Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Entropy-Based Assessment of Biodiversity, With Application to Ants' Nests Data

L. Altieri | D. Cocchi | M. Ventrucci

Department of Statistical Sciences, University of Bologna, Bologna, Italy

Correspondence: L. Altieri (linda.altieri@unibo.it)

Received: 30 October 2023 | **Revised:** 12 October 2024 | **Accepted:** 17 October 2024

Keywords: ants' nests data | Batty's entropy | entropy estimation | Leibovici's entropy | SpatEntropy | spatial entropy

ABSTRACT

The present work takes an innovative point of view in the study of a marked point pattern dataset of two ants' species, over an irregular region with a spatial covariate. The approach, based on entropy measures, brings new insights to the interpretation of the behavior of such ants' nesting habits, which can be exploited in the general area of biodiversity evaluation. We make proper use of descriptive entropy measures and inferential approaches, performing a comparative study of their uncertainty and interpretability in the context of biodiversity. For the first time in the study of these ants' nests data, all the available information is fully exploited, and interpretation guidelines are given for assessing both the observed and the latent biodiversity of the system, with a simultaneous consideration of spatial structures, covariate and interpoint interaction effects. Computations are supported by the new release of our R package SpatEntropy.

1 | Introduction

One relevant part of current challenges in environmental studies consists of monitoring the diversity of life on the planet, and how it is affected by the consequences of human activities, such as climate change issues, over-harvesting, pollution, habitat loss, and invasive species. Biodiversity studies are one of the bases of any ecological analysis, as biological diversity is fundamental for a wide variety of ecosystems services, such as "natural harvests, carbon sequestration, pollination, and soil formation" (Magurran 2004). The UN Convention on Biological Diversity[1] set a target to significantly reduce the rate of biodiversity loss, which has led to new developments in measuring biodiversity with appropriate syntheses (see, e.g., Hoskins et al. 2020; Drechsler 2020). Measurements of biological diversity may simply count species richness (Fisher et al. 2012) or measure the abundance of species in ecological communities (Scarnati et al. 2009), up to very recent and complex model-based approaches for counting individuals based on capture-recapture data (Altieri, Farcomeni, and Alunni-Fegatelli 2023). Entropy measures represent a largely used approach for properly measuring biodiversity (Leinster and Cobbold 2012).

Many data examples concerning biodiversity and the study of the behavior of animals and plants are available. One has been selected for this work, as it has raised a lot of interest in the literature, but presents both computational challenges and issues in methods, consistency, and interpretation of the results. The data consist of the spatial locations of nests built by two species of ants, *Messor wasmanni* (from now on *Messor*) and *Cataglyphis bicolor* (from now on *Cataglyphis*), recorded at a site in northern Greece and firstly described in Harkness and Isham (1983). Since nests usually have a single opening in the ground, they may be treated as a point pattern. The total number of nests is $n = 97$, divided into $n_1 = 29$ *Cataglyphis* nests and $n_2 = 68$ *Messor* nests. They are collected over a polygonal region, the *observation window* W, of

size | W |= 107·230.4 square feet. An environmental covariate C is also available, that is, the type of soil inside the region, with two levels: $c_1 = field$ and $c_2 = scrub$. The dataset is displayed in Figure 1.

The harvester ant *Messor* builds nests mainly by piling seed husks, which constitute its main food source; instead, *Cataglyphis* ants eat dead insects and other arthropods, which reflects on the material used for building nests. *Messor* ants, once dead, may be part of the *Cataglyphis* diet, but the two species do not have a predator–prey relation, as *Cataglyphis* ants do not kill *Messor* ants. This dataset is a very interesting example for ecological studies, regarding both computations and interpretation. Previous studies Harkness and Isham (1983), Isham (1984), Takacs and Fiksel (1986), Särkkä (1993), Högmander and Särkkä (1999), and Baddeley and Turner (2000) focused on rectangular subsets of the observation window (shown in Figure 1) for computational simplicity, and offered conclusions based on descriptive summary statistics or very simple models, often lacking proper checks; results are inconsistent across studies and scarce explanation is given about final findings. Moreover, such works only focused on the within- and between-species interaction, without deepening the role of the soil type and of other spatial effects in determining the ants' distribution. The presence of interaction within each species is expected in the form of repulsive behavior, due to competition for food. Also, association between the two species is believed to be positive, because of a preference for *Cataglyphis* ants in building nests close to *Messor* nests, in order to collect dead *Messor* ants. The evidence for interactions was investigated by the aforementioned works with different methods, with contrasting conclusions.

One contribution of our work is to approach the research questions with a different perspective, by using entropy-based indicators and models. Our aim is to show that particular entropy-based measures, those including spatial information, can provide new insights about the spatial distribution of nests and consider the role played by the covariate *field*/*scrub*. After reviewing the existing works on the same data, we show how our proposal overcomes the issues and inconsistencies in their results, by providing more exhaustive and interpretable conclusions. We believe that the ants case study provides an interesting example to illustrate the benefit of using spatial entropy-based indicators and how these can be applied more generally to the study of biodiversity assessment. We make the novel proposal of employing entropy-based descriptive approaches to investigate both the influence of the covariate and the presence and type of interpoint interaction. For instance, the spatial measure known as Batty's entropy (Batty 1974) has been largely employed over administrative boundaries, but has an unexplored potential in investigating the influence of environmental covariates. Similarly, the co-occurrence-based index known as Leibovici's entropy (Leibovici et al. 2014) is here reinterpreted to detect within/between species interaction at given distances. The two entropy-based approaches have been developed separately and have never been confronted; a further contribution of the present work is to show how a combination of the two measures is a useful exploratory way of evaluating what spatial components may play a role in determining the data behavior, and point toward the most appropriate model class for the data. A further innovation of this paper is that all descriptive measures are sided by simulation-based tests to assess the departure of the results from the reference situation of independence (intended as spatial randomness), thus providing an intuitive but well-founded tool for the interpretation of entropy indicators. In addition, the analysis is extended to more formal model fitting; for the first time over the ants' nests data, a model selection procedure is used to choose the best option, and the chosen model is properly interpreted and satisfactorily evaluated in terms of its goodness-of fit. We show how a model-based entropy approach can be used not only to evaluate the *latent* biodiversity of the system, by including the role of the covariate and of other spatial structures, but also to interpret the presence of interpoint interaction and give a comprehensive understanding of the ants' nests data. The *latent* biodiversity is intended as the underlying behavior of the unknown complex natural process, of which data represent a partial picture. This inferential conceptualization might provide interesting insights about the structure of the available data. The literature has already proposed entropy estimation methods, but the available estimators return a single number for the whole observation area, that is a good synthesis in very simple contexts, but oversimplifies a complex natural phenomenon. We illustrate how, under some conditions, entropy may be estimated locally: a measure that is allowed to vary over space may be informative about heterogeneity, as a trade-off between capturing the main data behavior and allowing for local variations. Furthermore, we show how to exploit the estimators' variances (or their upper bounds) to build intervals of plausible values and quantify uncertainty of the results. The present work also contributes to the existing software and computational tools, in order to make advanced statistical methods available for applied scientists. We introduce the novel version of our `R` package `SpatEntropy`,[2] stressing the related advantages in practical studies. We provide the complete code to reproduce the entire practical work of this paper as Supporting Information.

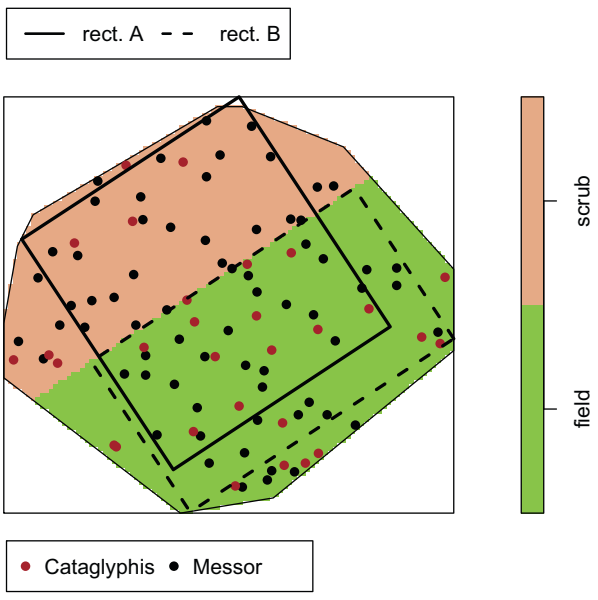

FIGURE 1 | Ants' nests data—A point pattern with nests of two species of ants over a polygonal area with two soil types. The two rectangles A and B mark the subsets of the region studied in the literature until now.

The paper is organized as follows: Section 2 summarizes the state of the art regarding both the ants' nests data and the computational tools for biodiversity measurements. Section 3 presents the necessary background for all the descriptive and inferential entropy-based measures. The practical Section 4 collects all the results about the ants' nests data. Lastly, Section 5 summarizes the paper contribution, and contains comparative and conclusive comments. An exhaustive R script with the practical work of the present paper is available as Supporting Information attached to the paper[3]; the case study can be downloaded in R, as part of the `spatstat` package, under the name `ants`.

2 | State of the Art and Motivation

Previous works on such ants' nests data, though providing interesting contributions, present limits due to the complexity of the data structure. A marked point pattern over an irregular observation window with the presence of a covariate has constituted a challenging data example up to recent years; the advanced tools that are needed and used in the present work have never been applied over the considered dataset. All previous works, except for Baddeley and Turner (2006), focus on a rectangular subset of the observation window, and the study of the species (i.e., the point pattern *marks*) interaction is always separated from a study of the covariate effect. A joint evaluation of the two aspects has never been faced in such ants' nests data studies. This Section presents a comparative review of the previous works on the ants' nests data. Moreover, it summarizes the available computational tools for the analysis of this type of data. The combination of recent methodological and computational progress in this field is exploited in the present work to contribute to the study of the ants' nests data.

2.1 | Literature on the Ants' Nests Data

Several papers have analyzed the ants' nests data of Figure 1: Harkness and Isham (1983), Isham (1984), Takacs and Fiksel (1986), Särkkä (1993), Högmander and Särkkä (1999), and Baddeley and Turner (2000, 2006). All contributions, apart from Baddeley and Turner (2006), reduce the analysis to a subset of the observation window. Harkness and Isham (1983) propose two overlapping rectangles, shown as A and B in Figure 1, where B follows the *field-scrub* border in an attempt to grasp influence of this covariate. The works by Isham (1984), Takacs and Fiksel (1986), Särkkä (1993), Högmander and Särkkä (1999), and Baddeley and Turner (2000) focus on rectangle A, discarding relevant information: only 61 out of 97 nests are considered, where the relative presence of the *Messor* nests is higher with respect to the original dataset. The reason for such choice is computational complexity: the first work presenting a software able to deal with irregularly-shaped window is Baddeley and Turner (2005), which presents the ants' nests data as a challenging case study. The only available study on the irregular window is in Baddeley and Turner (2006), and leads to different results than those obtained over rectangle A.

Harkness and Isham (1983) and Isham (1984) use the G, K, and L distance-based functions and Pearson's chi-square test (Diggle 2014) to capture significant deviations from *complete spatial randomness* (aka spatial homogeneity or stationarity). The first model-based approach is proposed in Takacs and Fiksel (1986), where a simple Gibbs process (Illian et al. 2008) is used to measure the interpoint interaction; no model diagnostics are run. Särkkä (1993) employs a Strauss, or hard-core, model (Bondesson and Fahlén 2003) to investigate repulsion, with an unsatisfying fit. In Högmander and Särkkä (1999), the interaction between the two species is assumed to have a direction, in order to moderately improve the goodness-of-fit; parameters are estimated with Maximum Likelihood, that is computationally demanding, and Maximum PseudoLikelihood, that the authors declare unreliable when the interaction is strong. Baddeley and Turner (2000) only show the implementation of the previous works in R, as part of a set of examples.

The above listed approaches lead to divergent results. For Harkness and Isham (1983), *Messor* ants have a repulsive within-species behavior, while *Cataglyphis* nests appear to be randomly scattered; interspecies interaction is detected with the G function, but not with the K function. Isham (1984) finds that there is no interaction in the pattern. According to Takacs and Fiksel (1986), instead, *Messor* nests are random, while *Cataglyphis* nests form a repulsive pattern. Särkkä (1993) and Högmander and Särkkä (1999) find a repulsive behavior both within and between species. Many results are unexpected, but no insights are provided. Baddeley and Turner (2006) find a substantial difference in the results both for the distance-based functions and for the model results when using the polygonal area instead of the rectangle. They propose hypothesis testing for the independence within and across species, and again results are inconsistent; in particular, when considering the polygonal window, the directional dependence found in Högmander and Särkkä (1999) disappears. A simple model for the covariate effect finds no significance for the *Messor* nests, but a significant preference of the *Cataglyphis* species to build nests over the *field* area, rather than the *scrub* one. Baddeley and Turner (2006) underline the overall inconsistency, the need for more formal methods and a more general approach able to consider other aspects, such as the covariate *soil type*. Though a question about the importance of the soil type is mentioned across the papers, no work up to Baddeley and Turner (2006) discusses such factor, nor any other possible influence in determining the ants' nesting behavior.

In the present paper, the specific questions on the ant case study are approached under a new perspective, using spatial entropy measures. Regarding dependence on the type of soil, we propose to use a partition-based approach (Batty 1974) detailed in Section 3.1, while to assess interactions we propose a distance-based approach (Leibovici et al. 2014), illustrated in Section 3.2. Uncertainty evaluation will be done via Monte Carlo simulations. In Section 3.3, we use a Bayesian model based approach (Altieri, Cocchi, and Ventrucci 2023) and argue that this may represent a unified framework to produce estimates of entropy, with the inclusion of spatial and covariate effects, and give a general understanding of the structure of the ants' nests data.

2.2 | Software for Biodiversity Evaluation

A comparative review of software and programming languages is out of the scope of this work. One excellent tool for statistical analysis with the major advantage of being open-source is the `R` software (R Core Team 2017). In this paper, we highlight the power and potential of the `R` software for biodiversity studies, and we overcome its difficulties by providing a ready-to-use script as Supporting Information.

In the field of biodiversity studies, a large number of `R` packages on CRAN[4] is available. One reference package is `vegan`, which contains both diversity indices and interesting data examples. Some packages focus on specific diversity measures, such as alpha and beta diversity (`abdiv`, `betapart`, `BAT`) or Hill's number (`hilldiv`, `hillR`). Others focus on specific applications, as forests (`fgeo`) and oceans (`robis`). Many packages are meant for biological data downloading, cleaning and visualization (`bdc`, `bdchecks`, `bdvis`, `galah`, `KnowBR`, `occCite`). There are packages for prediction of the number of species in a system (`DivE`, `SpadeR`), others give tools for easy computations of the frequencies of species (`divo`), or focus on phylogenetics (`adiv`). The number of packages is large, and many of them considerably overlap. The main goal in most packages is description, and basic inferential tools are used to associate uncertainty to descriptive measures (such as bootstrap samples or leave-one-out validation), while model-based approaches are hardly mentioned, probably because of their complexity. A further major drawback of these measures is the lack of spatial components; despite some indices may be computed separately for different areas or at different scales, space is not included in the available functions.

When the focus is on entropy measures, the most common `R` packages are `entropart`, `entropy` and `EntropyEstimation`. With any of them, Shannon's entropy may be computed and decomposed into the two terms known as mutual information and conditional entropy (Cover and Thomas 2006). The `entropart` package also provides computation of α, β, and γ diversity indices, together with Simpson's index. The package `entropy` contains functions for some popular entropy estimators, which are recalled in the remainder of the present work. The package `EntropyEstimation` includes functions which do the same job as the other packages, in addition to generalized Simpson's indices. There is no consideration of the spatial arrangement of data.

The first release of the `R` package `SpatEntropy`[5] was in 2018, with a related publication (Altieri, Cocchi, and Roli 2021); it provides functions for all the main available spatial approaches to entropy measures. The only overlap between the `SpatEntropy` package and the aforementioned ones consists of the function for computing Shannon's entropy (`shannon`); all other functions (`batty`, `battyLISA`, `oneill`, `leibovici`, `altieri`, and further related/auxiliary functions) constitute an original contribution of the present package to the `R` software community. The package has been progressively improved up to the newly released version 2.2-4 (November 2023), to become easier to approach by applied scientists. All functions work with point and gridded data, and return useful plots for easier interpretation of the results; the computational efficiency has been substantially improved. Some environmental data examples can be found in the `SpatEntropy` package, regarding both point and gridded data. When the input data are point patterns, the package relies on auxiliary functions from the `spatstat` package (Baddeley, Rubak, and Turner 2015).

3 | The Use of Entropy in Biodiversity Assessment

This Section summarizes the theoretical background for a solid approach to the evaluation of the biodiversity of the ants' data and a general understanding of their spatial structure. We show how the available descriptive approaches to spatial entropy may evaluate the influence of a single covariate, or the interaction between ants' species; they can be used as exploratory tools, and give useful suggestions when moving to inferential approaches.

The first diversity measure based on entropy diffused in ecological studies was proposed by Shannon (1948). Given a variable X with I possible categories, Shannon's entropy is:

$$H(X) = \sum_{i=1}^{I} p(x_i) \log \frac{1}{p(x_i)}, \tag{1}$$

where $p(x_i)$, for $i = 1, \ldots, I$, is the probability of occurrence of category i. The index $H(X)$ ranges in $[0, \log I]$, and expresses the average amount of heterogeneity across observations of a variable X. The probabilities $p(x_i)$ are usually estimated by the data relative frequencies n_i/n, where n is the total number of observations and n_i is the number of observations presenting the i-th category.

Shannon's entropy only considers the probabilities of the categories. In biodiversity studies, this may be a major limit, since often abiotic (e.g., environmental covariates) and biotic factors (such as interactions between or within species) affect the occurrence of a specific species in the sense that its distribution is not homogeneous in space. The indicator in Equation (1) does not assume any notion of spatial structure, thus it gives poor insights regarding the questions of interest in the ants case study. A dataset with randomly scattered observations over space should have a higher entropy than one with the same observations arranged according to a spatial structure; moreover, entropy can potentially vary over space. Shannon's entropy is not able to catch these differences and variations; such limit is widely discussed in the literature (see, e.g., Altieri, Cocchi, and Roli 2019). Considering the spatial structure in the data may improve the monitoring of the distribution of species.

3.1 | A Partition-Based Approach for the Dependence on Covariates

The partition-based approach to spatial entropy, briefly named Batty's entropy (Batty 1974, 1976, 2010), is originally meant for geographical applications, where the area of interest is partitioned by administrative boundaries, such as municipalities. Batty's entropy starts from the identification of a phenomenon

of interest F, for example the occurrence of nests of one specific ant species, and from a partition of the area into G sub-areas. Such entropy index assesses the distribution of the phenomenon F across sub-areas, and includes a measure of the sub-area size T_g, where $\sum_g T_g = T$. The *intensity* over a sub-area is $\lambda_g = p_g/T_g$, with p_g being the probability of having observations over area g, with $\sum_g p_g = 1$; in practice, p_g is the relative frequency of observations over area g with respect to the total number of observations over the window. Battys' entropy is:

$$H_B(F) = \sum_{g=1}^{G} p_g \log \frac{1}{\lambda_g} \tag{2}$$

and ranges in $\left[\log T_{g^*}, \log T\right]$, where g^* indicates the smallest sub-area. If the phenomenon is equally intense over the area, $\lambda_g = \lambda$ for $g = 1, \ldots, G$, and Batty's entropy is at its maximum; if observations are concentrated over one or few sub-areas, the entropy decreases, especially when observations occur over the smallest areas. Computational issues arise when the sub-areas' size is smaller than 1 measurement unit, since the logarithm in (2) takes negative values, while entropy is not allowed to be negative. One option is to tune the measurement unit, for example, by using meters instead of kilometers, so that the numbers involved in the computations are larger. If it is not possible to rescale the observation area prior to computations, `SpatEntropy` offers an automatic solution: the observation area is rescaled by a factor $c > 1$, so that each sub-area has a new size equal to $\left(T_g \times c\right) > 1$ for all g; Batty's entropy is computed over the rescaled area, and then $\log c$ is subtracted from the entropy result, to obtain the entropy referring to the original area.

To our knowledge, this approach has only been applied to area partitions based on administrative boundaries. We argue that such measure has a potential in determining whether the intensity of a phenomenon varies according to a spatial covariate: this can be done by partitioning the area into tiles according to the categories or values of the available covariate, and computing (2) on such partition. In Section 4, we implement such partition-based entropy to investigate the effect of the soil type on the ants' nests occurrence.

3.2 | A Distance-Based Approach for the Detection of Interaction

The second approach considers the distance between occurrences of the categories, and evaluates the level of heterogeneity of couples (co-occurrences) at some chosen distance; it is known as distance-based entropy, co-occurrence-based entropy or Leibovici's entropy (Leibovici 2009; Leibovici et al. 2014). The variable denoting species is X, with I categories, while the variable classifying the types of co-occurrences is named Z. The number of categories of Z is I^2, as ordered couples are considered; the probabilities of each co-occurrence category is $p(z_r)$ for $r = 1, \ldots, I^2$.

The first proposal in such direction is O'Neill's entropy (O'Neill et al. 1988) for gridded data and contiguous couples, that is, pixels sharing a border. Leibovici's entropy extends O'Neill's proposal by substituting the notion of contiguity with that of distance, that is, by considering couples placed *within a chosen distance d*, which can also apply to point data. The entropy is:

$$H_d(Z) = \sum_{r=1}^{I^2} p^{(d)}(z_r) \log \frac{1}{p^{(d)}(z_r)}, \tag{3}$$

where $p^{(d)}(z_r)$ is the probability of finding the co-occurrence category z_r, for $r = 1, \ldots, I^2$, at the fixed distance d (for a more formal definition, see Altieri, Cocchi, and Roli 2019). Measure (3) can be computed for any distance within the observation window, according to the researcher's choice. The index ranges in $\left[0, \log I^2\right]$, where the maximum is reached when all possible couples of categories of X are equally represented, while 0 is reached when one single category is present (Altieri, Cocchi, and Roli 2018).

In our opinion, this measure also has an unexplored potential for biodiversity studies. In Section 4, we show how it can be used to investigate possible presence of interaction between and within species.

3.3 | Model-Based Approaches for a Comprehensive Biodiversity Study

Inferential approaches start from the assumption that the observed data are a realization of an underlying stochastic process, with parameters governing the species probabilities. Estimating the parameters of the process allows to compute an estimate of the entropy of the system.

The existing literature about entropy estimation mainly consists of model-based approaches, both frequentist and Bayesian (Paninski 2003); a non-parametric approach is also among the main competitors for estimation (Zhang 2012). The common starting point is the maximum likelihood (*ML*) estimator, which substitutes the unknown probability distribution of interest with the observed relative frequencies. Such estimator is known to be negatively biased (for details, see section 3.3 in Paninski 2003). All proposals focus on improving the performance of the *ML* estimator as regards the estimator bias; they are based on the simplest model, where no auxiliary information is considered and, most importantly, independence among realizations is assumed. In the literature, it is not common to find models accounting for spatial dependence and addressing biodiversity estimation via entropy measures, as opposed to what is done in species distribution modeling, where the focus is on the relationship between species abundance and environmental factors or temporal/spatial effects (Ventrucci et al. 2020; Martinez-Minaya et al. 2018).

A recent proposal to entropy estimation, named *BMB* (Bayesian model-based), relaxes the independence assumption (Altieri, Cocchi, and Ventrucci 2023). In order to include data structures, a model-based approach is taken for the estimation of the main components of an entropy index, that is, the species probabilities; the entropy function is successively derived. Following a multinomial-logit model, the occurrence probabilities for each species i and spatial location u can be expressed as:

$$p_{ui} = \frac{\exp\left(\boldsymbol{c}'_{ui}\boldsymbol{\beta}_i + \phi_{ui}\right)}{\sum_i \exp\left(\boldsymbol{c}'_{ui}\boldsymbol{\beta}_i + \phi_{ui}\right)}, \quad i = 1, \ldots, I, \tag{4}$$

where the linear predictor $g_{ui} = \boldsymbol{c}'_{ui}\boldsymbol{\beta_i} + \phi_{ui}$ is a function of species-specific covariates $\mathbf{c}_{ui}$ and species-specific random effects $\phi_{u1}, \ldots, \phi_{ul}$ at location u; several priors can be assumed for the random spatial effects, to impose some degree of smoothing, can be applied. As examples of smooth spatial effects, we mention the Intrinsic Conditional AutoRegressive (*ICAR*) model, where the spatial influence is extended to the 4 nearest neighbors, and the Random Walk in 2 dimensions (*RW2d*), with a neighborhood system including 12 neighbors. These are two popular spatial smoothing priors belonging to the class of Gaussian Markov Random Fields (*GMRFs*); for the theory on *GMRF*s, we refer to Rue and Held (2005). The parameters of our models are estimated via a Bayesian approach, using the `R-INLA` package (Rue, Martino, and Chopin 2009). To fit a multinomial regression with `R-INLA`, the multinomial likelihood is transformed into a Poisson likelihood with extra parameters; such approach is known as the "multinomial-Poisson trick" and is detailed in Serafini (2019) and in Barmoudeh, Baghishani, and Martino (2022). For practical implementation of the *ICAR* and the *RW2d* models in `R-INLA`, see Lindgren and Rue (2013) and Bivand, Gómez-Rubio, and Rue (2015).

Finally, the entropy estimate, denoted as $\widehat{H}_u^{BMB}(X)$, is derived by sampling from the model posterior in two steps: first, sampling from the marginal posterior distribution of the quantities in Equation (4), that is, the $\boldsymbol{\beta}'_i$ and the random effects ϕ_{ui}, would give us a sample of the occurrence probabilities p_{ui} for each location and species; second, plugging-in the sampled p_{ui} in Equation (1), we obtain a sample from the posterior distribution of the entropy indicator. From this posterior sample, we can compute any summary (mean, median, quantiles, ...); assume $s = 1, \ldots, S$ indexes the draws from the posterior and let $\tilde{p}_{uis}$ be the s-th draw for the occurrence probability for species i at location u; we take the posterior mean as the estimated entropy:

$$\widehat{H}_u^{BMB}(X) = \frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{I}\tilde{p}_{uis}\log\frac{1}{\tilde{p}_{uis}}. \tag{5}$$

The estimate $\widehat{H}_u^{BMB}(X)$ varies with u and represents the local entropy. Therefore, for spatial data, a two-dimensional surface evaluating the latent biodiversity of the considered system may be obtained.

4 | An Entropy-Based Study of the Biodiversity of Ants' Nests

In this Section, we propose an alternative approach to those of the literature of Section 2 to analyze the ants' nests data and interpret their spatial arrangement in the context of biodiversity evaluation. The entropy measures presented in this work are given both in absolute and relative terms, where *relative* means that we divide the measure by its maximum value; this is meant to facilitate comparison across studies and also interpretation, since the level of heterogeneity may be given in proportional terms with respect to the maximum possible level. Along the Section, we mention some of the most relevant `R` functions, and we refer to the Supporting Information for details.

Traditional entropy-based measures are not accompanied by theoretical or empirical confidence intervals, therefore conclusions are very subjective. For a more formal evaluation of the results, we borrow standard tools of point process studies. In such field, it is common to compare empirical values with a large set of values computed over random generations under a homogeneous process with the same number of points as the case study; this allows to check whether there is a significant departure from the situation of spatial randomness. Thanks to the `SpatEntropy` package, such approach is now possible, with quickly available results.

4.1 | Non-Spatial Heterogeneity

Shannon's descriptive entropy is computed based on the global relative frequencies of the two ants' species: $f_1 = n_1/n = 0.299$ for *Cataglyphis* and $f_2 = n_2/n = 0.701$ for *Messor*. The entropy value is $H(X) = 0.61$, returned by the function `shannon(ants)` of the `SpatEntropy` package, and its relative value is $H_{rel}(X) = 0.61/\log 2 = 0.88$, meaning that the level of heterogeneity is equal to 88% of the maximum possible entropy. We can exploit the simulation-based approach to check whether this value marks a significant departure from the reference situation, which, in the case of Shannon's entropy, is the uniform distribution for the two types of ants. We generate 1000 replicates of point patterns with $n = 97$ points, where the two ants' species are randomly assigned to the points with equal probabilities. The resulting 95% confidence interval for Shannon's relative entropy of the ants species is [95.9%; 100%]; the data value is 88%, which indicates a significant departure from the uniform distribution, with a predominance of the *Messor* nests. Shannon's entropy has been used in the past to measure the level of biodiversity in the area; nevertheless, a spatial rearrangement of the nests would not affect the species' relative frequencies. Thus, Shannon's entropy has very limited information power for measuring biodiversity.

4.2 | Evaluation of the Covariate Influence

The environmental covariate is a binary variable with levels *field* and *scrub*, available in the auxiliary file `ants.extra` in the package `spatstat`. We choose a very fine resolution of $100 \times 100 = 10000$ pixels over the window enclosing rectangle, where values are 0 for *field* (fixed as the reference category) and 1 for *scrub*.[6]

We can compute Batty's entropy (2) on the overall dataset, as if they all were simply nests with no distinction, or on each of the two species separately. Once the phenomenon of interest is selected, the probabilities of occurrence p_g are computed for each sub-area defined by the environmental covariate ($G = 2$), using the data relative frequencies over the areas. Then, probabilities are divided by the sub-area size T_g to obtain the intensity for each sub-area $\lambda_g = p_g/T_g$; finally, Batty's entropy is computed over the intensities: the main quantities are reported in Table 1; the intensity is multiplied by 10^6 for easier reading, but the original number is stored for computations. The values are automatically reported in the output of the `R` function `batty` belonging to the `SpatEntropy` package and provided in the Supporting Information.

TABLE 1 | Batty's entropy and related quantities for ants' nests data.

Sub-area	Size	n_g	p_g	$\lambda_g 10^6$	Batty	Rel. Batty
Overall data						
Field	126881.3	59	0.608	2.397	12.969	0.99995
Scrub	87684.8	38	0.392	2.234		
Cataglyphis						
Field	126881.3	21	0.724	2.854	12.931	0.99705
Scrub	87684.8	8	0.276	1.573		
Messor						
Field	126881.3	38	0.559	2.202	12.967	0.99983
Scrub	87684.8	30	0.441	2.516		

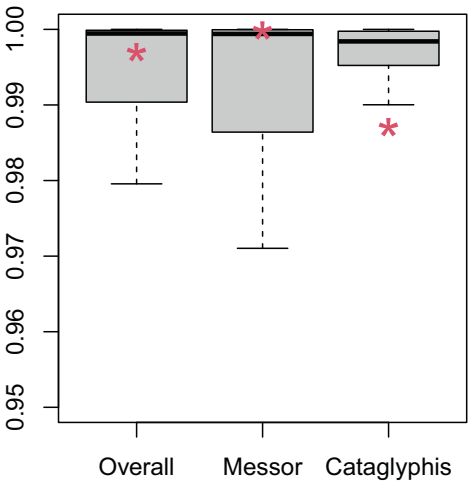

FIGURE 2 | Batty's relative entropy for the covariate-based area partition on 1000 simulated homogeneous datasets. The red dots mark Batty's entropy for the original data.

In order to assess the preference of the ants for a certain soil type, we need to understand whether the two intensity values linked to the two soil types, and thus Batty's entropy, are significantly different from the ones computed over a homogeneous pattern, that is, where points are randomly scattered irrespective of the covariate value. We generated 1000 homogeneous point patterns with the same number of points as the overall, *Messor* and *Cataglyphis* datasets, and computed Batty's relative entropy for all generations. Results are shown in Figure 2, where the interpretation is that if the empirical values lies within the boxplot, then its spatial arrangements is considered as not influenced by the covariate. As can be seen, the overall pattern and the *Messor* nests' distributions are identified as randomly scattered and not affected by the soil type, while the *Cataglyphis* pattern produces a value for Batty's entropy that lies outside the empirical interval computed over the 1000 simulations. This is an alternative, entropy-based way of assessing a significant dependence of the *Cataglyphis* nesting habits on the type of soil, with a preference for the *field* area, which has a higher intensity. Such results provide a motivation for including the covariate in a comprehensive evaluation of the ants' nesting behavior.

4.3 | Evaluation of the Interpoint Interaction

In traditional point pattern exploratory analysis, and in many of the papers mentioned in Section 2, the distance-based summary functions G, F, K and L (Diggle 2014) are a standard tool to point out the evidence of a spatial structure and the presence of interpoint interaction, as opposed to a spatially homogeneous (randomly scattered) pattern. For completeness, we have run several tests based on such functions over the data, and results are reported in the Appendix A. In general, the collection of the results in Figures A1 and A2 of the Appendix A shows that, with the current version of such tools computed over the polygonal window, there is no particular suggestion for interaction within or between ants species, as regards their nesting habits. The only evidence is obtained for the *Messor* species with the G and L functions: since the data curve goes slightly below the gray band, it is a weak indication of a repulsive intra-species behavior.

The alternative approach we propose makes use of the co-occurrence-based entropy measures. For Leibovici's entropy, the new variable Z identifying couples of ants over the window has $R = 4$ categories, that is, (*Cataglyphis, Cataglyphis*), (*Cataglyphis, Messor*), (*Messor, Cataglyphis*), and (*Messor, Messor*). The distance d is specified by the researcher and becomes the radius of a disc that is placed around each of the ants' nests to search for all its neighboring nests and build the co-occurrences; the same operation is repeated for all nests. Eventually, the overall counts of the co-occurrences at distance d, for each of the 4 categories, are available over the area, and their relative frequencies are used as $p(z_r|d)$ in (3). A plot showing the width of the disc is automatically produced by the `leibovici` function of `SpatEntropy`. Couples are counted right- and downward within the chosen distance, as in Leibovici et al. (2014).

Our idea is to use Leibovici's entropy as an alternative approach for evaluating the presence of interpoint interaction. Such measure considers the different types of nests' pairs at various distances, and may identify whether there is an over- or under-representation of one pair category. We can compare the entropy results over the data with those obtained by randomly scattering the same data over the area, to check whether there is a significant deviation from the results expected under a random spatial structure. An advantage of such measure with respect to the distance based functions is that it considers all types of

pairs simultaneously, and all pairs within distance d, not only one nearest neighbor as, for example, the G function; therefore it may detect any type of inter- or intra-species at the same time; deviations may be investigated by checking the pairs' relative frequencies and finding out which category is predominant.

Leibovici's relative entropy is computed for 20 distance values equally spaced between 10 and 200 ft; results are shown in Figure 3. As for the previously computed measures, the empirical results are supported by a band of plausible values under spatial randomness, obtained by computing entropy for 1000 generated datasets with the same number of *Messor* and *Cataglyphis* nests randomly scattered over the region. Leibovici's entropy lies within the bands for most distances, meaning that results are consistent with what expected under spatial randomness, and do not show evidence for interpoint interaction, except for two distances, that is, 70 and 100 ft, marked with red dots in the Figure. A low entropy value indicates a significant predominance of one specific couple of ants' nests. Table 2 reports the detailed results for the two distances with evidence for interpoint interaction. By comparing the frequencies with the one based on the random simulations, we found that the pair that is significantly predominant is *Messor-Messor* (marked by a star in the Table). Since this happens at large distances (70 and 100 ft), the interpretation is that there is a repulsive behavior in the nesting habits of the species *Messor*, that is, that *Messor* ants tend to build their nests at a large distance from other *Messor* nests. Such result overcomes the inconsistencies found in previous works (see Section 2), and is supported by the evidence for repulsion detected with the simulation-based tests over the distance based summary functions G and L (see the Appendix A). Using the proposed tool, we can be precise about the spatial scale at which repulsion may take place. In conclusion, we argue that Leibovici's entropy may provide richer information than the traditional distance based summary functions; for instance, if a significant predominance of *Messor-Messor* pairs were found at small distances, this would be interpreted as attraction instead of repulsion.

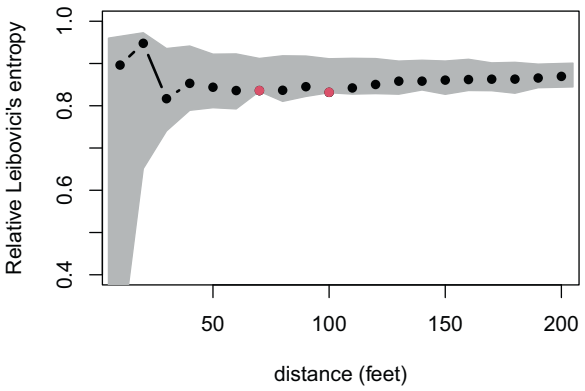

FIGURE 3 | Relative Leibovici's entropy for different co-occurrence distances. The red dots mark the distances where there is a deviation from the results expected under a random spatial scattering of the nests, based on 1000 simulations (gray band).

TABLE 2 | Co-occurrences distributions for Leibovici's entropy at distances 70 and 100 ft.

	Dist = 70 ft		Dist = 100 ft	
Couple	$n^{(70)}$	$f^{(70)}$	$n^{(100)}$	$f^{(100)}$
Cataglyphis-Cataglyphis	22	0.079	39	0.071
Cataglyphis-Messor	46	0.165	96	0.176
Messor-Cataglyphis	59	0.212	115	0.210
Messor-Messor	151	0.543*	297	0.543*
Total	278		547	

Note: The star marks the frequency that determines the signifcant departure from a uniform distribution shown in Figure 3.

4.4 | Inferential Biodiversity Evaluation

A selection of available entropy estimators is computed on the ants' nests data for comparison purposes. The first two follow a frequentist approach: the Maximum-Likelihood estimator (*MLH*) substitutes the probabilities in (1) with the relative frequencies of the ants' species; the Miller-Madow correction (*MMH*) is a well-known adjustment to the *MLH* to improve its bias. Then, we take Zhang's non parametric estimator (*ZhH*) and two Bayesian estimators with different priors to tune how much a uniform distribution for the ants' species is favored, that is, a Laplace's prior (*BLH*) and a Jeffrey's prior (*BJH*). For details, we refer to Paninski (2003); Altieri, Cocchi, and Ventrucci (2023). The existing estimation methods are implemented in `R` by using functions provided by the packages `entropy` and `EntropyEstimation`.

The *BMB* approach is applied to the data by discretizing the original point pattern into a grid of 100×100 cells, each of size 4.14×3.83 feet.[7] For each pixel, we have 0 or 1 nests, so we do not lose information about the exact number and species of the ants' nests. For the *BMB* estimation approach, there is no available package, and the necessary code can be found in the Supporting Information.

Results for a global (absolute and relative) value for each of the literature estimators are shown in black in Figure 4. The *MLH* coincides, by construction, with Shannon's descriptive entropy; the variations of the other estimation proposals produce very similar results, as is the case for most applications (differences may

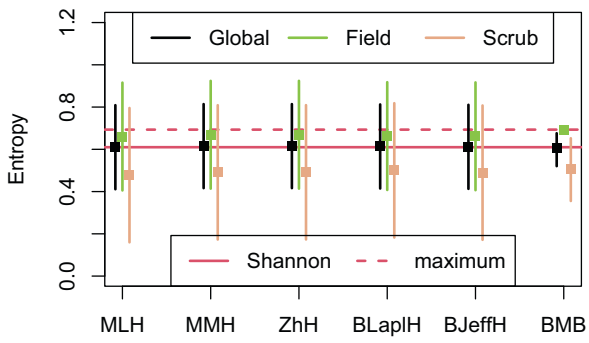

FIGURE 4 | Global entropy estimates and local estimates conditional on the covariate value—point estimates and confidence intervals.

be appreciated for a very large number of species I). The interpretation is analogous to the one of Shannon's entropy, since all measures are based on the species' relative frequencies. Therefore, despite the interesting efforts in improving the performance of entropy estimators, the resulting numbers do not add useful conclusions to the study at hand. The frequentist and non-parametric entropy estimators *MLH*, *MMH*, and *ZhH* are asymptotically normal, provided that the estimators for the probabilities may be considered normal (Paninski 2003). The standard criteria for the normality assumption (detailed in the Appendix A) hold here, therefore we can compute the upper bounds of the confidence intervals for the frequentist and non-parametric entropy estimators, which are the same for all estimators. For coherence and comparison, the upper bounds are also computed for the Bayesian estimators *BLH* and *BJH*. The upper limits of all intervals must be bounded by the entropy maximum $\log 2 = 0.693$.

A stimulating extension of the available approaches consists of estimating local entropies conditional on the covariate value. In this context, local computations are feasible, because the covariate has only two values, and each sub-area contains a lot of ants of both species; therefore, local relative frequencies may be considered reliable. Results for the local estimates are shown as green and orange lines in Figure 4, and show a difference between the biodiversity of the *field* area, higher, and the biodiversity of the *scrub* area, lower. According to the upper bounds for the confidence intervals, though, the difference is not significant between the two sub-areas. Remember that, being upper bounds, results are very conservative; they offer anyway one step forward in the analysis of differences in biodiversity across sub-areas, impossible with descriptive approaches.

The counterpart of the literature estimators in the *BMB* approach is the estimate coming from the independence model, where probabilities p_{ui} of Equation (4) only depend on a global species-specific intercept: the linear predictor is $g_{ui} = \beta_i$. By fitting such model to the data and then drawing $S = 1000$ posterior samples, following the full Bayesian approach outlined in Section 3.3, the *BMB* point estimator is equal to $H^{(ind)}_{BMB} = 0.605$, which, in relative terms, is 0.873. The number, as expected, is very similar to the available estimators. An advantage of the *BMB* approach is the possibility to compute the actual 95% credible interval[8] based on the simulations from the entropy posterior distribution, which is equal to $[0.52; 0.676]$; it does not need truncation because it is, by construction, below the entropy maximum. Note that an additional advantage of a model-based approach is the availability of model selection tools; information criteria[9] may be obtained by the software and are $DIC^{(ind)} = 565.986$ and $WAIC^{(ind)} = 11843.6$.

The model with a dependence on the soil covariate can be considered as the first departure from independence; the linear predictor for the probabilities in Equation (4) is $g_{ui} = c_u\beta_i$, where $c_u = 0$ for *field* and $c_u = 1$ for *scrub*. The estimates give evidence for a difference between the two sub-areas: the entropy for the *field* area is 0.692 [0.691; 0.693], while for the *scrub* area it is 0.508 [0.355; 0.652] with no overlap. Therefore, based on this model we could say that the biodiversity of the ants over the area is influenced by the soil type. For the covariate model, though, $DIC^{(cov)} = 566.441$ and $WAIC^{(cov)} = 11859.2$; since values are larger than those of the independence model, according to such criteria there is no sufficient motivation for a more complicated model.

A summary of the global and covariate-specific results for the literature estimators on ants' nests data, compared to the *BMB* approach, is displayed in Figure 4. The horizontal red line represents the global value of Shannon's entropy as a benchmark, while the dashed red line marks the maximum value for entropy, that is, $\log 2 = 0.693$; then, for each estimator, the black square is the global value, the green square is the entropy of the ants on the *field* area, while the brown square is the local entropy conditional on the *scrub* area. The segments are the upper bounds for the confidence intervals, apart from the *BMB* approach, where the credible intervals are reported. If we focus on the global values, which measure the overall latent biodiversity of the system under the assumption of independence, we see that all conclusions are very similar and close to the observed biodiversity (Shannon's line), but we can link the *BMB* estimator to a far smaller interval than the other ones. Let us now focus on the two local values for each estimator (which, in the case of the *BMB* approach, come from the covariate model). For all the literature estimators, considering the upper bounds for the confidence intervals shows that the difference between the two local values given the soil type is non-significant, while the *BMB* approach gives evidence for a difference in the latent biodiversity between the two sub-areas; moreover, all confidence intervals need to be truncated at the entropy maximum, while the *BMB* interval only covers realistic values and is more reliable in associating the uncertainty to the results.

Lastly, we exploited the complete model of Equation (4), including the covariate effect and a random spatial effect, and we tried the alternative priors *RW2d* and *ICAR* (Lindgren and Rue 2013) introduced in Section 3.3 on the vector collecting the spatial effects $\boldsymbol{\phi_i} = (\phi_{1i}, \dots \phi_{Ni})^T$, applied to each of the ants' species. According to the Information Criteria, the best performing one is the *RW2d* model applied to *Messor*. Therefore, a possible final model for the ants' nests data includes such spatial effect and a dependence on the soil covariate, and has the best performance in terms of $DIC^{(spat)} = 498.689$ and $WAIC^{(spat)} = 11243.61$. Results are shown in Figure 5. The top-left panel shows the entropy surface, whose values range from 0.394 to 0.692, which is pretty consistent to the results and intervals mentioned in the previously presented approaches. This panel shows the behavior of the potential biodiversity of the ants' nests, which depends both on the soil type and on a spatial structure. The other panels decompose and investigate results. The top-middle panel shows the estimate based on the fixed effect coefficients for the influence of the covariate soil type, which has a substantial effect on entropy, as there is a clear difference between *scrub* and *field* (coefficients are significant and confidence intervals do not overlap) Therefore, the soil type plays a role in the biodiversity of the system, with *field* increasing the heterogeneity of the ants and *scrub* decreasing it. By looking at the relative frequencies of ants conditional on the soil type, we see that $f_{Cat} \mid field = 0.373$ and $f_{Mes} \mid field = 0.627$, while $f_{Cat} \mid scrub = 0.184$ and $f_{Mes} \mid scrub = 0.816$. Matching such information with the model output, we can say that the soil type *field* is favored by ants of species *Cataglyphis* with respect to species *Messor*, and this affects the latent biodiversity in that sub-area. On the top-right panel, the spatial effect shows a linear trend that increases from right to left, suggesting that there

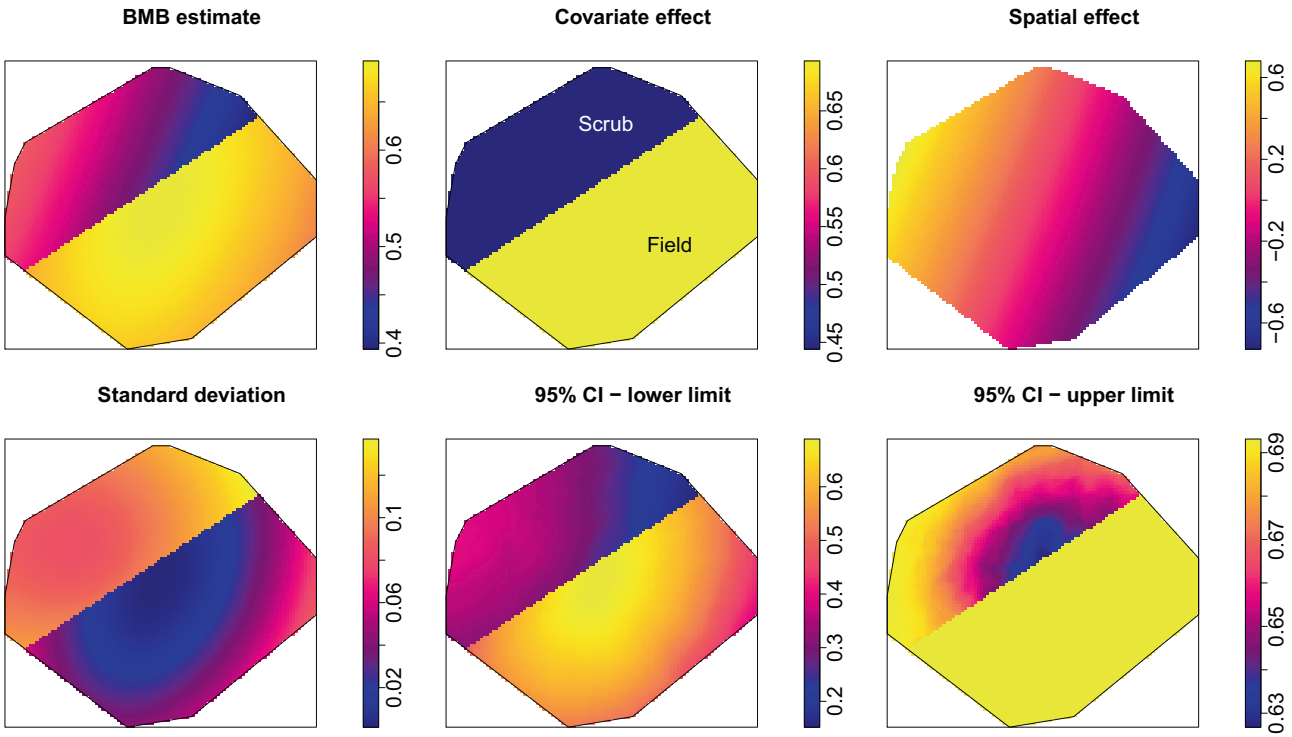

FIGURE 5 | *BMB* approach with covariate and spatial effect — upper panels: Entropy surface, covariate effect, spatial effect; lower panels: Standard deviation, extremes of the 95% credible interval.

may be other unobserved spatial factors influencing the diversity of the ants over the region. It would be interesting to study this further, but this will require collecting more data. However, using the *BMB* approach one is able to visualize spatially varying patterns of biodiversity and this may be advantageous in an exploratory analysis to generate hypotheses on the reasons behind such spatial variation in the entropy levels. Since the spatial effect is so smooth and almost linear, an alternative model with a linear trend on the spatial coordinates was tested, but returned worse values for the Information Criteria. On the bottom-left panel, we see the standard deviation, which gives an idea of the reliability of the estimates. High entropy values are more prone to estimation errors, and values in the center of the area are linked to a smaller error, because their estimate benefits from the information of the neighborhood, which is reduced at the area edges. The middle and right bottom panels show the lower and upper limit of the credible intervals that can be derived from the simulation from the entropy posterior distribution; they give us an idea about the extremes of the plausible values for the latent biodiversity of the system.

Our proposal must be accompanied by a goodness-of-fit test for the selected model. We rely on the standard χ^2 test (Illian et al. 2008): the window is divided into P equivalent areas; for each area, the local component of the test statistic evaluates the difference between the observed and the expected number of nests under the model. Under the hypothesis of a good model for the data, the statistic is distributed as a χ^2 with $(P - b)$ degrees of freedom, where b is the number of parameters. We ran the test with $P = 3,4,5,6$, both on the overall dataset and on each species, and we obtained consistent and satisfactory results about the goodness of the model, which is a further step forward with respect to the literature works summarized in Section 2.

5 | Concluding Remarks

In this paper, we used a marked point pattern dataset of two ants' species over an irregular region to illustrate an approach to biodiversity evaluation based on entropy measures, with novel use and extensions of the existing methods; the proposed approach can provide new insights on the considered case study and contribute to the general area of biodiversity assessment.

Harkness and Isham (1983) ants' nests data have been studied by several papers over the years; the methods used are now obsolete and require simplifications, such as the reduction of the observation window to a rectangle, or the use of basic models; moreover, conclusions over the data behavior in the literature are inconsistent and lack explanation. The present paper faces the research questions under a different perspective, and exploits recent methodology and computational tools to overcome the difficulties. Using Batty's entropy in a novel way, that is, by partitioning the area according to the soil type, we evaluated the influence of the covariate over the nesting habits. We propose a new simulation-based assessment of the departure from the situation of independence from the covariate, which allowed to show that *Cataglyphis* ants have a significant preference for the soil type *field*. Then, we proposed a new way of using Leibovici's entropy to evaluate interpoint interaction, the most discussed research question about the ants' data. We argue that

TABLE 3 | Summary of the approaches' characteristics: Each column shows whether they allows for covariate inclusion, space consideration, uncertainty assessment and checking of the underlying assumptions.

Entropy		Covariates	Space as variable	Uncertainty	Checking
Descr.	Batty	Yes, limited	No	No	No
	Leibovici	No	Yes, limited	No	No
Inf.	Literature	Yes, limited	No	Yes, limited	No
	BMB	Yes	Yes	Yes	Yes

Leibovici's entropy is a competitive alternative to the traditional distance-based functions for measuring interpoint interaction, as it is not limited to one or few nearest neighbors of a specific type, rather it considers all couples of categories jointly, within a specific distance. We found that there is no inter-species interaction, which is supported by the computation of the distance-based summary functions over the pattern (shown in Figure A1 of the Appendix A); this can be motivated by the very different food requirements of the two ants' species (Isham 1984) and by the absence of a predator–prey relationship, since *Cataglyphis* ants eat dead insects, but do not kill them. The novel proposal of a simulation-based evaluation of Leibovici's entropy allowed us to detect a repulsive behavior within the *Messor* species, most likely due to competition for food. Altogether, these findings highlight the need for a more formal approach based on modeling, able to grasp the general structure of the ants' nesting habits considering not only the relationship between the species, but also the influence of the soil type and of other unknown factors.

Our *BMB* approach allows to fit complex models with spatial effects and covariates, which improves our understanding of the entropy of the system, compared to descriptive approaches; such models have never been applied to the ants' nests data, and have substantial advantages, because their results can be explained in detail and all assumptions can be checked. We can conclude (Figure 5) that the ants' nests latent biodiversity covers a pretty wide range of values from average to very high, and depends on the soil type, with *scrub* decreasing the biodiversity level, as it sees a predominance of *Messor* over *Cataglyphis*. The model allows to show the behavior of the spatial effect that accounts for unmeasurable sources of spatial heterogeneity. For uncertainty evaluation, we plot the standard deviation of the estimates, that depends on the extent of the spatial neighborhood system set in the model, and the point-wise credible intervals to help the researcher understand the plausible values for the latent biodiversity of the ants' system. Moreover, after selecting the best model for the data with the Information Criteria, a goodness-of-fit test is run to check the adequacy of the model to the data. The whole set of diagnostics allows to safely disseminate the results, overcoming another weak aspect of previous works on the same data.

A few more general comments must be given, that extend the contribution of this work beyond the results over the ants' nests data. We offer a way to properly use descriptive and inferential approaches; we propose a joint use of separate descriptive approaches in the literature, that is, partition- and distance-based entropies, with new insights on their interpretation for biodiversity assessment. A combination of the two can provide an exhaustive perspective for the description of the heterogeneity and biodiversity of a system. We also propose a new, intuitive, but rigorous, way of interpreting the significance of the results thanks to simulations. In addition, we shed a new light on literature entropy estimators, which are currently used at a global level for the whole dataset under study. We have shown that they may be used locally and provide interesting information when applied to sub-areas defined by covariate levels/quantiles. We also show how the *BMB* approach outperforms the literature entropy estimators. A major limit of such measures is that they are computed under the strong assumption of independence, as the whole dataset contributes to the computation of global relative frequencies for the ants' species, that substitute the probabilities p_i as the main components of Shannon's entropy (1). Even when relative frequencies are computed conditional on the type of soil in order to estimate local entropies, the assumption is that, conditional on each covariate value, there is independence between nests. This appears as an unrealistic simplification in the study of ants' biodiversity, which is nevertheless taken in all the available approaches. The more flexible *BMB* approach is able to grasp the actual structure of a complex phenomenon where independence should not be assumed, by improving the way probabilities are estimated. A final comparative Table 3 summarizes the pros and cons of each entropy-based measure considered in this work.

The availability of advanced computational tools is a final important aspect of the present work. Thanks to the new release of our `SpatEntropy` package for the `R` software, user friendly computational tools are provided: for Batty's entropy, results are immediate; Leibovici's entropy may take a few minutes for large datasets (details about the computational time may be found in Altieri, Cocchi, and Roli 2023). The computational complexity may be the only drawback of the *BMB* approach, as it needs some more code-writing by the user. Nevertheless, the provided Supporting Information facilitates the reproduction of the main analyzes in the present work with a minimum effort by the user. Model-based entropy estimation is immediate for fixed effect models, and may take up to a few minutes for spatial models; thanks to INLA's computational efficiency, it is very easy to fit many model options and select the best performing one for the case study. We believe this might be of great help for applied scientists interested in well-grounded results in biodiversity monitoring.

Acknowledgments

Open access publishing facilitated by Universita degli Studi di Bologna, as part of the Wiley - CRUI-CARE agreement.

Data Availability Statement

The data that supports the findings of this study are available in the Supporting Information of this article.

Endnotes

[1] http://www.cbd.int.

[2] https://CRAN.R-project.org/package=SpatEntropy.

[3] For the model-based approach parameter estimation of Section 3.3, `R-INLA` is the reference software for working with INLA (Rue, Martino, and Chopin 2009). Such package can be downloaded from the website www.r-inla.org, where examples and tutorials are also available.

[4] https://cran.r-project.org.

[5] https://CRAN.R-project.org/package=SpatEntropy.

[6] Missing values are assigned to the pixels outside the polygonal boundary (stored as `NA` values in `R`).

[7] Note that we do not have to worry about building square cells, as the software deals with any type of cells. The chosen grid resolution is the same as the resolution of the covariate image.

[8] Credible intervals are based on posterior densities, and are the Bayesian counterpart of the more popular frequentist confidence intervals: they give a plausible range of values for the quantity of interest.

[9] Information Criteria are well-established methods for model selection, as they are a trade-off between good fitting to the data (likelihood value) and complexity (number of parameters). The rule for using Information Criteria for model selection is "the smaller, the better." *DIC* stands for *Deviance Information Criterion* and *WAIC* is short term for *Watanabe-Akaike Information Criterion*. See Casella and Berger (2021) for details.

References

Altieri, L., D. Cocchi, and G. Roli. 2018. "A New Approach to Spatial Entropy Measures." *Environmental and Ecological Statistics* 25, no. 1: 95–110.

Altieri, L., D. Cocchi, and G. Roli. 2019. "Advances in Spatial Entropy Measures." *Stochastic Environmental Research and Risk Assessment* 33, no. 4: 1223–1240.

Altieri, L., D. Cocchi, and G. Roli. 2021. "Spatial Entropy for Biodiversity and Environmental Data: The R-Package SpatEntropy." *Environmental Modelling and Software* 144: 105–149.

Altieri, L., D. Cocchi, and G. Roli. 2023. "Efficient Computation of Spatial Entropy Measures." *Entropy* 25: 1634.

Altieri, L., D. Cocchi, and M. Ventrucci. 2023. "Model-Based Entropy Estimation for Data With Covariates and Dependence Structures." *Environmental and Ecological Statistics* 30: 477–499.

Altieri, L., A. Farcomeni, and D. Alunni-Fegatelli. 2023. "Continuous Time-Interaction Processes for Population Size Estimation, With an Application to Drug Dealing in Italy." *Biometrics* 79: 1254–1267.

Baddeley, A. J., E. Rubak, and R. Turner. 2015. *Spatial Point Patterns: Methodology and Applications With R*. London: Chapman and Hall/CRC Press.

Baddeley, A. J., and R. Turner. 2000. "Practical Maximum Pseudolikelihood for Spatial Point Patterns." *Australian and New Zealand Journal of Statistics* 42: 283–322.

Baddeley, A. J., and R. Turner. 2005. "Spatstat: An R Package for Analyzing Spatial Point Patterns." *Journal of Statistical Software* 12, no. 6: 1–42.

Baddeley, A. J., and R. Turner. 2006. "Modelling Spatial Point Patterns in R." *Case Studies in Spatial Point Process Modeling* 185: 23–74.

Barmoudeh, L., H. Baghishani, and S. Martino. 2022. "Bayesian Spatial Analysis of Crash Severity Data With the INLA Approach: Assessment of Different Identification Constraints." *Accident Analysis and Prevention* 167: 106570.

Batty, M. 1974. "Spatial Entropy." *Geographical Analysis* 6: 1–31.

Batty, M. 1976. "Entropy in Spatial Aggregation." *Geographical Analysis* 8: 1–21.

Batty, M. 2010. "Space, Scale, and Scaling in Entropy Maximizing." *Geographical Analysis* 42: 395–421.

Bivand, R., V. Gómez-Rubio, and H. Rue. 2015. "Spatial Data Analysis With R-Inla With Some Extensions." *Journal of Statistical Software* 63, no. 20: 1–31.

Bondesson, L., and J. Fahlén. 2003. "Mean and Variance of Vacancy for Hard-Core Disc Processes and Applications." *Scandinavian Journal of Statistics* 30: 797–816.

Casella, G., and R. L. Berger. 2021. *Statistical Inference*. Boston, MA: Cengage Learning.

Cover, T. M., and J. A. Thomas. 2006. *Elements of Information Theory*. Second ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Diggle, P. J. 2014. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Third ed. Boca Raton: Taylor & Francis Group.

Drechsler, M. 2020. "Model-Based Integration of Ecology and Socio-Economics for the Management of Biodiversity and Ecosystem Services: State of the Art, Diversity and Current Trends." *Environmental Modelling and Software* 134: 104892.

Durrett, R. 2004. *Probability: Theory and Examples*. 3rd ed. New York, NY: Cambridge University Press.

Fisher, R., R. A. O'Leary, S. Low-Choy, K. Mengersen, and M. J. Caley. 2012. "A Software Tool for Elicitation of Expert Knowledge About Species Richness or Similar Counts." *Environmental Modelling and Software* 30: 1–14.

Harkness, R., and V. Isham. 1983. "A Bivariate Spatial Point Pattern of Ants' Nests." *Applied Statistics* 32: 293–303.

Högmander, H., and A. Särkkä. 1999. "Multitype Spatial Point Patterns With Hierarchical Interactions." *Biometrics* 55: 1051–1058.

Hoskins, A. J., T. D. Harwood, C. Ware, et al. 2020. "Bilbi: Supporting Global Biodiversity Assessment Through High-Resolution Macroecological Modelling." *Environmental Modelling and Software* 132: 104806.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. 2008. *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester: Wiley.

Isham, V. 1984. "Multitype Markov Point Processes: Some Approximations." *Proceedings of the Royal Society of London, Series A* 391: 39–53.

Leibovici, D. G. 2009. "Defining Spatial Entropy From Multivariate Distributions of Co-Occurrences." In *Cosit 2009*, Lecture Notes in Computer Science, edited by K. S. Hornsby et al., vol. 5756, 392–404. Berlin: Springer.

Leibovici, D. G., C. Claramunt, D. LeGuyader, and D. Brosset. 2014. "Local and Global Spatio-Temporal Entropy Indices Based on Distance Ratios and Co-Occurrences Distributions." *International Journal of Geographical Information Science* 28: 1061–1084.

Leinster, T., and C. Cobbold. 2012. "Measuring Diversity: The Importance of Species Similarity." *Ecology* 93, no. 3: 477–489.

Lindgren, F., and H. Rue. 2013. "Bayesian Spatial Modelling With R-INLA." *Journal of Statistical Software* 63, no. 19: 1–25.

Magurran, A. E. 2004. *Measuring Biological Diversity*. Oxford: Blackwell Publishing.

Martinez-Minaya, J., M. Cameletti, D. Conesa, and M. G. Pennino. 2018. "Species Distribution Modeling: A Statistical Review With Focus in

Spatio-Temporal Issues." *Stochastic Environmental Research and Risk Assessment* 32: 3227–3244.

O'Neill, R. V., J. R. Krummel, R. H. Gardner, et al. 1988. "Indices of Landscape Pattern." *Landscape Ecology* 1: 153–162.

Paninski, L. 2003. "Estimation of Entropy and Mutual Information." *Journal of Neural Computation* 15: 1191–1253.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rue, H., and L. Held. 2005. *Gaussian Markov Random Fields. Theory and Applications*. Boca Raton: Chapman and Hall.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society B* 71, no. 2: 319–392.

Särkkä, A. 1993. *Pseudo-Likelihood Approach for Pair Potential Estimation of Gibbs Processes*, Number 22 in Jyvaskyla Studies in Computer Science, Economics and Statistics. Finland: University of Jyvaskyla.

Scarnati, L., F. Attorre, A. Farcomeni, F. Francesconi, and M. D. Sanctis. 2009. "Modelling the Spatial Distribution of Tree Species With Fragmented Populations From Abundance Data." *Community Ecology* 10: 215–224.

Serafini, F. 2019. "Multinomial Logit Models With INLA." R-INLA Tutorial. https://inla.r-inla-download.org/r-inla.org/doc/vignettes/multinomial.pdf.

Shannon, C. 1948. "A Mathematical Theory of Communication." *Bell System Technical Journal* 27: 379–423. 623–656.

Takacs, R., and T. Fiksel. 1986. "Interaction Pair-Potentials for a System of Ants' Nests." *Biometrical Journal* 28: 1007–1013.

Ventrucci, M., D. Cocchi, G. Burgazzi, and A. Laini. 2020. "PC Priors for Residual Correlation Parameters in One-Factor Mixed Models." *Statistical Methods and Applications* 29: 745–765.

Zhang, Z. 2012. "Entropy Estimation in Turing's Perspective." *Journal of Neural Computation* 24: 1368–1389.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix A

A.1 | Summary Functions for Interpoint Interaction

In traditional point pattern exploratory analysis, and in many of the papers mentioned in Section 2, the distance-based summary functions G, F, K and L (Diggle 2014) are a standard tool to evaluate the spatial structure and the presence of interpoint interaction with respect to a spatially homogeneous (randomly scattered) pattern. Simulation based tests can be run to assess whether the pattern shows a clustering or repulsive behavior, which has been used to evaluate interaction within and between ants species (Harkness and Isham 1983). Figure A1 shows the results for the overall pattern and for the two sub-patterns separately for the four distance-based functions as computed by the `spatstat` package (see, e.g., `Fest` and `Gest`). Such plots are meant to investigate overall interpoint interaction (first column), and intra-species interaction (second and third column). Figure A2 shows the "recent cross" versions of the functions, which are specifically meant to evaluate interaction between the two species of ants, by measuring the distance between nests of the two species. The empirical functions based on the ants data (black line) always lie entirely, or almost entirely, within the gray bands, which represent the plausible range of values for a spatially homogeneous process.

A.2 | Criteria for Normality Assumption of the Entropy Estimators

The entropy estimators may be assumed as normal provided that the probabilities involved in the computations may be considered normal, that is, under the assumption of a sufficiently large number of observations for satisfying the Central Limit Theorem. Two criteria are commonly used in the literature (see, e.g., Durrett 2004; Casella and Berger 2021) to decide whether the normality assumption holds. The first one is $np > 5$ and $n(1-p) > 5$, with n being the sample size and p being the proportion of any of the categories. In the present study, if p is the proportion of *Messor* ants' nests, that is, $p = 0.701$, then $np = 68$ and $n(1-p) = 29$; if p refers *Cataglyphis* nests, namely $p = 0.299$, then $np = 29$ and $n(1-p) = 68$; all numbers are larger than 5. A second standard criterion is that the interval $\left[p - 3\sqrt{\frac{p(1-p)}{n}}; p + 3\sqrt{\frac{p(1-p)}{n}}\right]$ must lie wholly within the interval $[0, 1]$. For the *Messor* species the interval is $[0.562; 0.840]$, while for the *Cataglyphis* ants it is $[0.160; 0.438]$. Under the normality assumption, we can compute the upper bound for the estimators' variance, which, for $I = 2$ species, is $1/n = 1/97 = 0.01$.

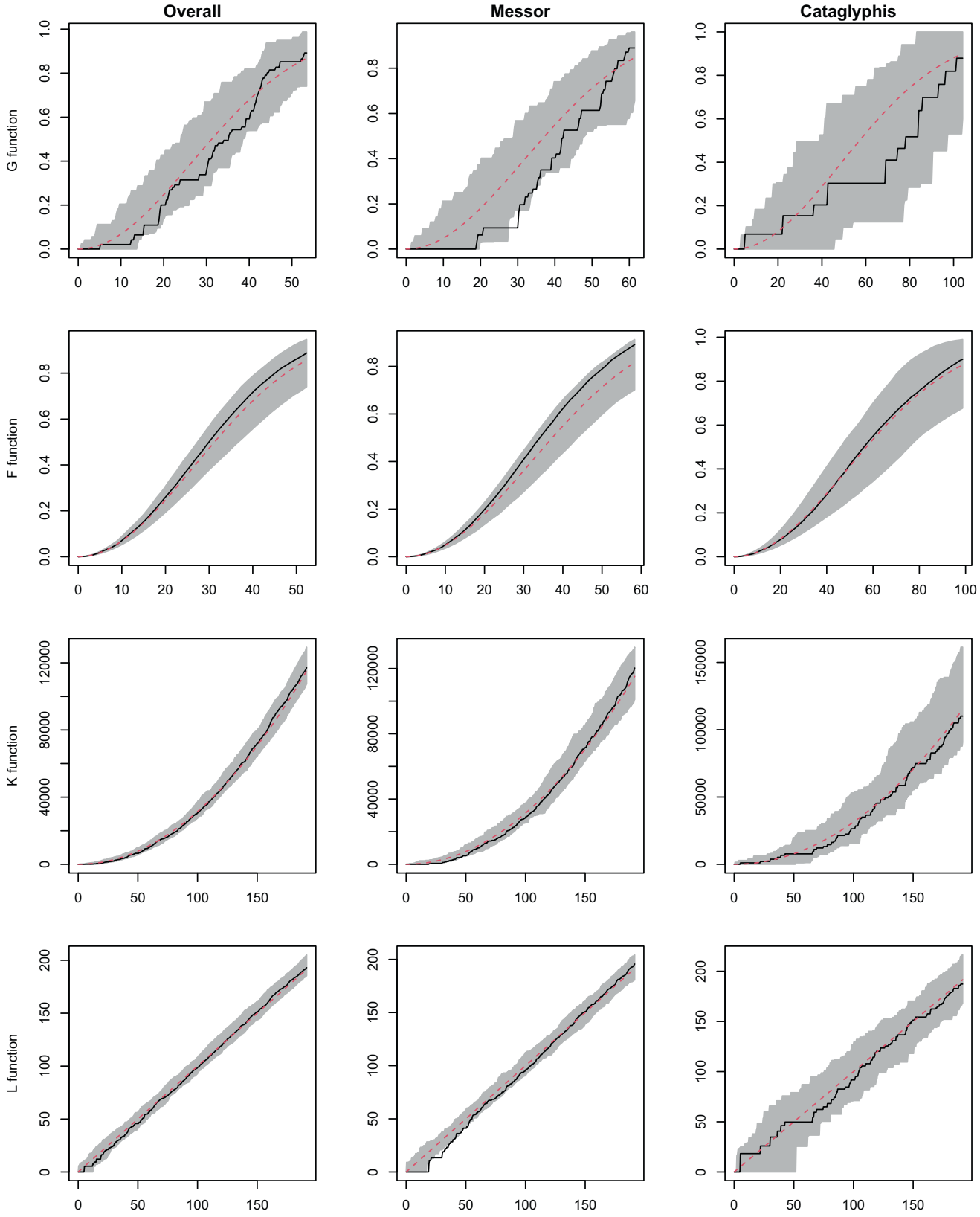

FIGURE A1 | Simulation-based tests for overall and intra-species interpoint interaction based on summary distance functions.

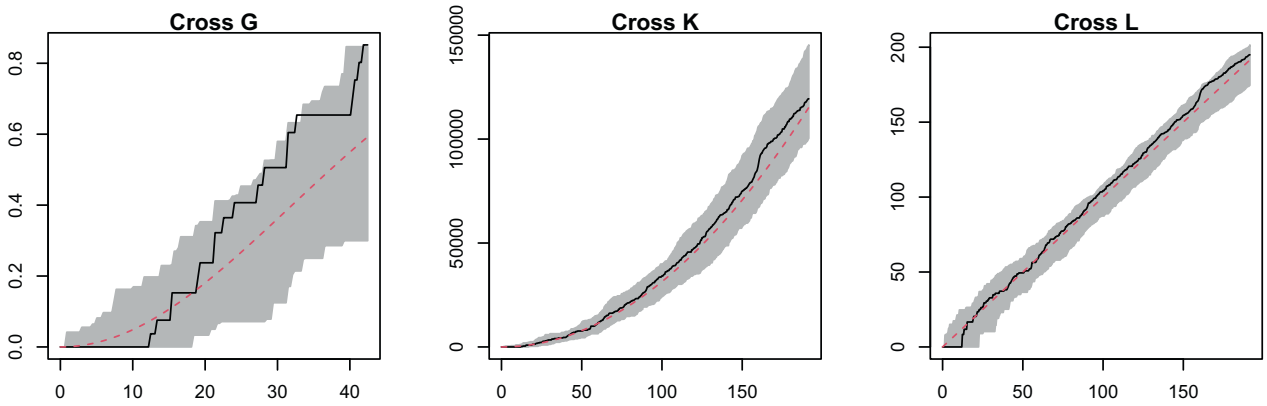

FIGURE A2 | Simulation-based tests for inter-species interpoint interaction based on summary distance functions.