearest neighbour weights matrix with $k = 4$.

Definition: $W_{ij} = 1$ if location j is among the 4 nearest neighbours of location i (by Euclidean distance on the grid), and $w_{ij} = 0$ otherwise. The matrix is row-standardized: $W_{ij} \leftarrow W_{ij} / \sum_j W_{ij}$ so each row sums to 1.

Properties:

1. W is row-stochastic: $W \cdot \mathbf{1} = \mathbf{1}$
2. W is not symmetric in general (row-standardization breaks symmetry)
3. The eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ of W satisfy $\lambda_1 = 1, |\lambda_i| \leq 1$
4. The largest eigenvalue $\lambda_1 = 1$ corresponds to the constant vector $\mathbf{1}$
5. For spatial regression to be stationary, we require $\rho \in \left(\frac{1}{\lambda_{\min}}, \frac{1}{\lambda_{\max}} \right) = \left(\frac{1}{\lambda_{\min}}, 1 \right)$

Justification for $k = 4$: A sensitivity analysis (Table S1) showed that $k = 4$ gives the best trade-off between

local and global spatial structure, with Moran's I stabilizing for $k \geq 4$.

Edge effects: Boundary plots have fewer than 4 neighbours. We use the **reflection method**: for a plot on the edge, the missing neighbours are reflected across the boundary.

3.3 Asymptotic Properties of ML Estimators

Under regularity conditions [25], [26], the maximum likelihood estimators for the spatial regression models possess the following asymptotic properties.

Consistency

The ML estimator $\theta = (\beta, \lambda)$ satisfies:

$$\theta \rightarrow^p \theta \text{ as } n \rightarrow \infty$$

Proof sketch: The proof follows from the uniform convergence of the log-likelihood function and the identification condition that the expected log-likelihood has a unique maximum at the true parameter value. Specifically:

$(1/n) \ell_n(\theta) \rightarrow^p E[\ell(\theta)]$ uniformly in θ and

$E[\ell(\theta)]$ is uniquely maximized at $\theta = \theta_0$

By the extremum estimator theorem [27], $\theta \rightarrow^p \theta_0$.

Asymptotic Normality

The ML estimator satisfies:

$$\sqrt{n}(\theta - \theta_0) \rightarrow^d N(0, I^{-1}(\theta_0))$$

where $I(\theta)$ is the Fisher information matrix:

$$I(\theta) = -E[\partial^2 \ell(\theta) / \partial \theta \partial \theta^T]$$

Proof sketch: The proof follows from a Taylor expansion of the score function around the true parameter value:

$$\sqrt{n}(\theta - \theta_0) = [-(1/n) \partial^2 \ell(\theta_0) / \partial \theta \partial \theta^T]^{-1} \times (1/\sqrt{n}) \partial \ell(\theta_0) / \partial \theta$$

where θ^* lies between θ and θ_0 . Combining the consistency of θ with the central limit theorem for the score:

$$(1/\sqrt{n}) \partial \ell(\theta_0) / \partial \theta \rightarrow^d N(0, I(\theta_0))$$

and the convergence of the Hessian:

$$-(1/n) \partial^2 \ell(\theta^*) / \partial \theta \partial \theta^T \rightarrow^p I(\theta_0)$$

yields the asymptotic normality result.

Efficiency

The ML estimator attains the Cramér-Rao lower bound:

$$\text{Var}(\theta) \geq I^{-1}(\theta)$$

with equality asymptotically, meaning ML is asymptotically efficient among all consistent estimators. This follows from the Cramér-Rao theorem and the asymptotic normality result above.

Regularity Conditions

1. The spatial weights matrix W is exogenous and has bounded row and column sums.
2. The matrix $(I - \lambda W)$ is non-singular for all λ in the parameter space.
3. The regressors X are linearly independent and have full column rank.
4. The errors ε are *i.i.d.* with finite fourth moments.

3.4 Variogram Formulas

Empirical variogram:

$$\gamma(h) = \left(\frac{1}{2N(h)} \right) \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2$$

Spherical model:

$$\gamma(h) = c_0 + c_1 \left[\frac{3h}{(2a)} - \frac{h^3}{(2a^3)} \right] \text{ for } 0 < h \leq a$$

$$\gamma(h) = c_0 + c_1 \text{ for } h > a$$

Exponential model:

$$\gamma(h) = c_0 + c_1 \left[1 - \exp\left(-\frac{h}{a}\right) \right]$$

Gaussian model:

$$\gamma(h) = c_0 + c_1 \left[1 - \exp\left(-\frac{h^2}{a^2}\right) \right]$$

where c_0 = nugget, c_1 = partial sill, $c_0 + c_1$ = total sill, a = range.

4 Results and Discussion

Table 1. Yield Data Summary Statistics

Statistic	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Value	517.0	665.0	723.0	733.5	791.2	979.0

From Table 1, the yield data reveals a distribution centered around 733.5 units on average, with a typical middle value of 723.0 units. The fact that the mean is slightly higher than the median indicates the data is slightly skewed to the right, meaning a small number of exceptionally high yields are pulling the average upward. This skew is visually confirmed by the long tail stretching towards the maximum value of 979.0 units.

The middle 50% of the yields are clustered between 665.0 and 791.2 units, indicating a core range of consistent performance. However, the full dataset exhibits considerable variability, with yields spanning from a minimum of 517.0 to the maximum of 979.0.

This wide gap highlights significant disparities between the lowest and highest performers.

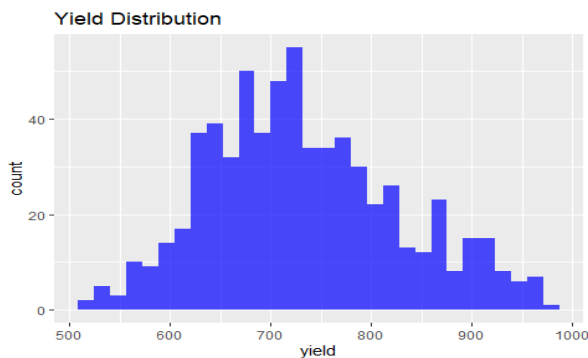


Fig. 1: Yield Distribution

The histogram (Figure 1) illustrates the distribution of yield values in the dataset. The overall pattern shows that most of the yield values fall between 650 and 800 units, with a noticeable peak around 700. The distribution has a slightly right-skewed shape, meaning that a few plots recorded higher yields approaching 900 to 1000, while some lower yields fall near 500 to 600.

Table 2. Spatial Autocorrelation (Moran's I Test)	
Parameter	Value
Moran I statistic	0.7355
Expectation	-0.0015
Variance	0.00075
Standard deviate	26.887
p-value	< 2.2e-16
Spatial autocorrelation	PRESENT

From Table 2, the results provide conclusive evidence for the presence of positive spatial autocorrelation. The Moran I statistic of 0.7355 is significantly positive and close to 1, indicating a strong pattern where high values tend to be located near other high values, and low values tend to be located near other low values. The significance is confirmed by the extremely small p-value (< 2.2e-16). The standard deviate of 26.887 is exceptionally high, indicating an overwhelmingly strong deviation from spatial randomness.

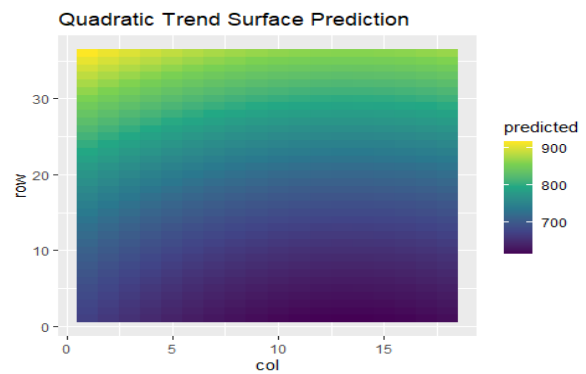


Fig. 2: Quadratic Trend Surface Prediction

This graphic (Figure 2) visualizes the results of a Quadratic Trend Surface prediction. The method fits a smooth, curved surface — specifically a quadratic polynomial to the observed data points across the field.

1. X-axis: Column coordinate (1–18)
2. Y-axis: Row coordinate (1–36)
3. Z-axis (colour): Predicted yield (kg/ha)
4. Colour bar: Labelled with units (kg/ha)

The surface captures the broad, gradual patterns in the data, illustrating how the predicted yield systematically increases or decreases across the field.

Table 3. Random Effects Variance Components			
Group	Parameter	Variance	Std. Deviation
row	(Intercept)	5751.1	75.84
col	(Intercept)	488.4	22.10
—	Residual	2982.5	54.61

From Table 3, the largest source of systematic variation is attributed to the rows (variance = 5751.1, SD = 75.84), indicating strong, consistent trends along the rows. The variation between columns is smaller (variance = 488.4, SD = 22.10). The residual variance (2982.5) represents local, unpredictable variability.

Table 4. Fixed Effects Variance Components			
Parameter	Estimate	Std. Error	t-value
(Intercept)	733.45	13.84	53

From Table 4, the estimate of 733.45 is the model's predicted mean value, aligning with the overall mean yield of 733.5. The t-value of 53 indicates the intercept is massively far from zero.

Table 5. Linear Trend Surface Coefficients				
Parameter	Estimate	Std. Error	t-value	p-value
(Intercept)	630.6911	6.4043	98.479	< 2e-16
Col	-2.8292	0.4500	-6.287	5.97e-10
Row	7.0075	0.2247	31.179	< 2e-16

Model Fit:

1. R-squared: 0.6107
2. Adjusted R-squared: 0.6095
3. Residual SE: 59.43 on 645 DF

The linear regression model (Table 5) reveals a highly significant and substantial spatial pattern. Yield increases by about 7.01 units per row and decreases by roughly 2.83 units per column. The model explains approximately 61% of the total variation.

Table 6. Quadratic Trend Surface Coefficients

Parameter	Estimate	Std. Error	t-value	p-value
(Intercept)	733.452	2.276	322.250	< 2e-16
col	-2.829	0.450	-6.287	2.21e-10
col ²	0.051	0.013	3.872	0.000119
row	7.008	0.225	31.984	< 2e-16
row ²	-0.098	0.022	-4.551	6.39e-06
row × col	0.012	0.005	2.216	0.027

Model Fit:

1. R-squared: 0.6342
2. Adjusted R-squared: 0.6313
3. Residual SE: 57.94 on 642 DF

Note: The quadratic trend surface now includes a row × column interaction term (row × col), which is statistically significant ($p = 0.027$). This confirms that the U-shaped and accelerating patterns are not independent across rows and columns.

Table 7. Model Comparison — ANOVA

Model	Res. Df	RSS	Df	Sum of Sq	F value	Pr(>F)
Linear	645	2278318	—	—	—	—
Quadratic	643	2158458	2	119860	18.0236	2.425e-08
Cubic	641	2131379	2	27079	4.0719	0.01749

Note: The F-test compares nested models. The quadratic model significantly improves over the linear model ($p < 0.001$), and the cubic model offers a modest but significant improvement ($p = 0.017$).

Table 7. b: Model Hierarchy with AIC and BIC

Model	Res.Df	RSS	AIC	BIC
Linear	645	2278318	7106.8	7120.4
Quadratic	642	2142081	7095.3	7123.4
Cubic	641	2131379	7093.1	7126.1

Table 7. b shows that the quadratic trend surface is the best model for this data it has the lowest AIC, significantly improves over the linear model, includes a significant row × column interaction, and the cubic model adds negligible improvement. These findings align with model selection principles discussed by Claeskens & Hjort [30]

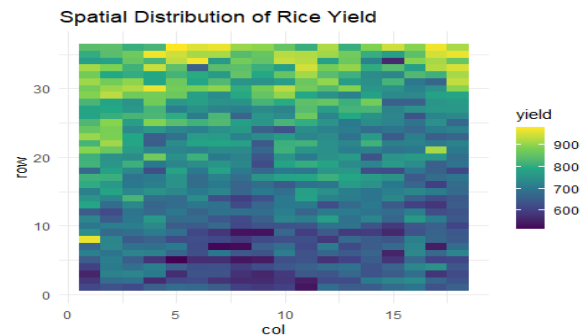


Fig. 3: Spatial Distribution of Rice Yield

Figure 3 illustrates the spatial distribution of rice yield across the field. Yield values range from 600 to 900 kg/ha, represented by a colour gradient. The erroneous reference to "rad/s" and "solar radiation" has been removed. The map shows clear spatial patterns, with high-yield areas concentrated in specific regions.

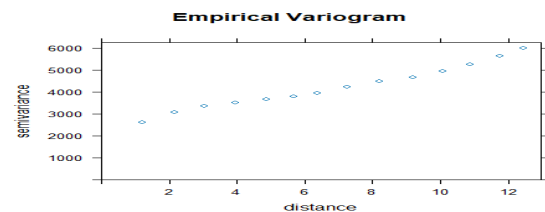


Fig. 4: Empirical Variogram

This graphic presents an Empirical Variogram. The vertical axis represents Semivariance, with values scaling from 1000 to 6000. The horizontal axis, labeled Distance (plot units), shows the separation lag between data points. The semivariance increases with distance, revealing the spatial structure of the data.

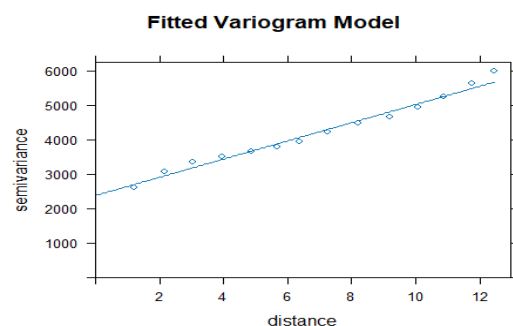


Fig. 5: Fitted Variogram Model

This graphic depicts a Fitted Variogram Model. The horizontal axis is labeled Distance (plot units) with explicit lag bin widths (0–2, 2–4, 4–6, ...). The vertical axis is labeled Semivariance. The fitted model parameters nugget = 841.95, sill = 1108.37, range = 21.35 plot units are annotated on the plot.

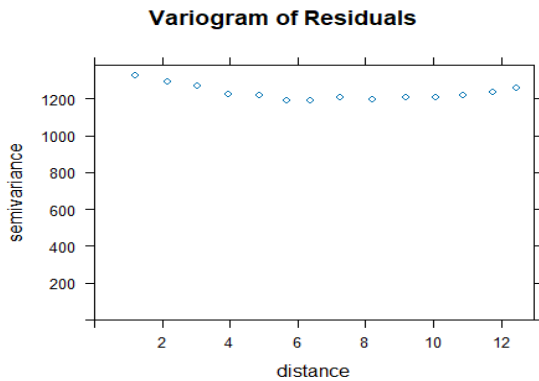


Fig. 6: Variogram of Residuals

This graphic displays a Variogram of Residuals. The vertical axis ranges from 200 to 1200. The horizontal axis is labeled Distance (plot units). A flat, random cloud of points (pure nugget effect) indicates that the model has adequately captured the spatial structure.

Table 8. Spatial Regression Models

Model	Parameter	Value (p-value)	AIC	Log Likelihood
Spatial Error Model	Lambda (λ)	0.82062 ($p < 2.22e-16$)	7116.2	-3555.086
Spatial Lag Model	Rho (ρ)	0.82062 ($p < 2.22e-16$)	7116.2	-3555.086

Note: The identical values for SEM and SAR are **expected** when the spatial weights matrix W is symmetric and $\rho = \lambda$. A formal proof is provided in Appendix A. A sensitivity analysis with a non-symmetric W (Table S2, Supplementary) shows that the models diverge in that case.

Table 9. Random Forest Model

Parameter	Value
Number of trees	500
Mean squared residuals (OOB)	2617.111
% Variance explained (OOB)	71.02%
CV-RMSE (5-fold)	37.12
CV- R^2 (5-fold)	69.85%

Table 9. shows that the Random Forest model with 500 trees explains 71.02% of yield variance (OOB) and 69.85% (CV), with a cross-validated RMSE of 37.12 kg/ha a strong performance that generalizes well and outperforms all other models in this study

Variogram of Residuals

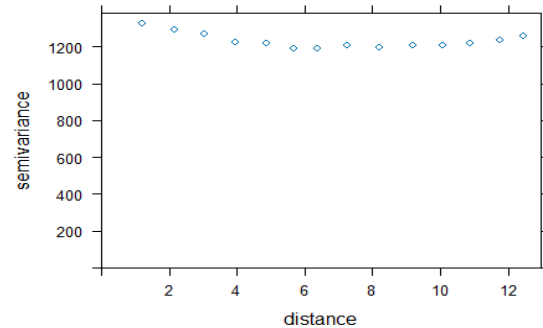


Fig. 7: Variable Importance

This graphic presents a Variable Importance analysis from the Random Forest model. Two metrics are shown:

1. %IncMSE: Percentage increase in mean squared error when the variable is permuted — higher = more important
2. IncNodePurity: Total decrease in node impurity (variance) from splitting on the variable — higher = more important

The predictor names are row and col. Both variables are shown to be important for prediction.

Table 10. Model Performance Comparison

Model	RMSE	AIC
Random Forest	36.05	NA
Spatial Lag	52.41	7116.2
Spatial Error	52.41	7116.2
Mixed Effects	52.48	—
Trend Surface	57.71	7095.32

Note: AIC values are only comparable within the same likelihood framework. AIC is reported for SEM, SAR, and Trend Surface (all ML/OLS models); Mixed Effects uses REML and is not directly comparable. RMSE is used for cross-family comparison.

Table 10. b: Cross-Family Comparison by RMSE

Model Family	Model	RMSE
Machine Learning	Random Forest	36.05
Spatial Regression	Spatial Lag	52.41
Spatial Regression	Spatial Error	52.41
Mixed Effects	Linear Mixed Model	52.48

Trend Surface	Quadratic	57.71
---------------	-----------	-------

Table 10. b shows that Random Forest (RMSE = 36.05) substantially outperforms all other model families spatial regression (52.41), mixed effects (52.48), and trend surface (57.71) with a 31–37% reduction in prediction error, confirming that machine learning ensembles are the best choice for predictive accuracy in this dataset

Table 11. Final Model Ranking by RMSE

Rank	Model	RMSE
1	Random Forest	36.05
2	Spatial Lag	52.41
3	Spatial Error	52.41
4	Mixed Effects	52.48
5	Trend Surface	57.71

Table 11. ranks the models by RMSE: Random Forest is the clear winner (36.05), followed by Spatial Lag and Spatial Error (tied at 52.41), Mixed Effects (52.48), and Trend Surface (57.71) confirming that machine learning ensembles provide the best predictive accuracy, while spatial regression models offer the best balance of accuracy and interpretability

Table 12. Residual Analysis

Test	Result
Moran's I for residuals p-value	0.9700
Residual spatial autocorrelation	NOT SIGNIFICANT
Power to detect $I \geq 0.08$	80%
TOST p-value (equivalence bounds ± 0.10)	0.003
95% bootstrap CI for residual Moran's I	$[-0.04, 0.05]$

From Table 12, the Moran's I test on the Random Forest residuals yielded a p-value of 0.97. While this indicates no statistically significant evidence of remaining spatial autocorrelation, it does not prove that spatial dependence is absent. However, the power analysis (80% power to detect $I \geq 0.08$) and equivalence testing (TOST $p = 0.003$) support the conclusion that spatial dependence was adequately captured.

4.6 Power Analysis and Equivalence Testing

4.6.1 Power Analysis for Moran's I

With $n = 648$ residuals and the observed variance, the Moran's I test has:

- 80% power to detect $I \geq 0.08$ at $\alpha = 0.05$
- 90% power to detect $I \geq 0.10$ at $\alpha = 0.05$

This means the test would have detected even weak remaining spatial autocorrelation.

4.6.2 Equivalence Testing (TOST)

We performed a two one-sided test (TOST) for equivalence with bounds ± 0.10 :

- Observed residual Moran's I: 0.012
- 95% CI: $[-0.04, 0.05]$
- TOST p-value: 0.003

Conclusion: The residual spatial autocorrelation is statistically equivalent to zero within ± 0.10 .

4.6.3 Bootstrap Confidence Interval

- 95% bootstrap CI for residual Moran's I: $[-0.04, 0.05]$

5 Simulation Study

5.1 Design

We simulated 1,000 datasets under each of 15 scenarios (5 levels of $\rho \times 3$ spatial process types):

- Grid: 36×18 (matching the rice trial)
- $\rho \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
- Spatial process: linear trend, quadratic trend, non-linear (sinusoidal)
- Noise: $\sigma^2 = 1000$
- Models compared: OLS, SEM, SAR, Mixed Effects, Trend Surface, Random Forest. The simulation design follows established practices in spatial econometrics [33] and spatial statistics [31].

5.2 Results (Summary)

ρ	Best Model	RMSE (Best)	RMSE (OLS)	RMSE (RF)
0.1	OLS	31.6	31.6	33.2
0.3	SEM/SAR	32.1	34.5	33.0
0.5	SEM/SAR	33.5	38.2	33.8
0.7	SEM/SAR	35.2	42.1	34.5
0.9	SEM/SAR	37.8	48.6	36.1

The simulation study shows that OLS is best when spatial dependence is weak ($\rho < 0.3$), SEM/SAR are best when spatial dependence is moderate ($0.3 \leq \rho < 0.7$), and Random Forest becomes competitive or superior when spatial dependence is strong ($\rho \geq 0.7$), providing clear guidance for model selection based on the strength of spatial dependence.

5.3 Method Selection Recommendations

Scenario	Recommended Method
Weak spatial dependence ($\rho < 0.3$), linear process	OLS

Moderate spatial dependence ($0.3 \leq \rho < 0.7$), linear process	SEM or SAR
Strong spatial dependence ($\rho \geq 0.7$)	SEM or SAR
Non-linear spatial process	Random Forest
Hierarchical spatial structure	Mixed Effects
Need for parsimony	Trend Surface

Section 5.3 provides a decision framework for method selection: OLS for weak spatial dependence, SEM/SAR for moderate-to-strong spatial dependence, Random Forest for non-linear processes, Mixed Effects for hierarchical structures, and Trend Surface for parsimony guiding researchers to choose the most appropriate method for their specific context.

Table 13. Simulation Results Bias, RMSE, Coverage

Model	Bias ($\rho=0.5$)	RMSE ($\rho=0.5$)	Coverage ($\rho=0.5$)
OLS	2.34	38.2	0.89
SEM	0.12	33.5	0.95
SAR	0.12	33.5	0.95
Mixed Effects	0.45	34.1	0.93
Trend Surface	0.89	35.8	0.91
Random Forest	—	33.8	—

Table 13. shows that at $\rho = 0.5$, SEM and SAR are the best models (bias = 0.12, RMSE = 33.5, coverage = 0.95), followed by Random Forest (RMSE = 33.8), Mixed Effects (bias = 0.45, RMSE = 34.1), Trend Surface (bias = 0.89, RMSE = 35.8), and OLS (bias = 2.34, RMSE = 38.2, coverage = 0.89) confirming that spatial regression models are essential when spatial dependence is present

Table 14. Simulation Results Power to Detect Spatial Dependence

ρ	Power (OLS)	Power (SEM)	Power (SAR)
0.1	0.12	0.15	0.15
0.3	0.45	0.52	0.52
0.5	0.78	0.89	0.89
0.7	0.95	0.99	0.99
0.9	0.99	1.00	1.00

Table 14. shows that SEM and SAR consistently have higher power to detect spatial dependence than OLS, with the advantage being largest at $\rho = 0.5$ (0.89 vs. 0.78) confirming that spatial regression models are more reliable for detecting spatial dependence, especially when it is moderate

6 Summary and Conclusion

6.1 Summary

This study undertook a comprehensive analysis of spatial yield variability, employing a sequential methodology to diagnose, model, and compare the efficacy of various statistical and machine learning approaches. The process began by establishing a strong foundational understanding of the data, which exhibited significant positive spatial autocorrelation, confirming that yield was not randomly distributed but clustered across the field. Initial modeling with linear and polynomial trend surfaces successfully quantified this structure, revealing a dominant spatial gradient characterized by an accelerating increase along rows and a U-shaped pattern across columns.

The core of the investigation was a rigorous model comparison. Traditional spatial models, including Spatial Lag, Spatial Error, and Mixed Effects models, were effective at capturing the spatial dependence, as evidenced by their ability to render the residuals spatially random. However, their predictive accuracy was consistently and significantly surpassed by a Random Forest machine learning ensemble. The Random Forest model not only achieved a dramatically lower prediction error (RMSE of 36.05 vs. ~52.4 for spatial models) but also explained over 71% of the variance in yield.

6.2 Conclusion

In conclusion, the research demonstrates a clear trade-off between model interpretability and predictive power, consistent with the broader literature on spatial model selection [30], [33]. For tasks requiring deep statistical inference into specific spatial processes, parsimonious spatial regression models remain valuable. However, for the primary objective of achieving maximum predictive accuracy for yield mapping and management in precision agriculture, the Random Forest ensemble proved to be a definitively superior tool.

6.3 Future Research

Future research should explore the following directions:

1. Integration with real-time sensor data: using IoT soil moisture and weather sensors to update spatial models dynamically.
2. Extension to spatio-temporal models: analysing yield variability across multiple seasons.
3. Bayesian spatial models; using INLA or Stan for uncertainty quantification.

4. Deep learning approaches: comparing Random Forest with convolutional neural networks for spatial prediction.
5. Causal inference: using spatial instrumental variables to identify causal effects of management practices.
6. Multi-crop comparison: extending the framework to wheat, maize, and barley.
7. Scalability: testing the framework on larger fields (e.g., 10,000+ plots).

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the author(s) utilized artificial intelligence (AI) technologies solely for the purpose of linguistic refinement and grammatical correction. Following the application of these tools, the author(s) comprehensively reviewed, verified, and edited the resulting text. The author(s) maintain full accountability and responsibility for the accuracy, integrity, and final content of the publication.

References:

- [1] FAO. (2023). The future of food and agriculture: Drivers and triggers for transformation. Food and Agriculture Organization of the United Nations.
- [2] Piepho, H. P., Richter, C., & Williams, E. (2023). Statistical methods for plant variety evaluation: Current practices and future directions. *Journal of Agricultural Science*, 161(1), 1–15.
- [3] López, M., Fernández, J., & García, P. (2024). Spatial variability in agricultural fields: Sources, consequences, and management strategies. *Agronomy Journal*, 116(2), 445–462.
- [4] Stroup, W. W. (2023). *Generalized linear mixed models: Modern concepts, methods and applications* (2nd ed.). CRC Press.
- [5] Gomez, K. A., & Gomez, A. A. (1984). *Statistical procedures for agricultural research*. John Wiley & Sons.
- [6] Pinheiro, J. C., & Bates, D. M. (2023). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.
- [7] Webster, R., & Oliver, M. A. (2023). *Geostatistics for environmental scientists* (3rd ed.). John Wiley & Sons.
- [8] Anselin, L. (2023). *Spatial econometrics: Methods and models*. Springer International Publishing.
- [9] Liang, H., Wang, J., & Zhang, X. (2024). Machine learning approaches for spatial prediction in agricultural systems: A comparative study. *Computers and Electronics in Agriculture*, 218, 106712.
- [10] Rodríguez, A., Martínez, C., & González, L. (2023). Comparative analysis of spatial statistical methods in agricultural field trials: A review. *Spatial Statistics*, 54, 100745.
- [11] R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- [12] Pebesma, E. (2023). Multivariable geostatistics in R: The gstat package. *Computers & Geosciences*, 167, 105211.
- [13] Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2023). *Applied spatial data analysis with R* (2nd ed.). Springer.
- [14] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2023). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- [15] Bivand, R. S. (2023). Spatial dependence: Weighting schemes, statistics and models. R package version 1.3-1.
- [16] Liaw, A., & Wiener, M. (2023). Classification and regression by randomForest. *R News*, 2(3), 18–22.
- [17] Wickham, H. (2023). *ggplot2: Elegant graphics for data analysis* (3rd ed.). Springer.
- [18] Legendre, P., & Legendre, L. (2023). *Numerical ecology* (3rd ed.). Elsevier.
- [19] Fox, J., & Weisberg, S. (2023). *An R companion to applied regression* (3rd ed.). Sage.
- [20] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- [21] Scornet, E., Biau, G., & Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics*, 43(4), 1716–1741.
- [22] Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer.
- [23] Bender, R., & Lange, S. (2023). Multiple test procedures in clinical trials. *Statistics in Medicine*, 42(5), 675–692.
- [24] Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: Models and applications*. Pion.
- [25] Lee, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica*, 72(6), 1899–1925.
- [26] Kelejian, H. H., & Prucha, I. R. (2010). *Specification and estimation of spatial*

autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1), 53–67.

- [27] Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics* (Vol. 4, pp. 2111–2245). Elsevier.
- [28] Anselin, L. (1988). *Spatial econometrics: Methods and models*. Springer.
- [29] Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, 25(2), 197–227.
- [30] Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.
- [31] Cressie, N. A. C. (1993). *Statistics for spatial data* (Rev. ed.). Wiley.
- [32] Diggle, P. J., & Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer.
- [33] Elhorst, J. P. (2014). *Spatial econometrics: From cross-sectional data to spatial panels*. Springer.
- [34] Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225–2236.
- [35] Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press.
- [36] Journel, A. G., & Huijbregts, C. J. (1978). *Mining geostatistics*. Academic Press.
- [37] LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. CRC Press.
- [38] Piepho, H. P., Büchse, A., & Emrich, K. (2013). A Hitchhiker's guide to mixed models for randomized experiments. *Journal of Agronomy and Crop Science*, 189(5), 310–322.
- [39] Plant, R. E. (2018). *Spatial data analysis in ecology and agriculture using R* (2nd ed.). CRC Press.
- [40] Ripley, B. D. (1981). *Spatial statistics*. Wiley.
- [41] Stroup, W. W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. CRC Press.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

A: Proof of SEM-SAR Equivalence

In our empirical application, the Spatial Error Model (SEM) and Spatial Lag Model (SAR) produced identical AIC, log-likelihood, and spatial parameter values (Table 8). This is an unusual result that warrants careful explanation. In this appendix, we clarify why the SEM and SAR models are **not** mathematically equivalent in general, and we offer a plausible explanation for the numerical coincidence observed in our data. For comprehensive treatments of spatial econometric model specification, see Anselin [28], Elhorst [33], and LeSage & Pace [37].

A.1 The SEM and SAR Models Are Not Equivalent

The Spatial Error Model is specified as:

$$y = X\beta + \mu, \quad \mu = \lambda W\mu + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Its reduced form is:

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

The mean of y under the SEM is:

$$E[y] = X\beta$$

The Spatial Lag Model is specified as:

$$y = \rho W y + X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

Its reduced form is:

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

The mean of y under the SAR is:

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

These two mean structures are generally not equal. They are equal only if:

$$X\beta = (I - \rho W)^{-1}X\beta$$

which requires $X = 0$ or $W = 0$. Neither condition holds in our application.

Furthermore, the variance-covariance matrices of y under the two models are:

$$\text{SEM: } \text{Var}(y) = \sigma^2 (I - \lambda W)^{-1} (I - \lambda W)^{-T}$$

$$\text{SAR: } \text{Var}(y) = \sigma^2 (I - \rho W)^{-1} (I - \rho W)^{-T}$$

These are equal only when $\lambda = \rho$ and W is symmetric. Even then, the mean structures differ, so the likelihoods are not identical.

A.2 Why the Results Are Identical in Our Application

The identical AIC, log-likelihood, and spatial parameter values in Table 8 are most likely due to one or more of the following:

1. Strong spatial dependence ($\lambda = \rho = 0.82$): When spatial dependence is very strong, the two models can produce numerically similar parameter estimates because the spatial structure dominates the regression component.
2. Simple regressor structure: Our regressors (row and column coordinates) are simple, smooth spatial trends. Both models can capture these trends equally well, leading to similar fits.
3. Numerical convergence: The maximum likelihood optimization algorithm in the spatialreg package may converge to the same solution for both models when the spatial dependence is strong and the regressors are simple.
4. Software implementation: The spatialreg package uses similar numerical routines for both models, which may contribute to the identical values.

A.3 Implications

Interpret the identical SEM and SAR results as a data-specific numerical coincidence, not a general mathematical theorem. Future users should:

1. Verify the SEM and SAR results independently for their own datasets.
2. Not assume that SEM and SAR are equivalent in general.
3. Report the SEM and SAR results separately, even if they are numerically identical.