Received: 10 August 2023 | Revised: 19 May 2024 | Accepted: 23 May 2024

DOI: 10.1002/env.2865

RESEARCH ARTICLE

WILEY

Functional zoning of biodiversity profiles

Natalia Golini[1] | **Rosaria Ignaccolo**[1] | **Luigi Ippoliti**[2] | **Nicola Pronello**[1,3]

[1] Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Turin, Italy

[2] Department of Economics, University "G. d'Annunzio", Pescara, Italy

[3] Department of Neuroscience, Imaging and Clinical Sciences, University "G. d'Annunzio", Chieti, Italy

Correspondence
Natalia Golini, Department of Economics and Statistics "Cognetti de Martiis", University of Turin, Lungo Dora Siena 100 A, 10153 Turin, Italy.
Email: natalia.golini@unito.it

Funding information
National Recovery and Resilience Plan (NRRP); European Union—NextGenerationEU—*Statistics for vegetation biodiversity: estimation and mapping (SVeBio)*, Grant/Award Number: P2022AW4LX – CUP B53D23029510001; European Union—NextGenerationEU—*Complex Environmental Data and Modeling (CoEnv)*, Grant/Award Number: 2022E3RY23 – CUP D53D23011080006

Abstract

Spatial mapping of biodiversity is crucial to investigate spatial variations in natural communities. Several indices have been proposed in the literature to represent biodiversity as a single statistic. However, these indices only provide information on individual dimensions of biodiversity, thus failing to grasp its complexity comprehensively. Consequently, relying solely on these single indices can lead to misleading conclusions about the actual state of biodiversity. In this work, we focus on *biodiversity profiles*, which provide a more flexible framework to express biodiversity through nonnegative and convex curves, which can be analyzed by means of functional data analysis. By treating the whole curves as single entities, we propose to achieve a *functional zoning* of the region of interest by means of a penalized model-based clustering procedure. This provides a spatial clustering of the biodiversity profiles, which is useful for policy-makers both for conserving and managing natural resources and revealing patterns of interest. Our approach is evaluated using a simulation study and discussed through the analysis of the *Harvard Forest Data*, which provides information on the spatial distribution of woody stems within a plot of the Harvard Forest.

KEYWORDS

biodiversity spatial mapping, diversity indices, Hill numbers, penalized model-based clustering, spatial functional data

1 | INTRODUCTION

Biodiversity, or biological diversity, is the scientific term indicating the variability among all living organisms in a given area (see DeLong, 1996). In recent years, there has been increasing concern about biodiversity loss, which not only leads to reductions in populations, genes, and ecosystems but also triggers irreversible environmental changes affecting human health and well-being (Cardinale et al., 2012; Díaz et al., 2006; Schmeller et al., 2020). In response to this decline numerous organizations, agencies, and commissions have established expert working groups or initiatives focused on monitoring, protecting, and restoring biodiversity (Díaz et al., 2015; European Commission, 2021; FAO, 2022; WHO Teams, 2020). At the core of these efforts lies the need for quantitative measurements that capture the complex nature of biodiversity and its spatial and temporal variations.

In the literature, a variety of mathematical functions known as *biodiversity indices* have been introduced to quantify biodiversity (Magurran, 2021; Pielou, 1975). Each index offers a unique perspective on biodiversity, capturing different aspects such as species richness or the distribution of species abundances. Due to these diverse approaches, there is currently no consensus on which indices offer the most accurate representation of biodiversity (Hurlbert, 1971; Purvis & Hector, 2000).

In the literature, a variety of mathematical functions known as *biodiversity indices* have been introduced to quantify biodiversity (Magurran, 2021; Pielou, 1975). Each index offers a unique perspective on biodiversity, capturing different aspects such as species richness or the distribution of species abundances. Due to these diverse approaches, there is currently no consensus on which indices offer the most accurate representation of biodiversity (Hurlbert, 1971; Purvis & Hector, 2000).

In this work, we consider the Hill numbers framework to measure species diversity, focusing on the *effective number of species* (Chao & Colwell, 2022; Hill, 1973). This framework offers a more robust and flexible analysis compared to traditional diversity indices. The Hill numbers refer to a family of species diversity indices defined for a parameter q, called *order* of the diversity, that gives information about the species abundance distribution. Mathematically, the Hill numbers can be represented as a positive, decreasing, and convex curve termed *Hill's biodiversity profile*. This curve can be treated as functional data and analyzed considering the functional data analysis (FDA) approach (Ferraty & Vieu, 2006; Ramsay & Silverman, 2005). This analytical approach was previously proposed by Gattone and Di Battista (2009), who used a functional linear regression model to assess the impact of habitat effects on diversity changes. Analysis of Hill's biodiversity profiles that take advantage of functional tools, such as derivatives, arc length and curvature, have also been proposed by Di Battista et al. (2016), Di Battista et al. (2017), Maturo and Di Battista (2018). In this paper, our focus shifts to clustering these biodiversity profiles, aiming to advance the concept of *functional zoning* in biodiversity analysis. Spatial clustering of biodiversity profiles offers a promising approach that aligns well with contemporary initiatives aimed at biodiversity conservation and natural resource management. By identifying the homogeneous zones based on species distributions and ecological patterns, this method provides a detailed spatial representation of biodiversity, allowing for targeted and informed conservation strategies. Such spatially explicit information can significantly enhance monitoring efforts by highlighting areas of high biodiversity value that require immediate attention or protection. Furthermore, it may aid in the efficient allocation of resources by directing conservation efforts toward areas with the greatest ecological significance. We illustrate the concept of *functional zoning* of biodiversity profiles using the Harvard Forest Data–a well-established dataset (Orwig et al., 2022) containing two censuses of all woody stems with a minimum diameter of 1 cm at breast height. Here, the study operates under the assumption of having complete census data for the entire biological population. While this may not always be feasible in larger study regions, the potential application of functional zoning in scenarios where only a sample of abundance vectors is available at specific sites will be discussed in Section 8. Furthermore, we note that this dataset was previously analyzed by Fortuna and Di Battista (2020), who employed a distance-based LISA map and both hierarchical and k-means algorithms.

From a methodological perspective, while clustering methods for general functional data have been extensively explored, there exists a relatively limited body of research specifically focused on clustering functional data with spatial structures—see, for example, the discussion in the recent review by Zhang and Parnell (2023). Proposals in the frameworks of hierarchical and dynamic clustering approaches, where the similarity between pairs of curves is based on the use of the variogram function, are given by Giraldo et al. (2012), Romano et al. (2015) and Romano et al. (2017). Other approaches based on the use of spatial heterogeneity measures and spatial partitioning methods were also proposed by Dabo-Niang et al. (2010), Secchi et al. (2013), and Fortuna and Di Battista (2020). A few proposals can also be found in the framework of model-based approaches. Vandewalle et al. (2021) and Wu and Li (2023), for example, incorporate longitude and latitude coordinates as regressors in a multinomial logistic regression model, which is employed to estimate the prior probabilities of a mixture model. On the other hand, Jiang and Serban (2012) and Liang et al. (2021) utilize Markov Random Fields and Gibbs distribution to account for spatial dependence in their clustering procedures.

In this paper, we also use a model-based approach for spatially correlated functional data. In particular, we consider a penalized model-based clustering procedure where a finite mixture of Gaussian distributions is used to model the expansion coefficients obtained from approximating the functional biodiversity profiles in a finite-dimensional space. To take care of the presence of spatial correlation, the procedure allows the modeling of the spatial distribution of the weights of the mixture such that observations corresponding to nearby locations are more likely to have similar allocation probabilities than observations that are far apart in space. The procedure represents a generalization of the approach proposed in Vandewalle et al. (2021), and implementation details are provided in Pronello et al. (2023).

The paper is structured as follows. In Section 2 we provide a brief description of the motivating example and the data used in this study. In Section 3 we summarize the key conceptual issues underlying the measurement of biodiversity, discuss some of the most commonly used biodiversity indices, their conversion to effective numbers and the derivation of biodiversity profiles. Section 4 introduces the functional representation of biodiversity profiles and proposes empirical variogram analysis to characterize their possible spatial dependence structure. Section 5 introduces the finite Gaussian Mixture Model (GMM) used for spatial clustering, followed by a performance evaluation against the *functional k-means algorithm* in Section 6 through a simulation study. Section 7 presents the results of the functional zoning of biodiversity profiles using the Harvard Forest Data. Finally, Section 8 concludes the paper with a discussion and suggestions for future research.

2 | THE MOTIVATING CASE STUDY

Forests play a crucial role in tackling biodiversity conservation and restoration. According to FAO and UNEP (2020), forests cover almost one-third of the global land area and harbour most of the terrestrial biodiversity. So it is essential to provide policymakers with a tool to prioritize forestry policies and implement plans that positively impact biodiversity at the population, genetic and ecosystem levels.

Harvard Forest is a vast laboratory and classroom of Harvard University, where observational studies and experiments are conducted to drive research and education on several topics. One of the most relevant is the study of biodiversity. An example of a dataset (data and metadata) for biodiversity studies is the *Harvard Forest CTFS-ForestGEO Mapped Forest Plot since 2014* (Number ID HF253, version 5, Orwig et al., 2022), where data were collected within the 35 ha plot located on Prospect Hill. This plot covers a rectangle area of size 500 m × 700 m, and it was designed to *include a continuous, expansive, and varied natural forest landscape* (Orwig et al., 2022), and it is a continuous grid of 875 cells of size 20 m × 20 m. HF253 is a collection of five datasets freely available for download at https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF253. In particular, we are interested in the most recent dataset, "hf253-05," consisting of 85,641 woody stems greater than 1 cm diameter at 1.3 m (at breast height) collected between May 2018 and January 2020 (second census). However, this census does not contain data from the swamp in the plot's central portion. Data collection in this area was supposed to take place during the winter of 2021 but was not carried out due to restrictions related to the COVID pandemic. Moreover, a winter census for the swamp area was not planned for 2022. Given the unique characteristics of the swamp area, we made the decision not to impute the missing data in this region by means of a statistical technique. Instead, we replaced the missing values with the 37,577 observations collected for the swamp area during the first census, which took place from June 2010 to March 2014. Figure 1 shows the available data within the Prospect Hill Tract long-term plot. In black are displayed the data collected during the second census, while in green we show the data collected during the first census in the swamp area.

Then, the complete dataset consists of 123,218 records providing information on each collected stem, identified by a unique identifier (`stem.id`) representing the primary key of the dataset. However, only some information is of interest for our analysis, specifically: the species mnemonic (the full Latin name, the family and other information on the species are available in the dataset "hf253-02"), the coordinates in meters (m) within the plot relative to the left-down corner of the area of interest, the diameter of the stem in centimeters (cm) and the status of the stem (alive, dead, lost stem, missing, prior). It is crucial to emphasize here that the terms "alive" and "dead" refer to the whole tree. If any stem remains alive, the tree is considered alive. The tree is deemed dead only when every single stem has perished. Given this information, we can calculate abundance data for each tree species within each of the 875 cells of the grid covering the Prospect Hill Tract long-term plot. From a spatial analysis standpoint, this dataset represents a typical example of lattice data.

In this application, we first perform a preprocessing step to focus on the stems that possess the "alive" status and have a diameter exceeding 5 cm, obtaining 34,287 woody stems. To retrieve the trees, we filtered the preprocessed stems dataset for unique rows based on the tree identifier (`tree.id`). This process resulted in a total of 31,153 individual trees, representing 37 different species that are mapped over the area of interest. Of these 31,153 trees, only 3140 have more than one stem.

The left panel of Figure 2 shows the absolute number of trees detected in each Prospect Hill Tract long-term plot cell. The most populated area is the one relative to the right-up corner of the Prospect Hill Tract long-term plot. In this area, it is also possible to note the higher species richness, that is, the absolute number of species present in each cell of the Prospect Hill Tract long-term plot (see right panel of Figure 2). Figure 3 shows that *Tsuga canadensis*, *Acer rubrum*, and *Betula alleghaniensis* are, among the other species, more present in this area. This information provides evidence of species evenness, that is, in a cell the community is perfectly even if every species is present in equal proportions and uneven if one species is dominant.

The swamp area records a few trees belonging to the same species, the *Acer rubrum* (acerru)—see Figures 2 and 3.

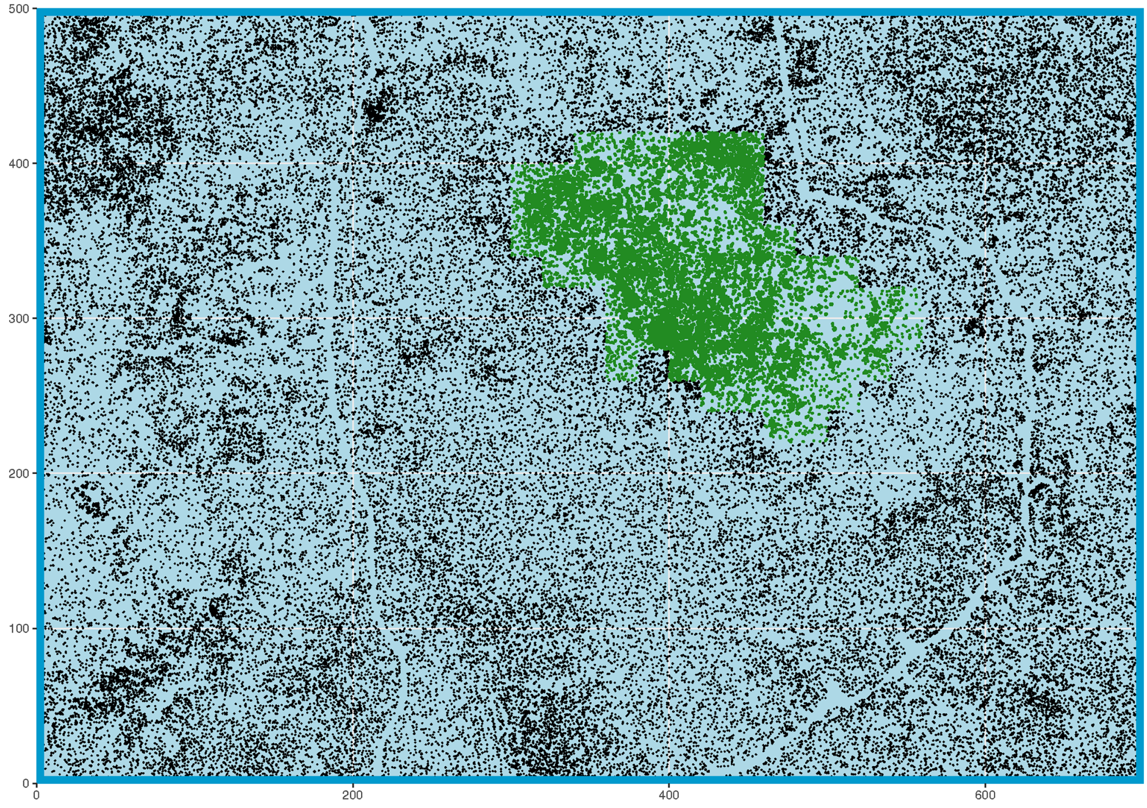

FIGURE 1 Distribution of the woody stems greater than 1 cm diameter at breast height collected within the Prospect Hill Tract long-term plot (500 m × 700 m). In black are the data collected during the second census (May 2018–January 2020); in green are the data collected during the first census (June 2010–March 2014) in the swamp area.

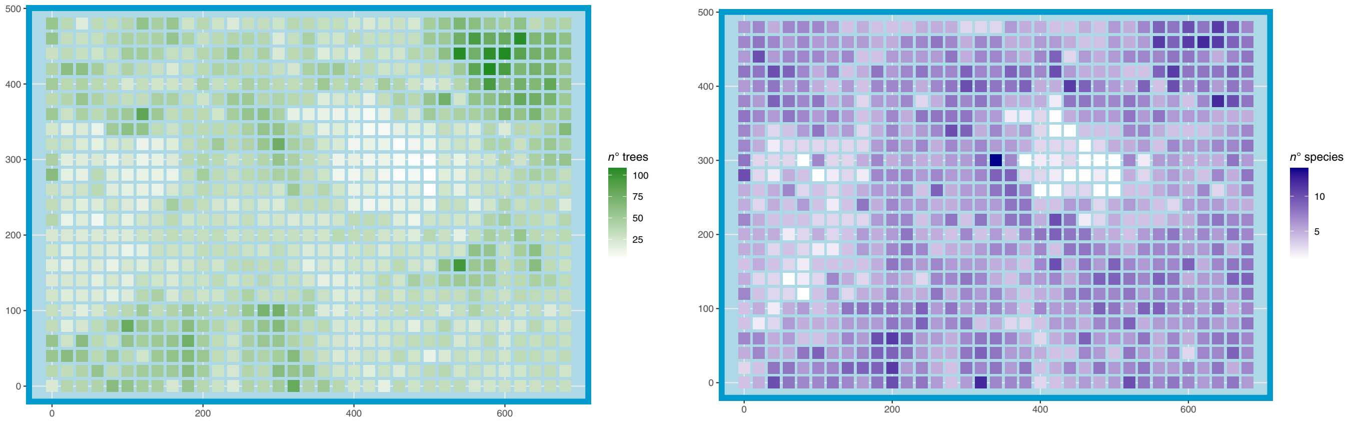

FIGURE 2 Absolute number of trees (left panel) and species (species richness, right panel) in each of the 875 cells of the Prospect Hill Tract long-term plot.

The descriptive analysis conducted on the Prospect Hill Tract long-term plot yields valuable insights into various aspects of biodiversity. It offers information on species richness, evenness, and the dominance of specific species, which are important indicators of biodiversity. However, it is important to note that no single measure can fully capture the complexity and entirety of biodiversity within this ecosystem. Biodiversity is a multifaceted concept that extends beyond solely considering the number and distribution of species. In the next section, we will thus delve into the challenge of measuring biodiversity and consider the use of biodiversity profiles as a method to address this complex issue.

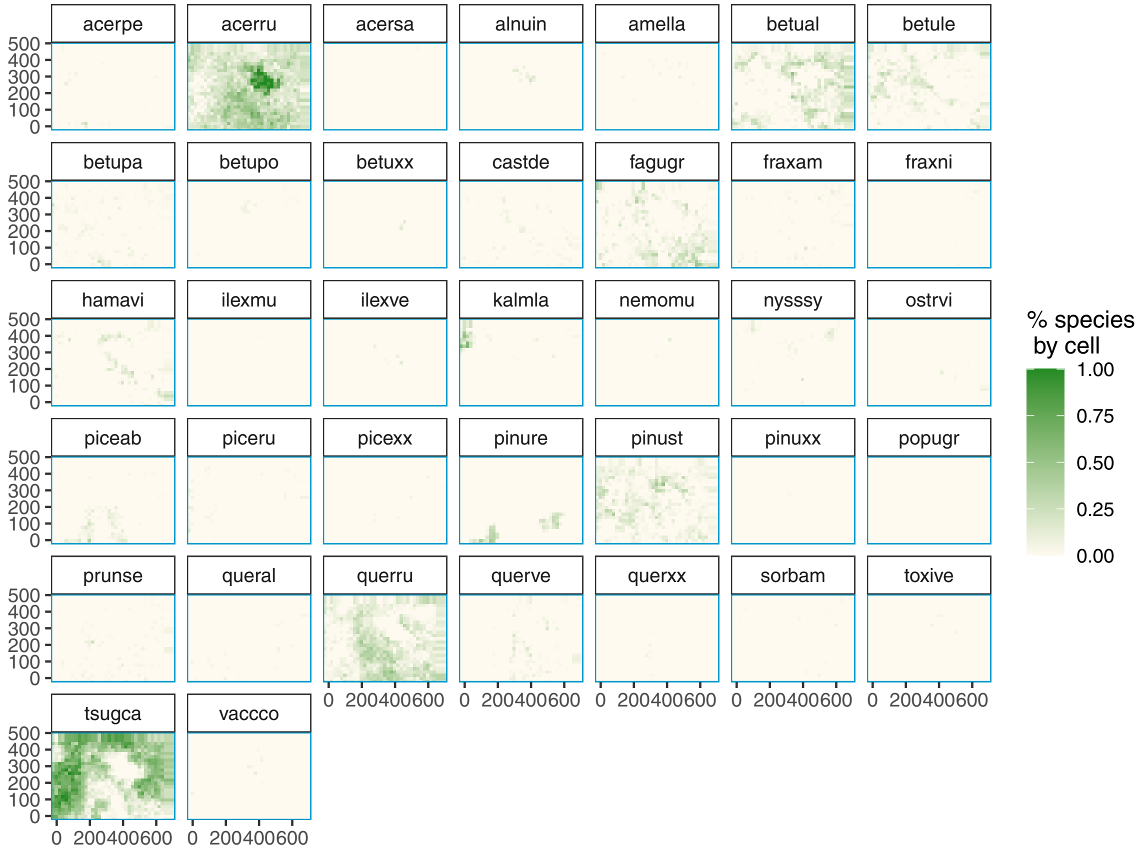

FIGURE 3 Spatial distribution of the relative abundance of species in each of the 875 cells of the Prospect Hill Tract long-term plot (species evenness).

3 | MEASURING BIODIVERSITY

In conservation ecology, information on the spatial distribution and composition of biological communities is essential in designing effective biodiversity conservation and management strategies. Site clustering and prioritization are crucial because resources for conservation are often limited, and it is essential to allocate them effectively to maximize conservation outcomes.

Biodiversity, primarily considered here in terms of the number and relative abundance of species in a community, that is, *taxonomic diversity*, can be measured in various ways depending on the study's specific objectives. Common measures of biodiversity include solely species richness or species evenness alone. However, biodiversity is a complex and multivariate concept, and attempting to measure it using a single index has its limitations. To encompass multiple biodiversity components simultaneously (e.g., species composition, abundance, and other ecological attributes), complexity and multivariate measures have been developed, such as the Shannon entropy (Shannon, 1948), and the Gini–Simpson diversity index (Gini, 1912; Simpson, 1949).

However, interpreting and comparing complex indices can be challenging due to variations in their measurement units and potential nonlinear formulations. But more importantly, these indices do not fulfil the *doubling propriety*, an essential requirement for the diversity measures. This propriety states that if two communities have equal diversity (measured using certain indices) and an equal number of individuals but do not share any species in common, then the diversity of the pooled community will be twice the diversity of either individual community.

To solve this problem, MacArthur (1965) proposed to convert the complexity measures to the *effective number of species*, that is the hypothetical number of equally abundant species that would produce the same value of a diversity measure as the observed community. By converting diversity measures into the effective number of species, researchers can quantify and compare diversity levels more accurately, accounting for differences in species richness and evenness. This approach helps to capture the underlying complexity of biodiversity and provides a more intuitive way to understand and interpret diversity values.

3.1 | Hill numbers and biodiversity profiles

The family of the Hill numbers is a family of diversity indices based on the concept of *effective number of species* that allows capturing both species richness and the evenness of species abundances within a community (cell). Hill numbers are expressed as a function of a parameter q, which determines the order of the Hill number (Hill, 1973). A direct way of forming a lattice is to specify the centroid of the ith cell $\boldsymbol{v}_i$ and to record its longitude and latitude coordinates (x_i, y_i). Assuming that each cell contains S_i, $i = 1, \dots, N$, species of trees, we denote with $\mathbf{p}_i = \mathbf{p}(\boldsymbol{v}_i) = \left(p_1(\boldsymbol{v}_i), \dots, p_s(\boldsymbol{v}_i), \dots, p_{S_i}(\boldsymbol{v}_i)\right)$ the cell-specific relative abundance vector of species, where $0 \leq p_s(\boldsymbol{v}_i) \leq 1$ and $\sum_{s=1}^{S_i} p_s(\boldsymbol{v}_i) = 1$. Then, the family of the Hill numbers is given by

$$H(q; \mathbf{p}_i) = \left(\sum_{s=1}^{S_i} p_s(\boldsymbol{v}_i)^q \right)^{1/(1-q)}, \quad \text{for} \quad q \in [0, +\infty) \setminus \{1\} \quad \text{and} \quad i = 1, \dots, N. \tag{1}$$

The order q of the Hill number determines the weight given to rare versus abundant species in the diversity evaluation. When $q = 0$, the Hill number represents the species richness. For $q = 1$ the Hill number is not defined, but the limit exists and gives the exponential of the Shannon entropy. When $q = 2$ the Hill number coincides with the inverse of the complement of the Gini–Simpson index. For all $q \geq 0$, Hill numbers satisfy the *doubly property* and have the same measurement unit as species richness.

To visualize the information captured by Hill numbers across different orders, a *biodiversity profile* can be created by plotting the Hill numbers on a single graph as a function of the parameter q. This profile shows how the Hill numbers change as the parameter q varies, providing a comprehensive view of biodiversity patterns and capturing the multivariate nature of biodiversity. In particular, the region of a biodiversity profile with small values of q provides insights into species richness and rare species since $H(q; \mathbf{p}_i)$ is influenced significantly by both common and rare species. Conversely, the tail of the biodiversity profile with large values of q sheds light on dominance and common species, as $H(q; \mathbf{p}_i)$ becomes less affected by rare species. The order parameter q represents, therefore, the *insensitivity* to rare species. As it grows, the perceived biodiversity $H(q; \mathbf{p}_i)$ drops.

4 | FUNCTIONAL DATA ANALISYS FOR HILL NUMBERS PROFILES

Let $\mathbf{p}_i = \mathbf{p}(\boldsymbol{v}_i) = \left(p_1(\boldsymbol{v}_i), \dots, p_s(\boldsymbol{v}_i), \dots, p_{S_i}(\boldsymbol{v}_i)\right)$, $i = 1, \dots, N$, denote the cell-specific relative abundance vector for S_i species and let $H(q; \mathbf{p}_i)$ be the corresponding biodiversity profile. These profiles can be perceived as samples of (spatially dependent) smooth curves which, in turn, can be viewed as realizations of an underlying biological process generating the abundance vectors $\mathbf{p}_i$. Following Gattone and Di Battista (2009), we examine biodiversity profiles, represented as functions of q and denoted by $H(q; \mathbf{p}_i)$, within the FDA framework. Despite the flexibility of the FDA approach, the inherent constraints of biodiversity profiles—specifically, their nonnegativity, monotonic decrease, and convexity—present modeling challenges. A conventional basis expansion of the curve might not adequately account for these specific characteristics, potentially introducing undesirable artifacts in the resulting functional data. To address this, we adopt the solution proposed by Ramsay (1998), as applied to biodiversity profiles by Gattone and Di Battista (2009) and Fortuna et al. (2020), and to Lorenz curves by Wu and Sickles (2018), which employs the following integral transformation:

$$H(q) = \int_0^q \exp\left(\int_0^x g(t) dt \right) dx, \tag{2}$$

where $g(\cdot)$ is a Lebesgue square integrable function. This parameterization provides a positive first derivative as $H'(q) = \exp\left(\int_0^x g(t) dt\right) > 0$ and so it guarantees the monotonicity restriction. The second derivative can be expressed as $H''(q) = H'(q)g(q)$ and this implies that $H''(q) \geq 0$ if $g(q) \geq 0$ for all q. Thus, the nonnegativity constraint on $g(\cdot)$ ensures convexity for $H(\cdot)$. Therefore, starting from Equation (2) one can derive the differential equation

$$\frac{H''(q)}{H'(q)} = g(q),$$

whose general solution is

$$H(q) = \xi_0 + \xi_1 \int_0^q \exp\left(\int_0^x g(t)dt \right) dx,$$

where ξ_0 and ξ_1 are arbitrary constants (see Thm. 1 in Ramsay, 1998). Following Wu and Sickles (2018), we parameterize $g(t)$ as a function of an unconstrained square-integrable function $\tilde{H}(\cdot)$, as $g(t) = g(\tilde{H}(t))$. This approach allows the constrained function H to be represented as a transformation of an unconstrained function $\tilde{H}$.

Considering each cell i, the function H can be written as

$$H(q; \mathbf{p}_i) = \xi_{0i} + \xi_{1i} \int_0^q \exp\left(\int_0^x g(\tilde{H}(t; \mathbf{p}_i))dt \right) dx, \qquad i = 1, \dots, N, \tag{3}$$

where ξ_{0i} and ξ_{1i} are arbitrary constants. The function g can be chosen as suggested by Wu and Sickles (2018), with a suitable choice represented by $g(q) = \frac{1}{2}q^2$. Being unconstrained, $\tilde{H}$ can be expanded as a linear combination of a finite set of basis functions $\phi_j(q), j = 1, \dots, J$, so that

$$\tilde{H}(q; \mathbf{p}_i) = \sum_{j=1}^{J} \alpha_{ji}\phi_j(q), \tag{4}$$

and each function $H(q; \mathbf{p}_i)$ can be represented by its vector of coefficients collected in the vector, $\boldsymbol{\beta}_i = (\xi_{0i}, \xi_{1i}, \alpha_{1i}, \dots, \alpha_{Ji})^T$, $i = 1, \dots, N$. For each profile, the parameters in $\boldsymbol{\beta}_i$ can be estimated in the framework of penalized regression (see Ramsay, 1998 or Wu & Sickles, 2018) and the fitted function takes the form

$$\hat{H}(q; \mathbf{p}_i) = \hat{\xi}_{0i} + \hat{\xi}_{1i} \int_0^q \exp\left(\int_0^x g\left(\sum_{j=1}^{J} \hat{\alpha}_{ji}\phi_j(t) \right) dt \right) dx, \qquad i = 1, \dots, N. \tag{5}$$

For each of the 875 cells in the Prospect Hill Tract long-term plot, Figure 4 displays the fitted curves for Hill number profiles with $q \in [0, 5] \setminus \{1\}$ and $J = 15$. This choice of $J = 15$ basis functions, based on the authors' experience, ensures a satisfactory fit of the profiles. As required, all the fitted curves are monotone decreasing, and they start from the maximum at $q = 0$, which coincides with the species richness. The 875 curves also intersect at various points, exhibiting different slopes and curvatures. To highlight potential similarities in the overall shape of the profiles and address their spatial distribution, we propose a suitable clustering procedure in the following sections.

4.1 | Assessing spatial dependence for functional data

Standard statistical techniques for modeling functional data primarily focus on independent functions. However, assuming independence appears unreasonable when observing samples of functions across different contiguous cells. Accordingly, when clustering biodiversity profiles in space, it is crucial to assess spatial dependence to understand the underlying spatial patterns and ensure the validity of the clustering results.

Analyzing the spatial variability of biodiversity profiles can be done using a trace-variogram for functions (Giraldo et al., 2011) defined as

$$2\gamma(\mathbf{h}) = E\left[\int_0^Q \left(H\left(q; \mathbf{p}_i(\boldsymbol{\nu}_i)\right) - H\left(q; \mathbf{p}_i(\boldsymbol{\nu}_i + \mathbf{h})\right) \right)^2 dq \right], \tag{6}$$

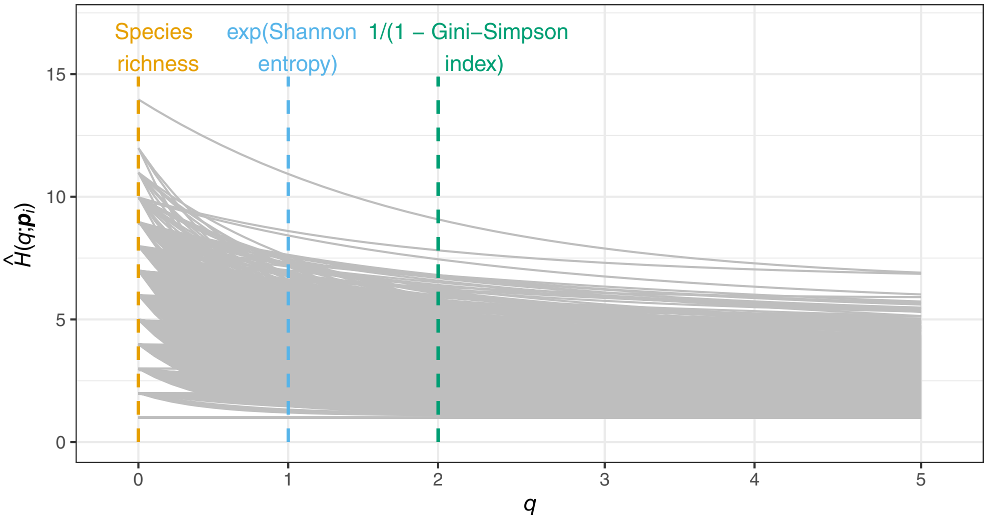

FIGURE 4 Fitted curves, one per each cell in the Prospect Hill Tract long-term plot.

over a vector distance **h**. Although usually defined for spatial processes whose index varies continuously over space, a variogram as a measure of spatial dependence can also be useful when working with lattice data (see Cressie, 1993, p. 401, and Wall, 2004).

An important assumption underlying the use of the L_2 distance in the trace-variogram in Equation (6) is that the length of the domain of the functions is fixed. Specifically, the latter assumption assumes perfect alignment of the functions, which is not a concern within the framework of biodiversity profiles.

Under the stationarity hypothesis, it is common practice to estimate the trace-variogram in Equation (6) by a mean value of samples grouped over an isotropic distance h:

$$2\hat{\gamma}(h) = \frac{1}{n(h)} \sum_{||\boldsymbol{v}_i - \boldsymbol{v}_r|| = h} \int_0^Q \left(\hat{H}(q;\mathbf{p}_i) - \hat{H}(q;\mathbf{p}_r)\right)^2 dq, \tag{7}$$

where $n(h)$ is the number of pairs $(\mathbf{p}(\boldsymbol{v}_i), \mathbf{p}(\boldsymbol{v}_r))$ at spatial distance h and $\hat{H}(\cdot)$ are as defined in Equation (5).

Figure 5 shows the (omni-directional) empirical trace-variogram as a function of separation distance h, for $Q = 5$. Each point on this plot thus represents an average over a number of pairs of estimated biodiversity profiles that are the same distance apart. The full curve in Figure 5 is the empirical LOESS fit to the estimated variogram. The variogram plot indicates a distinct finite range and sill, suggesting that nearby biodiversity profiles are more correlated and exhibit similar values. This observation underscores the significant role of the variogram in informing the definition of clusters.

5 | MODEL-BASED CLUSTERING FOR SPATIAL FUNCTIONAL DATA

By using the vector of coefficients $\boldsymbol{\beta} = (\xi_0, \xi_1, \alpha_1, \dots, \alpha_J)^T$, introduced in Equations (3) and (4), as representative data for a biodiversity profile, we propose a finite GMM with a L_1 penalized likelihood for functional clustering, named *Penalized model-based Functional Clustering* (PFC-L_1) in Pronello et al. (2023). If a latent variable $Z_i = \{Z_{i1}, \dots, Z_{iK}\}$ denotes the cluster membership of the ith curve to the kth group, the marginal density of $\boldsymbol{\beta}$ is a weighted combination of K (number of groups) Gaussian densities f_k with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, that is

$$f(\boldsymbol{\beta}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{v};\boldsymbol{\omega}) f_k(\boldsymbol{\beta};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k),$$

where $\pi_k(\boldsymbol{v};\boldsymbol{\omega})$ are spatially varying mixing proportions (changing with the spatial coordinate (x, y) of the cell $\boldsymbol{v}$ and such that $\sum_{k=1}^{K} \pi_k(\boldsymbol{v};\boldsymbol{\omega}) = 1$) depending on some parameters $\boldsymbol{\omega}$ that, a priori, give the probabilities of belonging to a group, that

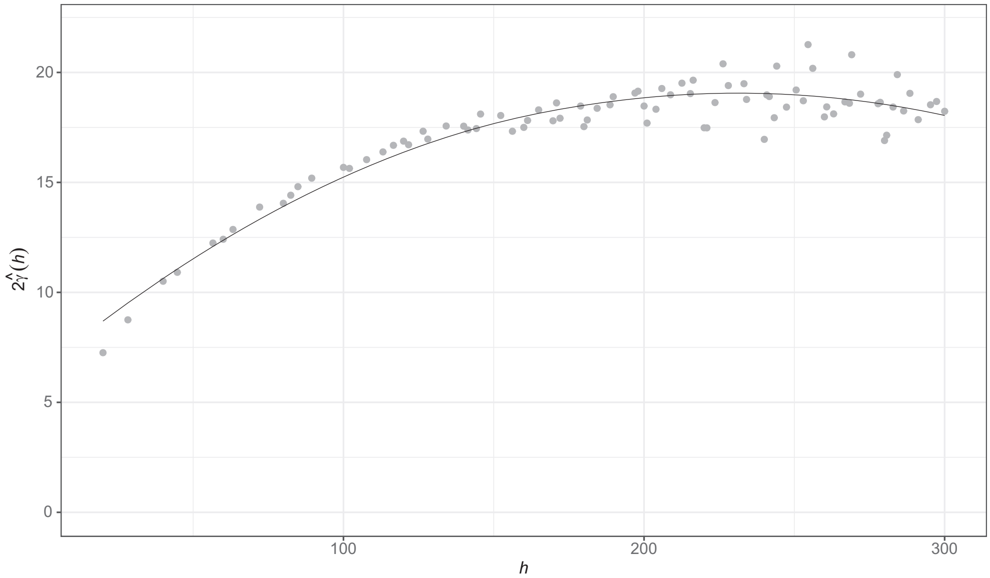

FIGURE 5 Empirical trace-variogram (in grey dots) and LOESS fit (in black solid line) for smoothed Hill number profiles.

is, $\pi_k(\boldsymbol{\nu};\boldsymbol{\omega}) = \mathbb{P}(Z_k(\boldsymbol{\nu}) = 1)$, $k = 1, \dots, K$, and $\pi_k(\boldsymbol{\nu};\boldsymbol{\omega}) > 0$ for each k. Then we can write the log-likelihood function as

$$l(\boldsymbol{\theta};\boldsymbol{\beta}) = \sum_{i=1}^{N} log\left[\sum_{k=1}^{K} \pi_k(\boldsymbol{\nu};\boldsymbol{\omega}) f_k(\boldsymbol{\beta}_i;\mu_k,\Sigma_k)\right],$$

where $\boldsymbol{\theta}$ is the set of all model parameters to be estimated, while $\boldsymbol{\beta}_i = (\xi_{0i}, \xi_{1i}, \alpha_{1i}, \dots, \alpha_{Ji})^T$ is the vector of coefficients of the ith biodiversity profile.

5.1 | Spatial modeling of mixing proportions

Spatially varying mixing proportions are introduced in the GMM model to take into account the spatial dependence among biodiversity profiles. We thus assume that observations corresponding to nearby locations are more likely to have similar allocation probabilities than observations that are far apart in space.

Considering the Kth group as a baseline, let

$$\zeta_k(\boldsymbol{\nu};\boldsymbol{\omega}) = \log(\pi_k(\boldsymbol{\nu};\boldsymbol{\omega})/\pi_K(\boldsymbol{\nu};\boldsymbol{\omega})), \qquad k = 1, \dots, K-1, \tag{8}$$

denote the log-odds spatial process. Also, let $\boldsymbol{V}$ be a valid $(N \times N)$ *generalized* variogram matrix (Chilès & Delfiner, 2012) and $\boldsymbol{U}$ a $(N \times 3)$ design matrix whose rows are defined as $\boldsymbol{u}_i = (1, x_i, y_i)^T$, where (x_i, y_i) are the spatial coordinates of the cell $\boldsymbol{\nu}_i$. Then, if we define the so-called *Bending Energy* matrix (Mardia et al., 1998) as

$$\mathbf{B} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{U}\left(\mathbf{U}'\mathbf{V}^{-1}\mathbf{U}\right)^{-1}\mathbf{U}'\mathbf{V}^{-1},$$

it can be shown—as a result of the Karhunen–Loéve (KL) theorem (Adler, 2010)—that the log-odds spatial process $\zeta_k(\boldsymbol{\nu};\boldsymbol{\omega})$ can be rewritten as a linear model through the following truncated KL expansion

$$\zeta_k(\boldsymbol{\nu_i};\boldsymbol{\omega}) = \sum_{l=1}^{L} \omega_{l,k}\ \psi_l(\boldsymbol{\nu}_i), \quad i = 1, \dots, N, \tag{9}$$

where $\omega_{l,k}$ are the elements of the vector $\boldsymbol{\omega}$ to be estimated, and the $\psi_l(\boldsymbol{\nu}_i)$ are basis functions defined as the eigenvectors obtained by the spectral decomposition $\mathbf{B} = \boldsymbol{\Psi}\mathbf{G}\boldsymbol{\Psi}'$, with $\mathbf{G} = \text{diag}(g_1, \dots, g_N)$ being the diagonal matrix of eigenvalues.

Since it can be shown that $\mathbf{BU} = \mathbf{0}$, it follows that the first three eigenvalues of $\mathbf{B}$ are equal to zero and the corresponding eigenvectors are given by the columns of $\mathbf{U}$.

In practice, the modeling of the log-odds spatial process is facilitated by the truncated KL expansion based on the property that, given any orthonormal basis functions, we can find some integer L so that $\zeta_k(\boldsymbol{\nu}; \boldsymbol{\omega})$ can be approximated by the finite weighted sum of basis functions. It can be shown (Mardia et al., 1996) that, when the variogram matrix is parametrized as follows

$$V(h_{i,r}) = \frac{1}{8\pi} h_{i,r}^2 \log(h_{i,r}),$$

where $h_{i,r} = ||\boldsymbol{\nu}_i - \boldsymbol{\nu}_r||_2$ and the basis functions $\psi_l(\boldsymbol{\nu}_i)$ are obtained through the spectral decomposition of $\mathbf{B}$ above, the spatial process $\zeta_k(\boldsymbol{\nu}; \boldsymbol{\omega})$ is modeled through a *Thin-plate spline.*

5.2 | Penalized likelihood

Allowing for different cluster means and covariance matrices the specified model can be over-parametrized, and to keep flexibility we avoid introducing any kind of constraints by, instead, considering two penalties that regularize parameter estimation in the log-likelihood function, as in Zhou et al. (2009). Thus, given the profile coefficients $\boldsymbol{\beta}_i$ with length $p = J + 2$, and conditional on the number of groups K, the penalized log-likelihood function can be written as

$$l_P(\boldsymbol{\theta}; \boldsymbol{\beta}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k(\boldsymbol{\nu}; \boldsymbol{\omega}) f_k(\boldsymbol{\beta}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right] - \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p} |\mu_{k,j}| - \lambda_2 \sum_{k=1}^{K} \sum_{j,q}^{p} |W_{k;j,q}|, \tag{10}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters to be suitably chosen (see Section 5.3), $\mu_{k,j}$ are cluster mean elements and $W_{k;j,q}$ are entries of the inverse of the cluster-specific covariance matrix $\mathbf{W}_k = \boldsymbol{\Sigma}_k^{-1}$. The name *Penalized model-based Functional Clustering* (PFC-L_1) in Pronello et al. (2023) is chosen because the penalty terms contain sums of absolute values, and so they are of L_1 (or LASSO) type. Indeed, the first penalty term facilitates the selection of basis functions appearing in the expansion of $\tilde{H}$ by keeping only the terms useful in separating groups. The second penalty term helps to shrink the elements $W_{k;j,q}$ and allows estimating—thanks to sparsity—large covariance matrices and avoiding possible singularity problems.

The model parameter estimation cannot be obtained by direct optimization of the log-likelihood function given in Equation (10) but, since Z is not observed, can be efficiently carried out using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The analytical solutions to update the cluster membership probabilities, the cluster mean elements and the cluster-specific precision matrices are detailed in Pronello et al. (2023). In particular, at each iteration the Graphical LASSO algorithm (Friedman et al., 2008) is used to obtain sparse cluster-specific precision matrices, whereas to estimate the spatially varying mixing proportions $\pi_k(\boldsymbol{\nu}; \boldsymbol{\omega})$ the multinomial logit model as specified in Section 5.1 needs to be fitted. Thus, the estimation of the parameters of the linear model in Equation (9) can be obtained at the $(d+1)$th iteration of the EM algorithm as the solution of the log-likelihood maximization of a weighted multinomial logit model, that is

$$\hat{\boldsymbol{\omega}}^{(d+1)} = \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{\tau}_k^{(d)}(\boldsymbol{\nu}_i) \log(\pi_k(\boldsymbol{\nu}_i; \boldsymbol{\omega})),$$

where $\hat{\tau}_k^{(d)}(\boldsymbol{\nu}_i)$ are the estimated posterior probabilities that a biodiversity profile i, summarized here by $\hat{\boldsymbol{\beta}}_i$, belongs to the kth group, and are computed through the iterations of the EM algorithm as

$$\hat{\tau}_k^{(d)}(\boldsymbol{\nu}_i) = \frac{\hat{\pi}_k^{(d)}(\boldsymbol{\nu}_i; \boldsymbol{\omega}) f_k(\hat{\boldsymbol{\beta}}_i; \hat{\boldsymbol{\mu}}_k^{(d)}, \hat{\boldsymbol{\Sigma}}_k^{(d)})}{\sum_{k=1}^{K} \hat{\pi}_k^{(d)}(\boldsymbol{\nu}_i; \boldsymbol{\omega}) f_k(\hat{\boldsymbol{\beta}}_i; \hat{\boldsymbol{\mu}}_k^{(d)}, \hat{\boldsymbol{\Sigma}}_k^{(d)})}. \tag{11}$$

5.3 | Model selection

One of the most difficult steps in clustering is to determine the optimal number of clusters, K, to group the data, and we know there is no "right" answer. In this paper, we perform a grid-search for model hyper-parameters and choose the triplet $\{K; \lambda_1; \lambda_2\}$ that allows for model selection based on information criteria. In particular, we consider likelihood-based measures of model fit that include a penalty for model complexity such as the Bayesian Information Criterion (BIC)

$$\mathrm{BIC}(K, \lambda_1, \lambda_2) = l(\hat{\boldsymbol{\theta}}_K; \hat{\boldsymbol{\beta}} | K, \lambda_1, \lambda_2) - \frac{\mathscr{C}}{2} \log(N),$$

and the Integrated Classification Likelihood (ICL) index (Baudry, 2015)

$$\mathrm{ICL}(K, \lambda_1, \lambda_2) = \mathrm{BIC}(K, \lambda_1, \lambda_2) + \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{\tau}_k(\boldsymbol{v}_i) \log \hat{\tau}_k(\boldsymbol{v}_i),$$

where $l(\hat{\boldsymbol{\theta}}_K; \hat{\boldsymbol{\beta}} | K, \lambda_1, \lambda_2)$ is the value of the maximized log-likelihood objective function with parameters $\hat{\boldsymbol{\theta}}_K$ estimated under the assumption of a model with K components, $\hat{\boldsymbol{\beta}}$ collects all $\hat{\boldsymbol{\beta}}_i$ and $\mathscr{C}$ measures the complexity of the model. While BIC has a penalty term only related to the number of observations N and the complexity measure $\mathscr{C}$, ICL also includes an additional term—that is the estimated mean entropy—to penalize clustering configurations with overlapping groups (this facilitates solutions with well-separated groups, i.e., with low entropy).

To use the above criteria it is necessary to clarify what is $\mathscr{C}$ in a penalized model. In our case, we consider

$$\mathscr{C} = \sum_{k=1}^{K} \sum_{j=1}^{p} I\left(\hat{\mu}_{k,j} \neq 0\right) + \sum_{k=1}^{K} \sum_{i \leq j} I\left(\widehat{\Sigma}_{k;j,q} \neq 0\right) + L(K-1),$$

where $I(\cdot)$ is the indicator function that applies to the (sparse) likelihood estimate of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, so that $\mathscr{C}$ is the number of nonzero entries in both the means and the upper half of the covariance matrices, plus the number of parameters for the spatial mixing proportions. In general, the model with the highest values of BIC or ICL could be selected as the desired model.

6 | SIMULATION STUDY

In this section, we conduct a simulation study to assess the performance of our proposed clustering procedure for biodiversity profiles. Additionally, we systematically compare our algorithm with the functional k-means algorithm (Abraham et al., 2003), a widely adopted method for clustering functional data. The comparison is carried out across three scenarios, each reflecting diverse spatial arrangements of the clusters over the spatial domain. The experimental setup enables us to evaluate the impact of integrating spatial priors into our mixed model framework and determine how effectively these priors enhance the performance of our clustering procedure compared to the functional k-means method (the latter also considered in a version with standardized coordinates of cells as a proxy of spatial information).

Specifically, we consider a (20×20) regular grid, where the cells are assigned to three spatial clusters, denoted as C_1, C_2, C_3. These clusters, with sizes $N_1 = 125$, $N_2 = 125$, and $N_3 = 150$, respectively, are illustrated in Figure 6. In the first scenario, cluster labels are randomly assigned to cells, resulting in biodiversity profiles that are spatially independent across the entire grid. In contrast, the second scenario introduces spatial structure, where each cluster of biodiversity profiles is confined to a specific region within the domain. Lastly, the third scenario maintains spatial structure but allows clusters C_1 and C_2 to be split across the grid.

To generate the biodiversity profiles, $H(q; \mathbf{p}_i)$, assigned to cells $\boldsymbol{v}_i$, $i = 1, \ldots, N$, we first randomly draw the number of species, S_i, by discretizing the realizations of a Normal distribution. Specifically, $S_i | (Z_i = k) = \lfloor T_k \rceil$ with $T_k \sim \mathcal{N}(\mu_k, \sigma_k)$, where Z_i is the cluster membership latent variable and the operator $\lfloor \cdot \rceil$ indicates rounding to the nearest integer. Then, we model the relative abundance vector, $\mathbf{p}_i$, for S_i species, using a Dirichlet distribution with parameter $\boldsymbol{\eta}_i$, that is, $\boldsymbol{p}_i \sim \mathrm{Dir}_{S_i}(\boldsymbol{\eta}_i)$. To induce variations in the curvature of the biodiversity profiles, we express $\boldsymbol{\eta}_i | (Z_i = k)$ as $S_i \times \boldsymbol{\delta}_k$, where $\boldsymbol{\delta}_k$ represents an S_i-dimensional vector of weights. These weights enable us to achieve different degrees of curvature in the

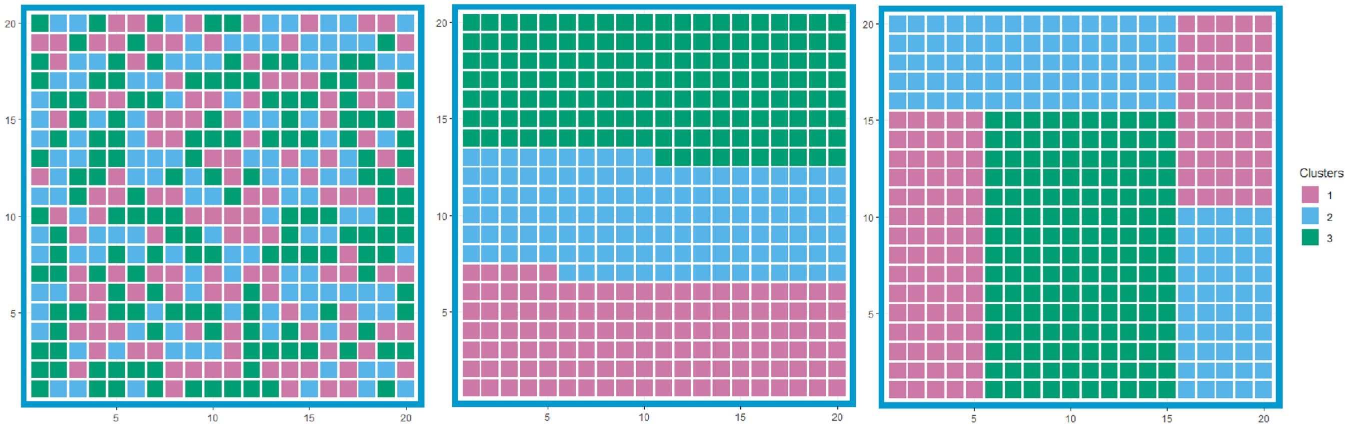

FIGURE 6 Spatial distributions of clusters across the grid: (left) Scenario 1, (middle) Scenario 2, (right) Scenario 3.

TABLE 1 Parametrization of the simulation setup: each row provides details about the generating process of the number of species within the kth cluster.

		$\boldsymbol{\delta_k}$	
$\boldsymbol{k}$	$\boldsymbol{T_k \sim \mathcal{N}(\mu_k, \sigma_k)}$	**Case 1**	**Case 2**
1	$T_1 \sim \mathcal{N}(6, 0.5)$	$\delta_1 = [2 \ \dots \ 2]$	$\delta_1 = [2 \ \dots \ 2]$
2	$T_2 \sim \mathcal{N}(12, 1.5)$	$\delta_2 = \left[2\ 2\ 2\ 2\ 2\ \frac{1}{5} \ \dots \ \frac{1}{5}\right]$	$\delta_2 = \left[2\ 2\ 2\ 2\ 2\ \frac{1}{4} \ \dots \ \frac{1}{4}\right]$
3	$T_3 \sim \mathcal{N}(10, 1.5)$	$\delta_3 = \left[2\ 2\ 2\ \frac{1}{2} \ \dots \ \frac{1}{2}\right]$	$\delta_3 = \left[2\ 2\ 2\ \frac{1}{2} \ \dots \ \frac{1}{2}\right]$

biodiversity profiles within each cluster C_k. The parameterization used for the simulation design, common to all three scenarios, is detailed in Table 1.

This parameterization enables the generation of diverse vectors of relative species abundance, thereby capturing various biodiversity profiles that represent grid cells where species are either evenly distributed (Case 1) or rare (Case 2). For instance, drawing from a $\text{Dir}_{S_i}(\boldsymbol{\eta}_i|(Z_i = 1)$, suggests that cluster C_1 represents an almost perfectly even and heterogeneous cell-community type. Conversely, drawing from $\text{Dir}_{S_i}(\boldsymbol{\eta}_i|(Z_i = 2))$ and $\text{Dir}_{S_i}(\boldsymbol{\eta}_i|(Z_i = 3))$, with $\boldsymbol{\eta}_i|(Z_i = 2)$ and $\boldsymbol{\eta}_i|(Z_i = 3)$ inducing skewness in the Dirichlet distribution, results in relative abundance vectors, $\boldsymbol{p}_i$, with increasingly dissimilar elements. This variation reflects the presence of relatively rarer species, which is particularly pronounced in the third cluster.

As an illustrative example, Figure 7 presents two sets of $N = 400$ biodiversity profiles, $H(q; \mathbf{p}_i)$, generated using the Dirichlet distribution parameterization detailed in Table 1. The two panels correspond to the parameterizations outlined in Case 1 (left panel) and Case 2 (right panel). The primary distinction between these cases lies in the generation of biodiversity profiles for Cluster 2 which, in Case 1, shows a steeper decline compared to those in Case 2, indicating a higher prevalence of rare species.

For this simulation study, we generated 100 samples, each comprising $N = 400$ biodiversity profiles. We then applied the proposed PFC-L_1 clustering procedure to each sample, selecting λ_1 and λ_2 through a grid-search over the set $\Lambda_1 \times \Lambda_2$, where $\Lambda_i = \{10, 30, 50, 70, 90\}, \ i = 1, 2$. For comparison purposes, both the functional k-means algorithm and the PFC-L_1 procedure are evaluated under the assumption of a known and fixed number of clusters, set at $K = 3$. Moreover, to evaluate the ability of both algorithms to recover accurately the cluster labels, Z_i, of the data, we utilized the Adjusted Rand Index (ARI).

The subplots in Figure 8 display the boxplots of the 100 ARI values obtained for Case 1 (upper panels) and Case 2 (lower panels) across all three scenarios. It is important to note the distinction between the first and second boxplots within each subplot: while both are related to the functional k-means algorithm, the second boxplot represents the distribution of ARI values when the spatial coordinates of the cells are provided as additional features. This inclusion allows the functional k-means algorithm to naively incorporate spatial information into the clustering process, potentially enhancing its ability to capture spatial structures within the data. As evident from the results, in Scenarios 2 and 3 where spatial structures

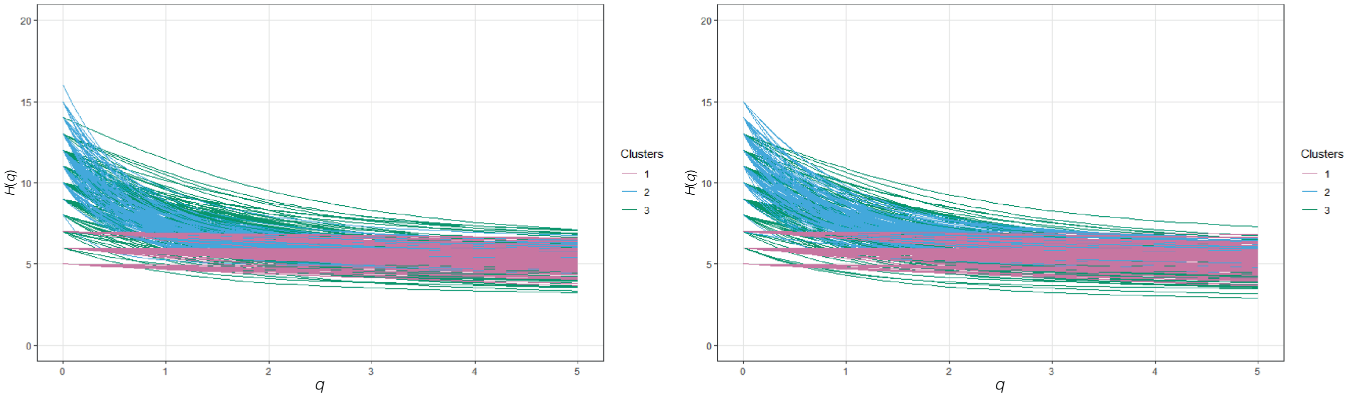

FIGURE 7 Samples of biodiversity profiles generated in Case 1 (left) and Case 2 (right).

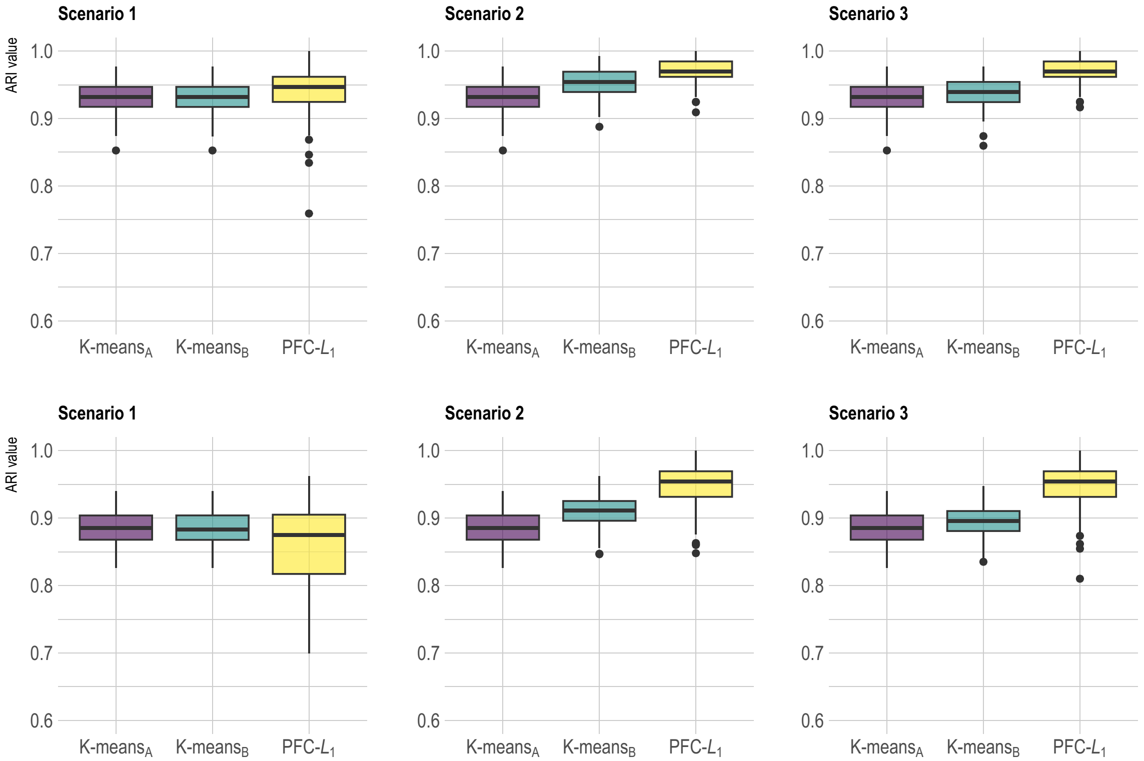

FIGURE 8 Boxplots of Adjusted Rand Index (ARI) values for three clustering methods across 100 simulations for Case 1 (upper) and Case 2 (lower).

are present, the proposed PFC-L_1 clustering procedure consistently outperforms both versions of the functional k-means algorithm. However, in Scenario 1 for Case 2, where the spatial structure is very weak or absent, the PFC-L_1 exhibits more variable ARI values. This scenario appears more challenging due to the overlapping of curves with similar curvature, making it more difficult to identify distinct partitions correctly.

Figure 9 also illustrates the PFC-L_1 algorithm capability in recovering cluster spatial distributions for Case 1 (left panel) and Case 2 (right panel). The algorithm seems to perform better in Case 1, offering more accurate maps of the estimated prior probabilities, $\hat{\pi}_k(\boldsymbol{v}; \boldsymbol{\omega})$, for each cluster. This setup thus allows us to assess the effectiveness of integrating spatial priors into our mixed model framework and to gauge their impact on enhancing clustering performance compared to functional k-means.

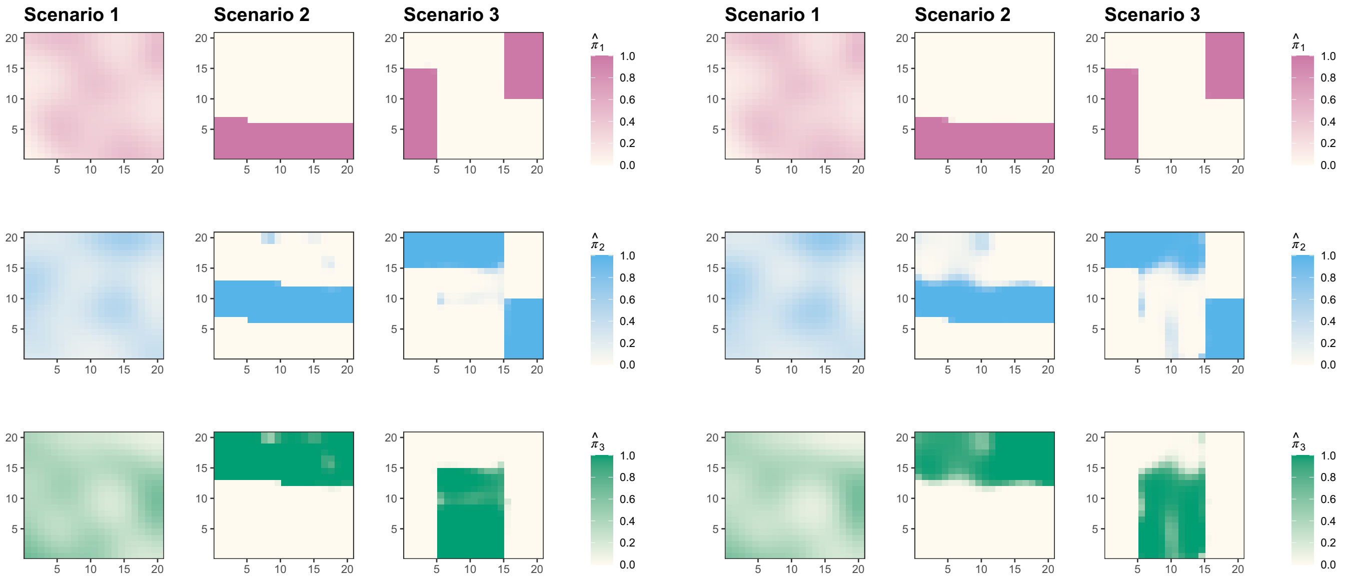

FIGURE 9 Maps of the estimated prior probabilities $\hat{\pi}_k(\boldsymbol{v};\boldsymbol{\omega})$ for each cluster of the grid for all scenarios in Case 1 (left) and Case 2 (right).

7 | APPLICATION TO THE HARVARD FOREST DATA

In this section, we extend the statistical analysis of the dataset discussed in Section 2 and present the results obtained from clustering the biodiversity profiles using the PFC-L_1 procedure. To take care of the spatial dependence among the profiles, we have considered a Thin-plate spline parameterization (see Section 5.1) with $L = 16 << N$ basis functions explaining about 91.50% of the spatial variability. The spatial patterns of the basis functions are shown in Figure 10 and, as expected, they show a decreasing order of smoothness. For example, the first basis function $\psi_1(\boldsymbol{v})$ is constant over all the domain of interest while $\psi_2(\boldsymbol{v})$ and $\psi_3(\boldsymbol{v})$ are linear trends of the longitude and latitude coordinates, respectively. More in general, higher-order functions correspond to larger-scale features while lower-order functions correspond to smaller-scale details.

By fixing $J = 15$ in Equation (5) and considering a discrete grid of values for the triplet $(K, \lambda_1, \lambda_2)$, the BIC and ICL criteria suggest that a GMM model with three spatial clusters should be considered (see Figure 11). BIC and ICL values closely align since the posterior probability estimates result in distinct partitions, where the clusters are well-separated with estimated mean entropy approaching zero. However, we are not aware of the original distribution which generated the data so, to validate the performance evaluation of the clustering process we also consider interpretation as an important part of model selection, especially from a knowledge discovery perspective. Interpretation can help us gain insights and guiding decisions based on our clustering procedure and for this, in the following, we favor the solution with $K = 4$ as it better highlights the group of cells with constant biodiversity profiles (see below) and for which the values of BIC and ICL are the "second best."

Figure 12 provides a spatial representation of the four clusters. In particular, the upper left panel illustrates the functional zoning of the Prospect Hill Tract long-term plot derived from these clusters, the upper right panel displays the behaviour of the estimated mean biodiversity profiles and the bottom panel exhibits the allocation of the individual biodiversity profiles $\hat{H}(q;\boldsymbol{p}_i)$ in each cluster. Due to the intersection of the estimated mean biodiversity profiles, direct comparisons among the four clusters are not feasible, as the profiles only offer a partial ordering of their biodiversities. Although this limitation cannot be entirely overcome, biodiversity profiles remain significantly more meaningful than univariate indices. In fact, even in cases where two communities (cells) are not directly comparable, examining where their biodiversity profiles intersect can reveal changes or variations in the composition of species.

Clusters 1 and 3 emerge as the most populated clusters, with 326 and 264 cells, respectively, whereas Cluster 2 includes 196 cells and, finally, Cluster 4 only contains 89 cells. All clusters display similar average species richness (when $q = 0$) despite different levels of variability and slope, as shown in the bottom panel of Figure 12. In particular, Cluster 4 exhibits the lowest average species richness among the clusters. Remarkably, the clusters exhibit diverse species compositions,

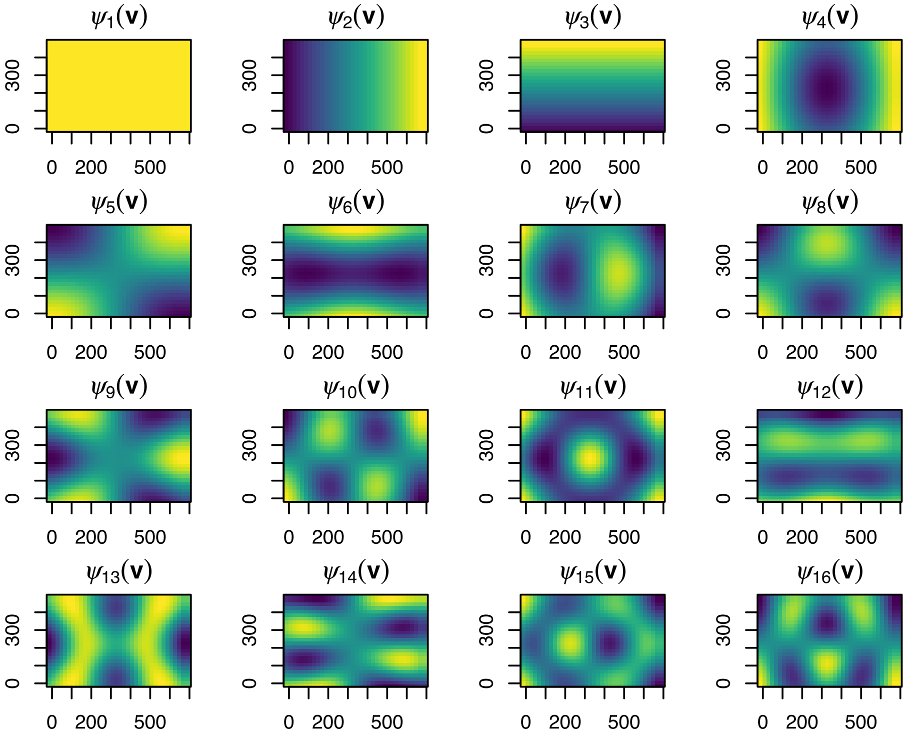

FIGURE 10 Spatial maps of the first 16 basis function ψ_l, $l = 1, \ldots, L$, obtained by the spectral decomposition of the *Bending Energy* matrix and used to model the spatial variability of the log-odds as in Equation (8).

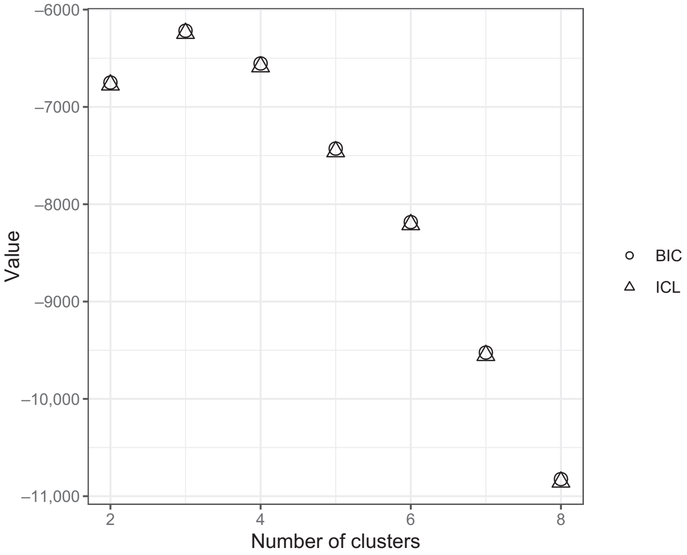

FIGURE 11 Bayesian Information Criterion (BIC) and Integrated Classification Likelihood (ICL) values for model selection. The plot maps the maximum BIC and ICL values achieved for the triplet $(K, \lambda_1, \lambda_2)$ according to the number of clusters K.

implying that they achieve similar average species richness by having unique sets of species in each cluster. For example, Cluster 1 includes solely one *Acer saccharum* tree, while this particular species is entirely absent in Cluster 3 as illustrated in Figure 13.

Although all clusters have similar average species richness, they show different values for average species abundance (when $q = 1$) and average species dominance (when $q = 2$). For example, compared with Clusters 1 and 2, Cluster 4 displays higher average species abundance and dominance resulting from estimated mean biodiversity profile intersections.

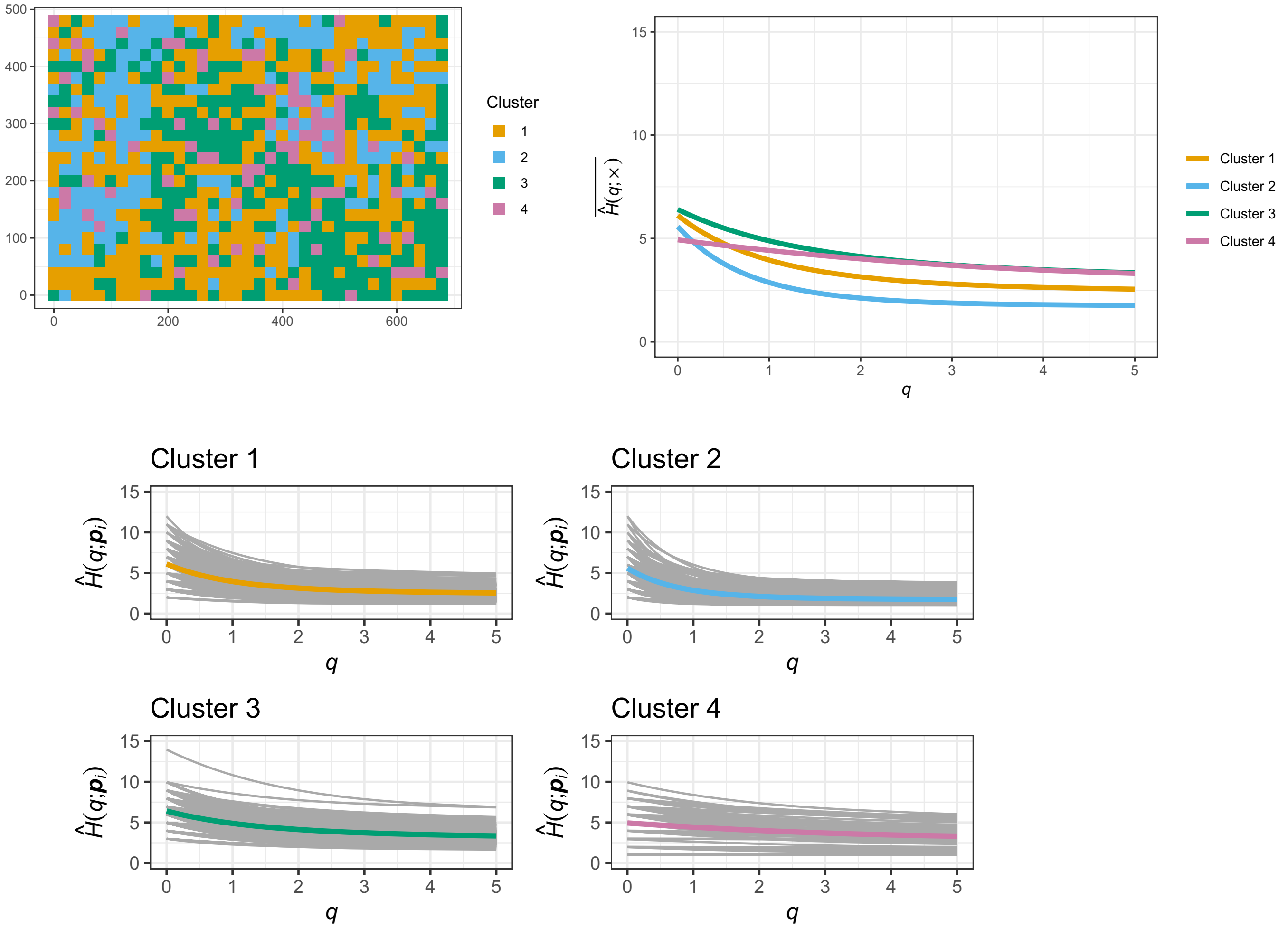

FIGURE 12 Upper left: Functional zoning results of the Prospect Hill Tract long-term plot with four clusters (each cell is assigned a specific colour based on its associated clustering label). Upper right: estimated mean biodiversity profiles $\overline{\hat{H}(q;\cdot)}$ in each cluster. Bottom: individual biodiversity profiles $\hat{H}(q;\boldsymbol{p}_i)$ in each cluster with superimposed estimated mean biodiversity profiles (thicker lines).

These findings emphasize the nuanced differences in species distribution and dominance within the identified clusters. The upper right plot of Figure 12 further confirms that for $0 \leq q \leq 2$, the biodiversity profiles are sufficient to characterize the *taxonomy diversity* in the Prospect Hill Tract long-term plot.

In general, the main contributing factor in differentiating between the clusters appears to be associated with the derivatives of the estimated Hill profiles. These derivative functions convey significant information and are consistent with the functional representation used in Equation (5). Clusters 1 and 2 are characterized by curves with steeper slopes, while Cluster 4 stands out with profiles that remain relatively constant regardless of the intercept level. This behaviour holds particular significance when interpreting the clustering results since, as demonstrated in the example from Section 3, a constant profile indicates a uniform distribution of species within the cell, while a more convex profile suggests an uneven distribution.

Figure 14 displays the spatial distribution of the estimated prior probabilities $\hat{\pi}_k(\boldsymbol{\nu};\boldsymbol{\omega})$ for each cluster. As it can be noticed, the distribution of the clusters clearly shows how the estimated posterior probabilities, $\hat{\tau}_k(\boldsymbol{\nu}_i)$, reflect the information about the spatial distribution of the weights of the mixture (see upper left panel Figure 12). As illustrated in Section 5, we note that clusters arise from a careful balance between geographical proximity and similarity among curves (biodiversity profiles). The values represented by $\hat{\pi}_k(\boldsymbol{\nu};\boldsymbol{\omega})$ provide valuable information about the spatial variability of the clusters. Consequently, the outcomes shown in Figure 14 serve as spatial predictions of the clustering labels, focusing solely on spatial information. These predictions enable us to divide the study area into distinct zones that highlight the prevalence of specific clusters, offering policymakers insightful guidance for crafting effective interventions. For instance,

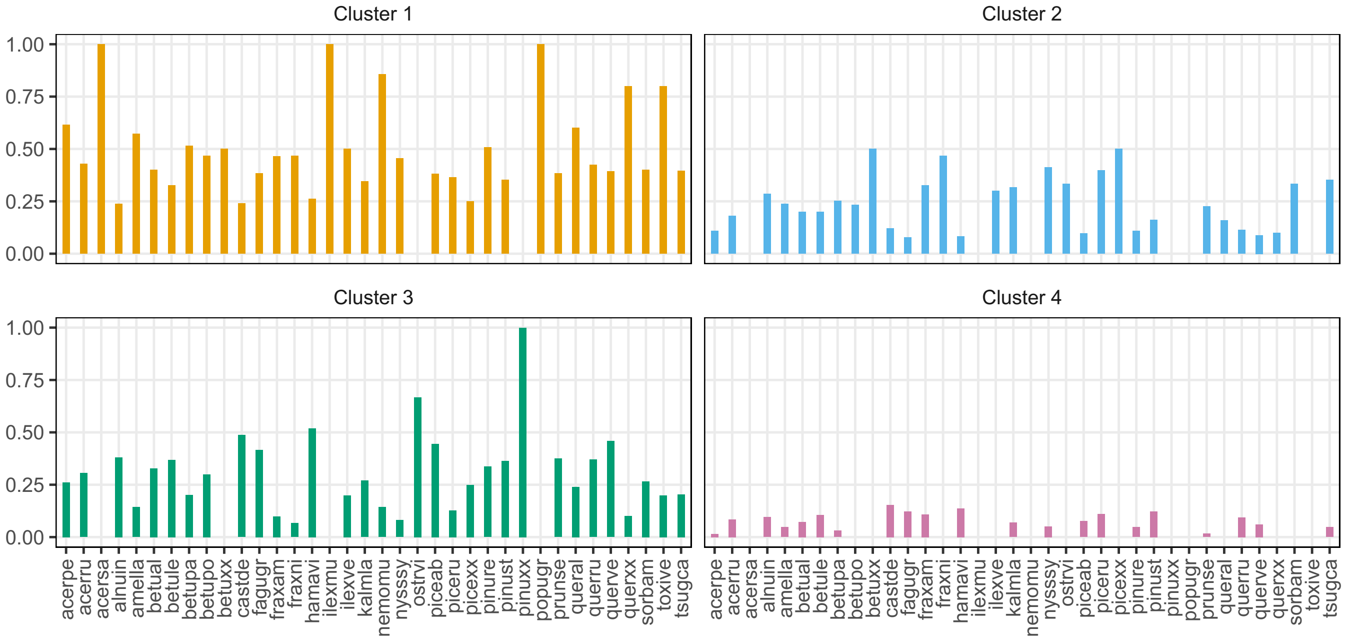

FIGURE 13 Distribution of species in each cluster.

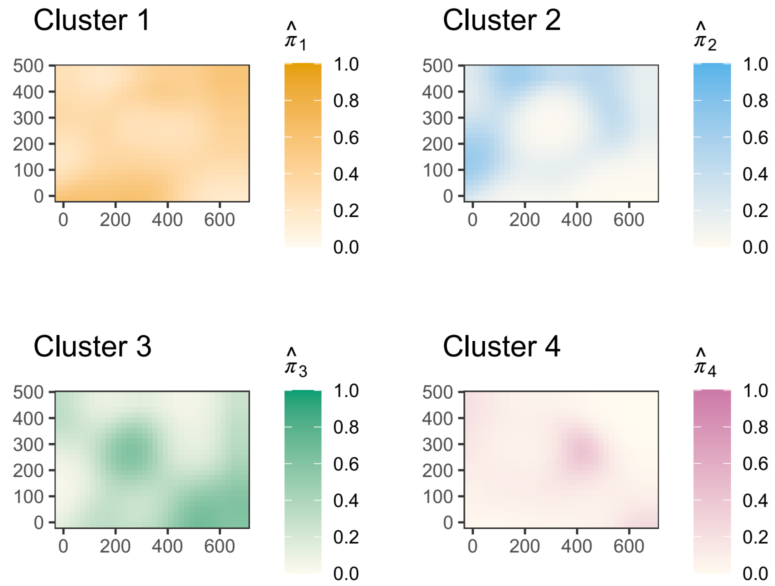

FIGURE 14 Maps of the estimated prior probabilities $\hat{\pi}_k(\boldsymbol{v}; \boldsymbol{\omega})$ for each cluster of the Prospect Hill Tract long-term plot.

policymakers could establish appropriate perimeters for areas at risk based on the clustering results and estimated prior probability maps, optimizing their decision-making process and resource allocation.

8 | DISCUSSION

Biodiversity profiles present a valuable tool for researchers to characterize and compare ecological communities by accounting for both abundant and rare species, thus recognizing the multidimensional aspects of diversity. In this study, following Gattone and Di Battista (2009), we have treated biodiversity profiles as non-negative and convex curves, amenable to analysis through the FDA approach. In particular, by considering the whole profiles as single entities, we have integrated FDA with spatial (model-based) clustering techniques to identify and delineate homogeneous zones based on spatial contiguity and shape similarity of the curves. This approach goes beyond traditional methods that may consider only individual abundance vectors and offers a more comprehensive understanding of biodiversity distribution, capturing the underlying patterns and variations across different regions. By focusing our study on a plot of the Harvard Forest, classification results indicate that our modeling approach can provide valuable information for policymakers, enabling them to make informed decisions regarding the conservation and management of natural resources.

However, given the nature of the available data and the absence of additional information, we recognize a few limitations in our taxonomic diversity study. Specifically, we acknowledge that working with complete census data is less common, particularly in larger study regions. In many realistic scenarios, in fact, only a sample of abundance vectors may be available at specific sites. In scenarios where complete abundance data is unavailable, achieving functional zoning of the domain presents several challenges. One approach involves interpolating the basis functions $\psi_l(\boldsymbol{\nu}_i)$ at new sites using kriging and subsequently predicting the posterior probability $\tau_k^{(d)}(\boldsymbol{\nu})$, as discussed in Pronello et al. (2023). Another alternative is to predict abundance vectors for missing sites. However, it is important to note that a single diversity functional profile can correspond to multiple abundance vectors. Given our focus on analyzing Hill's biodiversity profiles, our approach emphasizes predicting these functional profiles rather than directly estimating abundance vectors. This prediction can be accomplished by adopting a functional kriging model (Franco-Villoria & Ignaccolo, 2017; Giraldo et al., 2011; Ignaccolo et al., 2014; Mateu & Giraldo, 2021). Nevertheless, adopting a two-step approach—first predicting diversity functional profiles and then clustering them—could complicate the quantification of uncertainty in curve estimation and determining posterior probabilities of cluster membership. To mitigate these challenges, a model-based approach within the Bayesian framework could be advantageous. Within this integrated framework, potential measurement errors, curve prediction, and clustering can be addressed cohesively, offering improved control over error propagation across the hierarchy of conditional distributions. This unified approach would facilitate simultaneous handling of curve prediction and clustering complexities, while enhancing the quantification of uncertainty. We identify these areas as promising research topics for future work.

Another important consideration is that all species have been treated as equally distinct from one another, disregarding potential species differences in our study. In general, biodiversity extends beyond mere species diversity, encompassing a broader spectrum that includes phylogenetic, genetic, and functional diversity (Pielou, 1975). Relying solely on species names provides limited insights into the functions or evolutionary history of these species, which are instead crucial for understanding the underlying processes contributing to the observed levels of biodiversity. However, despite the acknowledged limitations, there are promising avenues to enhance our functional framework for biodiversity profiles. One approach involves incorporating pairwise similarities between species using a similarity matrix, leading to the *Leinster-Cobbold diversity* of order q as proposed by Leinster and Cobbold (2012). Alternatively, we can explore the unified framework proposed by Chao and Colwell (2022), which defines the *Hill-Chao numbers* of order q to assess biodiversity across multiple dimensions. By incorporating species trait similarities or adopting the more general framework of Chao and Colwell (2022), we can gain a more complete understanding of a community and improve predictions of ecosystem functions. These approaches represent promising directions for future research, aiming to provide a more nuanced and comprehensive perspective of biodiversity dynamics and their ecological significance.

ACKNOWLEDGMENTS

We are grateful to the Reviewers and the Associate Editor for their valuable comments and suggestions, which significantly contributed to improving the quality of this manuscript. Natalia Golini and Rosaria Ignaccolo acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender number 1409 published on September 14, 2022 by the Italian Ministry of University and Research (MUR), funded by the European Union—NextGenerationEU—*Statistics for vegetation biodiversity: estimation and mapping (SVeBio)*—P2022AW4LX—CUP B53D23029510001—Grant Assignment Decree No. 1378 adopted on September 1, 2023 by the Italian Ministry of Ministry of University and Research (MUR). All the authors acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on February 2, 2022 by the Italian Ministry of University and Research (MUR), funded by the European Union—NextGenerationEU—*Complex Environmental Data and Modeling (CoEnv)*—2022E3RY23 – CUP D53D23011080006-Grant Assignment Decree No. 1109 adopted on July 20, 2023 by the Italian Ministry of Ministry of University and Research (MUR). Open access publishing facilitated by Universita degli Studi di Torino, as part of the Wiley - CRUI-CARE agreement.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the Harvard Forest Data Archive at https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF253, reference number HF253. These data were derived from the following resources available in the public domain:—hf253-05: stems 2014; hf253-06: stems 2019, https://harvardforest1.fas.harvard.edu/exist/apps/datasets/showData.html?id=HF253. All the analyses are carried out by developing custom code within the R environment (R Core Team, 2023). The R code is available upon contacting the authors.

REFERENCES

Abraham, C., Cornillon, P.-A., Matzner-Løber, E., & Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics*, *30*(3), 581–595.

Adler, R. J. (2010). *The Geometry of Random Fields*. SIAM.

Baudry, J. (2015). Estimation and model selection for model-based clustering with the conditional classification likelihood. *Electronic Journal of Statistics*, *9*(1), 1041–1077.

Cardinale, B. J., Duffy, J. E., Gonzalez, A., Hooper, D. U., Perrings, C., Venail, P., & Naeem, S. (2012). Biodiversity loss and its impact on humanity. *Nature*, *486*(7401), 59–67.

Chao, A., & Colwell, R. K. (2022). *Biodiversity: Concepts, dimensions, and measures*. In *The ecological and societal consequences of biodiversity loss* (pp. 25–46). John Wiley & Sons, Ltd.

Chilès, J.-P., & Delfiner, P. (2012). *Geostatistics: Modeling spatial uncertainty*. John Wiley & Sons.

Cressie, N. (1993). *Statistics for spatial data* (2nd ed.). John Wiley.

Dabo-Niang, S., Yao, A., Pischedda, L., Cuny, P., & Gilbert, F. (2010). Spatial mode estimation for functional random fields with application to bioturbation problem. *Stochastic Environmental Research and Risk Assessment*, *24*, 487–497.

DeLong, D. C. (1996). Defining biodiversity. *Wildlife Society Bulletin*, *24*(4), 738–749.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, *39*, 1–38.

Di Battista, T., Fortuna, F., & Maturo, F. (2016). Environmental monitoring through functional biodiversity tools. *Ecological Indicators*, *60*, 237–247.

Di Battista, T., Fortuna, F., & Maturo, F. (2017). BioFTF: An R package for biodiversity assessment with the functional data analysis approach. *Ecological Indicators*, *73*, 726–732.

Díaz, S., Demissew, S., Carabias, J., Joly, C., Lonsdale, M., Ash, N., & Zlatanova, D. (2015). The IPBES conceptual framework—Connecting nature and people. *Current Opinion in Environmental Sustainability*, *14*, 1–16.

Díaz, S., Fargione, J., Chapin, F. S., III, & Tilman, D. (2006). Biodiversity loss threatens human well-being. *PLoS Biology*, *4*(8), e277.

European Commission. (2021). *EU biodiversity strategy for 2030: Bringing nature back into our lives* (Publications Office of the European Union Ed.). Directorate-General for Environment. https://data.europa.eu/doi/10.2779/677548

FAO and UNEP. (2020). *The state of the world's forests 2020. Forests, biodiversity and people*. Food and Agriculture Organization of the United Nations and UN Environment Programme. https://doi.org/10.4060/ca8642en.

FAO. (2022). *Action plan for mainstreaming biodiversity across agricultural sectors in Eastern Europe and Central Asia 2022–2023*. Food and Agriculture Organization of the United Nations. https://doi.org/10.4060/cc1159en.

Ferraty, F., & Vieu, P. (2006). *Nonparametric functional data analysis: Theory and practice* (Vol. *76*). Springer.

Fortuna, F., & Di Battista, T. (2020). Functional unsupervised classification of spatial biodiversity. *Ecological Indicators*, *111*, 106027.

Fortuna, F., Gattone, S. A., & Di Battista, T. (2020). Functional estimation of diversity profiles. *Environmetrics*, *31*(8), e2645.

Franco-Villoria, M., & Ignaccolo, R. (2017). Bootstrap based uncertainty bands for prediction in functional kriging. *Spatial Statistics*, *21*, 130–148.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, *9*, 432–441.

Gattone, S. A., & Di Battista, T. (2009). A functional approach to diversity profiles. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *58*(2), 267–284.

Gini, C. (1912). *Variabilità e mutabilità: Contributo allo studio delle distribuzioni e delle relazioni statistiche.[fasc. i.]*. Tipogr. di P. Cuppini.

Giraldo, R., Delicado, P., & Mateu, J. (2011). Ordinary kriging for function-valued spatial data. *Environmental and Ecological Statistics*, *18*, 411–426.

Giraldo, R., Delicado, P., & Mateu, J. (2012). Hierarchical clustering of spatially correlated functional data. *Statistica Neerlandica*, *66*(4), 403–421.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Hurlbert, S. H. (1971). The nonconcept of species diversity: A critique and alternative parameters. *Ecology*, *52*(4), 577–586.

Ignaccolo, R., Mateu, J., & Giraldo, R. (2014). Kriging with external drift for functional data for air quality monitoring. *Stochastic Environmental Research and Risk Assessment*, *28*, 1171–1186.

Jiang, H., & Serban, N. (2012). Clustering random curves under spatial interdependence with application to service accessibility. *Technometrics*, *54*, 108–119.

Leinster, T., & Cobbold, C. (2012). Measuring diversity: The importance of species similarity. *Ecology*, *93*, 477–489.

Liang, D., Zhang, H., Chang, X., & Huang, H. (2021). Modeling and regionalization of China's PM2.5 using spatial-functional mixture models. *Journal of the American Statistical Association*, *116*(533), 116–132.

MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews*, *40*(4), 510–533.

Magurran, A. E. (2021). Measuring biological diversity. *Current Biology*, *31*(19), R1174–R1177.

Mardia, K., Kent, J., Goodall, C., & Little, J. (1996). Kriging and splines with derivative information. *Biometrika*, *83*(1), 207–221.

Mardia, K., Redfern, E., Goodal, C., & Alonso, F. (1998). The Kriged Kalman filter. *Test*, *59*, 217–285.

Mateu, J., & Giraldo, R. (2021). *Geostatistical functional data analysis*. John Wiley & Sons.

Maturo, F., & Di Battista, T. (2018). A functional approach to hill's numbers for assessing changes in species variety of ecological communities overtime. *Ecological Indicators*, *84*, 70–81.

Orwig, D., Foster, D., & Ellison, A. (2022). *Harvard Forest CTFS-ForestGEO Mapped Forest Plot since 2014. Harvard Forest Data Archive: HF253 (v.5)*. Environmental Data Initiative.

Pielou, E. C. (1975). *Ecological diversity*. John Wiley.

Pronello, N., Ignaccolo, R., Ippoliti, L., & Fontanella, S. (2023). Penalized model-based clustering of complex functional data. *Statistics and Computing*, *33*(6), 122.

Purvis, A., & Hector, A. (2000). Getting the measure of biodiversity. *Nature*, *405*(6783), 212–219.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. https://www.R-project.org/

Ramsay, J. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *60*(2), 365–375.

Ramsay, J., & Silverman, B. (2005). *Functional data analysis*. Springer.

Romano, E., Balzanella, A., & Verde, R. (2017). Spatial variability clustering for spatially dependent functional data. *Statistics and Computing*, *27*, 645–658.

Romano, E., Mateu, J., & Giraldo, R. (2015). On the performance of two clustering methods for spatial functional data. *AStA Advances in Statistical Analysis*, *99*, 467–492.

Schmeller, D. S., Courchamp, F., & Killeen, G. (2020). Biodiversity loss, emerging pathogens and human health risks. *Biodiversity and Conservation*, *29*(11), 3095–3102.

Secchi, P., Vantini, S., & Vitelli, V. (2013). Bagging Voronoi classifiers for clustering spatial functional data. *International Journal of Applied Earth Observation and Geoinformation*, *22*, 53–64.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379–423.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*(4148), 688.

Vandewalle, V., Preda, C., & Dabo-Niang, S. (2021). *Clustering spatial functional data*. In J. Mateu & R. Giraldo (Eds.), *Geostatistical functional data analysis: Theory and methods* (pp. 155–174). John Wiley and Sons.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, *121*(2), 311–324.

WHO Teams. (2020). *Guidance on mainstreaming biodiversity for nutrition and health*. World Health Organization and Convention on Biological Diversity.

Wu, H., & Li, Y. F. (2023). Clustering spatially correlated functional data with multiple scalar covariates. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 7074–7088.

Wu, X., & Sickles, R. (2018). Semiparametric estimation under shape constraints. *Econometrics and Statistics*, *6*, 74–89.

Zhang, M., & Parnell, A. (2023). Review of clustering methods for functional data. *ACM Transactions on Knowledge Discovery from Data*, *17*(7), 1–34.

Zhou, H., Pan, W., & Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, *3*, 1473–1496.

How to cite this article: Golini, N., Ignaccolo, R., Ippoliti, L., & Pronello, N. (2024). Functional zoning of biodiversity profiles. *Environmetrics*, e2865. https://doi.org/10.1002/env.2865