

ORIGINAL ARTICLE

ML-SAS: A Machine-Learning-Based Spatial Autocorrelation Statistic for Mixed-Type Data

Woohyung Kim  | Sang-Il Lee 

Department of Geography Education, College of Education, Seoul National University, Seoul, Republic of Korea

Correspondence: Sang-Il Lee (si_lee@snu.ac.kr)

Received: 14 December 2025 | **Revised:** 21 September 2026 | **Accepted:** 30 September 2026

Keywords: machine-learning-based similarity | multivariate and mixed-type data | spatial autocorrelation statistic | unsupervised random forests

ABSTRACT

Existing measures of multivariate spatial autocorrelation often rely on covariance structures or fixed distances that are difficult to apply to nonlinear and mixed-type attributes. This paper proposes the Machine-Learning-Based Spatial Autocorrelation Statistic (ML-SAS), which combines a spatial weights matrix with an attribute-similarity matrix learned from unsupervised random-forest path proximities. Global and local statistics are defined within a generalized matrix-correlation framework, with analytical randomization moments as theoretical references and rank-based permutation tests for inference. Simulations assess correspondence with classical statistics, sensitivity to forest specifications and proximity definitions, null calibration, and local detection under continuous and nonlinear mixed-type processes. Path-based ML-SAS was stable across the tested forest settings and less sensitive than terminal-node proximity. Under the spatial null, the empirical probability of any rejection was 0.040. In a continuous multivariate design, ML-SAS was more selective than Multivariate Local Geary. In a nonlinear mixed-type design, adding a nominal attribute substantially increased recall and F1 relative to continuous-only ML-SAS, and leave-one-variable-out analysis identified that attribute as the largest contributor to changes in the local similarity surface. In Dutch municipalities, adding ordered urbanity changed 23 local classifications. ML-SAS provides a complementary framework for jointly analyzing continuous and categorical spatial attributes.

1 | Introduction

Spatial autocorrelation refers to the statistical dependence between attribute values and the spatial arrangement of observational units, such that nearby locations tend to exhibit systematically related—either more similar or more dissimilar—values than would be expected under spatial randomness (Cliff and Ord 1981; Hubert et al. 1981; Getis 2008). Global indicators such as Moran's I and Geary's c summarize the overall strength and direction of this dependence for a single variable (Moran 1950; Geary 1954), while local indicators of spatial association (LISA) decompose the pattern to identify clusters and outliers at the level of individual units (Getis and Ord 1992; Anselin 1995; Ord

and Getis 1995). These statistics underpin exploratory spatial data analysis and play an important role in model specification, diagnostics, and interpretation across many applied fields.

Multivariate local spatial autocorrelation requires an explicit choice of how attribute profiles are compared. Existing measures use linear covariance structures or fixed geometric distances, including Euclidean and Mahalanobis distances (Wartenberg 1985; S.-I. Lee 2001; M. Lee 2012; Anselin 2019). These choices can be useful for continuous attributes, but combining continuous and categorical variables requires decisions about encoding, scaling, and similarity. Nonlinear relationships can also make a single covariance summary incomplete. This study

focuses on the use of data-adaptive similarity in low-dimensional settings with continuous and categorical attributes and nonlinear relationships.

Random forests provide one possible route to such measures. As ensemble methods based on decision trees, random forests flexibly capture nonlinear relationships and interactions and naturally accommodate mixed data types (Breiman 2001). They also yield a proximity matrix that quantifies similarity between observations based on how often they follow similar paths through the trees (Breiman 2001). In the unsupervised random forest (URF) framework, a classifier is trained to distinguish observed data from synthetically generated reference data, and the resulting proximities reflect structure in the joint distribution of the variables (Shi and Horvath 2006). Recent work on path-based proximity refines this idea by measuring similarity through the overlap of entire decision paths rather than only terminal nodes, which reduces sensitivity to tree pruning and produces a denser similarity matrix (Kruber et al. 2019).

Although URF-based proximities have been used in clustering, anomaly detection, and in some cases for mapping spatial patterns in remotely sensed data (Peerbhay et al. 2015, 2016), they have not been fully integrated into the formal framework of spatial autocorrelation statistics. This paper proposes a Machine-Learning-based Spatial Autocorrelation Statistic (ML-SAS), which incorporates URF path proximity into a standard spatial autocorrelation framework. We construct an attribute similarity matrix from URF path proximities and combine it with a spatial weights matrix within a generalized matrix correlation setting to define a global statistic and corresponding local statistics.

Under randomization, we derive the expectation and variance of ML-SAS and interpret it as an alignment between spatial and learned attribute proximities. Under the specified input encoding, the forest can represent continuous and categorical variables jointly without selecting a fixed attribute-distance formula. The simulations and empirical application evaluate the resulting statistics in low-dimensional continuous and mixed-type settings, examining how attribute representation affects spatial pattern detection and how ML-SAS compares with existing measures.

Section 2 reviews multivariate spatial autocorrelation and unsupervised random-forest proximity. Section 3 defines ML-SAS and its inference, then specifies the simulation and empirical designs; analytical moments are given in Appendix A. Section 4 reports global benchmarks, forest-specification checks, and local-detection results. Section 5 reports the Dutch municipal application, and Section 6 discusses implications and limitations.

2 | Literature Review

Early spatial statistics introduced measures to quantify global spatial autocorrelation—the overall degree to which values of a variable are similar or dissimilar at nearby locations. The two classic indices are Moran's I (Moran 1950), a spatial analog of a correlation coefficient, and Geary's c (Geary 1954), which is based on squared differences between neighboring values. A major

limitation of these global measures is that they can mask local patterns by averaging over spatial heterogeneity. This motivated the development of local statistics, most notably the Local Indicators of Spatial Association (LISA) framework by Anselin (1995), which decomposes global statistics into their local counterparts to identify where significant clustering or outlier patterns occur.

Building on this foundation, subsequent research extended these concepts to a multivariate context. Wartenberg (1985) pioneered this effort by integrating principal component analysis with spatial autocorrelation, constructing spatial correlation matrices based on principal components. This line of work later evolved into what is now known as the concept of cross- or bivariate Moran statistics (Czaplewski and Reich 1993; Griffith 1993; S.-I. Lee 2001; Anselin et al. 2002). Subsequent developments included bivariate measures such as Lee's L (S.-I. Lee 2001, 2004, 2009, 2017), followed by local multivariate statistics based on the Mahalanobis distance-based measure (M. Lee 2012) and the Multivariate Local Geary (Anselin 2019). Together, these measures provide a general framework for analyzing spatial patterns across multiple variables, allowing the detection of multivariate clusters and outliers rather than patterns in a single attribute.

Multivariate spatial statistics differ in the attribute relationships they summarize. Covariance-based measures emphasize second-order dependence, whereas distance-based measures quantify geometric differences under a chosen scaling and metric (Wartenberg 1985; S.-I. Lee 2001; M. Lee 2012; Anselin 2019). A distance-based measure can detect local similarity even when the data-generating process contains nonlinear relationships. The relevant question is therefore how the chosen representation treats the continuous, count, and categorical attributes that occur together in multivariate areal data.

High-dimensional distance concentration provides one motivation for investigating alternative attribute similarities. Beyer et al. (1999) studied conditions under which the contrast between farthest and nearest distances diminishes as dimensionality increases. Aggarwal et al. (2001) showed that this behavior also depends on the distance metric. These findings motivate the investigation of alternative attribute representations. The present study examines learned similarity in continuous and mixed-type settings with at most four input attributes.

Mixed measurement types pose a related representation problem. Join-count statistics address categorical or binary configurations, while Lee's L (S.-I. Lee 2001), Mahalanobis distance-based measures (M. Lee 2012), and Multivariate Local Geary (Anselin 2019) offer different summaries for numerical attributes. Combining numerical and categorical information requires an explicit representation of both. We examine learned forest proximity as a joint representation of numerical and categorical attributes for spatial autocorrelation analysis.

An alternative perspective views the measurement of spatial autocorrelation as fundamentally an unsupervised learning problem: the goal is to identify spatial structure in a set of attributes without a predefined dependent variable. This motivates the use of data-adaptive similarity measures from machine learning, particularly those derived from random forests. Random Forests (RF) proximity (Breiman 2001) provides such a

measure. In its unsupervised mode, known as Unsupervised Random Forests (URF), a classifier is trained to distinguish the observed data from synthetically generated reference data, under the assumption that if the data contain inherent structure, they can be separated from randomness (Shi and Horvath 2006). The resulting proximity matrix summarizes how often pairs of observations traverse similar decision paths in the forests, yielding a nonlinear, interaction-aware similarity metric that naturally accommodates mixed data types.

URF proximity has been used for network intrusion detection (Zhang et al. 2008), semiconductor-chip analysis (Wu et al. 2007), and clustering, visualization, and data cleaning (Omar et al. 2022). In spatial applications, Peerbhay et al. (2015, 2016) combined proximity-based anomaly detection with remotely sensed data to map invasive plants. These applications provide precedents for using learned similarity to characterize multivariate structure. Building on this work, we integrate URF proximity into global and local spatial autocorrelation statistics.

Despite this growing body of work, the formal integration of URF-based proximity into the framework of spatial autocorrelation statistics remains limited. Existing applications tend to use the proximity matrix as an input for clustering or anomaly detection rather than as an explicit building block of a global or local spatial statistic. In particular, there is little work that combines a machine-learning-based attribute similarity matrix with a spatial weights matrix within a unified statistical framework and derives the distributional properties of the resulting statistic under randomization. The next section builds on these insights by using URF path proximity as a numerical similarity matrix and integrating it with a spatial proximity matrix to define the ML-SAS.

3 | The ML-SAS Framework and Study Design

This section defines ML-SAS and its inferential basis, then specifies the simulation experiments and empirical application. The framework combines learned attribute similarity with spatial adjacency.

3.1 | Unsupervised Random Forests and the Construction of the Proximity Matrix

The proximity is computed using a URF. To implement the URF, a synthetic dataset $\mathbf{X}_{\text{fake}}$ is first generated to be contrasted with the original data matrix $\mathbf{X}$. There are two common approaches to generate such synthetic data (Breiman 2002). In the Addcl1 approach, synthetic samples are drawn independently from the marginal distributions of each variable in the observed data, thereby preserving individual variable distributions while removing inter-variable dependence. In contrast, Addcl2 samples are drawn uniformly from a hyper-rectangular space defined by the range of each variable in the original data, which alters both the distributional shape and the underlying correlation structure. According to Shi and Horvath (2006), the Addcl1 method tends to outperform Addcl2 in practical applications. Consequently, this study adopts Addcl1 for the main analysis. The resulting proximity score reflects structural similarity—how often two

observations follow similar decision paths—rather than a direct geometric distance.

After the URF classifier is trained, the proximity between observations is calculated. The standard approach computes proximity as the proportion of trees in which two observations fall into the same terminal node. This measure can be sensitive to tree depth and stopping rules because these choices determine terminal-node co-membership and the sparsity of the resulting matrix (Kruber et al. 2019).

To address this limitation, this study employs Path Proximity, a measure of similarity proposed by Kruber et al. (2019). This technique considers the full paths of data points through the trees, not just the terminal nodes. The path for observation i in tree k is defined as the full root-to-terminal set of nodes $\mathcal{T}_{i,k}$, including both endpoints. To compare the paths of observations i and j , this technique uses the Jaccard index, which measures the similarity between their path sets. The final path-proximity value y_{ij} is the average of these Jaccard indices over all K trees in the forest, formulated as:

$$y_{ij} = \frac{1}{K} \sum_{k=1}^K \frac{|\mathcal{T}_{i,k} \cap \mathcal{T}_{j,k}|}{|\mathcal{T}_{i,k} \cup \mathcal{T}_{j,k}|} = \frac{1}{K} \sum_{k=1}^K \frac{|\mathcal{T}_{i,k} \cap \mathcal{T}_{j,k}|}{|\mathcal{T}_{i,k}| + |\mathcal{T}_{j,k}| - |\mathcal{T}_{i,k} \cap \mathcal{T}_{j,k}|}, \quad (1)$$

where $|\mathcal{T}_{i,k} \cap \mathcal{T}_{j,k}|$ represents the number of nodes shared by both observations, and the denominator represents the total number of unique nodes in their combined paths. The resulting proximity value y_{ij} lies within the unit interval $[0, 1]$, with larger values denoting greater overlap in decision paths. Even if two observations do not terminate in the same node, they retain positive proximity when their paths share nodes. Averaging over the forest yields the symmetric attribute-similarity matrix $\mathbf{Y} = (y_{ij})$. Before ML-SAS is calculated, its diagonal is set to zero so that perfect self-similarity does not enter the spatial association statistic.

3.2 | Definition and Inference of ML-SAS

Spatial autocorrelation can be represented as the alignment between two pairwise structures: a spatial proximity matrix and an attribute-similarity matrix (Mantel 1967; Hubert et al. 1981; Cliff and Ord 1981; S.-I. Lee 2004). This general matrix-correlation formulation applies to univariate, bivariate, and multivariate settings, provided that the attribute-side similarity is defined appropriately. In ML-SAS, attribute similarity is represented by the unsupervised random forest path-proximity matrix introduced in Section 3.1.

Based on this framework, the global ML-SAS is defined as follows:

$$ML-SAS = \frac{n(n-1) \sum_i \sum_{j \neq i} w_{ij} y_{ij}}{\left(\sum_i \sum_{j \neq i} w_{ij} \right) \left(\sum_i \sum_{j \neq i} y_{ij} \right)}, \quad (2)$$

where w_{ij} is the (i, j) element of the spatial proximity matrix $\mathbf{W}$, and y_{ij} is the corresponding element of the attribute-similarity matrix $\mathbf{Y}$. The diagonal elements of both matrices are set to zero. Because URF proximity assigns perfect self-similarity,

retaining diagonal elements would add no substantive information; self-adjacency is likewise excluded from $\mathbf{W}$.

The normalization in Equation (2) gives the global statistic an expected value of 1 under spatial randomization. Values larger than 1 indicate that spatially proximate observations are, on average, more similar in attribute space than expected under randomness, whereas values below 1 indicate weaker-than-expected local similarity. The analytical expectation and variance follow the generalized randomization framework for matrix-correlation statistics; the full expressions and notation are provided in Appendix A.

The local ML-SAS is obtained by decomposing the global measure as follows:

$$ML-SAS_i = \frac{n^2(n-1) \sum_{j \neq i} w_{ij} y_{ij}}{\left(\sum_i \sum_{j \neq i} w_{ij} \right) \left(\sum_i \sum_{j \neq i} y_{ij} \right)}. \quad (3)$$

This definition satisfies the additivity condition: the average of the local statistics equals the global ML-SAS (S.-I. Lee 2009). A large local value indicates that the focal observation and its spatial neighbors exhibit stronger attribute similarity than expected, whereas a small local value indicates local dissimilarity or a potential spatial outlier. Because URF path proximity measures structural similarity rather than covariance around a mean, local ML-SAS can identify clusters of high, low, or moderate values, as well as multivariate clusters involving nonlinear relationships and mixed data types.

For inference, both total and conditional randomization are available. Conditional randomization is preferred for local analysis because it fixes the focal observation and permutes the remaining observations, thereby generating a reference distribution for the focal unit's observed spatial context (Anselin 1995; S.-I. Lee 2009). Although analytical first and second moments can be derived under both schemes (Appendix A), local permutation distributions need not be well approximated by a normal distribution. We therefore use rank-based Monte Carlo permutation tests for the empirical analyses and control multiple testing with the false discovery rate (FDR) procedure (Caldas de Castro and Singer 2006; Anselin 2019).

The finite-sample behavior of this inference framework—including null calibration, robustness to URF hyperparameters and proximity definitions, and local detection performance—is evaluated directly in the simulation study in Section 4. This simulation-based validation provides the primary evidence for the practical adequacy of ML-SAS inference, while the analytical moments serve as theoretical reference results.

3.3 | Simulation Design

3.3.1 | Spatial Support and Inference

All experiments used a fixed 20×20 lattice of 400 equal-area hexagons. First-order shared-boundary contiguity defined a symmetric binary adjacency matrix with zero diagonal. Its row-standardized form was used for ML-SAS, Moran, Geary, and

other local-statistic calculations; the BB join-count benchmark used a binary matrix, consistent with its definition over undirected neighboring pairs.

Global reference arrangements reassigned complete attribute profiles among the 400 locations while holding the spatial weights fixed. Local ML-SAS used conditional randomization, holding the focal profile fixed while permuting profiles assigned to the other locations. Upper-tail inference was prespecified because the target was positive local similarity, represented by statistics larger than their conditional permutation expectations. Comparator directions were defined analogously; for Geary-type measures, positive local association corresponds to small raw squared differences. Local pseudo- p values were adjusted across the 400 units by the Benjamini–Hochberg procedure at 0.05, and a unit was classified as a cluster only when the adjusted p value was at most 0.05 and the statistic was in the prespecified similarity direction.

Cells are indexed by row r and column c , each ranging from 1 to 20, with $ID = 20(r-1) + c$. Hexagons have width 1 and height $h = 2/\sqrt{3}$; their centroids are at $c - 1/2 + 1/2 \cdot 1(r \text{ is even})$ horizontally and $(20-r)3h/4 + h/2$ vertically. Planted regions are defined using row–column indices, not distances between these centroids.

3.3.2 | Local Data-Generating Processes

In the distributions below, the second argument of N is the variance. Random draws are independent unless a shared latent variable is explicitly specified. DGP 1 provides a univariate check of local homogeneity at different attribute levels. Regions H , L , and M are row–column disks of radius 3.2 centered at (5, 5), (16, 16), and (10.5, 10.5), respectively. They contain 37, 37, and 32 cells. Values are generated by Equation (4). The moderate region has the same mean as the background but much lower variance, separating local homogeneity from concentration at an extreme level. Figure 5 shows the input, planted regions, and detection results.

$$\begin{aligned} X_i &\sim N(2.5, 0.15^2), & i \in H, \\ X_i &\sim N(-2.5, 0.15^2), & i \in L, \\ X_i &\sim N(0, 0.15^2), & i \in M, \\ X_i &\sim N(0, 1.25^2), & i \notin H \cup L \cup M. \end{aligned} \quad (4)$$

DGP 2 is a spatial null with no planted region. Three continuous attributes and a nominal attribute are generated independently across locations and variables according to Equation (5). The continuous specification uses X_1 – X_3 , and the mixed specification adds CAT1, denoted C below. This design evaluates false rejection under unstructured mixed inputs; an illustrative input map is provided in Figure B1.

$$X_{ik} \stackrel{\text{iid}}{\sim} N(0, 1), k = 1, 2, 3; \quad C_i \sim \text{Cat}(C1, C2, C3; 1/3, 1/3, 1/3). \quad (5)$$

DGP 3 uses the target region in Equation (6). It contains exactly 60 cells: columns 8–13 in rows 7 and 14, and columns 7–14 in rows 8–13. All other cells form the background.

TABLE 1 | Simulation scenarios and repeated-sample design.

Scenario	Structure	Variables	Primary comparison	Independent datasets
DGP 1	Three homogeneous regions at different levels	One continuous	Local Moran, Local Geary, G_i^* , ML-SAS	Representative only
DGP 2	Spatial null	$X_1 - X_3$; optional CAT1	Continuous vs. mixed ML-SAS	100
DGP 3	Additive central signal; one noise variable	$X_1 - X_3$	Multivariate Local Geary vs. path ML-SAS	100
DGP 4	Nonlinear central signal plus aligned category	$X_1, X_2, \text{CAT1}$	Multivariate Local Geary; continuous/mixed path; mixed terminal ML-SAS	100

Note: DGP 1 provides a representative construct check; DGPs 2–4 provide repeated operating-characteristic evidence.

$$\mathcal{T} = \{(r, c) : (c - 10.5)^2 + (r - 10.5)^2 \leq 4.5^2\}. \quad (6)$$

DGP 3 gives X_1 and X_2 the same region-specific mean shift and variance reduction, while X_3 remains unstructured noise (Equation 7). X_1 and X_2 are drawn independently conditional on target membership. This provides a continuous-data comparison in which local squared differences have a direct interpretation; it does not require nonlinear or categorical information. Figure 6 displays all three inputs.

$$\begin{aligned} X_{i1}, X_{i2} &\sim N(1.5, 0.35^2), & i \in \mathcal{T}, \\ X_{i1}, X_{i2} &\sim N(0, 1), & i \notin \mathcal{T}, \\ X_{i3} &\sim N(0, 1), & \text{all } i. \end{aligned} \quad (7)$$

DGP 4 uses the same target region $\mathcal{T}$ as DGP 3 and combines a curved continuous relationship with a nominal pattern aligned to the target. Within the target, the unstandardized continuous variables share a latent U as specified in Equation (8). In the background, both are independent $N(0, 1.20^2)$ draws. Each continuous variable is then standardized over all 400 cells using its sample mean and sample standard deviation s_k with denominator $n - 1$ (Equation (9)).

$$\begin{aligned} U_i &\sim N(0, 0.25^2), & \epsilon_{i1}, \epsilon_{i2} &\stackrel{\text{iid}}{\sim} N(0, 0.12^2), \\ \tilde{X}_{i1} &= 0.70 + U_i + \epsilon_{i1}, \\ \tilde{X}_{i2} &= 0.70 + 0.35(U_i^2 - 0.25^2) + \epsilon_{i2}, & i \in \mathcal{T}. \end{aligned} \quad (8)$$

$$X_{ik} = \left(\tilde{X}_{ik} - \overline{\tilde{X}_k} \right) / s_k, \quad k = 1, 2. \quad (9)$$

CAT1 is sampled independently of the continuous draws conditional on region membership, with the probabilities in Equation (10). Multivariate Local Geary and continuous-only ML-SAS receive X_1 and X_2 ; mixed-type ML-SAS receives X_1, X_2 , and CAT1. The product $X_1 X_2$ is only a derived diagnostic and is neither a forest input nor the rule defining the target. Figure 7 shows the three input distributions.

$$\begin{aligned} P(C_i = A, B, C | i \in \mathcal{T}) &= (0.90, 0.05, 0.05), \\ P(C_i = A, B, C | i \notin \mathcal{T}) &= (0.20, 0.40, 0.40). \end{aligned} \quad (10)$$

Within the target region, the shared latent construction induces nonlinear dependence between the two unstandardized continuous variables while their population covariance is zero. DGP 4

also shifts the marginal means and reduces the variances within the target, creating continuous homogeneity that distance-based local measures can detect. This design allows us to examine how adding aligned nominal information changes learned similarity and target recovery across common and uncommon categories. Table 1 summarizes the four simulation scenarios, their primary comparisons, and the repeated-sample design.

3.3.3 | Repeated Evaluation and Specification Checks

Representative maps and repeated experiments used 99,999 permutations. DGPs 2–4 each comprised 100 independently generated datasets. All repeated ML-SAS analyses used 100 trees, $m_{\text{try}} = 1$, and minimum node size = 1. Forests were fitted with the R package randomForest; CAT1 was supplied as an unordered factor, allowing splits between subsets of its categories. Paired method contrasts used the same dataset within each replication and 9999 dataset-level bootstrap resamples for 95% confidence intervals.

Local-detection performance was summarized by recall, the proportion of target cells detected; the background false-positive proportion (FPR), the proportion of nontarget cells classified as significant; precision, the proportion of detected cells that belonged to the target, conditional on at least one discovery; and F1, calculated as $\frac{2TP}{2TP + FP + FN}$, where TP, FP, and FN are true positives, false positives, and false negatives. For DGPs 3 and 4, F1 is zero when no cells are detected and is averaged over all 100 datasets; precision remains conditional on at least one discovery. Rank AUC is the probability that a randomly selected target cell is ranked ahead of a randomly selected background cell after orienting each statistic toward the prespecified cluster direction, with ties receiving half weight. Recall, FPR, precision, and F1 depend on the significance threshold and multiplicity adjustment, whereas rank AUC evaluates ordering without imposing that threshold.

The representative datasets use seeds 2026, 2027, 2028, and 2029 for DGPs 1–4. Repeated dataset r for DGP d uses seed $2,000,000 + 100,000d + r$. RF and permutation seeds are $3,000,000 + 10,000d + 10r + m$ and $4,000,000 + 10,000d + 10r + m$, respectively. Method codes m are 1 and 2 for the two DGP 2 specifications, 3 for Multivariate Local Geary, 4 for continuous path ML-SAS, 5 for mixed path ML-SAS, and 6 for mixed terminal ML-SAS. These contrasts pair datasets, not identical fitted forests.

Specification checks hold the full mixed DGP 4 representative dataset fixed. Path proximity is evaluated over 45 combinations of the number of trees (25, 50, 100, 250, or 500), mtry (1, 2, or 3), and minimum node size (1, 5, or 10), using 100 forest realizations per condition. Path and terminal proximity are also compared with 100 trees across the nine matched mtry/node-size settings.

A paired leave-one-variable-out (LOVO) contribution analysis uses the full mixed DGP 4 representative dataset. Each variable is omitted in turn, and its contribution is the mean absolute change in local ML-SAS across all locations relative to the full model. Full and reduced forests use matched random seeds across 100 paired forest realizations. This diagnostic quantifies the sensitivity of the fitted local similarity surface to omitting each input variable.

Post hoc diagnostics of the stored repeated local outputs examine the DGP 3 and DGP 4 performance patterns through small-*p*-value frequencies and detection and ranking by CAT1 category. The diagnostics retain the prespecified statistics and detection rules. Category-specific detection rates pool counts across the 100 datasets, whereas category-pair AUCs are averaged over datasets. Tables C1–C3 report these diagnostics.

3.3.4 | Global Benchmark Design

The benchmarks used fixed master samples and 1000 common spatial arrangements. For the continuous benchmark, a standard-normal vector of 400 values was reassigned to the lattice and evaluated by ML-SAS, Moran's *I*, and Geary's *c* using a row-standardized spatial weights matrix. For the binary benchmark, a balanced vector containing 200 B and 200 W observations was reassigned and evaluated by ML-SAS using a row-standardized matrix and the BB join count using a binary matrix. Holding the sample values and category frequencies fixed isolates the response of each statistic to spatial arrangement. Path proximity used 100 trees, mtry = 1, and minimum node size = 1, and each comparison was paired over the same 1000 arrangements.

3.4 | Empirical Data and Analytical Specification

We used the 2022 NLD_muni dataset distributed with the R package tmap (Tennekes 2018), which contains 345 Dutch municipalities and is derived from Statistics Netherlands' Key Figures for Districts and Neighborhoods 2022 (Statistics Netherlands 2024). Of these, 337 formed the main network of boundary-sharing municipalities and were included in the analysis. Eight municipalities in geographically separate island or border components were excluded. Queen contiguity was used to define neighboring municipalities, and the spatial weights were row standardized.

The analysis used four variables representing distinct socioeconomic and settlement domains: population, average assessed dwelling value, employment rate, and five-level ordered urbanity. The first three variables capture municipal scale, housing conditions, and labor-market participation; urbanity adds settlement context as an ordered categorical attribute. It was supplied to randomForest as an ordered factor with the dataset's five-level

ordering; the implementation uses threshold splits on the ordered level codes. Municipality identifiers and geographic labels were not used as attributes.

The primary specification included all four attributes. A nested continuous-only specification containing population, dwelling value, and employment rate was also estimated to compare the learned joint representations and spatial classifications produced by the complete three-attribute and four-attribute fitted specifications.

Following the procedure in Section 3.2, ML-SAS was calculated from a 1000-tree unsupervised random forest using row-standardized queen weights. Global significance was assessed using 99,999 total-label permutations. Local inference used 99,999 conditional permutations for each municipality, holding the focal profile fixed while reassigning the remaining profiles among the other locations. Because the empirical analysis did not prespecify whether a municipality would exhibit unusually high or unusually low local similarity, two-sided tests were used. This differs from the upper-tail inference in Section 4, which targeted planted positive-similarity clusters. Local *p* values were adjusted across the 337 municipalities using the Benjamini-Hochberg procedure at $q \leq 0.05$. Significant locations above their focal conditional permutation mean were classified as similarity clusters, whereas those below the mean were classified as dissimilarity/outliers. These categories describe the direction of multivariate similarity relative to the permutation distribution and are distinct from univariate High-High or Low-Low categories.

4 | Simulation Results

4.1 | Overview

The results address global correspondence with established statistics, sensitivity to forest specifications, and local detection under the four DGPs specified in Section 3.3. Representative maps show individual realizations; repeated experiments provide operating-characteristic evidence.

4.2 | Global Benchmark Against Established Statistics

The global benchmark evaluates whether ML-SAS varies consistently with established statistics across common spatial arrangements.

4.2.1 | Continuous Benchmark: Moran's *I* and Geary's *c*

Global ML-SAS was positively associated with Moran's *I* ($r = 0.473$; Figure 1a) and negatively associated with Geary's *c* ($r = -0.468$; Figure 1b). The latter direction is expected because stronger positive autocorrelation corresponds to smaller Geary values but larger ML-SAS values.

The moderate correlations indicate related but nonidentical measurement. Moran's *I* uses centered spatial cross-products,

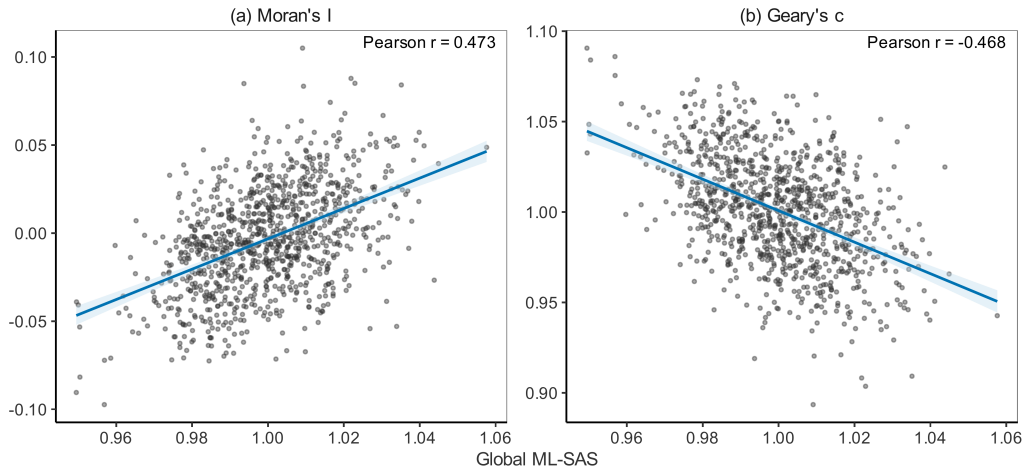


FIGURE 1 | Correspondence between global ML-SAS and classical continuous statistics. The 1000 common spatial arrangements use a fixed standard-normal sample on 400 hexagons. ML-SAS used path proximity with 100 trees and row-standardized weights matrix. Lines show linear fits with 95% confidence bands; (a) $r = 0.473$ for Moran's I and (b) $r = -0.468$ for Geary's c .

Geary's c uses neighboring squared differences, and ML-SAS relates adjacency to an attribute-similarity matrix learned from random-forest paths. Unlike the classical statistics, ML-SAS first transforms the original values into graded path-overlap similarities. Consequently, spatial arrangements that generate strong high-high or low-low cross-products, or small neighboring squared differences, need not receive proportionally identical similarities in the forest-induced attribute space. The observed correlations therefore represent a shared response to spatial arrangement rather than numerical equivalence among the three statistics.

4.2.2 | Binary Benchmark: Join-Count Statistic

Across the balanced binary arrangements, ML-SAS and the BB join count were strongly associated ($r = 0.874$; Figure 2), with more B-B adjacencies corresponding to larger ML-SAS values.

Both measures increase when identical binary labels are spatially concentrated, although the BB statistic directly counts B-B adjacencies whereas ML-SAS combines learned attribute similarity. This correspondence pertains to the balanced binary benchmark.

4.3 | Forest Specification Results

4.3.1 | Sensitivity to Forest Settings

Global path ML-SAS changed little across the tested forest settings. Medians remained near 1.13–1.14, and forests larger than 100 trees produced little visible change in central tendency or dispersion (Figure 3). These results support the 100-tree baseline for the reported experiments.

4.3.2 | Path and Terminal-Node Proximity

Path proximity was less sensitive than terminal-node proximity to the tested splitting and stopping settings. Path-based values remained near 1.13–1.14, while terminal-node values varied

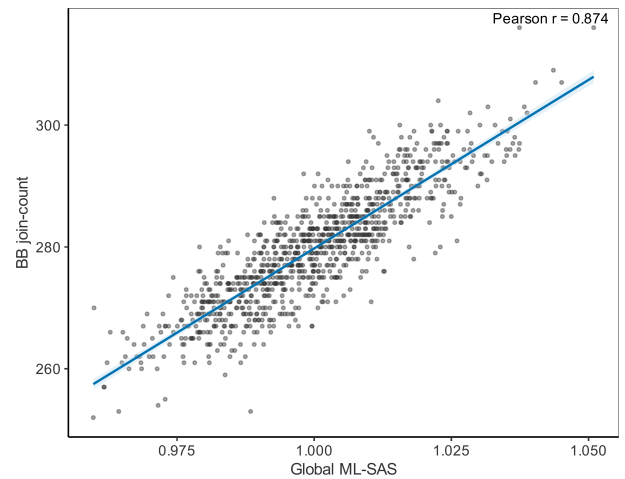


FIGURE 2 | Correspondence between global ML-SAS and the BB join-count statistic. A balanced sample with 200 B and 200 W observations was evaluated over 1000 common spatial arrangements. ML-SAS used row-standardized weights matrix and the BB join count used symmetric binary adjacency matrix; $r = 0.874$.

more strongly (Figure 4). Shared decision paths provide a graded similarity even without exact leaf co-membership. The repeated local comparison in Section 4.4.4 evaluates whether the two proximity definitions also differ in target recovery.

4.4 | Local Detection Performance

4.4.1 | DGP 1: Univariate Geary-Type Clusters

Local results distinguish illustrative spatial patterns from repeated operating characteristics. The DGP 1 representative dataset provides a univariate construct check: after 99,999 permutations and 5% false discovery rate control, Local Geary and ML-SAS recovered all three homogeneous regions, including the moderate central region, whereas mean-centered Local Moran and G_i^* emphasized the high- and low-level regions (Figure 5).

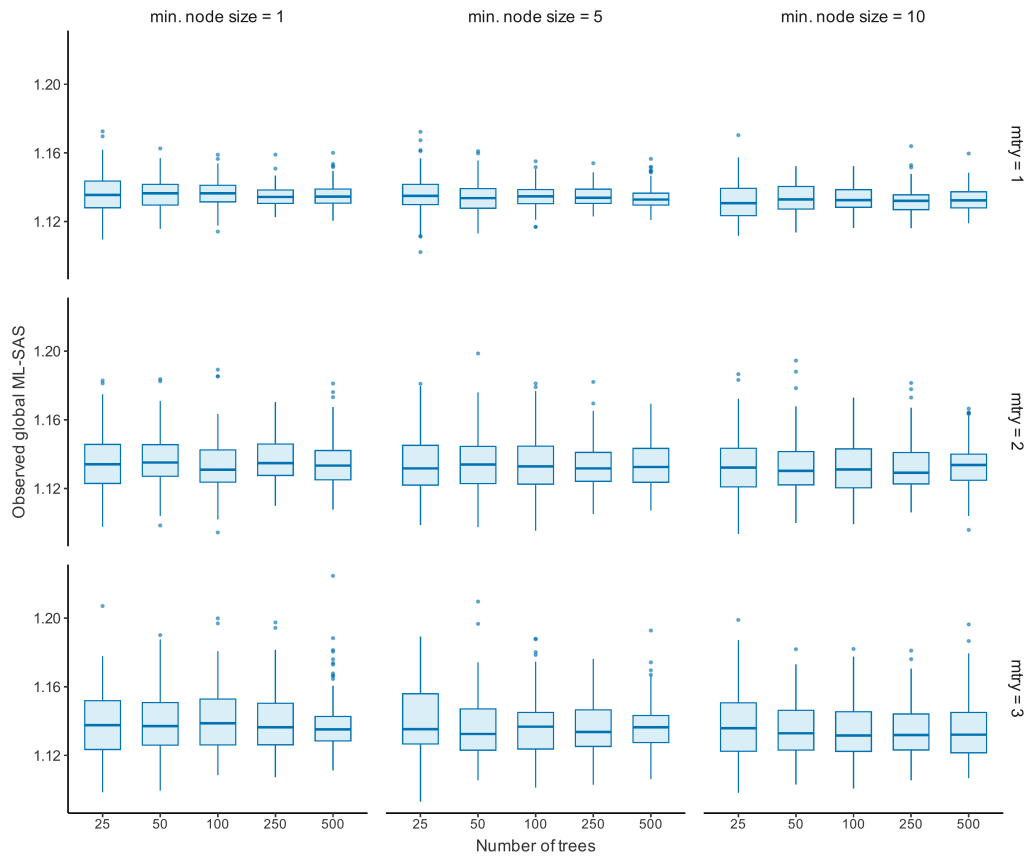


FIGURE 3 | Sensitivity of global path ML-SAS to forest specifications. The fixed mixed-type DGP 4 dataset was evaluated across 45 combinations of the number of trees (25, 50, 100, 250, or 500), mtry (1, 2, or 3), and minimum node size (1, 5, or 10), with 100 forest realizations per condition. Boxes show the distributions of observed global ML-SAS values.

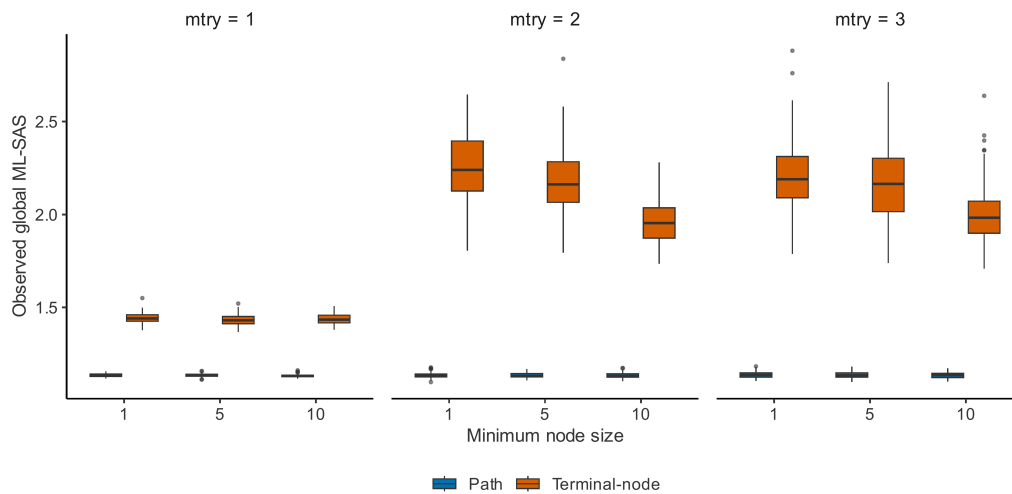


FIGURE 4 | Global ML-SAS under path and terminal-node proximity. The fixed mixed-type DGP 4 dataset was evaluated with 100 trees across nine combinations of mtry (1, 2, or 3) and minimum node size (1, 5, or 10), with 100 forest realizations per condition.

4.4.2 | DGP 2: Null Calibration

The repeated null experiment produced few false discoveries at the nominal 0.05 level. Each path specification rejected at least one location in 4 of 100 datasets, giving an empirical family-wise rejection probability of 0.040 (exact 95% confidence interval: 0.011–0.099; Table 2). Mean rejection counts were 0.040 for continuous inputs and 0.060 for mixed inputs, corresponding to

cellwise false-positive proportions of 0.00010 and 0.00015. The confidence interval quantifies the sampling uncertainty in the estimated rejection probability across the 100 null datasets.

4.4.3 | DGP 3: Continuous Multivariate Comparison

DGP 3 shows that better overall ranking need not yield more discoveries after multiplicity adjustment. Path ML-SAS ranked

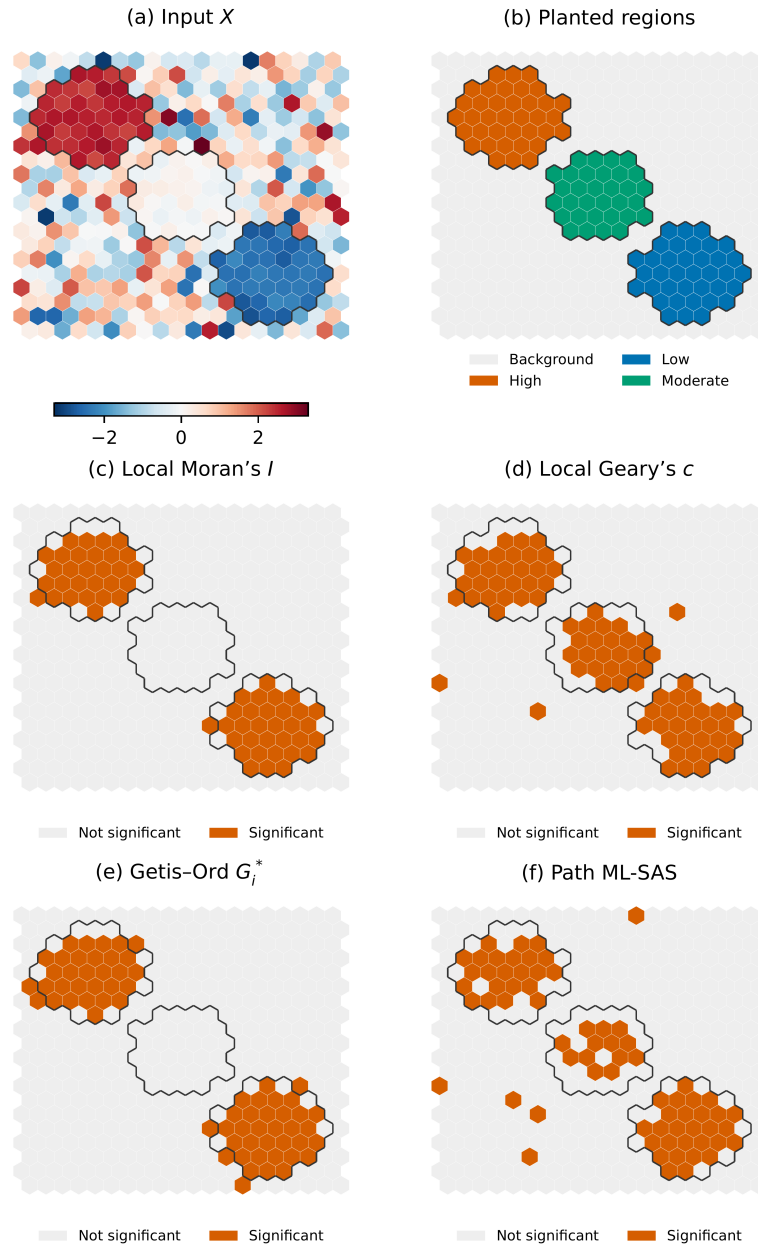


FIGURE 5 | Input data and representative local detection under DGP 1. Panels (a) and (b) show X and the three planted regions; panels (c)–(f) show Local Moran, Local Geary, Getis–Ord G_i^* , and path ML-SAS. Outlines mark the planted regions. All methods used 99,999 permutation draws and 5% false discovery rate control. ML-SAS used upper-tail conditional inference and required the observed statistic to exceed its focal permutation expectation. The representative input seed is 2026.

TABLE 2 | Repeated DGP 2 null operating characteristics.

ML-SAS specification	Datasets	Datasets with any rejection	Pr(any rejection) [95% CI]	Mean rejections
Path, $X_1 - X_3$	100	4	0.040 [0.011, 0.099]	0.040
Path, $X_1 - X_3 + \text{CAT1}$	100	4	0.040 [0.011, 0.099]	0.060

Note: Values are based on 100 independent datasets per specification and 99,999 conditional permutations per dataset.

target cells more accurately than Multivariate Local Geary (AUC 0.871 versus 0.780), but its recall and F1 were lower (Table 3). On average, it detected 16.77 target and 1.17 background cells, compared with 21.22 and 1.96 for Multivariate Local Geary. Background detections declined proportionally more than target detections, consistent with higher conditional precision, while

the reduction in target detections yielded lower recall. The paired recall difference was -0.074 (95% CI: -0.100 to -0.048), while the AUC difference was 0.091 (0.079 – 0.103).

The stored local p values locate this difference in the extreme tail relevant to BH rejection. At raw $p \leq 0.001$, 20.05% of target cells

TABLE 3 | Mean local-detection performance across repeated DGP 3 datasets.

Method/specification	Recall	FPR	Precision	F1	Rank AUC
Multivariate Local Geary, $X_1 - X_3$	0.354	0.006	0.920	0.499	0.780
Path ML-SAS, $X_1 - X_3$	0.280	0.003	0.939	0.410	0.871

Note: FPR is the background false-positive proportion. Precision is conditional on at least one discovery and is defined for 100 Multivariate Local Geary datasets and 99 path ML-SAS datasets. F1 includes a value of zero for the path specification's one dataset with no discoveries. Recall, FPR, F1, and rank AUC use all 100 datasets.

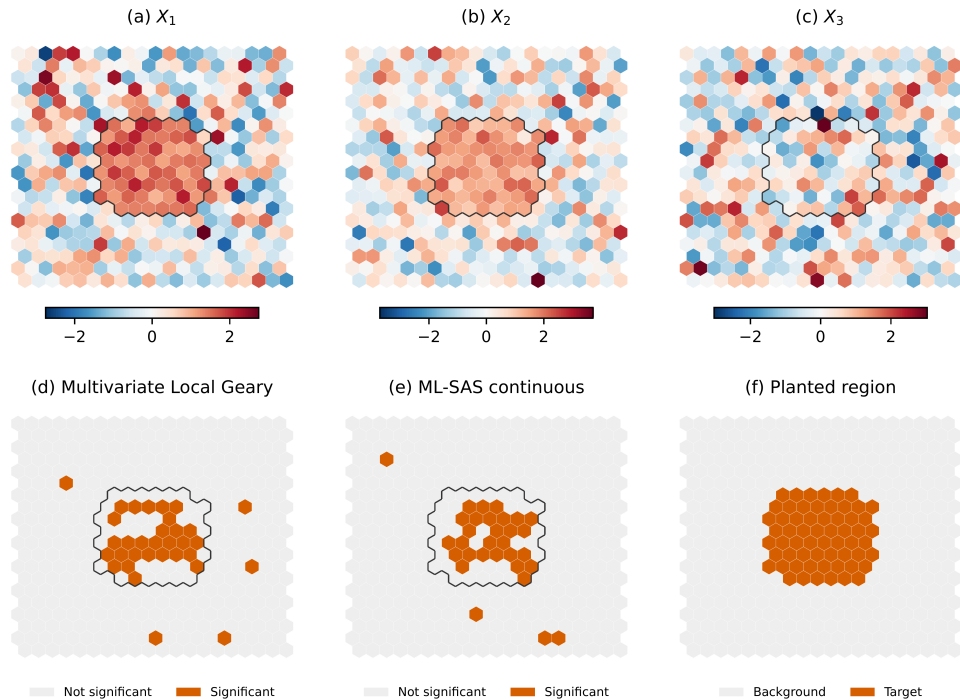


FIGURE 6 | Input data and representative local detection under DGP 3. Panels (a)–(c) show $X_1 - X_3$; panels (d) and (e) show Multivariate Local Geary and path ML-SAS; panel (f) shows the 60-cell target. Outlines identify the target in the other panels. The two methods identified 33 and 31 significant locations after 99,999 permutations and 5% false discovery rate control. The representative input seed is 2028.

TABLE 4 | Mean local-detection performance across repeated DGP 4 datasets.

Method/specification	Recall	FPR	Precision	F1	Rank AUC
Multivariate local geary, $X_1 - X_2$	0.675	0.007	0.948	0.787	0.975
Path ML-SAS, $X_1 - X_2$	0.395	0.005	0.940	0.546	0.968
Path ML-SAS, $X_1 - X_2 + \text{CAT1}$	0.618	0.007	0.942	0.744	0.934
Terminal ML-SAS, $X_1 - X_2 + \text{CAT1}$	0.601	0.007	0.939	0.730	0.911

Note: Values are based on 100 independent datasets. FPR is the background false-positive proportion. Precision is conditional on at least one discovery.

passed for ML-SAS versus 24.37% for Multivariate Local Geary; at $p \leq 0.05$, the corresponding proportions were 72.15% and 68.67% (Table C1). Thus, ML-SAS separated targets well across the broader range but placed fewer targets in the smallest- p -value range. Rank AUC uses the oriented raw statistic, whereas detection depends on each location's conditional reference distribution and the joint BH procedure.

Figure 6 shows one DGP 3 realization, while Table 3 summarizes performance across the 100 repeated datasets.

4.4.4 | DGP 4: Nonlinear Mixed-Type Cluster

Adding CAT1 improved thresholded recovery under DGP 4, but the benefit was not uniform across the target. Path ML-SAS

recall increased from 0.395 to 0.618 and F1 from 0.546 to 0.744 (Table 4). The paired gains were 0.223 (95% CI: 0.198–0.248) and 0.198 (0.174–0.223), respectively. Background FPR increased by 0.0020 (0.0008–0.0033), while conditional precision did not differ clearly. Multivariate Local Geary retained the highest mean recall and F1 for the continuous component, consistent with the mean and variance differences built into this DGP.

The lower mixed-model rank AUC, 0.934 versus 0.968, reflects a loss concentrated among target cells with uncommon categories. Across the 100 datasets, 5409 target cells had CAT1 = A and 591 had B or C. Their pooled detection rates changed from 39.66% to 68.44% for A, but from 38.24% to 0.68% for B/C. Category-pair rank diagnostics likewise improved for A targets and deteriorated for B/C targets (Tables C2 and C3). All these cells count as targets

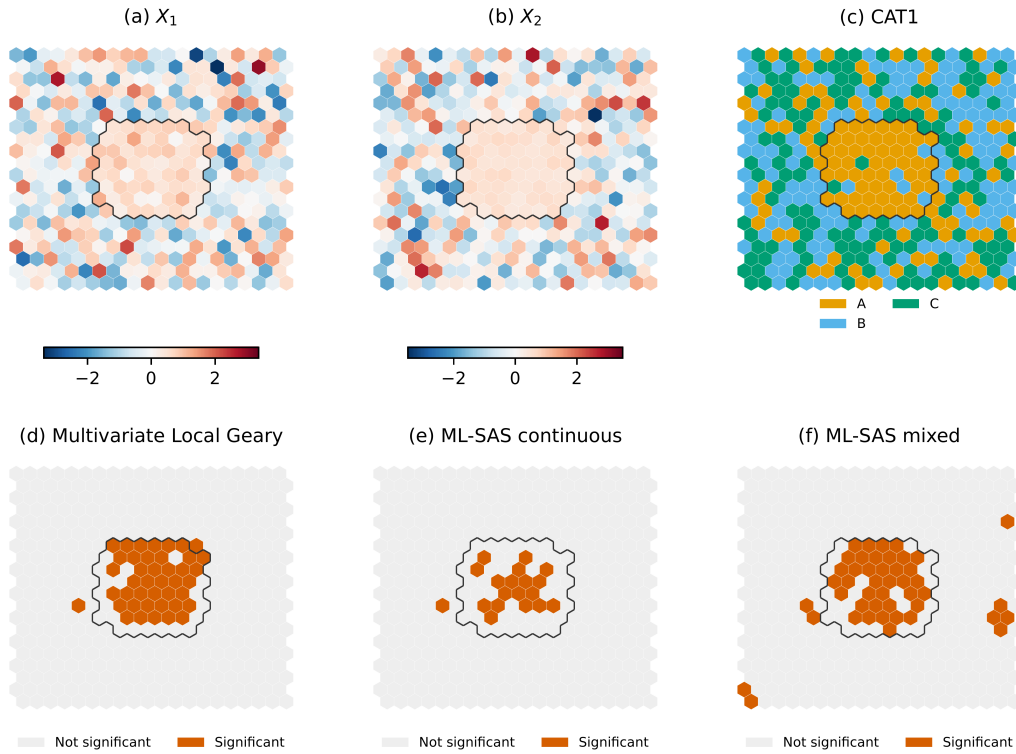


FIGURE 7 | Input data and representative local detection under DGP 4. Panels (a)–(c) show X_1 , X_2 , and CAT1; panels (d)–(f) show Multivariate Local Geary on $X_1 - X_2$, continuous-only path ML-SAS, and mixed-type path ML-SAS. Outlines mark the 60-cell target, including its minority-category cells. The methods identified 42, 19, and 46 significant locations after 99,999 conditional permutations and 5% false discovery rate control. The representative input seed is 2029.

under the spatial truth definition, although their categorical profiles differ.

The majority-category gain was large enough to increase overall recall and F1, whereas the minority-category rank loss lowered the average over all target–background pairs. CAT1 therefore changed which local profiles were treated as similar, improving recovery for the prevalent target category while reducing recovery and ranking quality for uncommon target categories.

Path mixed-type ML-SAS produced modestly higher recall, F1, and rank AUC than terminal-node ML-SAS. The paired path-minus-terminal differences were 0.017 for recall (95% CI: 0.008–0.025), 0.014 for F1 (0.008–0.021), and 0.023 for rank AUC (0.020–0.025). The corresponding differences in background false-positive proportion and conditional precision were not clearly distinguishable from zero.

Figure 7 shows one DGP 4 realization, in which the mixed-type map covers a broader portion of the target than the continuous-only ML-SAS map, consistent with the repeated gain in recall; Table 4 summarizes performance across the 100 repeated datasets.

The LOVO diagnostic also identified CAT1 as the largest contributor to changes in the fitted local similarity surface. Omitting CAT1 produced a mean absolute local change of 0.226 (SD = 0.010), followed by X_2 at 0.105 (SD = 0.014) and X_1 at 0.099 (SD = 0.012), across 100 paired forest realizations. The LOVO

results identify CAT1 as the attribute to which the fitted local similarity surface was most sensitive in this dataset (Figure 8).

Taken together, the local experiments yielded three main findings: DGP 2 produced an empirical familywise rejection probability of 0.040 for both specifications; under DGP 3, ML-SAS had lower background FPR, higher conditional precision, and higher rank AUC but lower recall and F1 than Multivariate Local Geary; and under DGP 4, incorporating nominal information directly improved thresholded target recovery relative to continuous-only ML-SAS, although rank AUC was lower.

5 | Empirical Results

The Dutch application examines spatial association in the four-attribute representation and the changes produced by adding ordered urbanity to the continuous baseline. The data and inferential specifications are given in Section 3.4.

5.1 | Attribute Distributions and Global Association

The continuous-only specification produced a global ML-SAS value of 1.245, whereas the mixed-type specification produced a value of 1.291. Both exceeded their corresponding permutation means of approximately 1.000 (two-sided $p = 0.00002$ in each specification). Thus, geographically adjacent municipalities were more similar in the learned attribute space than expected under

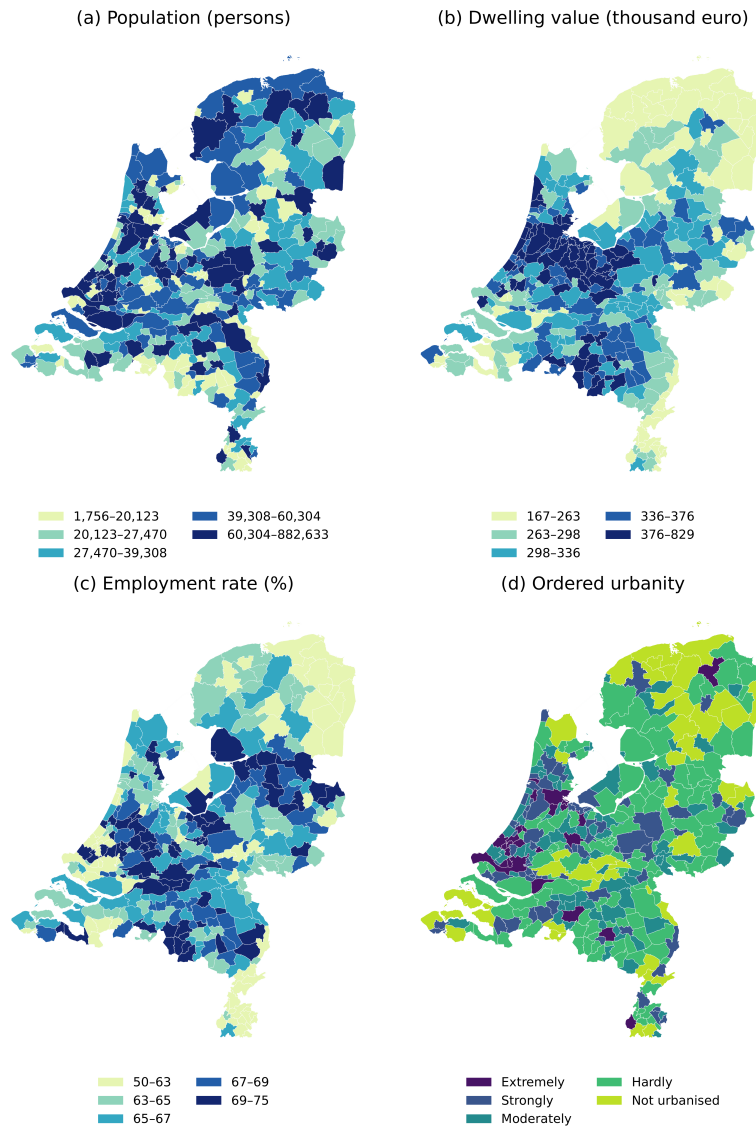


FIGURE 8 | Input attribute distributions for the 337 Dutch municipalities in the analytical network. Continuous variables use five quantile classes in their original units; urbanity retains its five ordered categories. These display classes are used only for mapping and do not replace the original values supplied to the forest.

random reassignment, first for the three continuous attributes and again when ordered urbanity was incorporated.

5.2 | Local Classifications and Specification Transitions

Figure 9 presents the local classifications and their transitions from the continuous-only to the mixed specification. After adjustment for multiple testing, the mixed specification classified 55 municipalities as similarity clusters, one as a dissimilarity/outlier, and 281 as nonsignificant. The significant locations extend across 11 provinces and form several geographically separated groups, with visible concentrations in the north, the western urban area, the east, Noord-Brabant, and Limburg.

A similarity cluster need not contain extreme values on any individual attribute. A municipality with moderate values on all four attributes may be significant when its learned joint profile is unusually similar to those of its neighbors.

The two specifications produced different local classification patterns. The continuous-only specification identified 44 similarity clusters, one dissimilarity/outlier, and 292 nonsignificant municipalities. Under the four-attribute mixed specification, 23 municipal classifications changed: 17 nonsignificant municipalities became similarity clusters and six previous similarity clusters became nonsignificant. The lower-tail municipality retained its classification in both specifications, while the 17 gains and six losses shifted the detected pattern in both directions.

Hellendoorn, Haarlemmermeer, Papendrecht, and Leidschendam-Voorburg, for example, were nonsignificant in the continuous-only analysis but were classified as similarity clusters in the mixed specification. By contrast, Zaltbommel, Montfoort, and Bodegraven-Reeuwijk moved from similarity clusters to nonsignificant classifications.

Midden-Delfland was the only significant lower-tail location in the mixed-type analysis (local ML-SAS = 0.608;

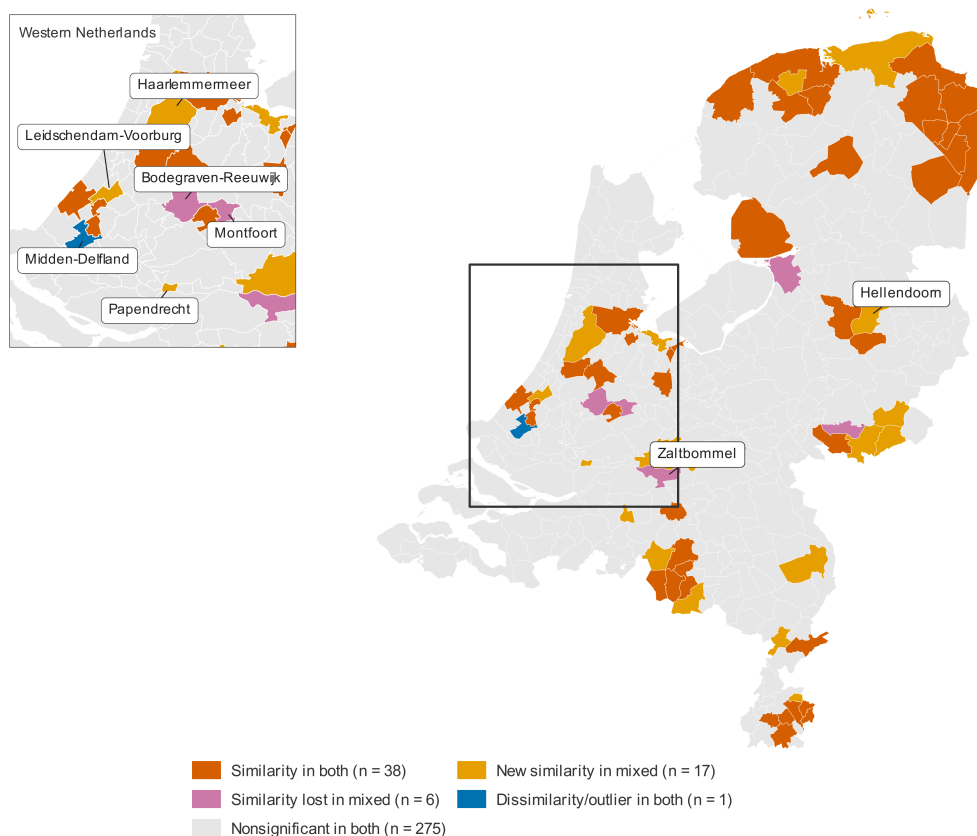


FIGURE 9 | Local ML-SAS classifications and specification transitions for Dutch municipalities. The analysis used population, average assessed dwelling value, employment rate, and ordered urbanity for the 337 municipalities in the main boundary-sharing network. Classes are based on 99,999 two-sided conditional permutations and Benjamini–Hochberg adjustment at $q \leq 0.05$. Dark orange denotes similarity clusters under both specifications, gold denotes new similarity clusters in the mixed specification, magenta denotes similarity clusters lost in the mixed specification, blue denotes the dissimilarity/outlier under both specifications, and gray denotes municipalities that were nonsignificant under both specifications.

BH-adjusted $p=0.0219$). Its joint profile was less similar to those of its neighbors than expected under focal conditional randomization. It was also the only lower-tail location in the continuous-only specification, retaining its classification in the mixed specification.

Together, the Dutch application showed stronger-than-random global spatial alignment in both specifications, geographically dispersed upper-tail departures, one lower-tail departure, and 23 local classification differences between the continuous-only and mixed-type specifications.

6 | Conclusion

This study proposed ML-SAS as a global and local measure of multivariate spatial autocorrelation based on similarity learned from an unsupervised random forest. The method combines path-based attribute proximity with a spatial weights matrix in a generalized matrix-correlation framework, and the global statistic equals the mean of the local statistics. Analytical expectations and variances under randomization provide theoretical reference results, while rank-based permutation tests support inference when finite-sample local distributions are not well approximated by a normal distribution. Under a specified encoding, the construction represents continuous and categorical attributes jointly through learned forest proximity.

The simulations characterized ML-SAS across global benchmarks, specification checks, a spatial null, and structured alternatives. Global ML-SAS varied consistently with Moran's I , Geary's c , and the BB join count across the benchmark arrangements, while path proximity was less sensitive than terminal-node proximity to the tested splitting and stopping specifications. In the repeated null experiment, both specifications had an empirical probability of any rejection of 0.040. Performance under structured alternatives depended on the data-generating process. In the continuous multivariate design, Multivariate Local Geary achieved higher recall and F1, whereas path ML-SAS was more selective and produced a lower background false-positive proportion, slightly higher conditional precision, and higher rank AUC. In the nonlinear mixed-type design, adding the nominal attribute substantially increased thresholded recall and F1 relative to continuous-only ML-SAS, accompanied by a small increase in the background false-positive proportion and lower rank AUC. The paired LOVO diagnostic identified the nominal attribute as the largest source of change in the learned local similarity surface. Within the mixed specification, path proximity produced modestly higher recall, F1, and rank AUC than terminal-node proximity. These findings position ML-SAS as a complementary statistic for analyses requiring a joint representation of mixed attributes.

The post hoc category diagnostics show how the effect of adding CAT1 varied across target categories in DGP 4: the prevalent target category gained detections, while uncommon target categories lost both detections and ranking quality. This distinction matters when a planted spatial region contains heterogeneous attribute profiles.

In the empirical study of Dutch municipalities, both the continuous-only and mixed specifications showed stronger global spatial association than expected under random reassignment. Relative to the continuous-only specification, the mixed specification incorporating ordered urbanity produced 23 local classification changes and yielded 55 similarity clusters, one dissimilarity/outlier, and 281 nonsignificant municipalities after multiplicity adjustment.

Several practical considerations follow from the construction of ML-SAS. The spatial weights matrix remains a substantive modeling choice: contiguity-based weights emphasize immediate neighborhood structure, whereas distance-based or other graph definitions may represent broader forms of spatial dependence. The learned proximity matrix allows a common analysis of continuous and categorical attributes. The LOVO analysis used here provides a variable-level sensitivity diagnostic, but it does not fully attribute individual pairwise similarities or separately isolate interaction effects; further diagnostics of forest decisions and local similarity contributions are therefore warranted. Storage of a full proximity matrix grows quadratically with the number of spatial units, which may become restrictive for large spatial systems.

Several limitations define the scope of the reported results. ML-SAS depends on the representation supplied to the forest, the construction of the synthetic reference data, and forest specifications. The simulations covered one lattice size and a limited set of continuous, binary, and mixed-type processes, leaving high-dimensional settings, other spatial supports, and broader attribute distributions for future evaluation. The empirical analysis was limited to one national municipal dataset and four variables.

Further research should examine additional spatial graphs, sample sizes, data-type combinations, missing-data patterns, and imbalanced categories. Broader comparisons with pre-specified mixed-type similarity measures and diagnostics of how individual attributes and forest decisions shape the proximity surface are also needed. For larger spatial systems, approximate or sparse proximities, block-wise computation, and low-rank representations may reduce computational and storage demands. The same matrix-correlation framework could also be evaluated with similarities derived from other ensemble methods, kernels, or learned representations. Applications in environmental monitoring, public health, socioeconomic inequality, and land-use change would provide further evidence on the behavior of ML-SAS across empirical settings.

ML-SAS thus extends spatial analysis to multivariate attributes that resist a single conventional distance or covariance structure while retaining explicit randomization-based inference.

Acknowledgments

The authors thank the editor and the anonymous reviewers for their constructive comments and suggestions, which helped improve the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available in tmap R package at https://r-tmap.github.io/tmap/articles/basics_facets. These data were derived from the following resources available in the public domain: Statistics Netherlands (CBS), Key Figures for Districts and <https://www.cbs.nl/nl-nl/cijfers/detail/85318NED>.

References

- Aggarwal, C. C., A. Hinneburg, and D. A. Keim. 2001. "On the Surprising Behavior of Distance Metrics in High Dimensional Space." In *Database Theory—ICDT 2001. Lecture Notes in Computer Science*, 1973, edited by J. Van den Bussche and V. Vianu, 420–434. Springer. https://doi.org/10.1007/3-540-44503-X_27.
- Anselin, L. 1995. "Local Indicators of Spatial Association—LISA." *Geographical Analysis* 27, no. 2: 93–115. <https://doi.org/10.1111/j.1538-4632.1995.tb00338.x>.
- Anselin, L. 2019. "A Local Indicator of Multivariate Spatial Association: Extending Geary's c." *Geographical Analysis* 51, no. 2: 133–150. <https://doi.org/10.1111/gean.12164>.
- Anselin, L., I. Syabri, and O. Smirnov. 2002. "Visualizing Multivariate Spatial Correlation With Dynamically Linked Windows." In *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*, edited by L. Anselin and S. Rey. Center for Spatially Integrated Social Science, University of California, Santa Barbara.
- Beyer, K. S., J. Goldstein, R. Ramakrishnan, and U. Shaft. 1999. "When Is 'Nearest Neighbor' Meaningful?" In *Database Theory—ICDT'99. Lecture Notes in Computer Science* 1540, edited by C. Beeri and P. Buneman, 217–235. Springer. https://doi.org/10.1007/3-540-49257-7_15.
- Breiman, L. 2001. "Random Forests." *Machine Learning* 45, no. 1: 5–32. <https://doi.org/10.1023/A:1010933404324>.
- Breiman, L. 2002. "Manual on Setting Up, Using, and Understanding Random Forests V3.1." https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf.
- Caldas de Castro, M., and B. H. Singer. 2006. "Controlling the False Discovery Rate: A New Application to Account for Multiple and Dependent Tests in Local Statistics of Spatial Association." *Geographical Analysis* 38, no. 2: 180–208. <https://doi.org/10.1111/j.0016-7363.2006.00682.x>.
- Cliff, A. D., and J. K. Ord. 1981. *Spatial Processes: Models & Applications*. Pion.
- Czaplewski, R. L., and R. M. Reich. 1993. *Expected Value and Variance of Moran's Bivariate Spatial Autocorrelation Statistic for a Permutation Test*. Research Paper RM-309, 13. U.S. Department of Agriculture, Forest Service, Rocky Mountain Forest and Range Experiment Station.
- Geary, R. C. 1954. "The Contiguity Ratio and Statistical Mapping." *Incorporated Statistician* 5, no. 3: 115–141. <https://doi.org/10.2307/2986645>.
- Getis, A. 2008. "A History of the Concept of Spatial Autocorrelation: A Geographer's Perspective." *Geographical Analysis* 40, no. 3: 297–309. <https://doi.org/10.1111/j.1538-4632.2008.00727.x>.

Getis, A., and J. K. Ord. 1992. "The Analysis of Spatial Association by Use of Distance Statistics." *Geographical Analysis* 24, no. 3: 189–206. <https://doi.org/10.1111/j.1538-4632.1992.tb00261.x>.

Griffith, D. A. 1993. "Which Spatial Statistics Techniques Should be Converted to GIS Functions?" In *Geographic Information Systems, Spatial Modelling and Policy Evaluation*, edited by M. M. Fischer and P. Nijkamp, 103–114. Springer. https://doi.org/10.1007/978-3-642-77500-0_7.

Hubert, L. J., R. G. Golledge, and C. M. Costanzo. 1981. "Generalized Procedures for Evaluating Spatial Autocorrelation." *Geographical Analysis* 13, no. 3: 224–233. <https://doi.org/10.1111/j.1538-4632.1981.tb00731.x>.

Kruber, F., J. Wurst, E. S. Morales, S. Chakraborty, and M. Botsch. 2019. "Unsupervised and Supervised Learning With the Random Forest Algorithm for Traffic Scenario Clustering and Classification." In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2463–2470. <https://doi.org/10.1109/IVS.2019.8813994>.

Lee, M. 2012. "Multivariate Spatial Cluster Analysis Using Mahalanobis Distance." *Journal of the Korean Cartographic Association* 12, no. 2: 37–46.

Lee, S.-I. 2001. "Developing a Bivariate Spatial Association Measure: An Integration of Pearson's r and Moran's I ." *Journal of Geographical Systems* 3, no. 4: 369–385. <https://doi.org/10.1007/s101090100064>.

Lee, S.-I. 2004. "A Generalized Significance Testing Method for Global Measures of Spatial Association: An Extension of the Mantel Test." *Environment and Planning A: Economy and Space* 36, no. 9: 1687–1703. <https://doi.org/10.1068/a34143>.

Lee, S.-I. 2009. "A Generalized Randomization Approach to Local Measures of Spatial Association." *Geographical Analysis* 41, no. 2: 221–248. <https://doi.org/10.1111/j.1538-4632.2009.00749.x>.

Lee, S.-I. 2017. "Correlation and Spatial Autocorrelation." In *Encyclopedia of GIS*, edited by S. Shekhar, H. Xiong, and X. Zhou, 360–368. Springer. https://doi.org/10.1007/978-3-319-17885-1_1524.

Mantel, N. 1967. "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research* 27, no. 2 pt 1: 209–220.

Moran, P. A. P. 1950. "Notes on Continuous Stochastic Phenomena." *Biometrika* 37, no. 1/2: 17–23. <https://doi.org/10.1093/biomet/37.1-2.17>.

Omar, Z. A., M. A. Abu Bakar, Z. H. Zamzuri, and N. M. Ariff. 2022. "Duplicate Detection Using Unsupervised Random Forests: A Preliminary Analysis." In *2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS)*, 66–71. <https://doi.org/10.1109/AiDAS56890.2022.9918724>.

Ord, J. K., and A. Getis. 1995. "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application." *Geographical Analysis* 27, no. 4: 286–306. <https://doi.org/10.1111/j.1538-4632.1995.tb00912.x>.

Peerbhay, K. Y., O. Mutanga, and R. Ismail. 2015. "Random Forests Unsupervised Classification: The Detection and Mapping of *Solanum mauritanum* Infestations in Plantation Forestry Using Hyperspectral Data." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8, no. 6: 3107–3122. <https://doi.org/10.1109/JSTARS.2015.2396577>.

Peerbhay, K. Y., O. Mutanga, R. Lottering, and R. Ismail. 2016. "Unsupervised Anomaly Weed Detection in Riparian Forest Areas Using Hyperspectral Data and LiDAR." In *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–5. <https://doi.org/10.1109/WHISPERS.2016.8071797>.

Shi, T., and S. Horvath. 2006. "Unsupervised Learning With Random Forest Predictors." *Journal of Computational and Graphical Statistics* 15, no. 1: 118–138. <https://doi.org/10.1198/106186006X94072>.

Statistics Netherlands. 2024. "Toelichting Wijk- en Buurtkaart 2022, 2023 en 2024." <https://www.cbs.nl/nl-nl/longread/diversen/2024/toelichting-wijk-en-buurtkaart-2022-2023-en-2024>.

Tennekes, M. 2018. "Tmap: Thematic Maps in R." *Journal of Statistical Software* 84, no. 6: 1–39. <https://doi.org/10.18637/jss.v084.i06>.

Wartenberg, D. 1985. "Multivariate Spatial Correlation: A Method for Exploratory Geographical Analysis." *Geographical Analysis* 17, no. 4: 263–283. <https://doi.org/10.1111/j.1538-4632.1985.tb00849.x>.

Wu, S. H., B. N. Lee, L. C. Wang, and M. S. Abadir. 2007. "Statistical Analysis and Optimization of Parametric Delay Test." In *2007 IEEE International Test Conference*, 1–10. <https://doi.org/10.1109/TEST.2007.4437626>.

Zhang, J., M. Zulkernine, and A. Haque. 2008. "Random-Forests-Based Network Intrusion Detection Systems." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38, no. 5: 649–659. <https://doi.org/10.1109/TSMCC.2008.923876>.

Appendix A: Randomization Moments of Global and Local ML-SAS

This appendix reports the analytical expectations and variances that support the inferential discussion in Section 3.2. The results follow the generalized randomization approach for matrix-correlation statistics (Mantel 1967; Cliff and Ord 1981; S.-I. Lee 2004, 2009).

Global ML-SAS

Under spatial randomization, the expected value of the global ML-SAS is 1. Its variance is

$$\text{Var}(ML-SAS) = \left[\frac{n(n-1)}{S_0 T_0} \right]^2 \left[\frac{S_1 T_1}{2n(n-1)} + \frac{(S_2 - 2S_1)(T_2 - 2T_1)}{4n(n-1)(n-2)} + \frac{(S_0^2 + S_1 - S_2)(T_0^2 + T_1 - T_2)}{n(n-1)(n-2)(n-3)} \right] - 1, \quad (A1)$$

where $S_0 = \sum_i \sum_{j \neq i} w_{ij}$, $S_1 = \frac{1}{2} \sum_i \sum_{j \neq i} (w_{ij} + w_{ji})^2$, $S_2 = \sum_i (w_i + w_i)^2$, and T_0 , T_1 , and T_2 are the corresponding quantities for the matrix $\mathbf{Y}$ (Cliff and Ord 1981; S.-I. Lee 2004).

Local ML-SAS Under Total Randomization

Under total randomization, all observed attribute profiles are exchangeable across locations. Building on the generalized local-randomization framework (S.-I. Lee 2009), the first two moments of the local ML-SAS are

$$E(ML-SAS_i) = \frac{nw_i}{S_0} \quad (A2)$$

and

$$\text{Var}(ML-SAS_i) = \left[\frac{n^2(n-1)}{S_0 T_0} \right]^2 \left[\frac{w_i^{(2)} T_1}{2n(n-1)} + \frac{(w_i^2 - w_i^{(2)})(T_2 - 2T_1)}{4n(n-1)(n-2)} \right] - \left(\frac{nw_i}{S_0} \right)^2, \quad (A3)$$

where $w_i = \sum_{j \neq i} w_{ij}$, $w_i^{(2)} = \sum_{j \neq i} w_{ij}^2$ and $w_i^2 = (\sum_{j \neq i} w_{ij})^2$. The expectation depends only on the spatial proximity structure and equals 1 when $\mathbf{W}$ is row-standardized.

Local ML-SAS Under Conditional Randomization

Under conditional randomization, the attribute profile at the focal location is held fixed and the remaining profiles are permuted. Building on the generalized local-randomization framework (S.-I. Lee 2009), the first two moments are

$$E(ML-SAS_i) = \frac{n^2 w_i y_i}{S_0 T_0} \quad (A4)$$

and

$$\text{Var}(ML-SAS_i) = \left(\frac{n^2}{S_0 T_0} \right)^2 \frac{1}{n-2} \left[(n-1) w_i^{(2)} - w_i^2 \right] \left[(n-1) y_i^{(2)} - y_i^2 \right], \quad (A5)$$

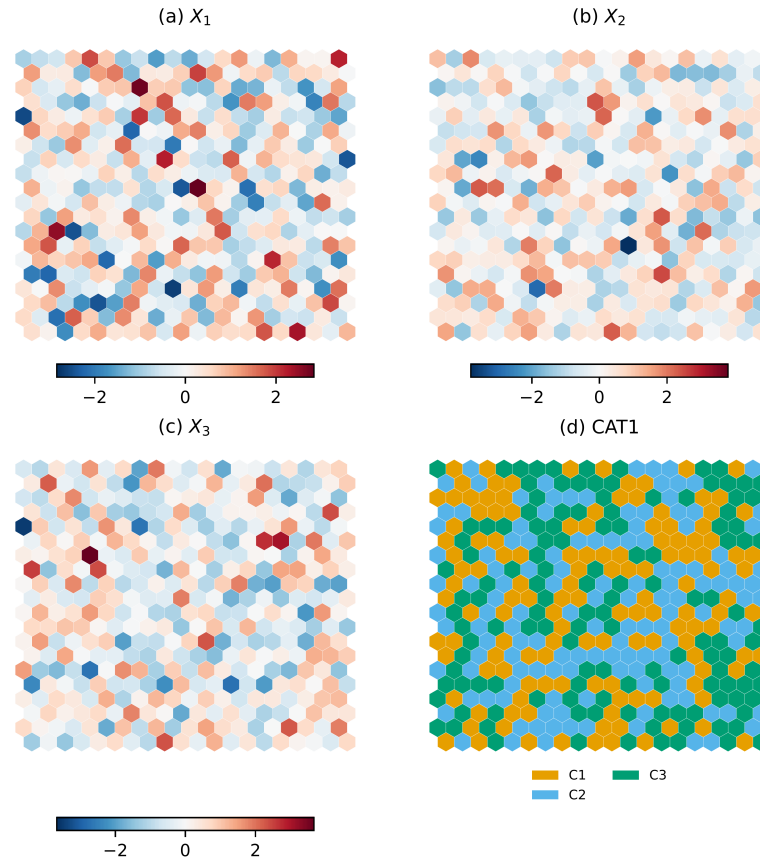


FIGURE B1 | Inputs for the representative DGP 2 dataset (seed 2027). $X_1 - X_3$ are independent standard-normal draws; CAT1 is sampled uniformly from three categories. No target region is planted. This map illustrates the generator and is separate from the 100 independent datasets used for null calibration.

TABLE C1 | Proportion of target cells with raw p values at or below each cutoff, averaged across 100 datasets.

Raw p cutoff	Multivariate Local Geary	Path ML-SAS
0.0001	0.09333	0.06700
0.0010	0.24367	0.20050
0.0030	0.35033	0.31933
0.0500	0.68667	0.72150

Note: Multivariate Local Geary target cells with $p \leq 0.003$ were all in its Positive class. The BH-adjusted recalls are 0.354 for Multivariate Local Geary and 0.280 for path ML-SAS.

where $y_{i.} = \sum_{j \neq i} y_{ij}$, $y_{i.}^{(2)} = \sum_{j \neq i} y_{ij}^2$ and $y_{i.}^2 = (\sum_{j \neq i} y_{ij})^2$. In this case, the expectation depends on both the spatial proximity and the numerical similarity structure, and it will be larger when a given area exhibits greater overall numerical similarity to other areas. If the two matrices are row-standardized, the expected value is 1.

These moments provide theoretical reference values. In the empirical and simulation analyses, statistical significance is assessed using rank-based Monte Carlo permutation tests because the finite-sample local distributions may exhibit skewness and kurtosis that make a normal approximation unreliable.

Appendix B: Spatial Null Input Example

Figure B1 shows the inputs for the representative DGP 2 dataset.

Appendix C: Diagnostics of the Repeated Local Results

The following post hoc diagnostics use the stored local outputs from the 100 repeated datasets for each DGP, applying the prespecified detection

TABLE C2 | Detection within the planted target, pooled over 100 datasets.

Target category	Cells	Continuous detected	Mixed detected
A	5409	2145 (39.66%)	3702 (68.44%)
B/C	591	226 (38.24%)	4 (0.68%)
All	6000	2371 (39.52%)	3706 (61.77%)

Note: The category probabilities are 0.90/0.05/0.05 inside the target and 0.20/0.40/0.40 outside it. Rates use the pooled category counts shown, not an equal-weight average of category rates from datasets with different category sizes. Continuous and mixed refer to path ML-SAS using $X_1 - X_2$ and $X_1 - X_2 - \text{CAT1}$, respectively.

rules. Raw- p -value thresholds below are diagnostic cutoffs, not additional significance decisions. Each dataset has 60 target cells.

DGP 3 Small p Values

Table C1 summarizes the frequencies of small raw p values among target cells under DGP 3.

DGP 4 Category-Specific Recovery

Weighting each stratum by its target-background pair count within each dataset recovers the overall AUC. For A targets, the mean contribution rises from 0.872824 to 0.888797; for B/C targets, it falls from 0.095438 to 0.045230. Their sum changes from 0.968262 to 0.934026. These are contributions to the overall AUC, not separate normalized AUCs. The minority-category loss outweighs the majority-category ranking gain even though the latter produces more detections overall.

TABLE C3 | Mean raw-statistic rank AUC within category-pair strata.

Target	Background	Continuous AUC	Mixed AUC
A	A	0.969500	0.995210
A	B/C	0.967842	0.983451
B/C	A	0.971088	0.637695
B/C	B/C	0.970090	0.396720

Note: In each dataset, target and background cells are compared within the indicated categories; AUCs are then averaged over datasets. A and B/C strata were present in every DGP 4 dataset.

Within DGP 4, these diagnostics link the increase in overall detections to gains for the prevalent target category and the decrease in overall rank AUC to losses for uncommon target categories.