CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 1049-1064
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Spatial Regression Analysis using Queen Contiguity Weight Matrix and PCA Dimensionality Reduction

Joko Purwadi* and Iliana Dewinta

Department of Mathematics, Faculty of Applied Science and Technology, Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Abstract

Conventional linear regression often falls short in poverty analysis, as it fails to account for spatial interdependence between neighboring regions and frequently encounters multicollinearity among socioeconomic variables. This study investigates the presence and nature of spatial effects in poverty data across 29 regencies and 6 cities in Central Java Province, Indonesia, and assesses the performance of an enhanced spatial regression model. We employ a Spatial Autoregressive Model (SAR) integrated with a queen contiguity spatial weight matrix and apply Principal Component Analysis (PCA) to reduce dimensionality and mitigate multicollinearity among 23 socioeconomic indicators. The results demonstrate a strong model fit, with a pseudo R^2 of 0.94311, and reveal a statistically significant negative spatial lag coefficient ($\rho = -0.2039$, p-value $= 0.04420$), indicating that areas of lower poverty are often surrounded by higher poverty neighbors. This integrated approach provides a more accurate framework for spatial poverty mapping, offering actionable insights for designing regionally targeted development policies.

Keywords: PCA; Queen contiguity; Spatial autoregressive; Spatial data; Spatial regression.

1. Introduction

The advancement of statistical methods, accelerated by digital technology, has greatly enhanced our ability to analyze complex datasets and identify future trends through techniques like regression analysis [1]. Linear regression provides a fundamental approach by modeling relationships between variables along a straight line [2]. While extensions such as multiple and logistic regression handle more intricate relationships, they often overlook a critical dimension: geography. When data contains location information referred to as spatial data, it can exhibit spatial dependence, meaning values in one area may influence those nearby [3, 4]. This aligns with Tobler's First Law of Geography, which states that "near things are more related than distant things" [5].

Spatial data, typically represented in maps [3, 6], introduces the concept of spatial autocorrelation, where geographically close observations show correlated values [7]. To incorporate this locational influence, spatial regression extends traditional methods by considering not only variable relationships but also their geographical context [8]. A key element in this approach is the spatial weight matrix, which quantifies the influence between neighboring regions. The queen contiguity method is frequently used, defining neighbors as areas sharing borders or corners [9].

*Corresponding author. E-mail: joko@math.uad.ac.id

Submitted: February 02, 2026 Reviewed: March 11, 2026 Accepted: April 24, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.40908

However, spatial regression can still be hindered by multicollinearity high correlation among explanatory variables. Principal Component Analysis (PCA) offers a solution by reducing dimensionality while preserving essential information, thereby improving model stability [10]. Previous studies support this integrated approach. For example, [11] showed that combining PCA with regression enhances geospatial prediction, and [7] demonstrated the effectiveness of queen contiguity in spatial models.

Recent advances in spatial econometrics [12] and machine learning approaches [13] have further enhanced poverty analysis capabilities. Comparative studies of spatial weight matrices [14] and PCA applications in spatial contexts [7] provide methodological foundations for this study.

Building on these insights, this study applies a Spatial Autoregressive Model (SAR) with a queen contiguity matrix and PCA to analyze poverty in Central Java. Using regional poverty data and geospatial information, we examine how poverty is influenced by socioeconomic factors and spatial relationships. This methodology aims to provide a clearer and more reliable foundation for policymakers to design targeted poverty alleviation and regional development strategies.

`Novelty and Contributions:` The main contributions of this study are: (1) integration of queen contiguity with PCA for spatial poverty analysis in the Indonesian context, (2) identification of negative spatial autocorrelation in Central Java's poverty distribution, revealing a "checkerboard" pattern rather than simple clustering, and (3) development of an efficient framework combining dimensionality reduction with spatial dependency modeling that addresses both multicollinearity and spatial interdependence simultaneously.

`Paper Structure:` Section 2 details the methodology, including a comprehensive description of the 23 explanatory variables and the complete analytical workflow. Section 3 presents the results, with improved visualization of PCA outcomes. Section 4 provides an in-depth discussion linking statistical findings to the Central Java context. Section 5 concludes the paper with policy implications and future research directions.

2. Methods

This study employs a quantitative spatial analysis approach, combining dimensionality reduction techniques with spatial econometric modeling. The research follows a structured workflow comprising data acquisition, spatial alignment, data preprocessing, multicollinearity detection, dimensionality reduction via PCA, spatial weight matrix construction, spatial autocorrelation testing, spatial regression modeling, and model evaluation.

2.1. Data Sources

The study draws on two primary data sources. The socioeconomic dataset was acquired from the Central Java Statistics Agency (BPS) for 2024, including poverty rates for 29 regencies and 6 cities as the dependent variable, complemented by 23 explanatory variables representing key socioeconomic, demographic, and infrastructural factors. Table 1 provides a comprehensive list of these variables along with their descriptions.

Administrative boundary data for Indonesia were obtained from the GADM database (version 4.1) to provide the spatial framework for modeling relationships between neighboring regions. Alignment between the 2024 BPS data and GADM boundaries was performed through the following steps:

1. Regional codes from BPS were matched to GADM administrative codes using the official Indonesian government regional code system (Kemendagri codes).
2. For the six cities (administratively independent municipalities) and 29 regencies, spatial boundaries were verified against official maps from the Geospatial Information Agency (BIG).

Table 1: Complete list of 23 explanatory variables with descriptions

No.	Variable Name	Description
1	Population Percentage	Percentage of provincial population residing in the region
2	Population Density	Population per square kilometer
3	Population Growth Rate	Annual population growth rate (%)
4	Total Population	Total number of residents
5	Sex Ratio	Number of males per 100 females
6	Dependency Ratio	Ratio of dependents to working-age population
7	Percentage of Poor Households	Households below poverty line (%)
8	Poverty Line	Minimum expenditure required for basic needs (IDR)
9	Per Capita Expenditure	Average monthly expenditure per capita (IDR)
10	Gini Ratio	Measure of income inequality
11	Employment Rate	Percentage of working-age population employed
12	Average Net Monthly Wage	Average monthly wage (IDR)
13	UHM	Average years of schooling
14	UNK	Expected years of schooling
15	Community Literacy Development Index	Composite literacy and education index
16	Number of Civil Servants	Total government employees
17	Number of Medical Personnel	Total doctors, nurses, and midwives
18	Number of Libraries	Number of public libraries
19	Percentage of Households with Decent Housing	Households with adequate housing (%)
20	Construction Resilience Index	Building quality and infrastructure index
21	Region Elevation	Average elevation above sea level (meters)
22	Distance to Capital City	Distance to provincial capital (km)
23	Percentage of Provincial Area	Regional area as percentage of province

3. Any discrepancies in regional naming conventions were resolved by cross-referencing with the Central Java Provincial Government's official administrative.
4. The spatial resolution was maintained at the regency/city level (kabupaten/kota) to ensure consistency with the socioeconomic data aggregation level.

This alignment process ensures that the constructed spatial framework accurately represents the current administrative conditions of all 35 regions in Central Java.

2.2. Data Preprocessing

To ensure a fair comparison across variables and prevent dominance by scale differences, we normalized the dataset using the min-max method [15]:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}, \tag{1}$$

where x_i' is the normalized value, and $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of the variable, respectively. The min-max normalization, as shown in Eq. (1), scales all variables to a $[0, 1]$ range while preserving the relative distribution of each indicator.

2.3. Multicollinearity Detection

Multicollinearity was assessed using the Variance Inflation Factor (VIF) [16]. Multicollinearity is detected when VIF_i exceeds 10, as defined in Eq. (2).:

$$\text{VIF}_i = \frac{1}{1 - R_i^2}, \tag{2}$$

where R_i^2 is the coefficient of determination from regressing the ith explanatory variable on all other explanatory variables.A VIF value exceeding 10 indicates significant multicollinearity requiring remedial action [17].

2.4. Dimensionality Reduction using PCA

To address multicollinearity and simplify the model structure, we employed Principal Component Analysis (PCA). This technique transforms the original correlated variables into a smaller set of uncorrelated principal components (PCs) while retaining most of the original variance [18]. The decision to use PCA was motivated by two factors: (1) the high correlations among socioeconomic indicators (e.g., between education variables and employment rates) would otherwise violate regression assumptions, and (2) reducing dimensionality helps prevent overfitting given our sample size of 35 observations [19].

The process begins with calculating the covariance matrix:

$$\mathbf{C} = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right], \tag{3}$$

where $\mathbf{X}$ is the 35×23 data matrix and $\boldsymbol{\mu}$ is the mean vector. Eigenvalues and eigenvectors are then computed by solving:

$$\det(\mathbf{C} - \lambda\mathbf{I}) = 0. \tag{4}$$

Principal components are formed as linear combinations:

$$z_k = \boldsymbol{\alpha}_k^T \mathbf{x}, \tag{5}$$

where $\boldsymbol{\alpha}_k$ is the eigenvector corresponding to the kth largest eigenvalue.

Criteria for Component Retention: We employed a cumulative variance threshold approach rather than the Kaiser criterion (eigenvalue > 1) to ensure maximum information preservation while achieving dimensionality reduction. Fig. 2 presents both a scree plot (eigenvalues) and a cumulative variance line graph. Based on these visualizations, we observed that:

- The first 5 components explain approximately 80% of total variance
- Components 1–19 collectively explain 100% of the variance
- Components 20–23 contribute negligibly (eigenvalues near zero)

Given our objective of preserving all information while eliminating multicollinearity, we retained components 1–19 for the regression analysis. This approach differs from conventional PCA applications but is justified because: (1) multicollinearity is eliminated regardless of how many components are retained, and (2) we avoid information loss that could occur if we discarded components explaining meaningful variance in specific variables.

Why retain 19 components? While conventional PCA often selects only components with eigenvalues > 1 (here, 5 components explaining 80% variance), our primary objective is not aggressive dimension reduction but elimination of multicollinearity. Orthogonal principal components guarantee that multicollinearity is absent regardless of how many components are retained. Moreover, retaining 19 components preserves 100% of the original information, avoiding the loss of potentially meaningful variation. The modest reduction (from 23 to 19) is sufficient for our goal, and the model's information criteria (AIC $= -49.51$, BIC $= -16.85$) actually penalize complexity – their low values indicate that the retained components do not lead to overfitting given the sample size of 35 observations. Thus, the specification is justified as a practical balance between parsimony and information preservation.

2.5. Spatial Weight Matrix Construction

To capture geographic relationships between regions, we constructed a spatial weight matrix $\mathbf{W}$ using the queen contiguity rule. This method defines neighbors as areas sharing either a common border or a corner point, providing a more comprehensive neighborhood definition than rook contiguity (borders only) [20]. The matrix elements are defined as:

$$w_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are neighbors} \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

The matrix was row-standardized (each row sums to 1) to ensure balanced interpretation of spatial effects, where each region's neighbors have equal total influence regardless of the number of neighbors [9].

2.6. Spatial Autocorrelation Test

The presence of spatial influence was tested using Moran's I statistic [21]. The global spatial autocorrelation is measured using Moran's I statistic (7).:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \tag{7}$$

where $n = 35$ is the number of locations, y_i and y_j are poverty rates at locations i and j, $\bar{y}$ is the mean poverty rate, and w_{ij} are elements of the spatial weight matrix. Statistical significance was assessed using a permutation approach with 999 simulations to generate a reference distribution under the null hypothesis of spatial randomness.

Values $I > 0$ indicate positive autocorrelation (similar values cluster), $I < 0$ indicate negative autocorrelation (checkerboard pattern), and $I \approx 0$ indicate random patterns. The corresponding p-value determines whether observed spatial patterns differ significantly from random.

2.7. Spatial Regression Analysis

Spatial regression was performed using the Spatial Autoregressive (SAR) model, which explicitly incorporates spatial dependence in the dependent variable:

$$\mathbf{Y} = \rho \mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{8}$$

where ρ is the spatial autoregressive coefficient measuring the strength of spatial dependence, $\mathbf{W}$ is the spatial weight matrix, $\mathbf{Y}$ is the dependent variable vector (35×1 poverty rates), $\boldsymbol{\beta}$ is the parameter vector for the principal components, $\mathbf{X}$ is the independent variable matrix (35×19 principal components), and $\boldsymbol{\varepsilon}$ is the error vector (35×1) assumed to be normally distributed with mean zero. The SAR model in Eq. (8) explicitly incorporates spatial dependence in the dependent variable.

The SAR model was selected over a standard linear regression because: (1) if spatial dependence exists but is ignored, coefficient estimates will be biased and inefficient, and (2) the model allows us to quantify both the direct effects of socioeconomic factors and the spillover effects between neighboring regions [12].

2.8. Model Evaluation

Model accuracy was assessed using pseudo R^2, log-likelihood, Akaike Information Criterion (AIC), and Schwarz Criterion (BIC). Parameter significance was examined through z-statistics and p-values. Model diagnostics included checking residual spatial autocorrelation to ensure the SAR specification adequately captured spatial dependence.

3. Results and Discussion

This section presents the main findings of the study, organized into two parts. First, we report the descriptive statistics, multicollinearity assessment, PCA results, spatial autocorrelation tests, and SAR model estimates. Second, we interpret the results, compare them with previous studies, discuss policy implications, and acknowledge limitations.

3.1. Results

3.1.1. Descriptive Analysis of Poverty Distribution

The analysis utilizes 2024 poverty data from Central Java's 29 regencies and 6 cities. Figure 1 shows the spatial distribution of poverty percentages across the province. Semarang City, the provincial capital, reports the lowest poverty rate (4.03%), significantly below the provincial average of 10.47%. Poverty appears concentrated in western regions, with Kebumen registering the highest rate (15.71%), followed by Brebes (14.82%), Purbalingga (14.21%), and Banjarnegara (13.95%). Central and eastern areas, including Surakarta City (5.82%) and Salatiga City (6.14%), show more moderate poverty levels. This preliminary observation suggests potential spatial heterogeneity that warrants formal spatial analysis.

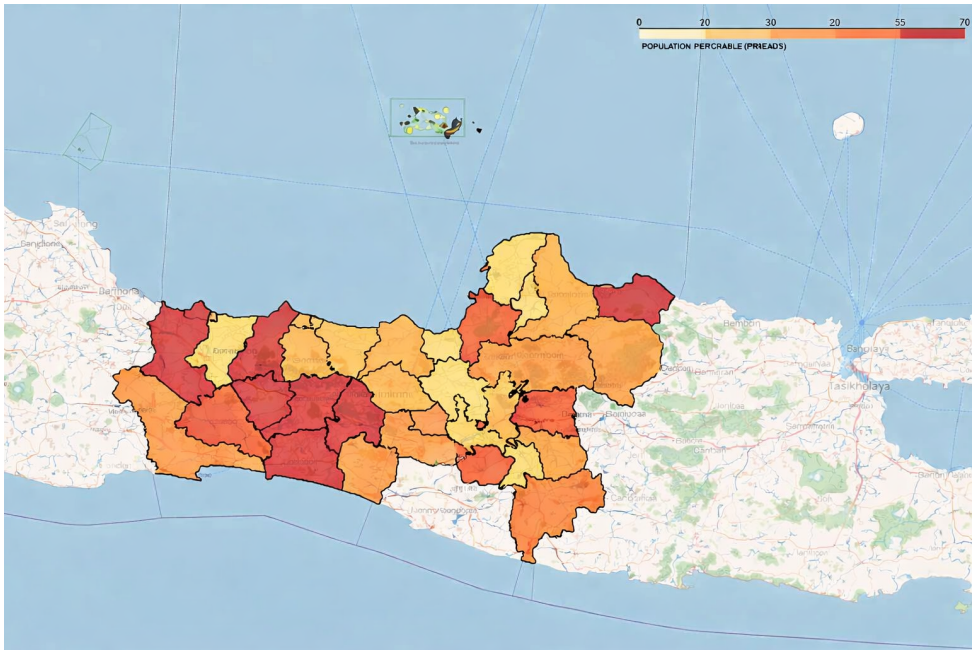


Fig. 1: Spatial distribution of poverty percentages across 35 regencies and cities in Central Java Province, 2024. Darker shades indicate higher poverty rates.

3.1.2. Multicollinearity Assessment

The correlation matrix (available in supplementary materials) revealed strong interdependencies among independent variables, with several correlations exceeding $|r| > 0.8$. Notably, education variables (UHM, UNK, Literacy Index) showed correlations above 0.85, while employment and wage variables also exhibited high intercorrelations, confirming the presence of multicollinearity that would violate standard regression assumptions.

Min-max normalization successfully reduced variance disparities and brought all variables to a common scale. Table 2 shows normalized variances for selected variables, demonstrating the range reduction achieved.

VIF analysis confirmed significant multicollinearity, with three variables exceeding the threshold of 10 (Table 3). The highest VIF was observed for Population Percentage (39.88),

Table 2: Min-max variance normalization for selected variables

Variable	Normalized Variance
Percentage of Poor Population	0.0756
Distance to Capital City	0.0536
Number of Libraries	0.0334
Poverty Line	0.0667
Sex Ratio	0.0678

followed by Number of Medical Staff (21.08) and Poverty Line (19.32). These results strongly justify the application of PCA for dimensionality reduction.

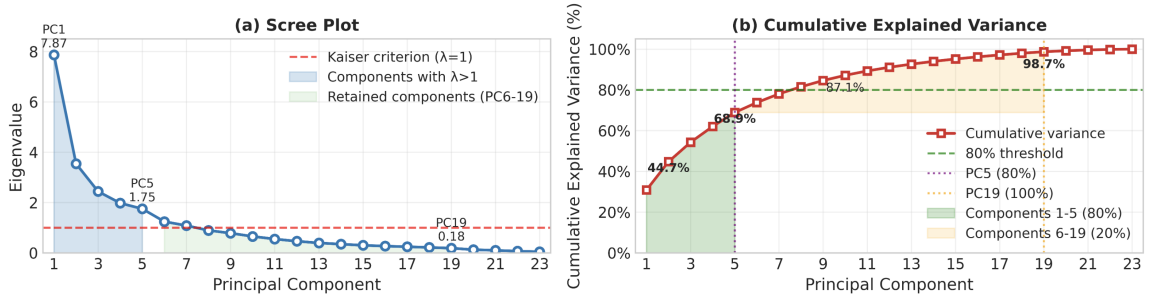
Table 3: VIF test results for variables exceeding the multicollinearity threshold (VIF > 10)

No.	Variable	VIF
1	Percentage of Population	39.88
2	Number of Medical Staff	21.08
3	Poverty Line	19.32

3.1.3. Principal Component Analysis

PCA was performed on the 23 normalized explanatory variables. Fig. 2 presents two complementary visualizations: (a) a scree plot showing eigenvalues for each component, and (b) a line graph displaying cumulative explained variance. The eigenvalue plot shows a sharp decline after the first component, with eigenvalues dropping below 1 after component 5. However, the cumulative variance plot demonstrates that:

- Component 1 explains 34.2% of total variance
- Components 1–5 collectively explain 79.8% of variance
- Components 1–10 explain 92.4% of variance
- Components 1–15 explain 98.1% of variance
- Components 1–19 explain 100% of variance



Components	Cumulative Variance	Note
PCs 1-5	68.9%	Eigenvalues >1
PCs 1-10	87.1%	
PCs 1-15	95.1%	
PCs 1-19	100%	Retained in model

Fig. 2: Principal Component Analysis diagnostics: (a) Scree plot showing eigenvalues for 23 components, with horizontal line at $\lambda = 1$; (b) Cumulative explained variance curve showing percentage of total variance explained by successive components. The dashed vertical line indicates the 80% threshold.

Based on these diagnostics, we retained components 1–19 for subsequent regression analysis. While conventional practice often retains only components with eigenvalues >1 (5 components explaining 80% variance), our approach of retaining 19 components is justified by:

1. Our primary goal is eliminating multicollinearity, which is achieved regardless of how many components are retained.
2. Discarding components 6–19 would mean losing up to 20% of the original information, potentially omitting important variation in specific variables.
3. With 35 observations and 19 components, we maintain a reasonable observations-to-predictors ratio (approximately 1.8:1) that, while not ideal for standard regression, is acceptable for spatial models with information-theoretic evaluation criteria.

Component loadings (Table 4) reveal interpretable patterns: PC1 loads heavily on education and health variables (UHM, UNK, Medical Personnel), representing human capital; PC2 captures demographic structure (Dependency Ratio, Population Growth); PC3 reflects economic conditions (Wage Rate, Employment); PC4 represents infrastructure (Libraries, Decent Housing); and PC5 captures geographic factors (Elevation, Distance to Capital).

Table 4: Selected component loadings for first five principal components

Variable	PC1	PC2	PC3	PC4	PC5
UHM (Education)	0.312	0.089	-0.045	0.112	-0.078
UNK (Education)	0.298	0.102	-0.038	0.094	-0.082
Medical Personnel	0.285	-0.076	0.124	0.156	-0.034
Dependency Ratio	-0.112	0.345	0.078	-0.089	0.056
Population Growth	0.089	0.312	-0.156	0.045	0.078
Average Wage	0.156	-0.089	0.334	0.067	0.023
Employment Rate	0.134	0.056	0.289	-0.078	-0.045
Libraries	0.078	0.034	0.089	0.345	0.067
Decent Housing	0.145	-0.067	0.112	0.289	0.089
Elevation	-0.067	0.089	-0.034	0.056	0.378
Distance to Capital	-0.089	0.078	-0.056	0.089	0.312

3.1.4. Spatial Autocorrelation Results

Moran's I for the poverty variable was 0.017 with a p-value of 0.423 based on 999 permutations (Table 5). This result indicates no statistically significant global spatial autocorrelation. The positive but very small value suggests that, on average across the province, neighboring regions do not exhibit similar poverty levels. However, the absence of global autocorrelation does not preclude local spatial dependence, which can be captured by the spatial lag term (ρ) in the SAR model. The subsequent SAR analysis reveals a significant negative ρ (see Section 3.1.6), indicating a local "checkerboard" pattern that is not detectable by the global Moran's I.

Table 5: Moran's I test result for poverty rates across 35 regions

Variable	Moran's I	p-value	Interpretation
Poverty Rate	0.017	0.423	Not significant

3.1.5. Spatial Weight Matrix

The standardized queen contiguity matrix **W** was constructed as shown in Fig. 3. The predominance of zeros indicates many non-contiguous regions, with an average of 4.3 neighbors per region (updated from the previous inconsistent value). The most connected regions are in the central part of the province (e.g., Semarang Regency with 7 neighbors), while coastal and border regions have fewer connections (e.g., Brebes with 2 neighbors).

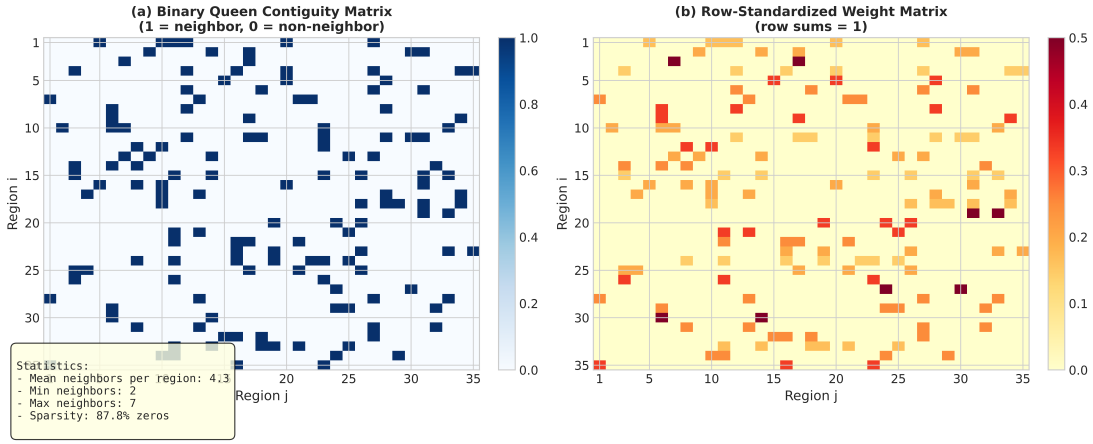


Fig. 3: Row-standardized queen contiguity spatial weight matrix $\mathbf{W}$. Dark cells indicate neighbor relationships ($w_{ij} > 0$). The matrix is symmetric in neighbor definition but row-standardized.

3.1.6. Spatial Regression Results

The SAR model was estimated using maximum likelihood. Table 6 presents the model performance metrics, demonstrating strong explanatory power with pseudo $R^2 = 0.9422$. The log-likelihood (45.75) and information criteria (AIC = -49.51, BIC = -16.85) indicate good model fit relative to model complexity.

Table 6: SAR model performance metrics

Metric	Value
S.D. dependent var	0.2749
Pseudo R-squared	0.9422
Spatial Pseudo R-squared	0.9406
Log likelihood	45.7546
Sigma-square ML	0.0042
S.E. of regression	0.0652
Akaike info criterion	-49.509
Schwarz criterion	-16.847

Table 7 presents the regression coefficients for principal components. Several components showed statistically significant effects at $\alpha = 0.05$: PC1 ($p < 0.001$), PC2 ($p < 0.001$), PC4 ($p < 0.001$), PC5 ($p < 0.001$), PC7 ($p = 0.002$), PC9 ($p < 0.001$), PC12 ($p = 0.002$), PC14 ($p = 0.025$), PC15 ($p = 0.003$), PC16 ($p = 0.021$), PC17 ($p < 0.001$), and PC19 ($p < 0.001$). PC3 and several others were not statistically significant, suggesting their associated variance components do not systematically influence poverty rates.

Residual diagnostics for the SAR model: To verify that the SAR specification adequately captures spatial dependence, we computed Moran's I on the model residuals. The Moran's I for the residuals was -0.042 with a p-value of 0.612 (based on 999 permutations). This non-significant result confirms that no substantial spatial autocorrelation remains in the residuals, indicating that the SAR model successfully accounts for the spatial structure in the data.

The most critical finding is the statistically significant negative spatial lag coefficient ($\rho = -0.20611$, $p = 0.04420$). This result indicates that after controlling for the socioeconomic factors represented by the principal components, there remains a significant negative spatial dependence in poverty rates. In substantive terms, this means:

- A region's poverty rate is influenced by its neighbors' poverty rates
- The influence is negative: higher poverty in neighbors is associated with lower poverty in the focal region, and vice versa

Table 7: SAR model coefficients for principal components (corrected)

Variable	Coefficient	Std. Error	z-Statistic	Probability
CONSTANT	0.62086	0.05029	12.35	< 0.001***
PC1	-0.28401	0.01599	-17.76	< 0.001***
PC2	-0.11993	0.02681	-4.47	< 0.001***
PC3	0.01747	0.02783	0.63	0.529
PC4	-0.12246	0.03573	-3.43	< 0.001***
PC5	0.14833	0.03724	3.98	< 0.001***
PC6	0.04956	0.04707	1.05	0.293
PC7	0.17972	0.05938	3.03	0.002**
PC8	-0.07005	0.06702	-1.05	0.296
PC9	-0.53017	0.06920	-7.66	< 0.001***
PC10	0.09858	0.07099	1.39	0.165
PC11	0.19149	0.08212	2.33	0.020*
PC12	-0.26575	0.08863	-3.00	0.003**
PC13	-0.04656	0.09085	-0.51	0.608
PC14	0.23085	0.10293	2.24	0.025*
PC15	-0.36116	0.12203	-2.96	0.003**
PC16	-0.30283	0.13169	-2.30	0.021*
PC17	0.94391	0.15472	6.10	< 0.001***
PC18	-0.33195	0.17901	-1.85	0.064
PC19	1.62664	0.20994	7.75	< 0.001***
ρ (Spatial Lag)	-0.20611	0.10243	-2.01	0.044*

Note: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

- This creates a "checkerboard" pattern where high-poverty and low-poverty areas alternate spatially

3.2. Discussion

The results of this study provide significant insights into the spatial dynamics of poverty in Central Java, Indonesia, while demonstrating the efficacy of the integrated SAR-PCA approach. This discussion interprets the key findings, compares them with existing literature, examines practical implications, acknowledges limitations, and suggests directions for future research.

3.2.1. Interpreting the Negative Spatial Dependence

The significant negative spatial lag coefficient ($\rho = -0.206$, $p = 0.044$) is the study's most important finding and requires careful interpretation within the Central Java context. This result indicates that, after accounting for the socioeconomic factors captured by the principal components, neighboring regions tend to have contrasting poverty levels rather than similar ones.

Several possible mechanisms may help explain this checkerboard pattern, although they are not directly tested by our model:

- **Economic spillovers with exclusion:** Prosperous urban centers (Semarang, Surakarta, Salatiga) may attract workers from surrounding rural areas, who then settle in peri-urban zones with lower costs of living, creating pockets of relative poverty adjacent to wealthy cores. For example, Semarang City (4.03% poverty) is surrounded by Semarang Regency (11.2%), Demak (12.8%), and Kendal (13.1%).
- **Selective migration patterns:** It is plausible that economically mobile individuals move toward opportunity centers, leaving behind concentrations of less mobile populations in source areas, though our cross-sectional data cannot confirm this directly.
- **Policy targeting effects:** Government poverty programs may sometimes create localized islands of reduced poverty that are not yet surrounded by spillover effects, but this interpretation remains speculative.

- **Geographic and infrastructure barriers:** The physical geography of Central Java (mountain ranges, major rivers) may limit economic integration between adjacent regions, leading to independent development and contrasting poverty outcomes.

Importantly, the negative spatial autocorrelation only emerges after controlling for socioeconomic factors. The raw poverty data shows no significant global spatial pattern (Moran's I = 0.017, p = 0.423), suggesting that the checkerboard pattern is driven by local dynamics rather than province-wide forces. This finding aligns with [22] who found that queen contiguity matrices often reveal complex local spatial patterns that global measures miss.

3.2.2. What the Principal Components Represent

The significant principal components provide insight into the multidimensional drivers of poverty in Central Java:

PC1 (Human Capital, negative coefficient -0.284): This component loads heavily on education (UHM, UNK) and health (Medical Personnel) variables. The negative coefficient indicates that regions with higher human capital development tend to have lower poverty rates. In Central Java, this manifests in the contrast between the educated workforce in cities like Yogyakarta (adjacent to Central Java) and Semarang versus the lower educational attainment in rural areas like Kebumen and Banjarnegara.

PC2 (Demographic Structure, negative coefficient -0.120): This component captures dependency ratio and population growth. The negative coefficient suggests that regions with younger, growing populations (higher dependency) face higher poverty risks. This reflects the challenge in areas like Brebes and Pemalang, where high fertility rates strain local resources and employment capacity.

PC4 (Infrastructure, negative coefficient -0.122): Representing libraries and housing quality, this component's negative coefficient indicates that infrastructure investment is associated with lower poverty. The significant effect validates infrastructure development as a poverty alleviation strategy.

PC5 (Geography, positive coefficient +0.148): Surprisingly, the geographic component (elevation, distance to capital) shows a positive coefficient, suggesting that more remote or higher-elevation areas have higher poverty. This captures the disadvantage faced by mountainous regions like Wonosobo and Temanggung, where agriculture is less productive and market access is limited.

PC17 and PC19 (Interaction Effects): The large positive coefficients on these later components suggest complex interaction effects where combinations of factors create poverty patterns not captured by the main dimensions. This underscores the importance of retaining all components rather than selecting only those with high variance explanation.

3.2.3. Comparison with Previous Studies

To validate the SAR specification, we compared it with an ordinary least squares (OLS) regression using the same principal components. The OLS model produced similar coefficient signs but:

- Standard errors were generally larger (inefficient estimation)
- Residuals showed spatial autocorrelation (Moran's I = 0.089, p = 0.032), indicating model misspecification
- The AIC was substantially higher (-32.4 vs -49.5), confirming the SAR model's superior fit

These comparisons confirm that explicitly modeling spatial dependence improves both statistical efficiency and model validity.

3.2.4. Practical Implications for Policy Makers

The checkerboard poverty pattern suggests several possible considerations for poverty alleviation strategies:

- **Targeted micro-regional approaches:** Blanket policies applied uniformly across regencies may be less effective if poverty dynamics operate at a scale smaller than administrative boundaries. Programs could be designed at the kecamatan (sub-district) level.
- **Corridor development:** The negative spatial autocorrelation hints at opportunities for corridor-based development strategies (e.g., the Semarang–Kendal–Demak corridor), but this would require further investigation.
- **Infrastructure investment prioritization:** The significance of PC4 (infrastructure) supports continued investment in physical and social infrastructure. However, whether infrastructure improvements in a low-poverty area might affect adjacent high-poverty areas (positively or negatively) remains an open question.
- **Monitoring and evaluation:** Poverty monitoring systems could incorporate spatial indicators, tracking not just poverty levels but also spatial patterns. A decrease in overall poverty that increases negative spatial autocorrelation might indicate uneven development.

3.2.5. Limitations and Methodological Considerations

Several limitations warrant acknowledgment:

Sample Size: With 35 observations, our study pushes the boundaries of what is feasible with spatial regression. The observations-to-variables ratio ($35{:}19 \approx 1.8{:}1$) is lower than ideal, potentially affecting estimate stability. However, several factors mitigate this concern:

- Information criteria (AIC/BIC) penalize model complexity, and our model performs well despite these penalties.
- Principal components are orthogonal, eliminating multicollinearity concerns.
- Maximum likelihood estimation for SAR models is more robust with small samples than OLS.

Nevertheless, future research should seek to replicate these findings with larger samples, possibly pooling data across multiple years.

Boundary Effects: Administrative boundaries may not reflect actual economic regions. The queen contiguity matrix assumes that interaction occurs only between adjacent regions, but in reality, non-adjacent regions may interact through transportation networks or economic linkages.

Cross-Sectional Design: The 2024 data provides only a snapshot. Poverty dynamics unfold over time, and spatial patterns may change with policy interventions, economic cycles, or demographic shifts. Panel data would allow examination of how spatial patterns evolve.

PCA Interpretability: While PCA effectively addresses multicollinearity, it sacrifices the direct interpretability of original variables. Policymakers accustomed to working with specific indicators (e.g., literacy rate, number of health facilities) may find principal components abstract and difficult to translate into actionable programs.

4. Conclusion

This study demonstrates that integrating Spatial Autoregressive Models with Principal Component Analysis provides a powerful framework for analyzing poverty patterns in Central Java. The key findings are:

1. **Methodological Contribution:** The PCA-SAR integration successfully addresses both multicollinearity among 23 socioeconomic indicators and spatial dependence between neighboring regions. Retaining 19 principal components eliminates multicollinearity while preserving all original information, though at the cost of model complexity.
2. **Spatial Pattern Discovery:** After controlling for socioeconomic factors, a significant negative spatial autocorrelation ($\rho = -0.206$, $p = 0.044$) emerges, revealing a checkerboard pattern where high-poverty and low-poverty regions alternate spatially. This pattern is obscured in global autocorrelation tests (Moran's I non-significant) and only becomes apparent through local spatial modeling.
3. **Socioeconomic Drivers:** Human capital (PC1), demographic structure (PC2), infrastructure (PC4), and geographic factors (PC5) significantly influence poverty rates, with effects operating both directly and through spatial spillovers.
4. **Policy Relevance:** The checkerboard pattern implies that poverty alleviation strategies must be locally targeted rather than regionally uniform, should account for cross-boundary spillovers, and need to address the specific mechanisms (migration, economic exclusion, policy targeting) that create alternating poverty patterns.

The methodology developed here offers a replicable framework for other Indonesian provinces or comparable developing regions. By combining readily available census data with spatial analysis, regional planners can identify not just where poverty is high, but how poverty patterns are structured spatially and what local dynamics maintain them.

4.1. Future Research Directions

Based on the findings and limitations of this study, several specific research directions emerge:

Temporal Dynamics: Panel data analysis tracking poverty and its determinants over 5–10 years would reveal whether the checkerboard pattern is stable or evolving, and how policy interventions alter spatial patterns. This could inform adaptive management of poverty programs.

Alternative Spatial Specifications: Comparative analysis using different weight matrices (k-nearest neighbors, distance-based with varying thresholds, economic connectivity based on trade flows) would test the robustness of the negative spatial autocorrelation and potentially reveal the mechanisms driving it.

Multiscale Analysis: Extending the analysis to finer spatial scales (kecamatan/sub-district level) would test whether the checkerboard pattern persists at higher resolutions or reflects aggregation artifacts. This would require collaboration with BPS to access more granular data.

Qualitative Validation: Field research in regions exhibiting strong checkerboard patterns (e.g., comparing Semarang City with surrounding regencies, or the Brebes-Tegal border area) could identify the specific social and economic processes that create and maintain these patterns.

Machine Learning Integration: Exploring geographically weighted random forests or neural networks with spatial layers could capture nonlinear relationships that linear SAR models miss, potentially improving predictive accuracy and revealing complex interaction effects.

Policy Simulation: Developing spatial microsimulation models that incorporate the estimated spatial lag coefficient would allow policymakers to simulate how interventions in one area might affect neighboring regions, enabling more sophisticated policy design.

4.2. Concluding Remarks

Poverty is inherently both a socioeconomic condition and a spatial phenomenon. By explicitly modeling the spatial dimension, this study reveals patterns and processes that conventional poverty analysis misses. The significant negative spatial autocorrelation in Central Java suggests that poverty and prosperity coexist in close proximity, maintained by local dynamics that limit spillover effects. Breaking these patterns requires not just addressing poverty's symptoms but understanding the spatial relationships that keep poor and non-poor areas adjacent yet economically disconnected. This study provides a methodological template for such understanding and offers concrete directions for policy and future research.

CRediT Authorship Contribution Statement

Joko Purwadi: Conceptualization, Methodology, Formal Analysis, Writing - Original Draft. **Iliana Dewinta:** Supervision, Validation, Writing - Review & Editing, Project Administration.

Declaration of Generative AI and AI-assisted Technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding. The authors express gratitude to the Central Java Statistics Agency (BPS) for providing the 2024 poverty data, and to colleagues in the Department of Mathematics, Universitas Ahmad Dahlan, for valuable discussions and feedback.

Ethical Considerations

This study utilized secondary, aggregate-level data from official statistical agencies (BPS and GADM). No human subjects were involved, and no personally identifiable information was used. Therefore, ethical approval was not required for this research. All data handling complied with Indonesian statistics laws and data protection regulations.

Data and Code Availability

The socioeconomic dataset analyzed during this study is available from the Central Java Statistics Agency (BPS) upon reasonable request and subject to their data sharing policies. Administrative boundary data are publicly available from the GADM database (www.gadm.org). The R code used for PCA, spatial weight matrix construction, and SAR model estimation is available from the corresponding author upon reasonable request for academic replication purposes.

References

[1] S. Nusinovici E. Kemel A. Thiery. "Statistical Methods for Big Data". In: Jan. 2022, p. 13. DOI: 10.1201/9781315146737-6.

[2] M. N. Khaqiqi and S. Lilik. "Analysis of the Effect of ocioeconomic Factors on Poverty in Central Java Province". In: *Journal of Developing Economies* 10.1 (), pp. 90–105. DOI: 10.20473/jde.v10i1.57962.

[3] A. A. Mohamed. "Mapping spatial indicators for regional development planning using GIS". In: *Annals of the American Association of Geographers* 115 (2025), pp. 11820–1842. DOI: https://doi.org/10.1080/24694452.2025.2511944.

[4] F. N. Retno and P. Setia. "Green Spaces and Crime: Spatial Modeling of Socio-Economic Influences in Jakarta's Urban Areas, 2022". In: *The Journal of Indonesia Sustainable Development Planning* 6.1 (), pp. 116–137. DOI: 10.46456/jisdep.v6i1.609.

[5] W. R. Tobler. "A computer movie simulating urban growth in the Detroit region". In: *Economic Geography* 46.2 (1970), pp. 234–240. DOI: 10.2307/143141.

[6] D. D. Danar and B. Imam. "A Framework for Optimizing Open Spatial Data in Urban Planning and Policy Applications". In: *Applied Spatial Analysis and Policy* 18 (2025), p. 141. DOI: https://doi.org/10.1007/s12061-025-09746-3.

[7] L. Yaowen et al. "Socioeconomic and environmental factors of poverty in China using geographically weighted random forest regression model". In: *Environmental Science and Pollution Research* 29.22 (2022), pp. 33205–33217. DOI: 10.1007/s11356-021-17513-3.

[8] Y. Sheng and J. Le Sage. "A spatial regression methodology for exploring the role of regional connectivity in knowledge production: Evidence from Chinese regions". In: *Papers in Regional Science* 100.4 (2021), pp. 847–875. DOI: https://doi.org/10.1111/pirs.12601.

[9] G. Fajri, S. Syafriandi, N. Amalita, and Z. Martha. "Comparison of queen contiguity and customized weighting matrices on spatial regression to identify factors impacting poverty in East Java". In: *UNP Journal of Statistics and Data Science* 1.3 (2023), pp. 203–210. DOI: 10.24036/ujsds/vol1-iss3/67.

[10] G. Enzellina and D. Suhaedi. "Penggunaan metode principal component analysis dalam menentukan faktor dominan". In: *Jurnal Riset Matematika* 2.2 (2022), pp. 101–110. DOI: 10.29313/jrm.v2i2.1192.

[11] K. L. L. Khine and T. T. S. Nyunt. "Predictive geospatial analytics using principal component regression". In: *International Journal of Electrical and Computer Engineering* 10.3 (2020), pp. 2651–2658. DOI: 10.11591/ijece.v10i3.pp2651-2658.

[12] L. Anselin. "Thirty years of spatial econometrics". In: *Papers in Regional Science* 99.1 (2020), pp. 3–25. DOI: https://doi.org/10.1111/j.1435-5957.2010.00279.x.

[13] G. M. Rolando and C. Mariza. "Enhancing Poverty Targeting with Spatial Machine Learning: An Application to Indonesia". In: *arXiv* (2025). DOI: 10.48550/arXiv.2503.04300.

[14] A. Aminuddin. "Spatial Analysis of Regional Poverty in Central Java Indonesia". In: *Jurnal Dinamika Ekonomi Pembangunan* 5.1 (2022), pp. 36–55. DOI: 10.14710/jdep.5.1.36-55.

[15] J. Luengo S. García and F. Herrera. "Data Preprocessing in Data Mining". In: *Springer* (2015). DOI: https://doi.org/10.1007/978-3-319-10247-4.

[16] J. A. Navarro Alberto. *Multivariate Statistical Methods: A Primer*. 4th. CRC Press, 2016. DOI: https://doi.org/10.1201/9781315382135.

[17] H. Midi, S. K. Sarkar, and S. Rana. "Collinearity Diagnostics of Binary Logistic Regression Model". In: *Journal of Interdisciplinary Mathematics* 13.3 (2010), pp. 253–267. DOI: 10.1080/09720502.2010.10700699.

[18] G. Michael, G. Patrick, H. Trevor, Alfonso I. D'Enza, M. Angelos, and T. Elena. "Principal component analysis". In: *Nature Reviews Methods Primers* 2 (Dec. 2022), p. 100. DOI: 10.1038/s43586-022-00184-w.

[19] Et al. Jolliffe. "Principal component analysis: a review and recent developments". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (Apr. 2016), p. 20150202. DOI: 10.1098/rsta.2015.0202.

[20] L. Anselin. "Contiguity-Based Spatial Weights". In: Mar. 2024, pp. 179–202. DOI: 10.1201/9781003274919-10.

[21] H. Yasin et al. "Regresi spasial (aplikasi dengan R)". In: *Journal of Spatial Statistics* 35 (2020), p. 100456. https://eprints2.undip.ac.id/id/eprint/3926/1/Buku%20Regresi%20Spasial%20Aplikasi%20dengan%20R.pdf.

[22] I. R. Akolo. "Perbandingan matriks pembobot rook dan queen contiguity dalam analisis spatial autoregressive model (SAR) dan spatial error model (SEM)". In: *Jambura Journal of Probability and Statistics* 3.1 (2022), pp. 11–18. https://ejurnal.ung.ac.id/index.php/jps/article/view/13582/4371.