OPEN ACCESS
Journal of Multidisciplinary Applied Natural Science Vol. 6 No. 2 (2026) https://doi.org/10.47352/jmans.2774-3047.345
Check for updates
Research Article

Spatial Clustering with Autocorrelation-Based Weighting for Regional Socio-Economic Pattern Analysis: A Case Study of East Java

Rahma Fitriani*, Eni Sumarminingsih, and Luthfatul Amaliana

Received : July 2, 2025	Revised : September 21, 2025	Accepted : November 19, 2025	Online : January 10, 2026

Abstract
Clustering, an unsupervised machine learning technique, categorizes objects into groups based on shared characteristics. When applied to spatial data, the assumption of independence is often violated due to similarities among adjacent regions—a phenomenon known as spatial autocorrelation. To address this, spatial clustering incorporates both non-spatial attributes (e.g., socio-economic indicators) and spatial attributes (e.g., geographic location), with spatial attributes weighted based on their influence in defining clusters. In regional economic development, creating clusters that are both spatially coherent and socio-economically homogeneous is critical for effective policy design. Strong interactions among neighboring regions can promote more integrated and balanced growth. This study proposes a spatial clustering framework that optimizes spatial attribute weighting according to the degree of spatial autocorrelation. A simulation study using 2023 data from East Java's 38 regencies/municipalities determines optimal weights under varying spatial dependence levels. The results show that optimal spatial weights increase with the number of clusters and vary according to the strength of spatial autocorrelation. Applied to East Java, the method produced clusters with higher socio-economic homogeneity than official zones, though with reduced spatial contiguity. These findings highlight the importance of adaptive, autocorrelation-aware clustering to improve regional planning and support more evidence-based development strategies.

Keywords: policy planning, regional economic development, simulation study, spatial dependence, unsupervised learning

1. INTRODUCTION

Cluster analysis is an unsupervised machine learning technique that groups similar objects while separating dissimilar ones. In spatial clustering, this method is applied to geographic regions, incorporating location as a key attribute. Unlike traditional clustering, which assumes independence among observations, spatial clustering accounts for spatial dependencies, making it a distinct approach within multivariate statistics [1]. By integrating geographical positioning, spatial clustering refines conventional clustering methods to better capture spatial relationships [2]-[4]. Spatial clustering is widely used in fields that rely on spatial data, such as economic development [5], epidemiology [6], and public health [7]. Since spatial data includes geographic references, observations are often interdependent, in which variables observed at one location tend to resemble those in nearby locations, or be influenced by shared regional factors. This spatial similarity or interaction are statistically measured through spatial autocorrelation.

Clustering techniques can be categorized into partitional methods (e.g., K-Means, K-Medoids), hierarchical clustering, and locality-based methods (e.g., density-based clustering, random distribution clustering) [8][9]. These techniques rely on distance -based algorithms to measure similarity or dissimilarity among objects. However, classical clustering methods do not inherently preserve spatial contiguity, making them less effective for spatial applications. To address this, early studies adjusted classical clustering results to enforce spatial continuity, but this approach often led to an excessive number of clusters [10][11]. Later research introduced alternative methods that incorporate spatial attributes, such as geometric centers or geographic coordinates, into clustering algorithms [12][13]. This shift led to the development of spatial clustering, where geographic location plays a critical role in cluster formation. Like classical clustering, spatial clustering can follow a hierarchical or partitional approach. Hierarchical methods use contiguity constraints to form clusters, including zoning, regionalization, spatially constrained clustering, and the p-region

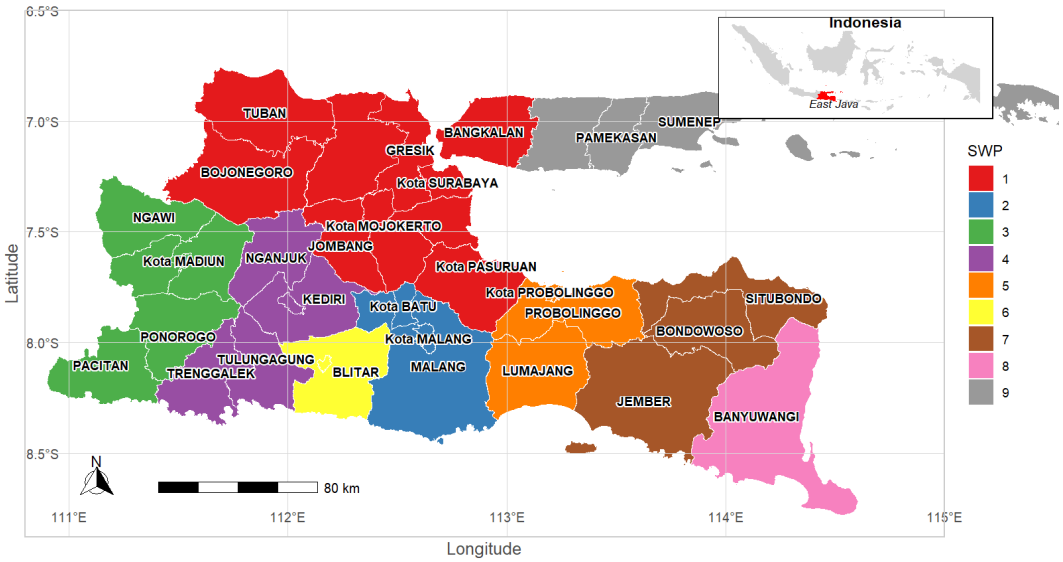

Figure 1. Map of development zoning divisions (*Satuan Wilayah Pengembangan* - SWP) of East Java, based on Provincial Spatial Planning Regulation No. 5 of 2012.

problem. These approaches divide n spatial units into p contiguous regions while maximizing intra-region similarity [14][15]. In partitional spatial clustering, methods such as the automatic zoning problem and max-p region problem allocate n spatial units into p larger, contiguous zones that exhibit internal homogeneity [16][17].

Unlike cross-sectional data, where observations are independent, spatial data exhibits spatial dependence, meaning attributes in one unit are influenced by nearby units. This phenomenon, known as spatial autocorrelation, is commonly measured using Moran's I, a widely used index for detecting spatial dependence [18][19]. A higher Moran's I value indicates a stronger spatial relationship. While Moran's I identifies spatial clustering patterns, it does not specify their locations. To address this, Anselin developed the local indicator of spatial association (LISA), which assesses spatial dependence at individual locations relative to global autocorrelation [20]. The local Moran's I index, a localized version of the global Moran's I, serves as a key tool in spatial clustering analysis.

Despite extensive research in spatial clustering, prior studies have primarily focused on forming clusters while maintaining spatial contiguity, without explicitly considering the role of spatial autocorrelation in cluster formation. Peeters et al. [4] used Getis-Ord's G statistic, a measure of spatial autocorrelation, to define clusters based on local G-values of key clustering variables. However, they did not explicitly integrate geographic positioning into their clustering approach. Incorporating spatial autocorrelation in clustering is also relevant to Stojanova et al., who applied predictive clustering trees (PTs) rather than spatial clustering techniques [21].

In regional economics, clustering geographically contiguous regions with homogeneous socio-economic attributes is crucial for effective zoning. These clusters consist of administratively adjacent areas that share similar socio-economic characteristics and exhibit strong economic interactions, reflected in significant positive spatial autocorrelation. Policymakers leverage these patterns to promote balanced and accelerated economic growth. For example , Provincial Spatial Planning Regulation No. 5 of 2012 designates administrative divisions into Development Units (*Satuan Wilayah Pengembangan* - SWP) in East Java (see Figure 1). Each SWP consists of a core development region and surrounding regions, maintaining geographic contiguity, which aligns with spatial clustering principles. The regulation intends for core regions to drive economic balance and acceleration, but their effectiveness depends on the presence of significant positive spatial autocorrelation, ensuring strong economic interconnectivity within the SWP.

Given this background, spatial autocorrelation is fundamental in defining clusters within spatial data,

particularly when clustering objectives rely on spatial interactions, such as in the formation of SWPs in East Java. Effective cluster determination must account for both geographic location and the degree of spatial autocorrelation. However, spatial autocorrelation has typically been considered at the descriptive or post-clustering evaluation stage, rather than being explicitly incorporated into the clustering process itself, particularly in the calibration of spatial attribute weights. This presents a gap in understanding the functional role of spatial dependence in influencing cluster structure. To address this gap, we hypothesize that higher levels of spatial autocorrelation correspond to a greater optimal weight for spatial attributes in clustering algorithms. The rationale is that strong spatial autocorrelation indicates a dominant role of spatial proximity in shaping data patterns, and should therefore be weighted more heavily in cluster formation. In contrast, with weaker spatial autocorrelation, non-spatial attribute similarity may be more decisive.

To test this hypothesis, a simulation-based approach is employed using synthetically generated data with varying levels of spatial autocorrelation. The 2023 economic growth of 38 regencies/cities in East Java serves as the basis for data generation. By varying levels of spatial autocorrelation, the study examines how the optimal spatial attribute weight shifts in response. Building on the simulation findings, the study further develops a spatial clustering approach that optimally incorporates spatial attributes, guided by the level of spatial autocorrelation observed in the data. Spatial autocorrelation helps determine how much influence geographic proximity should have in forming clusters, ensuring that regions grouped together are not only similar in socio-economic terms but also spatially coherent. This method is then applied to form spatial clusters for East Java's regencies and cities based on economic growth-related variables. The resulting clusters will be compared with existing SWPs to assess their alignment and effectiveness. Ultimately, this research provides a quantitative foundation for policy formulation, supporting more precise and evidence-based strategies for regional economic planning and development.

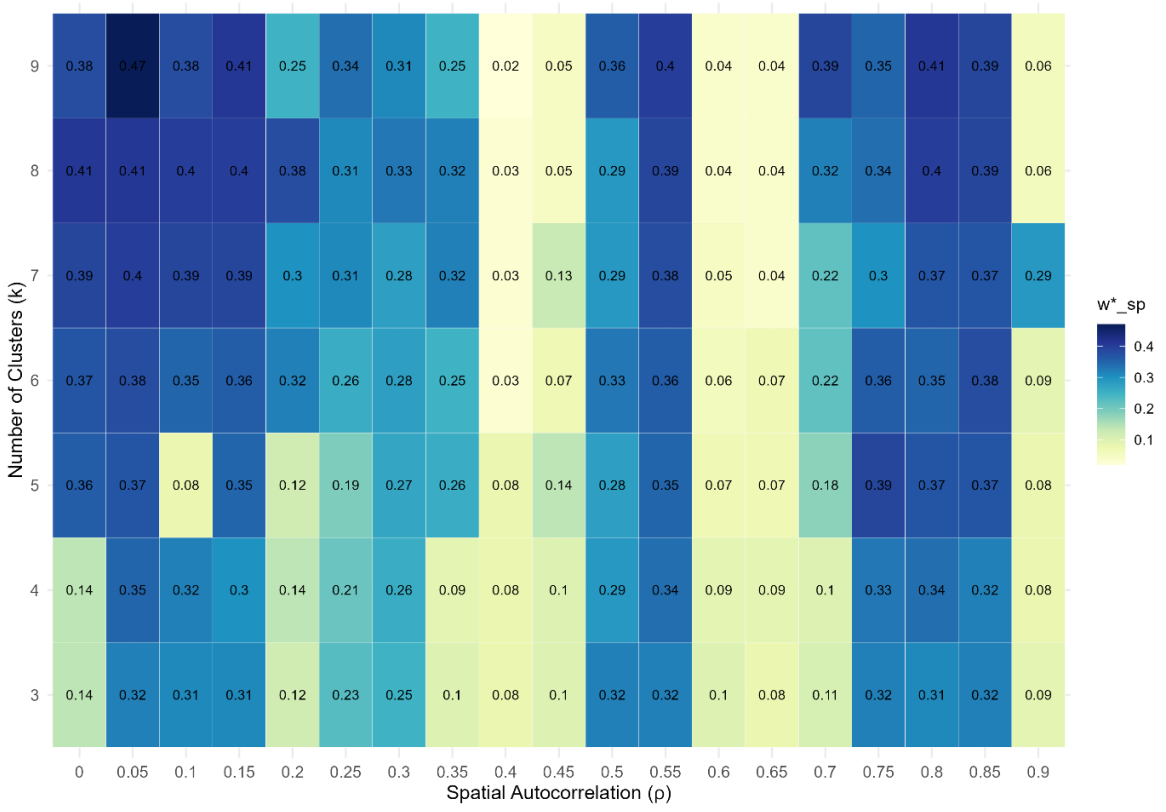

Figure 2. The optimal spatial weight (w^*_{sp}) avross varying degrees of spatial autocorrelation (ρ) and number of clusters (k).

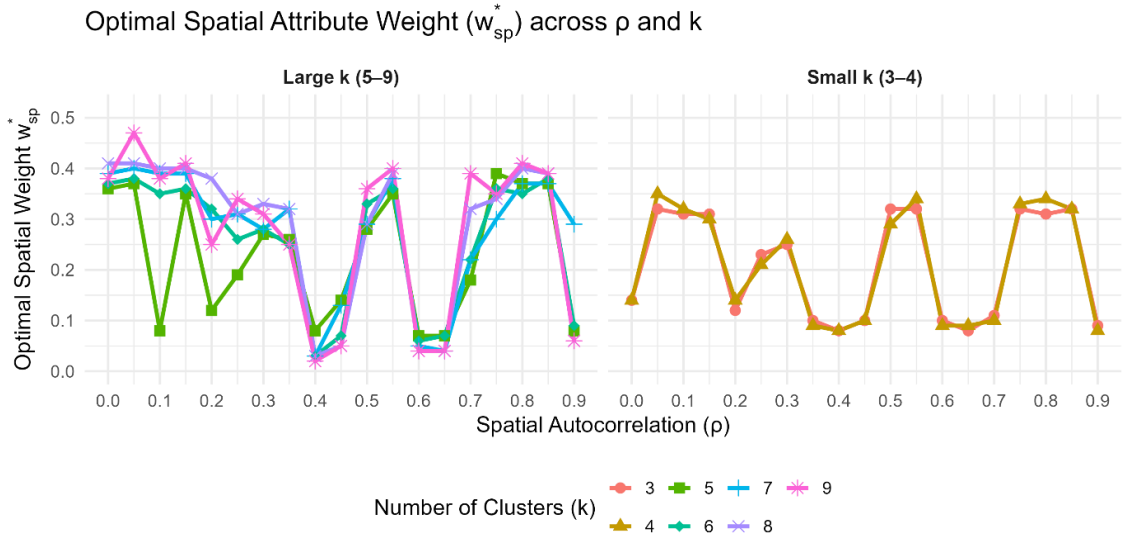

Figure 3. The optimal spatial weight (w^*_{sp}) avross varying degrees of spatial autocorrelation (ρ) and cluster regimes.

2. MATERIALS AND METHODS

2.1. East Java Dataset

The study uses a dataset of socioeconomic indicators from the year 2023, covering all 38 regencies/municipalities in East Java, obtained from Statistics Indonesia (*Badan Pusat Statistik* – BPS). The non spatial variables included in the analysis are GDP Growth (GGDP in %), workforce participation rate (WPR in %) and human development index (HDI, score from 0 to 1). These variables represent key indicators of regional economic development and serve as the basis for determining similarity across regions. In addition to socio-economic data, a digital map of East Java is used to extract the geographic coordinates (longitude and latitude) for each regency/ municipality. These coordinates are used as spatial attributes in the clustering process.

2.2. Generated of Spatially Dependent Data

To evaluate the impact of spatial autocorrelation on clustering outcomes, a simulation based data generation process was implemented. The procedure uses East Java's dataset as the input, to generate three variables which each has spatial dependence structure. The structure follows a spatial autoregressive (SAR) model, where the observed values inform the mean vector and variance parameters for each generated variable. In this study, the SAR model is chosen over alternative spatial models such as the spatial error model (SEM) or conditional autoregressive (CAR) models because of its alignment with the purpose of assessing spatial autocorrelation in clustering variables, rather than estimating regression parameters. Specifically, SAR captures the spatial dependence directly in the outcome variable without requiring additional predictor variables. This feature is crucial in the context of spatial clustering, where the primary concern is whether the variables used for clustering exhibit spatial autocorrelation. Unlike SEM, which focuses on spatial dependence in the error term, or CAR models that are typically used in Bayesian hierarchical frameworks, SAR is more appropriate for measuring the endogenous spatial interaction among the observed values of a single variable [22].

Following the SAR specification, the variance of each variable (σ^2_j, j =1,2,3) is used to construct the variance covariance matrix $(\sigma^2_j, (\boldsymbol{I} - \rho \boldsymbol{W})^{-T} (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$, where ρ represents the degree of spatial autocorrelation and $\boldsymbol{W}$ is the queen contiguity based spatial weights matrix derived from the geographic adjacency of regions. For each variable, an n x 1 vector is generated from a multivariate normal distribution, with the original data serving as the mean vector and the spatially structure variance covariance matrix as the dispersion component. In this study n = 38 corresponding to the number of

regencies/municipalities ini East Java. The simulation are conducted across a range of ρ values: 0,0.05,0.1,0.15, ... , 0.9, to represent scenarios with varying degrees of spatial autocorrelation. For each value ρ, 1000 datasets are generated to support a robust analysis.

2.3. Spatially Weighted K-Means Clustering

To ensure a balanced influence between spatial and non spatial variables in the clustering process, this study employs a weighted K-Means algorithm. The algorithm adjusts the contribution of each attribute type using a weighted scheme, enabling spatial attributes (i.e. geographical coordinates) and non spatial attributes to be incorporated proportionally according to a predefined spatial weight.

2.3.1. Calculating Weight for Spatial and Non-Spatial Attributes

Let p be the total number of attributes, including both spatial (geographic coordinates) and non-spatial (economic variables) attributes. In this study p = 5, comprising three non spatial variables and two spatial variables (longitude and latitude). Let w_{sp} represent the assigned weight for each spatial attribute. The remaining total weight, 1 - 2w_{sp}, is distributed equally across the p - 2 non spatial variables. Thus the weight for each non spatial variables, w_{nsp} is defined as Equation (1).

$$w_{nsp} = \frac{(1 - 2w_{sp})}{p - 2} \quad (1)$$

This formulation ensures that the total weight across all attributes sums to 1. When w_{sp} = 0, clustering is based solely on the non spatial variables. Conversely when w_{sp} = 0.5, clustering is driven entirely by geographic coordinates. Intermediate value of w_{sp} allow for varying degrees of spatial influence in the clustering outcome, enabling the analyis to identify an optimal balance between attribute types under different spatial autocorrelation scenarios. The resulting weight vector, $\boldsymbol{W}$ is constructed as Eq. 2.

$$\boldsymbol{w} = [w_{nsp}, w_{nsp}, \dots, w_{nsp}, w_{sp}, w_{sp}] \quad (2)$$

where the first p - 2 elements correspond to non spatial attributes, the last two elements correspond to spatial attributes.

The use of fixed weighted scheme is motivated by its computational simplicity, interpretability, and prior use in spatial clustering literature (e.g., [23] [24]), where similar linear weighting schemes have been applied to balance spatial and attribute features. While alternative schemes such as entropy -based or variance-based adaptive weighting (e.g., [25][26]) could offer data-driven weight optimization, such approaches are more suitable for supervised settings or contexts where classification performance is the main goal. Since the current study focuses on the influence of spatial autocorrelation on clustering results rather than automated weight learning, the fixed weighting scheme offers a clear experimental control. This enables systematic evaluation of clustering performance across a range of predefined spatial weights and spatial autocorrelation levels.

2.3.2. Applying Weights on Standardized Dataset

Before clustering, all variables are standardized to have zero mean and unit variance. The standardized attribute vector is denoted as $\boldsymbol{X'}$ The element-wise weighted attribute vector $\boldsymbol{X'}_{weighted}$ is then computed as Eq. 3.

$$\boldsymbol{X'}_{weighted} = \boldsymbol{X'} \circ \boldsymbol{W} \quad (3)$$

where $\circ$ denotes the Hadamard (element-wise) product. This step ensures that each variable contributes to the clustering process in proportion to its assigned weight.

2.3.3. Clustering with Weighted Attributes

After weighting, the K-Means algorithm is applied to partition the data into k clusters. The algorithm uses Eucledian distance computed from the weighted attribute vectors. The quality of the resulting clusters is evaluated using the within-cluster sum of squares (WCSS) defined as Eq. 4.

$$WCSS = \sum_{j=1}^{k} \sum_{i \in C_j} \|\boldsymbol{X'}_i - \boldsymbol{\mu}_j\|^2 \quad (4)$$

Where k is the number of clusters, C_j represents cluster j, $\boldsymbol{X'}_i$ the weighted attribute vector of data point i in cluster j, $\boldsymbol{\mu}_j$ is the centroid (mean vector) of cluster j, and $\| \boldsymbol{X'}_i - \boldsymbol{\mu}_j\|^2$ is the squared Euclidean Distance between data point iii and its cluster

Table 1. Summary of optimal spatial weights (w^*_{sp}) based on spatial autocorrelation (ρ) and number of clusters (k).

w^*_{sp}	k						
ρ	3	4	5	6	7	8	9
0.00	0.14	0.14	0.36	0.37	0.39	0.41	0.38
0.05	0.32	0.35	0.37	0.38	0.40	0.41	0.47
0.10	0.31	0.32	0.08	0.35	0.39	0.40	0.38
0.15	0.31	0.30	0.35	0.36	0.39	0.40	0.41
0.20	0.12	0.14	0.12	0.32	0.30	0.38	0.25
0.25	0.23	0.21	0.19	0.26	0.31	0.31	0.34
0.30	0.25	0.26	0.27	0.28	0.28	0.33	0.31
0.35	0.10	0.09	0.26	0.25	0.32	0.32	0.25
0.40	0.08	0.08	0.08	0.03	0.03	0.03	0.02
0.45	0.10	0.10	0.14	0.07	0.13	0.05	0.05
0.50	0.32	0.29	0.28	0.33	0.29	0.29	0.36
0.55	0.32	0.34	0.35	0.36	0.38	0.39	0.40
0.60	0.10	0.09	0.07	0.06	0.05	0.04	0.04
0.65	0.08	0.09	0.07	0.07	0.04	0.04	0.04
0.70	0.11	0.10	0.18	0.22	0.22	0.32	0.39
0.75	0.32	0.33	0.39	0.36	0.30	0.34	0.35
0.80	0.31	0.34	0.37	0.35	0.37	0.40	0.41
0.85	0.32	0.32	0.37	0.38	0.37	0.39	0.39
0.90	0.09	0.08	0.08	0.09	0.29	0.06	0.06

centroid. A lower WCSS indicates tighter cluster, suggesting a more effective clustering structure.

2.4. Application of Spatial Clustering on East Java Dataset

The analysis began with an exploratory spatial data assessment to evaluate the degree of spatial autocorrelation present in the 2023 East Java dataset. Specifically, the Global Moran's I statistics was computed for three key socio-economic indicators: GGDP, WPR, and HDI. The Moran's I values were calculated using a Queen contiguity-based spatial weight matrix, offering insights into the extent of spatial dependence among neighbouring regions. These results served as an empirical foundation for selecting the appropriate range of spatial weights w_{sp} in the clustering process.

Building on this, the spatially weighted K-Means algorithm was applied to the same dataset, integrating both non-spatial attributes (GGDP, WPR, HDI) and spatial coordinates. The algorithm was executed using a range of spatial weights derived from the earlier simulation study, with values tailored to match the observed spatial autocorrelation levels. In this study, the number of clusters (k = 9) was intentionally set to align with the existing development zones officially defined by the East Java's SWP. Rather than determining the optimal number of clusters using internal validation techniques such as the elbow method or silhouette score, this study adopts a comparative analysis approach. Similar methodological choices are found in prior studies. Grassi et al. fixed the number of clusters equal to the *known class labels* in benchmark datasets to enable one-to-one comparison between clustering algorithms and established classifications, rather than estimating k purely from internal validity indices [27]. Likewise, Watson emphasized that geographically defined clusters in applied research are often determined by administrative or operational boundaries, such as

census tracts or policy zones, when the analytical purpose concerns efficiency or alignment within those units [28]. Accordingly, by fixing k at 9, the resulting spatial clusters can be directly compare to the establish SWPs. This enables an assessment of how well the clusters formed using the proposed technique replicate or diverge from the current development zones. The focus, therefore, is on evaluating the alignment between the data driven spatial clusters and policy driven regional divisions, which is more relevant to the practical aims of this research.

The resulting spatial clustering configuration was then compared against the official SWP zoning from four key perpectives: alignment with administrative boundaries, geographic conitguity, internal socio-economic homogeneity, and potential mismatches suggesting opportunities for zonation refinement. This comparison allows for policy-relevant validation of the clustering model, offering evidence for whether data-driven spatial clusters can support or enhance existing regional development strategies.

3. RESULTS AND DISCUSSIONS

3.1. Optimal Spatial Attribute Weight Across Varying Levels of Spatial Autocorrelation and Cluster Sizes

The proposed spatially weighted K-Means algorithm generalizes the standard K-Means method by introducing a tunable spatial weight parameter w_{sp}. When $w_{sp} = 0$, the model reduces to a purely attribute-based K-Means clustering, enabling internal baseline comparisons without the need for a separate method. To systematically evaluate the model's responsiveness to spatial structure, a simulation framework was designed to vary the spatial autocorrelation parameter ρ systematically from 0 (no spatial dependence) to 0.9 (strong spatial dependence). For each level of ρ, the algorithm determines the optimal spatial weight w_{sp} by maximizing intra-cluster homogeneity and coherence. This design effectively functions as a built in parameter sensitivity analysis, highlighting how clustering outcome adapts to varying degrees of spatial autocorrelation.

The simulation results reveal how the optimal spatial attribute weight (w^*_{sp})in weighted K-Means clustering is influenced by both the degree of spatial autocorrelation (ρ) and the number of clusters (k). The patterns depicted in Figure 2 and Figure 3 challenge the assumption of a monotonic relationship and instead suggest a more complex interaction between spatial dependence and cluster sizes.

The plots in Figure 3 further clarify this interaction by separating trends for small-cluster and large cluster regimes. The number of clusters emerges as a dominant factor in shaping the value of w^*_{sp}, with clear regime differences between low k (3-4 clusters) and high k (5-9 clusters). At low k, the clustering algorithm consistently assigns lower spatial weights. This implies that when only a few clusters are needed, non-spatial attributes play a greater role in distinguishing between groups. Very low spatial autocorrelation ($\rho = 0$) and very high autocorrelation ($\rho = 0.9$) both result in particularly low spatial weights—often below 0.20. Surprisingly, medium spatial autocorrelation levels ($\rho = 0.4$ and $\rho = 0.6$) also yield low optimal w^*_{sp}, even though some spatial structure is present. This pattern indicates that at small number of clusters, the non spatial attributes alone are sufficient to create broad, meaningful clusters and adding spatial constraints provide little additional value to clustering performance. For the remaining ρ values, optimal spatial weight rises moderately into 0.21-0.34 range, reflecting some benefit of geographic information depending on the alignment of spatial and attribute structures.

As the number of clusters increases, the algorithm generally assigns higher spatial attribute weights, reflecting a growing need to enforce spatial coherence among a greater number of smaller, localized clusters. However, the relationship between spatial autocorrelation (ρ) and the optimal spatial weight (w^*_{sp}) is not strictly increasing, and varies depending on how well spatial structure complements the socio-economic attributes. At higher values of ρ particularly 0.4, 0.6, and most notably at 0.9, the optimal spatial weights decline when many clusters are specified (i.e., higher k). This suggests that spatial proximity alone does not always align with non spatial attributes grouping needs when many clusters are required. Rather than reflecting a failure to detect spatial structure, these lower weights may indicate

that forcing spatial contiguity at high ρ can conflict with forming distinct, attribute-based clusters—particularly when regions with similar non spatial attributes profiles are not spatially adjacent.

To better understand this pattern, it is helpful to revisit the simulation design underlying these results. In the simulation setup, each variable is generated using a SAR model with varying levels of spatial autocorrelation (ρ). Although ρ controls the strength of spatial dependence, it does not guarantee that the generated socio-economic patterns (used in clustering) will align neatly with geographic proximity — especially when clustering into many small groups (e.g., high value of k). For example, at high ρ (e.g., ρ = 0.9), the SAR process strongly pulls neighboring units to have similar values. But due to the random noise in simulation, non-adjacent units may still end up having very similar values for the clustering variables, while adjacent units may differ more than expected. If the clustering process were to enforce strong spatial coherence (i.e., apply high spatial weight w_{sp}), it would group units based on proximity even when their generated attributes differ — reducing clustering quality. Therefore, in such cases, the optimal w_{sp} is lowered by the algorithm to allow attribute-based grouping to prevail over proximity, avoiding spatially forced, incoherent clusters.

The algorithm in this study demonstrates results consistent with previous findings in the spatial clustering literature. Prior research has highlighted that trade-off between spatial contiguity and attribute similarity is a fundamental challenge in spatial clustering. Duque et al. demonstrated that excessive spatial constraints can produce spatially contiguous but internally incoherent regions in their Max-p-regions formulation [16]. Similarly Chavent et al. in their *ClustGeo* framework, proposed a parameter to balance attribute based and spatial dissimilarities, showing that increasing spatial weight may degrade attribute cohesion when inherent spatial dependence is already high [29]. For the remaining levels of spatial autocorrelation (other than ρ = 0.4,0.6,0.9), the optimal spatial weight (w^*_{sp}) is considerably higher, ranging between 0.22 and 0.47. This reflects scenarios in which geographic proximity and attribute similarity reinforce each other, improving clustering performance. Overall, the upper bound of spatial weight is significantly greater when a larger number of clusters is specified, confirming the increasing need to preserve spatial coherence as the number of clusters grows.

These findings confirm the hypothesis stated in the introduction: that higher levels of spatial autocorrelation tend to correspond with greater optimal spatial weights in clustering. The simulation shows that stronger spatial dependence often increases the influence of spatial attributes in cluster formation. However, this relationship is not strictly monotonic (i.e., ρ = 0.9), spatial weights decrease due to conflicts between spatial proximity and attribute-based grouping—suggesting that spatial information should not be overemphasized when it misaligns with attribute similarity. Additionally, across all levels of spatial autocorrelation, the optimal spatial weight tends to increase with the number of clusters. This reflects a greater need to enforce geographic cohesion in smaller cluster sizes. However, the interaction between spatial autocorrelation and the number of clusters is nuanced. A high ρ value does not always justify a high spatial weight, especially when doing so would compromise attribute-based homogeneity.

The simulation results, in terms of optimal spatial weights (w^*_{sp}) across values of ρ and k, are presented in Table 1. This table serves as a practical reference for setting spatial parameters in empirical clustering tasks. Rather than recalibrating the spatial weight for each new dataset, practitioners can consult the table to approximate a reasonable starting value of w_{sp} based on the observed spatial

Table 2. Results of Moran's I test for East Java's socio-economic variables.

Variable	Moran's I	E(I)	Var(I)	p_value
GGDP	0.3549	-0.027	0.0156	0.0011
WPR	0.3374	-0.027	0.0172	0.0027
HDI	0.3910	-0.027	0.0172	0.0007

autocorrelation (e.g., from Moran's I) and the target number of clusters. For example, if a dataset exhibits moderate spatial dependence (e.g., $\rho \approx 0.3$) and aims to form 7 clusters, the table suggests setting w_{sp} near 0.28 as a well-supported initial value. This not only streamlines the model tuning process but also helps maintain consistency between simulation insights and real-world applications.

These findings have practical relevance for regional clustering tasks, such as the delineation of East Java's SWPs. When targeting a specific number of clusters (e.g., $k = 9$), spatial information should be weighted more heavily, especially if the spatial autocorrelation is moderate. However, planners should remain cautious at extremely high ρ values, where spatial coherence may paradoxically conflict with socio-economic variability, leading to counterintuitive clustering results if spatial weights are overemphasized.

3.2. Application of Spatially Weighted K-Means Clustering on the East Java Dataset

Following the simulation study, the spatially weighted K-Means clustering method was applied to the actual 2023 socio-economic dataset of 38 regencies/municipalities in East Java. The number of clusters is set to $k = 9$, corresponding to the nine official zones (SWP) specified in East Java's spatial development plan (RTRW 2011–2031). A critical input to the clustering model is the weight assigned to spatial attributes, which governs the influence of geographic proximity in the formation of clusters. The weight is chosen based on the results of simulation study, which determined the optimal spatial weight for each level of spatial autocorrelation.

To define the appropriate weight range for East Java's dataset, the degree of spatial autocorrelation of each variable is evaluated using Moran's I statistic. Moran's I quantifies the extent to which values in one region are correlated with values in neighboring regions. The statistical significance of Moran's I was assessed using the analytical distribution under the null hypothesis of spatial randomness. This approach evaluates the observed Moran's I against its expected value and variance, assuming that the observed variable is randomly distributed across spatial units. This procedure is consistent with standard practices in spatial econometrics and regional analysis, where Moran's I is commonly employed to diagnose spatial dependence before constructing spatial weight matrices or performing clustering. For instance, Anselin [20] and Cliff and Ord [30] established Moran's I as a diagnostic measure for spatial autocorrelation, and more recent works such as Duque et al. [16] and Grassi et al. [27] applied Morans's I to evaluate spatial dependence when defining spatially constrained clusters. Accordingly, the use of Moran's I in this study provides a theoretically grounded and empirically supported basis for determining the range of spatial weights that reflect the inherent degree of spatial dependence in the observed data.

The analysis revealed statistically significant positive spatial autocorrelation across all three socio-economic indicators (see Table 2). Specifically, the HDI recorded the highest Moran's I value at 0.3910, followed by GGDP at 0.3549 and WPR at 0.3374. These values indicate moderate spatial autocorrelation across the region, suggesting that while socio-economic indicators exhibit spatial clustering, the patterns are not strongly dictated by geography alone. This reinforces the rationale for using spatial weighting in clustering, where attribute similarity can still play a central role alongside spatial proximity. Accordingly, the optimal spatial attribute weight is selected from the corresponding simulation results, falling between 0.25 and 0.31 (see Table 1). This weight range ensures a balanced consideration of both attribute similarity and spatial proximity in the clustering process. The resulting clusters are then evaluated and compared against the existing SWP configuration from four key perspectives. These values indicate moderate spatial autocorrelation across the region, suggesting that while socio-economic indicators exhibit spatial clustering, the patterns are not strongly dictated by geography alone. This further justifies the use of spatial weighting in clustering while allowing attribute similarity to play a central role.

3.2.1. Alignment with Official SWP Zoning

The first aspect of evaluation focuses on the alignment between the spatial weighted clusters and the official SWP zoning. The side-by-side

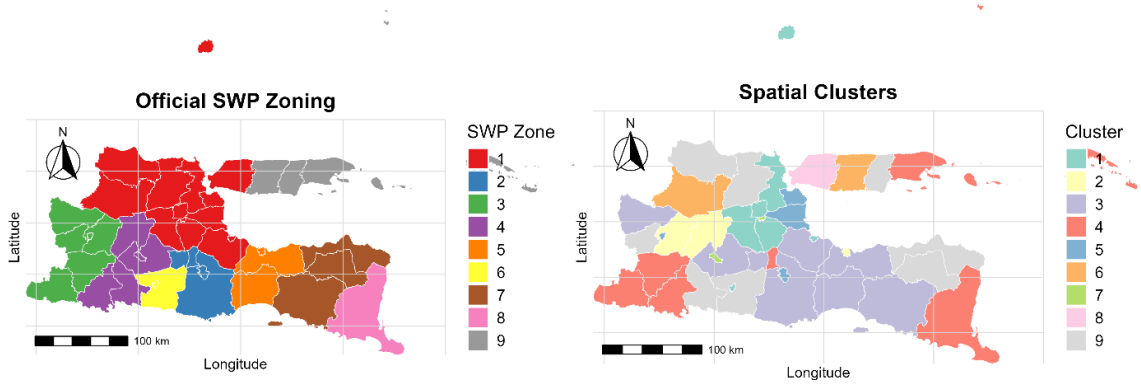

Figure 4. Comparison of spatially weighted clusters and official SWP zoning in East Jawa.

comparison in Figure 4, reveals that many of the spatial clusters differ substantially from the official SWP configuration. Several regencies are assigned to different clusters than their designated SWPs, indicating misalignments between administrative zoning and actual socio-economic patterns. These misalignments may reflect underlying economic characteristics—such as GDP growth, workforce participation, or human development levels— that connect certain regions more closely to areas outside their assigned SWP. This is particularly evident in the central and eastern parts of East Java, where spatial clusters span multiple SWP boundaries. These findings indicate that while the existing SWP zoning captures broad spatial logic, it may not fully reflect the current economic structure of the province. The spatial clustering results offer empirical insights that can inform more adaptive and data-driven regional planning, particularly for enhancing economic integration across development zones.

3.2.2. Contiguity Comparison between Spatial Weight Cluster and Official SWP

Another critical aspect of evaluating spatial clustering outcomes is assessing the geographic contiguity of resulting clusters. This study quantifies contiguity using the contiguity consistency index (CCI), calculated as Eq. 5.

$$CCI_c = \frac{1}{n_c} \quad (5)$$

where n_c is the number of disconnected polygonal components (i.e., separate contiguous areas) within a given cluster *c*. A CCI value of 1.00 indicates perfect spatial contiguity, meaning all regencies/municipalities in that cluster form a single connected region, while lower values (near to zero) reflect fragmentation across multiple disconected subregions. The left panel in Figure 4 displays the official SWP zoning map, designed with administrative and spatial coherence in mind. In contrast, the right panel in Figure 4 shows the spatial clustering result, derived using the spatially weighted K-Means algorithm with a moderate spatial weight (optimally chosen between 0.25–0.31) as determined by the dataset's spatial autocorrelation profile.

The official SWP zones generally exhibit higher contiguity. Out of the 9 SWPs, five demonstrate perfect contiguity (CCI = 1.00). For example, SWP 3, 4, 5, 6, and 8 all form single, unified geographic regions. SWP 12, although covering 12 regencies, is only modestly fragmented into three components (CCI = 0.333). SWP 9, which includes island regions (e.g., Madura and Bawean), shows the most extreme fragmented with a CCI of 0.0303, due to its archipelagic nature. In contrast, the spatial weighted clusters exhibit lower contiguity. Only cluster 8 exhibits perfect contiguity (CCI = 1.00). Several other clusters, such as: cluster 4 despite containing only 6 regencies/municipalities, is split across 34 disconnected geographic components (CCI = 0.03), indicating severe fragmentation. This extreme fragmentation is primarily due to the inclusion of island and non-contiguous areas, such as Madura and Bawean, which are geographically isolated; cluster 9 spans 8 regencies but is split into 6 disconnected geographic components (CCI =0.17); clusters 1 and 3 each consist of 4 separate geographic components (CCI = 0.25), suggesting moderate fragmentation. These findings reflect a

lower spatial cohesion in the spatial weight clusters.

3.2.3. Socio-Economic Homogeneity Within Clusters

To evaluate the internal consistency of each cluster, intra-cluster variation was measured using the standard deviation of three key socio-economic indicators: GGDP, HDI, and WPR. The bar chart in Figure 5 compares the standard deviation for each variable between the spatial weighted clusters and the official SWP zones. This figure clearly shows that spatial weighted cluster have lower standard deviations, particularly for HDI and WPR, indicating greater internal homogeneity. This is especially evident in clusters such as 2, 3, and 5, where the spatially weighted K-Means method groups together regions with more consistent socio-economic characteristics. In contrast, the official SWP zones exhibit higher variability within zones—highlighting a broader spread in regional development levels among the grouped areas.

3.3. Policy Implications

The results of this study offer several valuable insights for regional development policy, particularly in East Java. The misalignments observed between the spatially weighted clusters and the official SWP zoning suggest that existing administrative boundaries may not fully reflect the current socio-economic realities of the region. In many cases, the clustering results group together regencies that, despite being administratively separate, share similar economic development patterns. This highlights the potential for data-driven refinement of development zones that align more closely with on-the-ground socio-economic structures.

These findings emphasize the need for a more adaptive and evidence-based approach to regional planning. Similar patterns have been observed in other studies, where data-driven clusters based on socio-economic similarities often diverge from official administrative boundaries. For instance, Wicht et al. found that functional regions derived from economic indicators exhibit greater internal homogeneity than predefined administrative zones, suggesting that policy frameworks could benefit from incorporating such functional delineations [31]. Similarly, Fang et al. demonstrated that urban area boundaries derived from multi-source geospatial data provide a more accurate representation of socio-spatial structures than traditional zoning approaches [32]. In the context of East Java, several regions that appear fragmented across SWP boundaries in the clustering results show stronger socio-economic similarities with areas within their data-defined clusters than with their formally designated SWPs. This suggests opportunities for cross-regional collaboration in areas such as infrastructure, employment, and human development initiatives. Policy frameworks could be designed to facilitate cooperation between these economically aligned regions, regardless of their formal zoning, thus enhancing regional integration and reducing development disparities.

Moreover, the analysis of cluster contiguity reveals a meaningful trade-off between geographic coherence and socio-economic similarity. The official SWPs, which were designed with spatial continuity in mind, generally exhibit higher contiguity. In contrast, the spatially weighted clusters, shaped by moderate spatial weighting, tend to prioritize socio-economic homogeneity, leading to more fragmented but internally consistent groupings. Although some clusters (e.g., Cluster 4) show very low contiguity (CCI=0.03), this does not necessarily render them meaningless. Instead, it may reflect the presence of similar regions that are geographically dispersed but share common socio economic development characteristics. From a practical standpoint, policymakers in East Java could use these results to re-evaluate SWP boundaries. For instance, if regencies like Pacitan, Ponorogo, Trenggalek, Banyuwangi and Sumenep, consistently cluster together despite belonging to different SWPs, policymakers might consider coordinated planning efforts across these economically aligned regencies. However, this trade-off deserves further exploration. Future studies could consider incorporating minimum contiguity thresholds or spatial compactness penalties (e.g., as suggested by Duque et al. [16]) to better align clusters with practical implementation needs. The study also demonstrates the utility of incorporating spatial autocorrelation—measured through Moran's I—into the clustering process. This allows planners to strike a balance between spatial proximity and attribute similarity when

forming development zones. By periodically updating spatial weights based on evolving autocorrelation patterns, regional planning can become more dynamic and responsive to socio-economic changes.

3.4. Limitation and Future Research Direction

This study has several limitations that warrant consideration and offer opportunities for future research. First, the clustering analysis relies on only three socio-economic indicators—GGDP, WPR, and HDI. While these are critical measures of regional development, they do not capture the full spectrum of factors influencing regional dynamics, such as industrial specialization, environmental conditions, infrastructure access, or education outcomes. To strengthen the relevance and precision of future clustering analyses, it is recommended that a more comprehensive set of variables be included. This would allow for more robust policy insights and better-aligned regional planning strategies.

Second, the current methodology does not impose spatial contiguity constraints, which can result in fragmented clusters with very low contiguity indices. Although this flexibility allows the algorithm to prioritize attribute homogeneity, it may reduce practical interpretability and coherence from a policy perspective. Future research could explore the integration of contiguity constraints to improve the spatial usability of the results in administrative and planning contexts. Third, while the proposed method generalizes K-Means by introducing a spatial weight parameter (which defaults to K-Means when w_{sp} = 0), it has not yet been benchmarked against alternative clustering models such as hierarchical or density-based algorithms. Incorporating such comparisons in future work will provide further validation of the method's adaptability and effectiveness across various spatial configurations.

4. CONCLUSIONS

This study developed an unsupervised machine learning approach for spatial data, that integrates optimal spatial weighting informed by spatial autocorrelation levels. The simulation results demonstrate that both the spatial autocorrelation coefficient (ρ) and the number of clusters (k) influence the optimal spatial weight (w_{sp}). When spatial and attribute structures align, higher weights help enforce geographic contiguity, especially as the number of clusters increases. However, when these structures conflict, optimal weights may decrease even under high ρ, reflecting a trade-off between spatial and attribute coherence. Application to East Java's 2023 data showed that the proposed approach produced clusters with stronger socio-economic similarity than the official SWP zones, albeit with lower spatial contiguity. These results suggest the value of data-driven

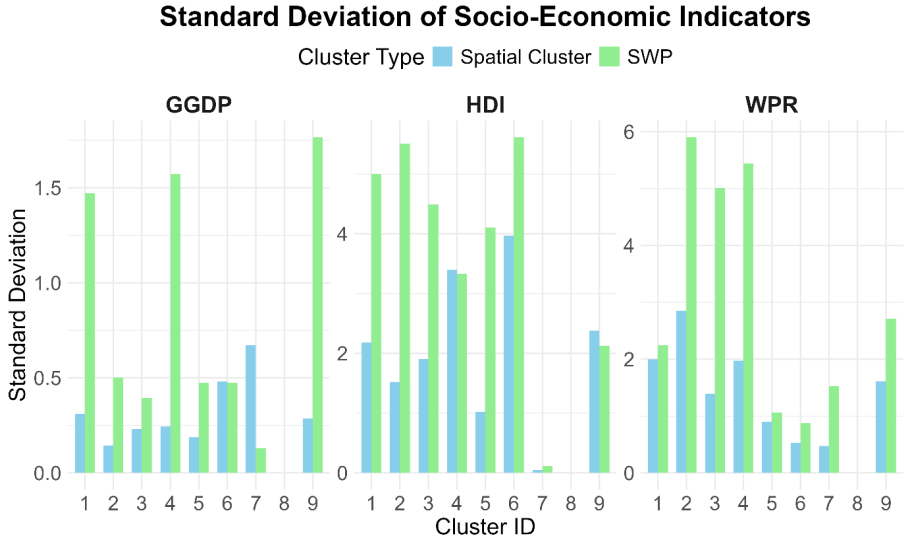

Figure 5. Comparison of intra-cluster homogeneity (standard deviation) between spatial clusters and SWP zones in East Java for socio-economic variables.

zoning refinements to better align development planning with economic realities. Future research could expand the variable set to capture multidimensional development patterns and consider adding spatial contiguity constraints to improve geographic cohesion—important in policy contexts. Further benchmarking against alternative clustering algorithms would also strengthen robustness and applicability. Overall, this approach offers a flexible framework for regional analysis and adaptive planning, balancing spatial and non-spatial structures in practical, data-informed ways.

AUTHOR INFORMATION

Corresponding Author

Rahma Fitriani — Department of Statistics, Universitas Brawijaya, Malang-65145 (Indonesia);
orcid.org/0000-0002-6478-7661
Email: rahmafitriani@ub.ac.id

Authors

Eni Sumarminingsih — Department of Statistics, Universitas Brawijaya, Malang-65145 (Indonesia);
orcid.org/0000-0003-4283-2852

Luthfatul Amaliana — Department of Statistics, Universitas Brawijaya, Malang-65145 (Indonesia);
orcid.org/0000-0002-6624-891X

Author Contributions

Conceptualization, Methodology, Resources, Writing – Original Draft Preparation, Visualization, R. F.; Supervision, and Funding Acquisition, R. F.; Software, Formal Analysis, and Investigation, R.F. and E.S.; Validation, and Writing – Review & Editing, R. F., E. S., and L. A.; Data Curation, and Project Administration, L. A.

Conflicts of Interest

The authors declare no conflict of interest.

ACKNOWLEDGEMENT

This study is supported by the 2024 UB Doctoral Research Grant (Hibah Penelitian Doktor UB).

DECLARATION OF GENERATIVE AI

During the preparation of this work, the authors used ChatGPT (developed by OpenAI) in order to assist with correcting errors in R codes, editing text for clarity and coherence, and refining the focus and structure of the manuscript content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

[1] R. A. Johnson and D. W. Wichern. (2002). "Applied Multivariate Statistical Analysis". Prentice Hall, Upper Saddle River, NJ.

[2] E. Kolatch. (2001). "Clustering Algorithms for Spatial Databases: A Survey".

[3] Q. Liu, M. Deng, Y. Shi, and J. Wang. (2012). "A Density-Based Spatial Clustering Algorithm Considering Both Spatial Proximity and Attribute Similarity". *Computers & Geosciences.* **46** : 296-309. 10.1016/j.cageo.2011.12.017.

[4] A. Peeters, M. Zude, J. Käthner, M. Ünlü, R. Kanber, A. Hetzroni, R. Gebbers, and A. Ben -Gal. (2015). "Getis-Ord's Hot- and Cold-Spot Statistics as a Basis for Multivariate Spatial Clustering of Orchard Tree Data". *Computers and Electronics in Agriculture.* **111** : 140-150. 10.1016/j.compag.2014.12.011.

[5] N. Yu, M. De Jong, S. Storm, and J. Mi. (2012). "Transport Infrastructure, Spatial Clusters and Regional Economic Growth in China". *Transport Reviews.* **32** (1): 3-28. 10.1080/01441647.2011.603104.

[6] T. E. Carpenter. (2001). "Methods to Investigate Spatial and Temporal Clustering in Veterinary Epidemiology". *Preventive Veterinary Medicine.* **48** (4): 303-320. 10.1016/S0167-5877(00)00199-9.

[7] S. Wang and J. Wu. (2020). "Spatial Heterogeneity of the Associations of Economic and Health Care Factors with Infant Mortality in China Using Geographically Weighted Regression and Spatial Clustering". *Social Science &*

Medicine. **263** : 113287. 10.1016/j.socscimed.2020.113287.

[8] S. Landau, M. Leese, D. Stahl, and B. S. Everitt. (2011). "Cluster Analysis". John Wiley & Sons.

[9] L. Kaufman and P. J. Rousseeuw. (2009). "Finding Groups in Data: An Introduction to Cluster Analysis". John Wiley & Sons.

[10] S. Openshaw. (1973). "A Regionalisation Program for Large Data Sets". *Computer Applications.* **3** (4): 136-147.

[11] S. Openshaw, P. J. Taylor, and N. Wrigley. (1979). "Statistical Applications in the Spatial Sciences". Pion, London. 127-144.

[12] A. T. Murray and T. K. Shyy. (2000). "Integrating Attribute and Space Characteristics in Choropleth Display and Spatial Data Mining". *International Journal of Geographical Information Science.* **14** (7): 649-667. 10.1080/136588100424954.

[13] R. Webster and P. A. Burrough. (1972). "Computer-Based Soil Mapping of Small Areas from Sample Data: I. Multivariate Classification and Ordination". *Journal of Soil Science.* **23** (2): 210-221. 10.1111/j.1365-2389.1972.tb01654.x.

[14] A. T. Murray and T. H. Grubesic. (2002). "Identifying Non-Hierarchical Spatial Clusters". *International Journal of Industrial Engineering.* **9** : 86-95.

[15] J. C. Duque, R. Ramos, and J. Suriñach. (2007). "Supervised Regionalization Methods: A Survey". *International Regional Science Review.* **30** (3): 195-220. 10.1177/0160017607301605.

[16] J. C. Duque, L. Anselin, and S. J. Rey. (2012). "The Max-p-Regions Problem". *Journal of Regional Science.* **52** (3): 397-419. 10.1111/j.1467-9787.2011.00743.x.

[17] J. C. Duque, R. L. Church, and R. S. Middleton. (2011). "The p-Regions Problem". *Geographical Analysis.* **43** (1): 104-126. 10.1111/j.1538-4632.2010.00810.x.

[18] L. Anselin. (1988). "Spatial Econometrics: Methods and Models". Kluwer Academic Publishers, Dordrecht; Boston. 10.1007/978-94-015-7799-1.

[19] A. D. Cliff and J. K. Ord. (1972). "Testing for Spatial Autocorrelation among Regression Residuals". *Geographical Analysis.* **4** (3): 267-284. 10.1111/j.1538-4632.1972.tb00475.x.

[20] L. Anselin. (1995). "Local Indicators of Spatial Association (LISA)". *Geographical Analysis.* **27** (2): 93-115. 10.1111/j.1538-4632.1995.tb00338.x.

[21] D. Stojanova, M. Ceci, A. Appice, D. Malerba, and S. Džeroski. (2013). "Dealing with Spatial Autocorrelation When Learning Predictive Clustering Trees". *Ecological Informatics.* **13** : 22-39. 10.1016/j.ecoinf.2012.10.006.

[22] J. P. LeSage and R. K. Pace.(2009)." Introduction to Spatial Econometrics". CRC Press, Boca Raton, FL. 10.1201/9781420064254.

[23] J. Z. Huang, M. K. Ng, H. Rong, and Z. Li. (2005). "Automated Variable Weighting in K -Means Type Clustering". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **27** (5): 657-668. 10.1109/TPAMI.2005.95.

[24] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. (2010). "Understanding of Internal Clustering Validation Measures". *Proceedings of the IEEE International Conference on Data Mining*. 911-916. 10.1109/ICDM.2010.35

[25] A. Jain and D. Zongker. (2002). "Feature Selection: Evaluation, Application, and Small Sample Performance". *IEEE Transactions on Pattern Analysis and Machine Intelligence.* **19** (2): 153-158. 10.1109/34.574797.

[26] C. Y. Tsai and C. C. Chiu. (2004). "A Purchase-Based Market Segmentation Methodology". *Expert Systems with Applications.* **27** (2): 265-276. 10.1016/j.eswa.2004.02.005.

[27] K. Grassi, É. Poisson-Caillault, A. Bigand, and A. Lefebvre. (2020). "Comparative Study of Clustering Approaches Applied to Spatial or Temporal Pattern Discovery". *Journal of Marine Science and Engineering.* **8** (9): 713. 10.3390/jmse8090713.

[28] S. I. Watson. (2022). "Efficient Design of Geographically Defined Clusters with Spatial Autocorrelation". *Journal of Applied Statistics.* **49** (13): 3300-3318. 10.1080/02664763.2021.1941807.

[29] M. Chavent, V. Kuentz-Simonet, A. Labenne, and J. Saracco. (2018). "ClustGeo: An R Package for Hierarchical Clustering with Spatial Constraints". *Computational Statistics.* **33** (4): 1799-1822. 10.1007/s00180-018-0791-1.

[30] A. D. Cliff and J. K. Ord. (1981). "Spatial Processes: Models and Applications". Pion Limited, London.

[31] A. Wicht, P. Kropp, and B. Schwengler. (2020). "Are Functional Regions More Homogeneous than Administrative Regions?". *Papers in Regional Science.* **99** (1): 135-165. 10.1111/pirs.12471.

[32] C. Fang, L. Zhou, X. Gu, X. Liu, and M. Werner. (2025). "A Data-Driven Approach to Urban Area Delineation Using Multi-Source Geospatial Data". *Scientific Reports.* **15** (1): 8708. 10.1038/s41598-025-93366-x.