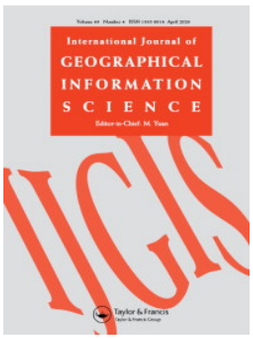

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

An entropogram-based Random Field model for categorical geospatial data prediction

Wen-Bin Zhang, Yong Ge , Xuan Wan , Shengjie Lai & Peter M. Atkinson

To cite this article: Wen-Bin Zhang, Yong Ge , Xuan Wan , Shengjie Lai & Peter M. Atkinson (30 Mar 2026): An entropogram-based Random Field model for categorical geospatial data prediction, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2026.2650365

To link to this article: https://doi.org/10.1080/13658816.2026.2650365

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2026.2650365

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

An entropogram-based Random Field model for categorical geospatial data prediction

Wen-Bin Zhang[a,b,c], Yong Ge[b], Xuan Wan[b], Shengjie Lai[a] and Peter M. Atkinson[c]

[a]WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton, UK; [b]State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China; [c]Lancaster Environment Center, Faculty of Science and Technology, Lancaster University, Lancaster, UK

ABSTRACT

Categorical geospatial data underpin applications from biodiversity monitoring to land-use planning, yet existing approaches often fail to recover rare classes while preserving realistic patch structures. We introduced an Entropogram-based Random Field (ERF) model that integrates intrinsic randomness from local class probabilities with entropogram-derived spatial dependence, balancing local class proportions with global neighborhood associations. Using a 10-class, 1-km land-cover map of Northern Ireland, we compared ERF against Indicator Kriging (IK), multi-phase Indicator Kriging (MIK), Compositional Data Analysis (CoDA) and a spatial multinomial logistic (SMLM) model. ERF matches IK and MIK in overall accuracy but achieves higher recall and F1 scores for minority classes, reducing the loss of small, coherent patches. While CoDA ensures compositional validity, it underperforms on rare classes and increases spatial aggregation; MIK improves rare-class recovery but still favors dominant types. SMLM performs comparably to ERF but with far higher computational demand. Landscape metrics showed that ERF and SMLM best preserved patch diversity and realistic geometry, whereas IK and CoDA produced more aggregated patterns. Together, these results highlight ERF as a computationally efficient, scalable and balanced solution for categorical mapping, particularly in applications where minority-class recovery and spatial realism are critical for biodiversity monitoring, habitat connectivity and land-use planning.

ARTICLE HISTORY
Received 10 September 2025
Accepted 21 March 2026

KEYWORDS
Entropogram; categorical geospatial data; geostatistics

CONTACT Wen-Bin Zhang wb.zhang@soton.ac.uk

1. Introduction

1.1. Background and motivation

Spatial data modeling is essential across a wide range of disciplines, including ecology, the earth sciences, geography, epidemiology and the environmental sciences. Well-defined and accurate spatial modeling is needed to support inference, prediction and decision-making (Koldasbayeva *et al.* 2024), where characterizing the underlying spatial structure is essential. While many established models exist for continuous geospatial data, eg most notably Gaussian random fields (Bailey and Krzanowski 2012, Hristopulos 2020), with their associated prediction and simulation tools, such as Kriging and conditional simulation (Hoshiya 1995, Emery 2007, Chilès and Desassis 2018), respectively, the same level of methodological maturity is lacking for the handling of categorical geospatial data, and tools for categorical spatial prediction remain relatively limited. Besides, in practice, prediction of categorical fields poses major challenges, particularly when the data exhibit severe class imbalance, fine-scale heterogeneity, or small, spatially coherent patches. Rare or minority classes, however, are often environmentally significant, eg small wetlands, riparian buffers, or narrow woodland corridors may comprise a small proportion of total area yet play disproportionately important roles in biodiversity, ecosystem functioning, risk monitoring and conservation planning (Tulloch *et al.* 2016, Hunter *et al.* 2017). Spatial methods that over-smooth rare classes or merge them into dominant surrounding categories can therefore distort ecological or planning decisions.

1.2. Literature review

Several approaches have been proposed to handle categorical geospatial data, which present unique challenges compared to continuous numerical data. Unlike temperature or elevation, categorical variables (such as soil types or land use classes) lack a natural ordering or numerical scale, making standard statistical tools derived from the mean or variance inapplicable directly. One of the most widely used is a variant of Indicator Kriging (IK) which transforms each category into a binary indicator layer and applies traditional Kriging independently to each (Chiang *et al.* 2014). Although the IK method leverages the well-established geostatistical variogram, the IK technique was not designed originally for categorical data but rather as an extension of the traditional variogram to accommodate non-Gaussian and skewed continuous variables by thresholding them into binary indicators at various cutoffs (Journel 1983). The above variant of IK becomes increasingly inefficient in multi-class problems (Li and Zhang 2019), as it requires multiple interpolations of binary variables, one for each of the categories relative to others.

A well-known limitation of IK is its tendency to underestimate minority or rare classes, leading to oversmoothed maps in which small, spatially coherent patches are often suppressed. To address this problem, Soares (1992) proposed a modified IK procedure for multi-phase structures, ie multi-phase IK (MIK), that adjusts the estimation to better preserve underrepresented classes. While this approach offers improvement for rare-class estimation, it enforces mutual exclusivity for each single category,

without capturing inter-class relationships or joint spatial structure. As a result, there is no unified analogue to a 'categorical variogram' that captures holistically the spatial dynamics of the entire categorical field. By contrast, in multivariate geostatistics, joint spatial dependence is commonly formalized through valid cross-covariance (or cross-variogram) structures (Genton and Kleiber 2015).

An alternative approach is offered by probabilistic models. Markov Chain Random Fields (MCRFs) model spatial dependencies through the conditional probability of each class given the states of neighboring observations (Li 2007, Carle and Fogg 2024). This framework aligns well with the discrete nature of categorical data and naturally supports both prediction and simulation (Cao *et al.* 2011b, Li *et al.* 2015, Yang *et al.* 2023). However, typical MCRF implementations often rely on simplistic or heuristically defined transition rules that may not fully capture complex spatial patterns (Zhang *et al.* 2024a). Additionally, when conditioning on all surrounding observations, MCRFs can suffer from a shadowing effect, where predictions are overly influenced by the spatially dominant or nearest conditioning locations, rather than reflecting the full configuration of surrounding data. This effect can lead to overly deterministic or oversmoothed predictions, as the information from other less proximate observations is marginalized (Li and Zhang 2019). Another general probabilistic framework is Bayesian Maximum Entropy (BME), which integrates diverse data sources and constraints into a unified posterior through entropy maximization (Christakos 1990, He and Kolovos 2018). While originally developed for non-Gaussian continuous fields, BME has been adapted to nominal data by imposing class-specific constraints (Bogaert 2002). However, defining these constraints can be complex, and it does not inherently address multi-class spatial interactions as seamlessly as a dedicated categorical measure.

More recently, information-theoretical measures have been explored for spatial modeling. Methods based on conditional entropy, such as the histogram via entropy reduction (HER) approach and related non-parametric geostatistical models (Thiesen *et al.* 2020, Thiesen and Ehret 2022), offer alternatives to variogram-based techniques and provide tools for assessing local and spatial uncertainty. In parallel, the entropogram (Zhang *et al.* 2023, Zhang *et al.* 2024b) uses mutual information to quantify shared information between spatially lagged categorical variables, offering a holistic measure of multi-class spatial dependence that jointly reflects within-class coherence and cross-class transitions. However, the theoretical properties, admissibility conditions and optimal parametric models for entropograms and related information-theoretic measures remain active areas of research.

1.3. Contributions

In this study, we propose an Entropogram-based Random Field (ERF) model for spatial prediction of categorical fields. The core idea is to use the entropogram (Zhang *et al.* 2023) to summarize multi-class spatial dependence and then embed this information into a random-field style predictor. First, we estimate and parametrically fit the entropogram, obtaining lag-dependent mutual information and class-by-class joint probabilities that capture how categories co-occur across distances. These estimates are then used in two ways. We used the entropogram to construct an Indicator-Kriging-like prior, yielding local soft class probabilities that account for spatial correlation across all categories. We

then update this prior using entropogram-based joint probability matrices along the neighborhood lag distances, with mutual information controlling the influence of each neighbor on the posterior class distribution. The resulting ERF predictor combines a kriging-style spatial prior with an information-theoretic correction that explicitly encodes multi-class co-occurrence and transition structure. Compared to standard indicator kriging and multi-phase approaches, ERF provides better balance between majority and minority classes and more faithful reproduction of patch configuration and landscape heterogeneity, while being conceptually simple and scalable to large categorical maps.

2. Method

In this section we describe the proposed Entropogram-based Random Field (ERF) predictor. We first introduce the entropogram as an information-theoretic measure of spatial dependence for categorical random fields and define how it is estimated from data. The empirical entropogram are then summarized by a simple parametric model, which yields lag-dependent mutual information and class–class joint probability tables (Section 2.1). Next, we show how the fitted entropogram is used to construct the ERF predictor (Section 2.2): these are combined with local neighborhood information to form a kriging-style prior over classes and an entropogram-based correction that updates this prior to obtain posterior class probabilities.

2.1. Entropogram

The entropogram $\tau(\mathbf{h})$ was introduced by Zhang *et al*. (2023) as an information-theoretic analogue of the variogram for categorical (and discretized continuous) second-order stationary random field. Instead of squared differences, it is built from the mutual information between the random variables at two locations separated by lag $\mathbf{h}$. For a random field $Z(\mathbf{s})$ taking categorical classes from a set of n classes $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$, the entropogram at lag $\mathbf{h}$ is defined as

$$\begin{aligned}\tau(\mathbf{h}) &= H(Z(\mathbf{s})) + H(Z(\mathbf{s}-\mathbf{h})) - H(Z(\mathbf{s}), Z(\mathbf{s}-\mathbf{h})) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}(\mathbf{h}) \ln\left(\frac{p_{ij}(\mathbf{h})}{p_i p_j}\right)\end{aligned} \tag{1}$$

where $H(Z(\mathbf{s}))$ is the Shannon entropy of the categorical variable $Z(\mathbf{s})$, $p_{ij}(\mathbf{h})$ refers to $p\big(Z(\mathbf{s}) = c_i, Z(\mathbf{s}-\mathbf{h}) = c_j\big)$ is the probability that the two locations are in classes c_i and c_j, and p_i and p_j are the corresponding marginal probabilities. Because the entropogram $\tau(\mathbf{h})$ is de facto mutual information, it inherits several well-known properties: it is non-negative, symmetric in the two locations and equals zero if and only if $Z(s)$ and $Z(s-\mathbf{h})$ are independent at lag $\mathbf{h}$. Under second-order stationarity, dependence summaries are functions of the lag vector $\mathbf{h}$. In this work we further assume isotropy, so dependence depends only on lag distance $h = \| \mathbf{h} \|$; consequently, we use omni-directional pairings to estimate $p_{ij}(h)$ and $\tau(h)$, which increases the effective sample size per lag bin and stabilizes estimation. If directional effects are present, the same construction can be applied by conditioning on the direction (angle) of $\mathbf{h}$.

The entropogram can be normalized by dividing the information of the other locations given the reference location,

$$\overline{\tau}(h) = \frac{\tau(h)}{H(Z)} = -\frac{\tau(h)}{\sum_{j=1}^{n} p_j \ln p_j} \in [0, 1] \tag{2}$$

In this way, the normalized entropogram in Equation (2) measures the information share of a location that can be provided by the observed reference location. For a stationary categorical random field, mutual information satisfies $0 \leq \tau(h) \leq H(Z)$, so $0 \leq \overline{\tau}(h) \leq 1$ whenever $H(Z) > 0$. In the idealized case of coincidence locations (ie comparing $Z(\mathbf{s})$ with itself), $\tau(0) = H(Z)$ and thus $\overline{\tau}(0) = 1$. In practice, however, a non-negligible nugget may exist due to boundaries and microscale variation, leading to a $\overline{\tau}(0)$ less than 1.

2.1.1. Monte Carlo estimation of joint probability

Empirical estimation is performed separately for each target lag distance h, as in classical variogram construction; this is an estimation strategy and does not assume independence of the underlying process across lags. To estimate the joint probability distribution for a specific lag distance h, we first randomly sampled several locations (indexed by $\mathbf{s}_i$) from the total data. Each location $\mathbf{s}_i$ could be sampled multiple times, and the full set of sampled locations was treated as a reference point for which co-occurrences were evaluated. For every other location $\mathbf{s}_j$ (including the sampled location $\mathbf{s}_i$), the distance from $\mathbf{s}_j$ to $\mathbf{s}_i$ was computed. This distance was then compared against the target lag distance h, and a Gaussian kernel was used, where the weight is defined as $w_{ij}(h) = \exp\left[-\frac{\left(d(\mathbf{s}_i, \mathbf{s}_j) - h\right)^2}{2\sigma^2}\right]$. If the distance $d(\mathbf{s}_i, \mathbf{s}_j)$ between locations $\mathbf{s}_i$ and $\mathbf{s}_j$ is close to the target lag distance h, $w_{ij}(h)$ would be relatively large, indicating that this pair contributes more strongly to the joint probability for lag distance h, and *vice versa*. The weighted contributions, $w_{ij}(h)$, were accumulated by counting how often each pair of classes co-occurred in the samples, forming a raw measure of co-occurrences that was specifically tuned to lag distance h.

After summing the weights, we obtained a non-symmetric matrix where rows refer to the class probability distribution for the sampled location $\mathbf{s}_i$, and columns refer to the class probability distribution for the other location $\mathbf{s}_j$,

$$P^h = \begin{bmatrix} p_{11}(h) & \cdots & p_{1n}(h) \\ \vdots & \ddots & \vdots \\ p_{n1}(h) & \cdots & p_{nn}(h) \end{bmatrix} \tag{3}$$

where $p_{ij}(h)$ refers to the class pairs c_i and c_j at locations $\mathbf{s}_i$ and $\mathbf{s}_j$, respectively. Transition probabilities P^h characterize how categorical classes change with lag distance (Li 2006). The matrix was then normalized to produce a joint probability matrix, such that the sum of probabilities of all class pairs equals to 1.

2.1.2. Model fitting

To further smooth out random fluctuations and obtain the underlying trends with limited sparse samples, we fitted simple parametric models to both the distance-

weighted joint probability estimates in Equation (3) and the sample entropogram. Although spatial association generally decays with increasing distance, some class-pair joint probabilities can increase over short ranges, eg for classes that tend to occur together along edges. Thus, we used a flexible function,

$$f(h) = a - \frac{k}{1 + \exp(b * h)} \tag{4}$$

for all the curves $p_{ij}(h)$, where a, b and k are parameters. This form is bounded and can accommodate both monotone decay and mild non-linear curvature with distance. In practice, we constrain the fitted curves to remain non-negative and within the range implied by the empirical joint probabilities (and, for the entropogram, by the marginal entropy), truncating any small negative values that may arise from numerical fitting. The resulting smooth curves provide more interpretable summaries than raw distance-specific estimates and yield an entropogram model that captures how class co-occurrences evolve from near-adjacent distances to large separations, while respecting the basic constraints inherited from mutual information.

2.2. Entropogram-based Random Fields

Let Ω denote the spatial domain of interest, partitioned into locations $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_m\}$ that each could assume a class label from a finite set $\mathcal{C} = \{c_1, c_2, ..., c_n\}$. A subset of these locations $\{\mathbf{s}_1, ..., \mathbf{s}_k\} \subseteq \Omega$ was observed, while the remaining locations $\{\mathbf{s}_{k+1}, ..., \mathbf{s}_m\}$ were unobserved and needed to be predicted. The task was to find a configuration of classes $\mathbf{z} = \{z_1, ..., z_m\}$, with each $z_i \in \mathcal{C}$, that approximated the maximum *a posteriori* (MAP) solution under a spatial Bayesian framework.

Each unobserved location $\mathbf{s}_u$, $u \in \{k+1, ..., m\}$, was assigned with a class probability distribution defined by

$$P\big(Z(\mathbf{s}_u)\big) = [1 - \overline{\tau}(0)]\pi_{Z(\mathbf{s}_u)} + \overline{\tau}(0)P\big(Z(\Omega\ \{\mathbf{s}_u\})|Z(\mathbf{s}_u)\big) \tag{5}$$

where $\pi_{Z(\mathbf{s}_u)}$ is the intrinsic randomness at $\mathbf{s}_u$, calculated via Kriging system where covariance was filled by the corresponding normalized entropogram values. The intrinsic randomness reflects the baseline variability that cannot be explained by spatial dependence, which can be viewed as analogous to the nugget effect in geostatistics. The second component $P\big(Z(\Omega\ \{\mathbf{s}_u\})|Z(\mathbf{s}_u)\big)$ is the conditional probability, encoding spatial dependence using the joint probability over different spatial lags. However, the exact joint conditional distribution $P\big(Z(\Omega\ \{\mathbf{s}_u\})|Z(\mathbf{s}_u)\big)$ cannot be uniquely reconstructed from pairwise entropogram information alone without specifying a full higher-order probability model. We therefore adopt a conditional-independence approximation, treating neighbors as conditionally independent given the class at $\mathbf{s}_u$ and weighing each neighbor's contribution by $\overline{\tau}(h)$. This leads to approximation,

$$P\big(Z(\Omega\ \{\mathbf{s}_u\})|Z(\mathbf{s}_u)\big) = \prod\nolimits_{\mathbf{s}_j \in Z(\Omega\ \{\mathbf{s}_u\})} P\big(Z(\mathbf{s}_j)|Z(\mathbf{s}_u)\big)^{\overline{\tau}\left(h = d(\mathbf{s}_u, \mathbf{s}_j)\right)} \tag{6}$$

where $P\big(Z(\mathbf{s}_j)|Z(\mathbf{s}_u)\big)$ can be estimated via the joint probability matrix in Equation (3) through $\frac{p_{ij}}{\sum_j p_{ij}}$. Here, Equation (6) should be interpreted as a pseudo-likelihood approximation that uses the entropogram to modulate the influence of each observed

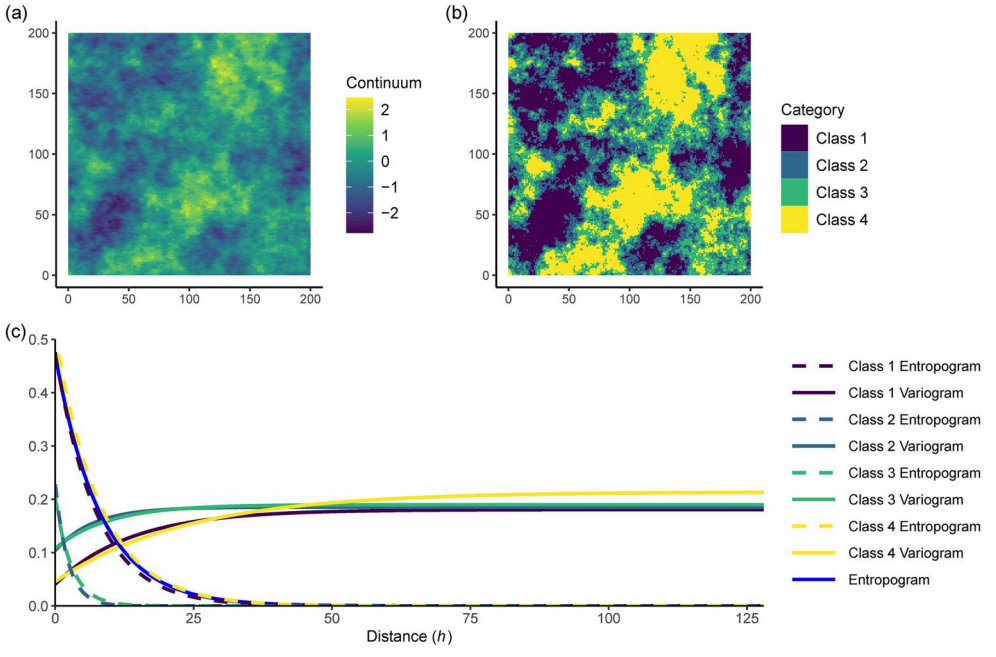

Figure 1. (a) Simulated Gaussian Random Field generated with a known variogram. (b) The same continuous data discretized into four classes of equal proportion. (c) The entropogram and class-specific variograms and entropograms.

neighbor, rather than as an exact expression for the true joint conditional distribution. This approximation is what makes the ERF predictor computationally tractable while still exploiting the multi-class spatial dependence encoded in the entropogram.

3. Results

3.1. Comparison of the entropogram and variogram

Figure 1 demonstrates the process of generating and analyzing simulated categorical geospatial data to illustrate some of the key concepts of, and differences between, the variogram and entropogram. In Figure 1(a), we first generated a simulated Gaussian RF with a known variogram. Then, we converted these continuous geospatial data into categorical data by dividing them into four distinct classes of equal proportions based on quantiles, as shown in Figure 1(b). After this categorization, clear spatial patterns emerged, with class 1 and class 4 exhibiting more pronounced clustering, whereas class 2 and class 3 were distributed in a more fragmented pattern, as expected.

Figure 1(c) compares the variograms and entropograms derived from the categorical field. Both the variograms and entropograms displayed a clear nugget effect, indicating additional uncertainty introduced by the classification process. At lag zero, the overall entropogram reached only about 0.5 rather than the theoretical maximum of 1, reflecting intrinsic uncertainty. Put in other words, even at infinitesimal distances, there was only a 50% chance that a given class remained unchanged. The

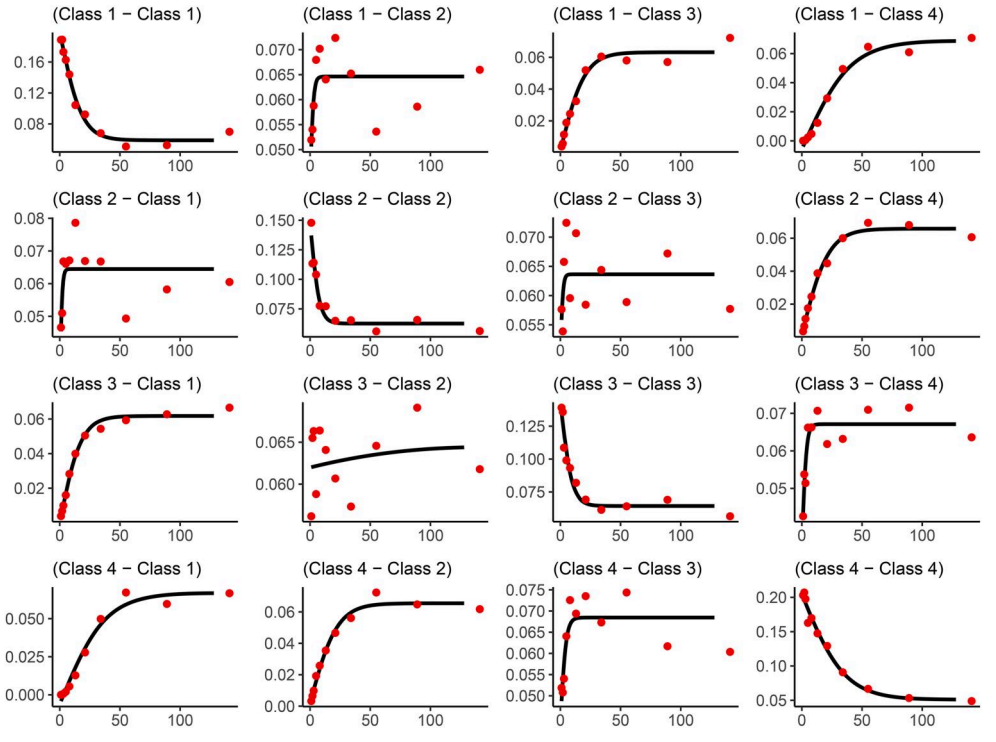

Figure 2. Joint probability matrices (represented by the red dots) and the curves fitted to these data across spatial lags which are used to compute the entropogram. The fit equation is Equation (4).

class-specific entropograms further revealed contrasts in spatial association where classes 1 and 4 exhibited higher variability, while classes 2 and 3 showed greater internal coherence. In this way, the entropogram directly highlights within-class spatial uncertainty, complementing the information provided by variograms, which emphasize aggregate variation across class boundaries.

Figure 2 provides a detailed illustration of the joint probabilities used to derive the entropogram curves shown in Figure 1(c). To enhance computational efficiency and provide accurate estimation across various spatial lags, we summarized the joint probabilities at eight representative spatial lag distances that were defined by powers of two. These spatial lag distances spanned from the smallest meaningful distance between data points to the largest representative separation within the dataset. Generally, the joint probabilities of identical class pairs, represented along the diagonal of Figure 2, started high at the shortest lag distances and gradually decreased with increasing spatial separation. This decline occurred because spatial association and, consequently, joint class occurrence diminished as locations became further apart. In contrast, the joint probabilities for different class pairs (off-diagonal plots) typically increased with distance, reflecting progressively weakened spatial dependence. At larger spatial lags, class occurrences became increasingly independent, thus, approaching the probabilities reflective of random spatial distributions.

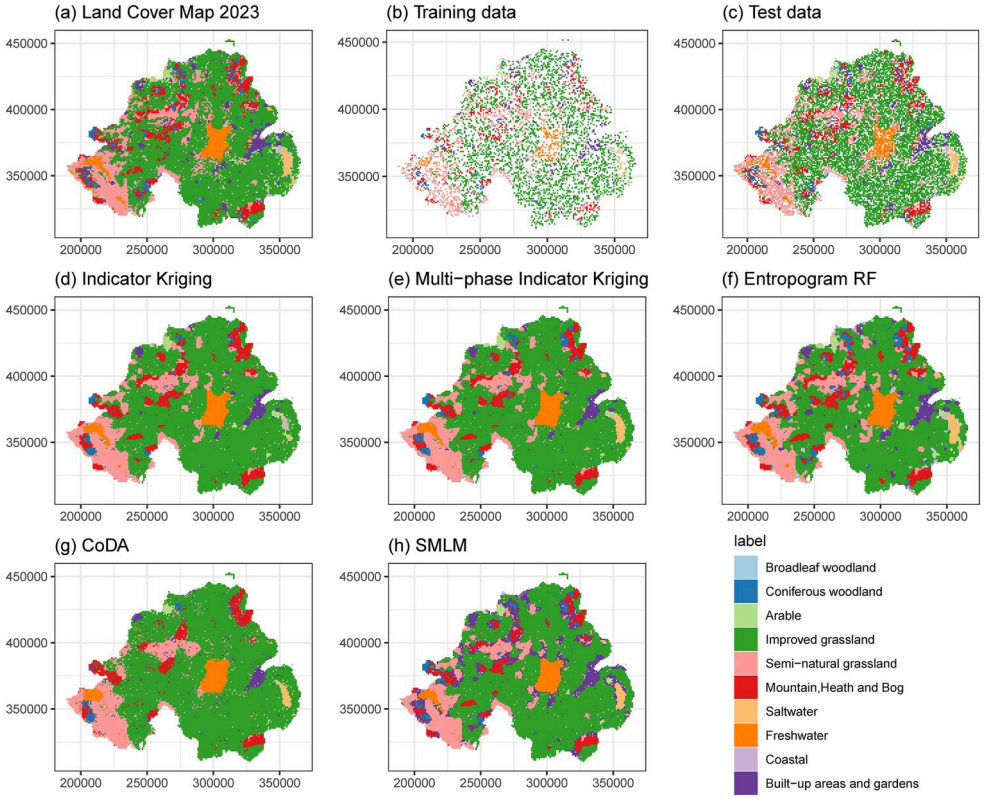

Figure 3. Land cover classification and prediction results. (a) UKCEH Land Cover Map 2023 with a spatial resolution of 1 km and 10 aggregated classes. (b) Training dataset comprising 30% of grid cells. (c) Test dataset comprising the remaining 70% of grid cells. (d) Predictions obtained using Indicator Kriging. (e) Predictions obtained using Multi-phase Indicator Kriging. (f) Predictions obtained using the entropogram-based random field model. (g) Predictions obtained using Compositional Data Analysis (CoDA; ilr-kriging). The CoDA approach applies isometric log-ratio (ilr) transformation to class proportions and Kriging of ilr components before back-transformation to probabilities. (h) Predictions obtained using the spatial multinomial logistic mixed (SMLM) model.

3.2. Spatial prediction

In this section, we used a land cover raster dataset with a spatial resolution of 1 km as a reference to examine the performance of the entropogram-based random field (ERF) model, in comparison with ordinary Indicator Kriging (IK), the multi-phase Indicator Kriging (MIK, Soares 1992), Compositional Data Analysis (CoDA, Greenacre 2021) and a spatial multinomial logistic mixed model (SMLM, Cao *et al.* 2011a). CoDA was implemented via isometric log-ratio (ilr) transformation (Oh *et al.* 2024).

Figure 3(a) shows the land surface of Northern Ireland, classified using the aggregate target class schema. The aggregate class scheme comprises 10 aggregated classes that are groupings of the 21 UKCEH land cover classes based upon Biodiversity Action Plan broad habitats. Figure 3(b) illustrates the 30% of sampled locations ($N = 4290$) that were used to evaluate the entropogram. The remaining 70% of locations ($N = 10{,}011$) in Figure 3(c) were used to validate prediction accuracy.

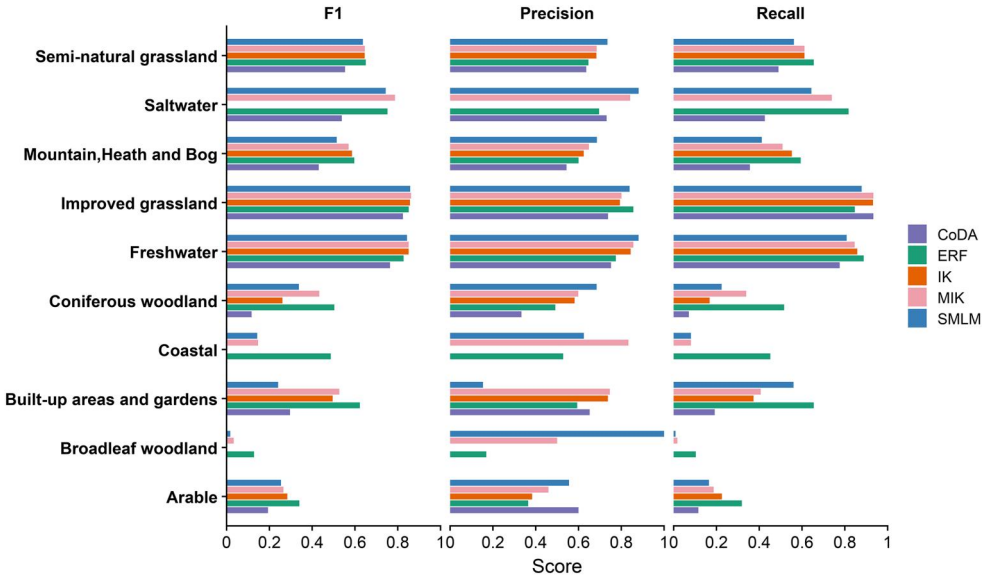

Figure 4. Comparison of per-class predictive performance (Precision, Recall, F1) for five methods: Indicator Kriging (IK), the multi-phase Indicator Kriging (MIK), Entropogram-based Random Field (ERF), Compositional Data Analysis (CoDA; ilr-kriging) and the spatial multinomial logistic mixed model (SMLM).

IK produced a visually smoother categorical map with diffuse patch boundaries and extensive dominance of Improved grassland, see Figure 3(d). However, it completely failed to predict several rare classes present in the test data, such as Broadleaf woodland (1.1% of training points) and Coastal (0.6%). MIK (Figure 3(e)) reduced some of this bias and sharpened a few patch boundaries but still tended to favor the majority classes and under-represent the sparsest types. In contrast, the ERF reproduced most of the small, spatially coherent patches observed in the sample and successfully recovered a broader range of land-cover classes as in Figure 3(f), including extremely sparse classes such as Coastal and Broadleaf woodland. Although CoDA inherently preserved the compositional constraint ensuring that predicted probabilities were valid and interpretable, Figure 3(g) shows that the practical classification performance of CoDA was not markedly superior to IK, particularly in the recovery of minority land-cover types. Finally, the SMLM in Figure 3(h) produced predictions visually comparable to ERF, with improved representation of minority classes and more coherent patch boundaries compared to IK, multi-phase IK and CoDA.

3.3. Model evaluation

Figure 4 compares class-wise precision, recall and F1 scores for Indicator Kriging (IK), the multi-phase Indicator Kriging (MIK), the Entropogram-based Random Field (ERF), Compositional Data Analysis (CoDA) and the spatial multinomial logistic mixed (SMLM) model. For dominant classes, all five methods were performed accurately. Improved grassland achieved F1 scores of 0.86 with IK, 0.86 with MIK, 0.85 with ERF, 0.82 with

Table 1. Classification performance of Indicator Kriging (IK), the multi-phase Indicator Kriging (MIK), Entropogram-based Random Field (ERF) models, Compositional Data Analysis (CoDA; ilr-kriging) and the spatial multinomial logistic mixed (SMLM) model.

Metric	IK	MIK	ERF	CoDA	SMLM
Overall accuracy	0.76	0.77	0.76	0.71	0.72
Kappa	0.56	0.56	0.60	0.45	0.53
Precision	0.72	0.75	0.75	0.67	0.77
Recall	0.76	0.77	0.76	0.71	0.72
F1	0.73	0.75	0.75	0.67	0.72

Note: The number of observations used to evaluate the modes is 10,011.

Table 2. Landscape-level indices comparing the reference land cover map (Ground Truth) with Indicator Kriging (IK), the multi-phase Indicator Kriging (MIK), Entropogram-based Random Field (ERF) models, Compositional Data Analysis (CoDA; ilr-kriging) and the spatial multinomial logistic mixed (SMLM) model predictions.

Metric	Ground truth	IK	MIK	ERF	CoDA	SMLM
Cohesion	97.3	99.0	99.0	98.2	99.4	98.5
Contagion	51.4	66.5	68.5	56.7	73.7	63.5
Landscape division	0.73	0.59	0.59	0.74	0.45	0.68
Landscape patch index	51.1	63.8	63.3	50.8	74.0	56.1

CoDA and 0.86 with SMLM, while Freshwater reached 0.85, 0.85, 0.83, 0.76 and 0.84, respectively.

Clearer differences emerged for minority classes, where ERF generally achieved the strongest or near-strongest performance. For Coniferous woodland, F1 increased from 0.26 (IK), 0.43 (MIK), 0.12 (CoDA) and 0.34 (SMLM) to 0.50 with ERF. For the rarest forest type, Broadleaf woodland, IK and CoDA failed completely (zero recall), and MIK and SMLM achieved only exceptionally low F1 (0.03 and 0.02), whereas ERF reached 0.13 by trading some precision for markedly higher recall (0.10). Semi-natural grassland and Mountain, Heath and Bog showed only modest between-model differences ($F1 \approx 0.55$–0.65), but ERF still obtained the highest F1 (0.65 and 0.60, respectively), slightly improving on IK, MIK and SMLM. Built-up areas and gardens also benefited substantially, with ERF raising F1 to 0.62 compared with 0.50–0.53 for IK and MIK, 0.30 for CoDA and 0.24 for SMLM. For Saltwater and Freshwater, MIK achieved the highest F1 (0.79 and 0.85), but ERF remained competitive (0.75 and 0.83) and offered the highest recall for Saltwater (0.82), providing more complete recovery. Overall, these results indicate that ERF provides the most balanced recovery of rare classes across the legend, substantially improving on IK and CoDA and generally matching or surpassing the stochastic alternatives (MIK and SMLM), while retaining computational scalability.

In general, the five approaches achieved broadly comparable performance, with MIK and ERF both reaching weighted F1 scores of 0.75 (see Table 1). MIK also delivered the highest overall accuracy (0.77) and recall (0.77), while ERF achieved the strongest agreement with the reference (kappa $=$ 0.60) and similarly high precision and recall (0.75 and 0.76). IK remained a strong baseline, with slightly lower F1 (0.73) but the same overall accuracy as ERF (0.76) and kappa comparable to MIK (0.56). SMLM performed competitively, attaining the highest weighted precision (0.77) and a weighted F1 score of 0.72 despite slightly lower overall accuracy (0.72). CoDA underperformed on all weighted metrics, with the lowest F1 score (0.67) and kappa (0.45),

indicating weaker agreement with the reference despite enforcing compositional validity.

In addition to classification performance, the landscape indices in Table 2 highlight that the ERF produced spatial patterns most closely matching the reference. ERF achieved values almost identical to the ground truth for landscape division (0.74 vs. 0.73) and dominant patch index (50.8 vs. 51.1) and only slightly higher contagion (56.7 vs. 51.4). SMLM also reproduced the reference landscape reasonably well, with intermediate contagion (63.5) and division (0.68) and a dominant patch index (56.1) closer to the ground truth than IK, MIK or CoDA. IK and MIK yielded more aggregated landscapes, with higher contagion (66.5 and 68.5) and lower division (both 0.59), indicating overexpansion and coalescence of dominant classes. CoDA exhibited the strongest aggregation effects, with the highest contagion (73.7), the largest dominant patch index (74.0) and the lowest division (0.45), consistent with the visual impression of overly smoothed maps.

4. Discussion

4.1. Per-pixel predictive performance

Across the five approaches, global accuracy and weighted F1 were broadly similar, but the way each model balanced majority versus minority classes differed markedly. Standard Indicator Kriging (IK) remained a strong baseline and achieved high scores for dominant classes such as Improved grassland and Arable, but it strongly underrepresented rare categories and completely missed some of them in classification. Its improved model, the multi-phase Indicator Kriging (MIK), explicitly targets this weakness. MIK reduced the complete omission of rare classes and improved recall for several minority categories compared to IK. In our experiment, it achieved slightly higher weighted F1 and recall than IK and showed clearly better recovery of infrequent landcover types, while preserving reliable performance on dominant classes. However, its gains for rare types were still more modest than those obtained by the entropogram-based and multinomial models.

The Entropogram-based Random Field (ERF) matched the best methods in terms of overall accuracy and weighted F1 in this study but redistributed performance more evenly across classes. It substantially improved recall for minority categories such as Coniferous woodland, Mountain, Heath and Bog and Coastal, without sacrificing performance on the major classes. The spatial multinomial logistic mixed (SMLM) model also performed competitively, with particularly high precision and strong F1 for some rare classes, reflecting its fully joint treatment of the categorical field. Compositional Data Analysis (CoDA), despite enforcing compositional constraints by construction, consistently underperformed on both dominant and minority classes in this highly imbalanced setting.

Overall, IK and MIK provide solid baselines, with MIK clearly mitigating the classic minority-class underestimation of standard IK. ERF and SMLM go further: they achieve comparable or better global metrics while delivering the strongest and most consistent recovery of rare classes, which is often the primary concern in ecological and planning applications.

4.2. Landscape structure

Landscape-level indices show that similar per-pixel metrics can mask crucial differences in spatial configuration. ERF produced spatial patterns that most closely resembled the reference map, preserving both fragmentation and patch geometry. Its contagion, fragmentation (division) and dominant patch index remained close to the ground truth, suggesting that ERF neither over-aggregates nor over-fragments the land-cover mosaic.

SMLM also yielded realistic landscapes and captured patch structure reasonably well, with indices generally between those of ERF and the kriging-based methods. MIK, while improving the representation of minority classes relative to IK, still tended to produce more aggregated patches of dominant classes and somewhat larger contiguous regions than observed in the reference. IK behaved similarly, with both kriging-based predictors showing a tendency for dominant categories to expand and coalesce. CoDA amplified this effect most strongly, yielding the smoothest maps, the largest dominant patches and the highest contagion.

For applications concerned with connectivity, edge effects and small but coherent habitat fragments, these differences are non-trivial (Keeley *et al.* 2021, Keller and Sullivan 2023). ERF and SMLM not only recover rare categories better but also preserve the spatial mosaics they form, whereas the variogram-based methods, even in their multi-phase version, tend to simplify the landscape into larger homogeneous blocks.

4.3. Computational considerations

The computational cost of ERF has two main components: (1) estimation and parametric fitting of the entropogram and associated joint probability curves and (2) local neighborhood prediction. The entropogram is computed once on the training data and then reused for all subsequent predictions. In our experiments, entropogram estimation remained modest even as the training set grew. For training sample sizes of 500, 1000, 2000 and 5000 points, computation times were 39–48 s, indicating computation time scales approximately sub-linearly with the size of training dataset.

Prediction cost is dominated by local neighborhood operations. To quantify scaling, we varied both the number of neighbors k and the number of prediction locations n_{pred}. With $k = 5$, runtimes increased roughly linearly with n_{pred}: about 6.2 s for 100 locations, 32.4 s for 500 and 61.9 s for 1000. At fixed n_{pred}, runtimes grew approximately proportionally with k: for 500 prediction locations, computation times were about 32.4 s for $k = 5$, 59.3s for $k = 10$ and 123.8 s for $k = 20$. These patterns are consistent with the expected $O(n_{\text{pred}} \cdot k)$ scaling of the ERF predictor, which relies only on small local systems rather than inversion of a global covariance matrix.

In the case study, the one-off entropogram estimation cost was small relative to the time spent on per-pixel prediction, and the overall runtime of ERF was of the same order as indicator kriging and CoDA for comparable neighborhood sizes, while remaining substantially lower than that of the SMLM implementation. These results support our claim that ERF is computationally feasible and scalable for large categorical maps, provided that a moderate neighborhood size is chosen.

4.4. Relationships to other approaches

ERF builds on one particular dependence measure, the entropogram, but sits within a broader ecosystem of tools for categorical spatial modeling. Indicator variograms, indicator cross-variograms and the associated indicator covariance and cross-covariance functions provide a rich second-order description of multi-class fields. In principle, the full set of these auto- and cross-structures contains more information about class relationships than a single entropogram curve, which aggregates all joint behavior into a scalar function of lag. Transiograms further encode directional transition probabilities between classes and are particularly well suited to Markov-chain-based modeling.

The entropogram trades some of this detailed structure for parsimony and robustness. By aggregating over all class pairs, it summarizes within-class persistence and between-class transitions into a single mutual-information-based measure at each lag. This inevitably discards some pair-specific information, but it stabilizes estimation when the number of classes is large and provides a natural, scalar weighting function that reflects how informative neighbors are, on average, at each distance. ERF then reintroduces class-specific structure via fitted joint probability matrices, which are used to approximate conditional probabilities within a neighborhood.

Information-theoretic approaches based on conditional entropy, such as histogram via entropy reduction (HER) and related non-parametric geostatistical methods, offer another family of tools. These methods have been developed for continuous variables and focus on prediction and uncertainty quantification by optimizing a conditional entropy criterion. By contrast, the entropogram uses mutual information to characterize dependence in categorical or discrete fields, and in ERF, it is employed to weight and modulate local class distributions. ERF therefore complements HER-type approaches, although they are both rooted in information theory, by targeting different data types and embed entropy or mutual information at distinct stages of the modeling process.

Finally, ERF does not aim to replace covariance or transiogram-based modeling. Rather, it offers a pragmatic alternative for situations where modeling the full indicator cross-covariance structure is computationally or practically infeasible, yet one still wishes to exploit multi-class neighborhood structure through an explicit, interpretable dependence measure.

4.5. Methodological limitations and future directions

Conceptually, ERF is an approximation inspired by Bayesian updating rather than a fully specified probabilistic generative model. The mixing formulas combine a local prior (derived from entropogram-based kriging weights) with entropogram-informed conditional information under a product-of-experts style assumption across neighbors. This construction is motivated by mutual information and conditional probability rules but is not derived from an explicit joint likelihood. It therefore should be viewed as an information-theoretic smoothing and weighting scheme with a Bayesian flavor, rather than a complete hierarchical model. A more formal underpinning, for instance via composite likelihoods or Markov random field theory, is an important direction for future research.

A second limitation concerns the parametric modeling of the entropogram and joint probability curves. We impose simple constraints, such as non-negativity, monotonic decay with distance and upper bounds defined by marginal entropy and truncate small negative artefacts of numerical fitting. However, a full characterization of admissible entropogram functions, eg analogous to Bochner's theorem for covariance models, is not yet available. In this work, we therefore treat the fitted entropogram primarily as a smooth, bounded summary used to construct weights.

Third, although the entropogram could, in principle, be used as a target for simulation, the ERF implementation developed here is prediction oriented. It does not guarantee reproduction of the empirical entropogram under repeated realizations, nor does it produce conditional simulations in the classic geostatistical sense. Extending ERF to a proper simulation framework, such as by designing samplers that explicitly honor the entropogram, is a promising direction.

Finally, our empirical analysis is based on a single, moderately complex land-cover dataset. The observed advantages of ERF and SMLM in rare-class recovery and landscape realism are consistent with theoretical considerations, but broader generalization will require further case studies across different biomes, spatial resolutions, anisotropic structures and levels of class imbalance.

5. Conclusion

We proposed an Entropogram-based Random Field (ERF) model for categorical spatial prediction and evaluated it alongside Indicator Kriging (IK), multi-phase Indicator Kriging (MIK), Compositional Data Analysis (CoDA) and a spatial multinomial logistic mixed (SMLM) model. MIK successfully mitigated some of the classic weaknesses of standard IK by improving the representation of minority classes, but ERF and SMLM went further, combining strong global metrics with substantially better rare-class recovery and more realistic patch structure. Although ERF is an approximate, prediction-focused framework rather than a full simulation model, its balance between accuracy, rare-class retention, landscape realism and scalability makes it a practical choice for applications where both minority classes and fine-scale spatial patterns are critical, including biodiversity monitoring, habitat connectivity analysis and land-use planning.

Author contributions

CRediT: **Wen-Bin Zhang**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft; **Yong Ge**: Conceptualization, Supervision, Writing – review & editing; **Xuan Wan**: Data curation, Writing – review & editing; **Shengjie Lai**: Investigation, Writing – review & editing; **Peter M. Atkinson**: Conceptualization, Methodology, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the National Natural Science Foundation of China (No. 42230110, Y.G.), Horizon Europe (UKRI Guarantee 10041831, S.L.) and the U.S. National Science Foundation (DMS-2327797, S.L.).

Notes on contributors

Wen-Bin Zhang is a Senior Research Fellow at WorldPop, University of Southampton. His research interests include GIScience and complexity. Applications concern health geography, population dynamics and discovering the facts of the world.

Yong Ge is a Full Professor in the Institute of Geographical Science and Natural Resources Research, Chinese Academy of Science. Her research interests include spatial statistics and spatial data science including machine learning. Applications concern poverty, land-use and land-cover change detection and scaling Earth science data.

Xuan Wan is a PhD student in the Institute of Geographical Science and Natural Resources Research, Chinese Academy of Science. Her research interests include spatial statistics and spatial data science, with applications in eco-hydrological process simulation, the evaluation of ecosystem responses to hydrological changes and the scaling of Earth science data.

Shengjie Lai is a Principal Research Fellow in WorldPop at the University of Southampton, specializing in spatial epidemiology and demography. His research integrates diverse data sources and quantitative methods to understand population dynamics and their implications for public health and environmental risks. He has contributed extensively to international collaborative projects and the development of data-driven tools that support evidence-based decision-making.

Peter M. Atkinson is Distinguished professor of Spatial Data Science at Lancaster University, where he was previously Executive Dean of the Faculty of Science and Technology (2015–2025). Peter's research focus is on spatial data science, including the development of geostatistical, machine learning and AI models and the use of Earth observation for application to a wide range of environmental and epidemiological science questions. He is founding Editor-in-Chief of *Science of Remote Sensing* and Associate Editor of *Environmetrics*.

ORCID

Wen-Bin Zhang http://orcid.org/0000-0002-9295-1019
Shengjie Lai http://orcid.org/0000-0001-9781-8148
Peter M. Atkinson http://orcid.org/0000-0002-5489-6880

Data and codes availability statement

The data and codes that support the findings of this study are available via figshare: https://doi.org/10.6084/m9.figshare.30103093.

References

Bailey, T.C., and Krzanowski, W.J., 2012. An overview of approaches to the analysis and modeling of multivariate geostatistical data. *Mathematical Geosciences*, 44 (4), 381–393.

Bogaert, P., 2002. Spatial prediction of categorical variables: the Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 16 (6), 425–448.

Cao, G., Kyriakidis, P.C., and Goodchild, M.F., 2011a. A multinomial logistic mixed model for the prediction of categorical spatial data. *International Journal of Geographical Information Science*, 25 (12), 2071–2086.

Cao, G., Kyriakidis, P.C., and Goodchild, M.F., 2011b. Combining spatial transition probabilities for stochastic simulation of categorical fields. *International Journal of Geographical Information Science*, 25 (11), 1773–1791.

Carle, F.S., and Fogg, E.G., 2024. Conditional simulation of hydrofacies architecture: a transition probability approach. *In: Applied spatiotemporal data analytics and machine learning*. IntechOpen. https://doi.org/10.5772/intechopen.114883

Chiang, J.L., *et al.*, 2014. A feature-space indicator kriging approach for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52 (7), 4046–4055.

Chilès, J.-P., and Desassis, N., 2018. Fifty years of kriging. *In*: B. Daya Sagar, Q. Cheng, and F. Agterberg, eds. *Handbook of mathematical geosciences: fifty years of IAMG*. Cham: Springer, 589–612. https://doi.org/10.1007/978-3-319-78999-6_29

Christakos, G., 1990. A Bayesian/maximum-entropy view to the spatial estimation problem. *Mathematical Geology*, 22 (7), 763–777.

Emery, X., 2007. Conditioning simulations of Gaussian random fields by ordinary kriging. *Mathematical Geology*, 39 (6), 607–623.

Genton, M.G., and Kleiber, W., 2015. Cross-covariance functions for multivariate geostatistics. *Statistical Science*, 30 (2), 147–163.

Greenacre, M., 2021. Compositional data analysis. *Annual Review of Statistics and Its Application*, 8 (1), 271–299.

He, J., and Kolovos, A., 2018. Bayesian maximum entropy approach and its applications: a review. *Stochastic Environmental Research and Risk Assessment*, 32 (4), 859–877.

Hoshiya, M., 1995. Kriging and conditional simulation of Gaussian field. *Journal of Engineering Mechanics*, 121 (2), 181–186.

Hristopulos, D.T., 2020. *Random fields for spatial data modeling*. *In: Advances in geographic information science*. Dordrecht: Springer. https://doi.org/10.1007/978-94-024-1918-4

Hunter, M.L., Jr., *et al.*, 2017. Conserving small natural features with large ecological roles: a synthetic overview. *Biological Conservation*, 211, 88–95.

Journel, A.G., 1983. Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15 (3), 445–468.

Keeley, A.T., Beier, P., and Jenness, J.S., 2021. Connectivity metrics for conservation planning and monitoring. *Biological Conservation*, 255, 109008.

Keller, J.K., and Sullivan, P.J., 2023. The importance of patch shape at threshold occupancy: functional patch size within total habitat amount. *Oecologia*, 203 (1-2), 95–112.

Koldasbayeva, D., *et al.*, 2024. Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications*, 15 (1), 10700.

Li, W., 2006. Transiogram: a spatial relationship measure for categorical data. *International Journal of Geographical Information Science*, 20 (6), 693–699.

Li, W., 2007. Markov chain random fields for estimation of categorical variables. *Mathematical Geology*, 39 (3), 321–335.

Li, W., *et al.*, 2015. Bayesian Markov chain random field cosimulation for improving land cover classification accuracy. *Mathematical Geosciences*, 47 (2), 123–148.

Li, W., and Zhang, C., 2019. Markov chain random fields in the perspective of spatial Bayesian networks and optimal neighborhoods for simulation of categorical fields. *Computational Geosciences*, 23 (5), 1087–1106.

Oh, J., *et al.*, 2024. Using isometric log-ratio in compositional data analysis for developing a groundwater pollution index. *Scientific Reports*, 14 (1), 12196.

Soares, A., 1992. Geostatistical estimation of multi-phase structures. *Mathematical Geology*, 24 (2), 149–160.

Thiesen, S., *et al.*, 2020. Histogram via entropy reduction (HER): an information-theoretic alternative for geostatistics. *Hydrology and Earth System Sciences*, 24 (9), 4523–4540.

Thiesen, S., and Ehret, U., 2022. Assessing local and spatial uncertainty with nonparametric geostatistics. *Stochastic Environmental Research and Risk Assessment*, 36 (1), 173–199.

Tulloch, A.I., *et al.*, 2016. Understanding the importance of small patches of habitat for conservation. *Journal of Applied Ecology*, 53 (2), 418–429.

Yang, H.Q., *et al.*, 2023. Stochastic simulation of geological cross-sections from boreholes: a random field approach with Markov Chain Monte Carlo method. *Engineering Geology*, 327, 107356.

Zhang, B., Li, W., and Zhang, C., 2024a. Sensitivity analysis of the MCRF model to different transiogram joint modeling methods for simulating categorical spatial variables. *Computational Geosciences*, 28 (4), 697–714.

Zhang, W.B., *et al.*, 2023. Spatial association from the perspective of mutual information. *Annals of the American Association of Geographers*, 113 (8), 1960–1976.

Zhang, W.B., *et al.*, 2024b. Scaling geospatial data from the perspective of complexity: exploring the scaling behavior of the entropogram. *Annals of the American Association of Geographers*, 114 (10), 2264–2280.