Model-based entropy estimation for data with covariates and dependence structures

Linda Altieri[1] [ORCID] **· Daniela Cocchi[1] · Massimo Ventrucci[1]**

Abstract
Entropy is widely used in ecological and environmental studies, where data often present complex interactions. Difficulties arise in linking entropy to available covariates or data dependence structures, thus, all existing entropy estimators assume independence. To overcome this limit, we take a Bayesian model-based approach which focuses on estimating the probabilities that compose the index, accounting for any data dependence and correlation. An estimate of entropy can be constructed from the model fitted values, returning an observation-specific measure of entropy rather than an overall index. This way, the latent heterogeneity of the system can be represented by a curve in time or a surface in space, according to the characteristics of the survey study at hand. An empirical study illustrates the flexibility and interpretability of our results over temporally and spatially correlated data. An application is presented about the biodiversity of spatially structured rainforest tree data.

Keywords Bayesian multinomial regression · Biodiversity · Correlated data · Entropy estimation · Rainforest trees

1 Introduction

Ecology can be seen as a quantitative discipline: ecologists use data to "understand how populations and communities are structured and to make predictions about how ecological systems might change over space or time, or in response to external forces" (Magurran 2004). Ecological issues can be faced from a spatial and/or

---

✉ Linda Altieri
linda.altieri@unibo.it

Daniela Cocchi
daniela.cocchi@unibo.it

Massimo Ventrucci
massimo.ventrucci@unibo.it

1 Department of Statistical Sciences, University of Bologna, via Belle Arti 41, 40126 Bologna, Italy

temporal point of view, by documenting patterns, but also by considering the under-lying processes. A current challenge is assessing the extent of short-term and local changes in ecological communities, and scientists have a limited understanding of such processes (Magurran 2004). Studies on biological diversity are a fundamental part of every ecological survey, for biodiversity plays a crucial role in the deliv-ery of a range of ecosystem services such as natural harvests, carbon sequestration, pollination and soil formation (Frosini 2004). At the same time, it is threatened by climate changes, over-harvesting, pollution, habitat loss and invasive species, due to human activities. Widespread concern about the extent of habitat and species loss led the United Nations to declare 2010 as its International Year of Biodiversity, rec-ognizing that the conservation of biological diversity is a major policy issue. The UN Convention on Biological Diversity set a target to significantly reduce the rate of biodiversity loss, and this has stimulated the development of a large set of diver-sity indices. One of the most popular descriptive measures is entropy, as defined by Shannon (1948); in the remainder of the paper, the term "entropy" follows Shan-non's definition.

Shannon's entropy is a successful measure in many fields, able to synthesize sev-eral concepts in a single number: entropy, information, heterogeneity, surprise, con-tagion. The flexibility of such index and the possibility to describe any kind of data, including categorical variables, motivate its diffusion across applied fields such as geography, ecology, biology and landscape studies, whenever synthetic measures of heterogeneity may be of interest (Frosini 2004). In biodiversity studies, the advan-tages of Shannon's entropy lie on its ability to synthesize both information about the number of species over an area and information about their distribution: the index quantifies the uncertainty in predicting the species identity of an individual that is taken at random from the dataset. Entropy increases when species richness (the number of species) and/or species evenness (the even distribution of species) increase, while it decreases when one or few species are predominant; it is com-monly used to evaluate α diversity, defined as the mean diversity of species in differ-ent sites or habitats within a local scale (Anderson et al. 2011).

Typically, in environmental and ecological applications, the amount of data for different categories (species) depends on several factors, such as environmental covariates, spatial location, temporal structures; a major drawback of Shannon's entropy is that it cannot account for auxiliary information or data dependence. Recently, scientists have shown interest in the topic of data structures and entropy; for example, ecological studies attempt to find a relationship between quantities related to entropy, such as species abundance and richness, and environmental fac-tors or temporal/spatial effects using regression or mixed models (Ventrucci et al. 2020; Martinez-Minaya et al. 2018). Such studies, though providing interesting con-tributions, do not consider the relationship between the occurrence probabilities of each species and covariates or effects, besides, they do not account for absent spe-cies and for the possibility that species are not the same across observation sites. Besides, Gelfand and Shirota (2019) and Gelfand (2022) raise the issue of pres-ence/absence data that ignore the abundance on the spatial unit. A further limit of recent studies trying to compute local entropy measure is that they may be unreli-able if they are based on very few local data. In addition, failing to consider data in

neighbouring areas (or time points) produces a noisy set of entropy values, which is usually non-informative and difficult to interpret. Since entropy measures are often employed to describe spatial data, the inclusion of spatial information has been the target of intensive research (Batty 1976; Leibovici et al. 2014; Altieri et al. 2018, 2019). Gelfand (2022) reviewed recent literature on the distribution of species, highlighting the need to incorporate spatial dependence in indices such as entropy: very little of the literature is spatial, despite the evident dependence within site as well as across sites. In biodiversity studies, space is investigated with β diversity indices, which look for variation in the identities of species among sites, but Shannon's entropy is not one of them. Anderson et al. (2011) highlight the need for diversity measures able to detect latent processes driving observed patterns.

Latent processes regard a different branch of studies, which estimate the entropy of a phenomenon. The standard approach relies on the Maximum Likelihood (ML) estimator, which substitutes the unknown probability distribution of interest with relative frequencies and performs well when independence is an acceptable assumption (see Antos and Kontoyiannis (2001) and details in Appendix A). To our knowledge, no study faces the task of estimating entropy for data presenting dependence on available covariates, spatial/temporal association or other types of dependence. In the literature of entropy estimation, the assumption of independence is never relaxed; on the other side, spatial entropy studies never consider the aspect of inference.

The aim of this paper is to present a new, model-based approach to entropy estimation. We take a new perspective that moves from an overall measure of entropy under the assumption of independence, which is untenable in most survey studies, to provide an estimate of entropy that can vary in time, space, or according to available covariates. We build a Bayesian hierarchical model for multinomial data, where the species occurrence probabilities are modelled based on available information, such as environmental covariates and/or spatial coordinates. The proposed models can be conveniently implemented in R-INLA (Integrated Nested Laplace Approximation, Rue et al. (2009)), which is particularly useful from a user perspective, as it allows to handle a wide range of spatial and temporal random effects to model dependence. The use of a multinomial response is a key novelty of our work, that exploits the information about the abundance of each category (species) in real data studies, and avoids reducing data to presence/absence, with related issues raised by Gelfand and Shirota (2019). After obtaining a posterior distribution for all parameters, the posterior distribution of entropy is straightforward, since the entropy formula only depends on the probabilities (see Equation (1)). Any diversity index based on probabilities can also be derived, such as Hill's number (Hill 1973). A point entropy estimator can be, e.g., the mean of the entropy posterior distribution. Our approach also provides standard errors to quantify the estimation uncertainty, addressing one of the present challenges of spatial ecological analysis (Gelfand 2022). For time series data, the estimate can be represented by a curve, that takes into account the temporal neigbourhood of each observation point, smoothes out any random variation in the data and allows to grasp the general behaviour of entropy. In the spatial context, the estimation output is a spatial surface for the entropy over the area under study.

The proposed framework is suitable for different types of ecological surveys and of data. When biodiversity studies involve an observation area, they may be based on three types of spatial data. Sometimes the researcher does not have the precise location of individuals and only has "patches" (irregular areas or a grid) with data aggregated as local abundances: an example is raster data in discrete space. In other studies, the researcher fixes the observation points, collects data at those points only and then needs to interpolate the information over the rest of the area; these are geostatistical data studies. Other works are based on data collected over the whole observation area and record their exact location: these are point process data, where points are random and part of the response. A usual issue with point process data is computational complexity. The most common option is to discretize space with a very fine grid: the approximated model is known to converge to the true process (Moller and Waagepetersen 2007) and results are very accurate in most applications. Many papers on point process models, though continuous in theory, turn to very fine gridding when it comes to practical computations (Gelfand and Shirota 2019; Moller and Waagepetersen 2007), and the widest and most up-to-date R package for point processes, spatstat (Baddeley et al. 2015) largely relies on grid approximations. In Sect. 4, we show an application with the gridding approach, which might be a satisfying trade-off between the complexity and accuracy of the results. Moreover, gridding overcomes the issues of spatial misalignment due to the fact that covariates are available at a grid resolution (Gelfand 2022). Our approach solves the grid limit of reducing local abundances to presence/absence data: we are able to take into account the exact amount of individuals for each species. In addition, if wished, the proposed model includes a sophisticated option that avoids gridding, by estimating models with the Stochastic Partial Differential Equation (SPDE) approach, available with INLA (Krainski et al. 2019).

Once we obtain entropy estimates, we focus on the interpretation for biodiversity data. Biological diversity is related to the number of species: if a lot of individuals are present, but they all belong to the same species, there is no biodiversity; if there are few individuals, but they all belong to different species, biodiversity is larger. In descriptive studies, entropy and biodiversity are considered as synonyms; when it comes to evaluation of the underlying process, the two concepts are not the same thing, though strongly related. Entropy estimates evaluate heterogeneity, which is not analogous to the *observed* biodiversity of a dataset, rather it describes the *latent* biodiversity of the process under study. When entropy is low, one (or few) of the species probabilities is predominant, i.e. no matter how many individuals are present, they tend to belong to one (or few) species. Therefore, low entropy corresponds to low biodiversity, both observed and latent. When entropy is high, species probabilities are more evenly distributed; the *observed* biodiversity may be low if very few observations are available, but the *latent* biodiversity of the system is high. Comments along the paper follow this interpretation of entropy.

The paper is organized as follows. Section 2 presents the methodology for deriving the entropy value, curve or surface for both binary and multinomial data. Section 3 illustrates our proposal over a selection of scenarios and highlights its advantages in information and interpretability. The real data example is presented in Sect. 4. Finally, Sect. 5 summarizes the work and contains concluding comments.

Additional details about both the state of the art in entropy estimation and the empirical study are in the Appendices.

2 Models and methods

Let X be a categorical variable of interest with I categories (species), $i = 1, \dots, I$. Shannon's entropy of X is (Shannon 1948):

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

where $p(x_i)$ is the probability of occurrence of the i-th category, and $\boldsymbol{p}_X = (p(x_1), \dots, p(x_I))'$ is the probability mass function of X. Entropy ranges in $[0, \log I]$, where 0 is obtained when $p(x_{i^*}) = 1$ for some category i^* and $p(x_i) = 0$ for $i \neq i^*$, and $\log I$ derives from $p(x_i) = 1/I$ for all i. Shannon's entropy is a particular case of the increasingly popular Hill's number (Hill 1973), which weights each species by its relative abundance: $D(X) = \left(\sum_{i=1}^{I} p(x_i)^q \right)^{1/1-q}$, where q defines the weights: the limit for $q = 1$ is the exponential of Shannon's entropy. Our results hold for any Hill's number; we refer to Sect. 5 for additional comments.

Under the perspective of estimation, a stochastic process is assumed to generate data according to an unknown probability function; one realization of the process is observed and employed to estimate the underlying entropy. In this Section, we show how to relax the assumption of independence by building a flexible model for the estimation of the main components of an entropy index $p(x_i)$ for $i = 1, \dots, I$; then, we derive our entropy estimator.

2.1 Bayesian multinomial regression

For simplicity of presentation, let us start with the binary case $I = 2$ (e.g. species A and B), and assume $X_u \sim \text{Ber}(p_u)$ with observations indexed by $u = 1, \dots, n$, where p_u is the probability of occurrence of species A at u. If we are dealing with observations in space or time, we can think of u as the spatial or temporal location index. Let us consider the associated binomial variable $Y_u \in \{0, \dots, N_u\}$ representing the number of individuals of species A among the total number of individuals N_u observed at u. We assume the model:

$$Y_u \mid g_u \sim \text{Bin}\left(N_u, p_u = \frac{\exp(g_u)}{1 + \exp(g_u)} \right), \tag{2}$$

where g_u is the probability of occurrence in the logit scale, modelled as $g_u = z_u' \boldsymbol{\beta}$, where z_u is a vector of observed covariates at location u.

When the number of species $I > 2$, an extension to the multinomial logit model arises naturally. Let $Y_{u1}, \dots, Y_{uI}$ be the number of individuals of each species at site u, that may be equal or greater than 0 with the constraint $\sum_{i=1}^{I} Y_{ui} = N_u$, and

probabilities $p_{u1}, \ldots, p_{uI}$, with $\sum_{i=1}^{I} p_{ui} = 1$. The generic species occurrence probability can be expressed as:

$$p_{ui} = \frac{\exp(z'_{ui}\boldsymbol{\beta_i} + \phi_{ui})}{\sum_i \exp(z'_{ui}\boldsymbol{\beta_i} + \phi_{ui})}, \quad i = 1, \ldots, I, \tag{3}$$

where the linear predictor is modelled as a function of covariates and random effects. Precisely, z'_{ui} are species-specific covariates observed at u and $\phi_{u1}, \ldots, \phi_{uI}$ are species-specific random effects at location u; some modelling options for the random effects are discussed in Sect. 2.2. Note that both the covariates and the fixed effects in Equation (3) are species-specific; for identifiability purposes, the linear constraints $\sum_{i=1}^{I} \beta_i = 0$ and $\sum_{i=1}^{I} \phi_{ui} = 0$ are needed. In ecology studies, it is common to work with environmental and climate variables that only vary at observation level (e.g. space or time); thus, we can assume $z'_{ui} = z'_{u}$ in Equation (3).

Fitting a multinomial model with R-INLA (Rue et al. 2009) needs additional care, as detailed in Serafini (2019). INLA requires that each response depends on the linear predictor through a linear combination of its elements, a condition that does not hold for (3). We exploit the Multinomial-Poisson transform (Baker 1994), which re-expresses the multinomial likelihood into a Poisson likelihood at the cost of adding model parameters. For a tutorial on the implementation of the multinomial-Poisson transform in INLA, see Serafini (2019); for further details on the required constraints in multinomial regression and how these can be implemented in INLA, see Barmoudeh et al. (2022).

2.2 Modelling temporal and spatial dependence

Temporal/spatial random effects can be introduced in the linear predictor g_{ui} in (3) to capture variations over time/space of the species-specific occurrence probabilities, which lead to the evaluation of the temporal/spatial heterogeneity in the latent entropy. INLA allows user-friendly implementation of various types of spatial and temporal random effects, with the inclusion of constraints to ensure model identifiability. In the following, we outline standard available options to model temporal and spatial variations in entropy levels.

Let us assume that u is the time index in the linear predictor in (3), so that $\boldsymbol{\phi_i} = (\phi_{1i}, \ldots, \phi_{ni})'$ becomes the vector of temporal effects for species i. Temporal dependence in the species occurrence probabilities can be modelled using first or second-order random walk models, i.e. RW1 and RW2 (Rue and Held 2005). A RW1 is suitable when the interest is to penalize deviations from an overall constant level. Under a RW1, the joint density of $\boldsymbol{\phi}$ (dropping the species index i in the right-hand side of (4) for simplicity of notation) is:

$$\pi(\phi|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{u=1}^{n-1}\left(\phi_{u+1} - \phi_u\right)^2\right), \tag{4}$$

where τ is the precision parameter controlling/penalizing the amount of deviation from the overall constant. The RW2 permits to penalize second order differences between temporally-adjacent random effects, hence it is more suitable when smoothness is anticipated. Under a RW2, the joint density of $\boldsymbol{\phi}$ is (dropping the species index i):

$$\pi(\phi|\tau) \propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\sum_{u=1}^{n-2}\left(\phi_u - 2\phi_{u+1} + \phi_{u+2}\right)^2\right). \tag{5}$$

If an intercept is added, the constraint $\sum_u \phi_u = 0$ is needed for identifiability.

When u is the spatial location index, spatial dependence in the species occurrence probabilities can be modelled using Conditional Auto Regressive (CAR) (Cressie 2015) models. A CAR model is generally defined by a Normal multivariate distribution (dropping the species index i):

$$\boldsymbol{\phi} \sim MVN_n(\boldsymbol{\mu = 0}, \boldsymbol{\Sigma} = [\tau(\boldsymbol{D} - \rho\boldsymbol{A})]^{-1} \tag{6}$$

The random effects covariance matrix $\boldsymbol{\Sigma}$ in (6) depends on the precision parameter τ, a dependence parameter $\rho \in [-1, 1]$ quantifying the strength of correlation between spatial units, and a known adjacency matrix $\boldsymbol{A}$ encoding spatial relationships between locations. $\boldsymbol{D}$ is a diagonal matrix, with the row sums of $\boldsymbol{A}$. The adjacency matrix $\boldsymbol{A}$ encapsulates neighbouring relationships between areas, reflecting the assumption that ϕ_u is influenced by what happens at surrounding locations, i.e. those values $\phi_{u'}$ where $u' \in \mathcal{N}(u)$, the set of neighbours of u: $\boldsymbol{A} = \{a_{uu'}\}_{u,u'=1,\ldots,n}$ is a square $n \times n$ matrix such that $a_{uu'} = 1$ when units u and u' are neighbours, $a_{uu'} = 0$ otherwise, with diagonal elements all zero by default.

Different neighbourhood rules give rise to different types of CAR. For instance, a standard neighbourhood system for grid data is called 'nearest neighbours' (Cressie 2015), i.e. each unit is influenced by the pixels sharing a border along the cardinal directions; another one is the '12 nearest neighbours', including two pixels in each cardinal direction plus the four ones along the diagonals. A model implementing the nearest neighbour dependence is the Intrinsic CAR (ICAR) model (Rue and Held 2005), where the density of $\boldsymbol{\phi}$ (dropping the species index i for simplicity of notation) is:

$$\pi(\boldsymbol{\phi}|\tau) \propto \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2}\sum_{u' \in \mathcal{N}(u)}(\phi_u - \phi_{u'})^2\right). \tag{7}$$

Model (7) corresponds to setting $\rho = 1$ in Equation (6), as well as adopting the nearest neighbours rule to build $\boldsymbol{A}$. The extension of this model to 12 nearest neighbours for grid data has recently become of standard use thanks to its implementation with INLA under the name of RW2d effect (Rue and Held 2005).

2.3 Estimation of entropy components

Once the probabilities are estimated for each category and observation, the estimate for entropy is straightforward, as it is a function of the probabilities, which returns values in $[0, \log I]$. The point entropy estimator is:

$$\widehat{H}_u^{BMB}(X) = \sum_{i=1}^{I} \widehat{p}_{ui}^{BMB} \log \frac{1}{\widehat{p}_{ui}^{BMB}} \tag{8}$$

where BMB stands for Bayesian Model-Based estimator, and $\widehat{H}_u$ is the local entropy estimate at u. To obtain $\widehat{p}_{ui}^{BMB}$, samples are drawn from the probability posterior distribution for each category i and observation u, and each local posterior mean becomes the basis for computing $\widehat{H}_u$; by sampling from the posterior distribution, standard errors may be associated to the estimator value to give an idea of the estimation uncertainty.

By taking this approach for estimation, the entropy estimate $\widehat{H}_u^{BMB}(X)$ varies with u. According to the structure of the linear predictor in (3), we may have one local entropy value for each unit u, or aggregate units into subsets with a common entropy value. For example, if data depend on a binary covariate, they can be split in two subsets according to the covariate value; each subset will have one estimated entropy value. If, instead, we have a model with two covariates, say one binary and one taking five values, we obtain $2 \times 5 = 10$ combinations of values for the covariates that influence the probabilities. Thus, data can be split into 10 groups, each having its own estimated probabilities and its own entropy estimate, and we obtain 10 local values for $\widehat{H}_u^{BMB}(X)$. For a continuous covariate, temporal or spatial effects, we have n different entropy estimates, one for each unit u. We obtain a curve for a continuous covariate or temporal data, and a surface for georeferenced data.

Two main improvements must be highlighted wrt a local (unit-specific) computation of any of the existing estimators. Firstly, with the BMB approach, estimates can be computed even when very few observations of X (even one or none) are available for each unit. Secondly, when random effects are included, estimates take the neighbourhood into account (according to the specific dependence structure) and we obtain smooth curves/surfaces that catch the main behaviour of the data under study and remove the noise. Even when the model only includes covariates and does not have random effects, the whole dataset is exploited to estimate the fixed effect coefficients and returns results with solid ground. The resulting entropy value/curve/surface is obtained by exploiting all the information.

3 Empirical illustration of the BMB approach

The present Section is aimed at illustrating the benefit of using a model-based approach for entropy estimation in scenarios where data are collected over space, time or relate to covariates. Bayesian models are well known in the literature of temporal and spatial statistical studies as regards properties and performance and evaluating

them is out of the scopes of this work (see, e.g. Cressie (2015)). We focus on conclusions and interpretation about the latent biodiversity of the system under study.

We show results for a multinomial variable, i.e. the abundance of species at each location; a preliminary study on presence/absence data can be found in Altieri and Cocchi (2021). We build two standard scenarios in ecological surveys, i.e. temporal and spatial dependence; the reference case of independence and the dependence on a covariate are in Appendix B. For each scenario, we generate $S = 1000$ replicated sequences of $n = 2500$ observations from the variable X representing species, taking values in $\{A, B, C\}$ with category- and observation-specific probabilities p_{uA}, p_{uB}, p_{uC}, where $u = 1, \ldots, 2500$ is the observation index. We simulate a multinomial variable with one trial for each observation point, so that the response variable is a 2500×3 binary matrix, where each row sums to 1 (see the example Table 1, where the matrix is transposed to save space). This produces the maximum uncertainty (Serafini 2019): estimates become more stable as the number of trials for each observation point increases; uncertainty may be accurately evaluated for all results. For the spatial scenario, we also test a second option with more trials.

The BMB estimator is derived with a fully Bayesian approach. For each replicate, we estimate the model with INLA. Then, we simulate 500 samples from the probabilities' posterior distribution and we obtain a posterior distribution for entropy; its mean is chosen as a point entropy estimate. The same procedure is used to estimate the standard error.

We focus the comment on estimated values for entropy and practical consequences for real data studies. For comparative comments, we choose a selection of existing estimators (detailed in Appendix A): the ML estimator (ML), the Miller-Madow correction (MM), Zhang's non-parametric estimator (Zh), and the Bayesian estimator with a Laplace prior (BLapl) and with Jeffrey's prior (BJeff). If the interest lies in a comparative evaluation of the estimators' accuracy, the Mean Square Error is suitable (MSE, presented in Appendix B). All computations are run on the R software. The ML estimator is computed using the data relative frequencies as probability estimates in the entropy formula. The MM, BLapl and BJeff estimators are computed using the `entropy` package (https://CRAN.R-project.org/package=entropy), while Zh is obtained using `EntropyEstimation` (https://CRAN.R-project.org/package=EntropyEstimation). The code for the present proposal is produced by the authors, with the support of the `INLA` package (https://www.r-inla.org) for parameter estimation, and is available as Supplementary Material.

	u	1	2	3	4	5	6	…
Table 1 Data example for 6 out of 2500 observations, one trial for each observation point	A	1	0	0	0	1	0	…
	B	0	1	1	0	0	1	…
	C	0	0	0	1	0	0	…

3.1 Data with temporal dependence

In the temporal scenario, data are generated with a category-specific intercept and a smooth temporal effect in the linear predictor $g_{ui} = \beta_i + I(i = A)\phi_u$, where $\beta = \{1, 4, 3\}$ and ϕ is modelled as an Autoregressive Integrated Moving Average (ARIMA) model. The indicator function $I(i = A)$ ensures the identifiability of the model by applying the effect to category A (Serafini 2019). The fitting model has a simple RW1 temporal effect (other options do not change results substantially).

Figure 1 shows the time-correlated probabilities (left panel), entropy together with its BMB estimate (middle panel), and the estimation error (right panel). The dashed red line represents the posterior estimate, averaged over the 1000 replicates (grey band), and is a smooth version of the true entropy, which captures the main behaviour and removes local variations, despite the difference between generating and estimating model.

A few highlights for this scenario follow:

- The BMB estimate captures the temporal structure of the data, illustrating that the species' probabilities vary smoothly over time;
- The system's entropy is not constant: it increases over time (as the predominance of specie *B* decreases), reaches a peak around time 2000, when probabilities are more evenly distributed and the system has the maximum latent biodiversity, and then decreases again;
- A plot of the estimatet uncertainty sides the resulting entropy, to make scientists more aware of the validity of the conclusions, and shows what time points are linked to higher or lower uncertainty;
- With only one trial per observation point, uncertainty is at its maximum: it is expected to decrease as species' abundance increases;
- Despite the complexity of the procedure, results are achieved in a few seconds with R and INLA, therefore the computational burden is not an issue;
- The model can be tested against others to choose the most suitable option.

Table 2 shows the global estimates for the existing approaches, and the global entropy value $\overline{BMB} = \frac{1}{n}\sum_u \hat{H}_u^{BMB}$. We report the standard deviation for precision evaluation. The available estimators return a number lacking associated uncertainty and interpretability, since they are based on the assumption of
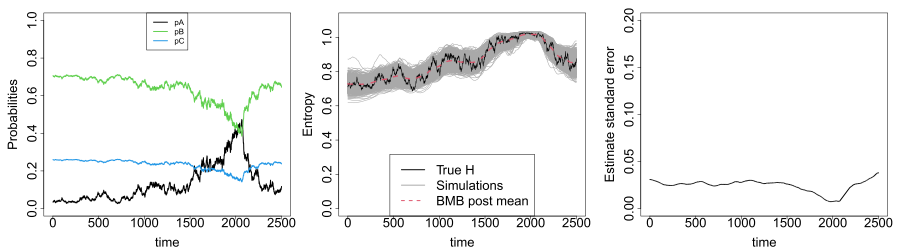

Fig. 1 Temporal effect—probabilities (left), entropy (middle), uncertainty (right)

Table 2 Temporal effect-entropy estimation	ML	MM	Zh	BLapl	BJeff	$\overline{\text{BMB}}$ (sd)
$\hat{H}$	0.8964	0.8969	0.8966	0.8970	0.8967	0.8577(0.0104)

independence. Conversely, the global $\overline{\text{BMB}}$ value takes the temporal dependence into account and is a more accurate estimator (MSEs in Appendix B).

3.2 Data with spatial dependence

We build a square observation area with 50×50 cells, to have $n = 2500$ observation sites. We generate varying probabilities for the three categories over the area, by setting the category-specific intercepts as $\beta = (1.5, 1, 0.5)$ and by adding a spatial structure to category A as a Thomas Process, i.e. a Poisson cluster process (Moller and Waagepetersen 2004). The true entropy surface is shown in Fig. 2, left panel; it varies smoothly, taking low values in the centre, where category A is predominant over the others, and approaching its maximum ($\log 3$) at the corners, where the probability distribution is almost uniform.

We then generate data over each observation site with two options: option 1, with one trial per site, and option 2, with a number of trials per site randomly generated between 1 and 20, to produce a similar scenario to the application of Sect. 4 and to check the performance of ML estimators with enough data. The trials are assigned to category A, B, C based on the probabilities. One data replicate for option 1 is shown in Fig. 2, right panel.

The model for BMB estimation includes a category-specific intercept and a spatial effect in the linear predictor $g_{ui} = \beta_i + I(i = A)\phi_u$. For ϕ, we tried different Intrinsic CAR and Besag models and ran model selection via WAIC (Widely Applicable Information Criterion, Watanabe (2013)). The best option was the Random Walk in two dimensions (RW2d, see Sect. 2) applied to category A. Despite the estimation model not being the same as the generating one as regards the spatial effect, our model is able to produce accurate estimates over both options and is displayed in Fig. 3, along with the estimation standard error (in gray scale). Both entropy surfaces, even the left-hand-side one produced by option 1, prove to be very smooth and accurate estimates. The estimation standard error is very small and decreases when increasing the number of trials.

Table 3 shows the global estimators. A few highlights for this last scenario:

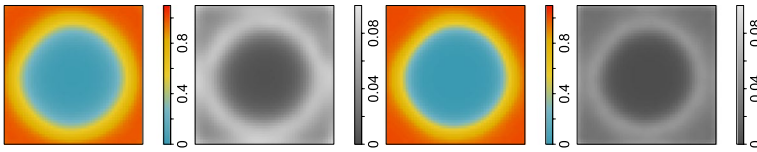Fig. 2 Spatial effect—generating entropy (left), data with 1 trial per site (right)
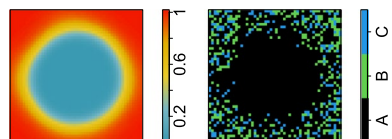
Fig. 3 Spatial effect—BMB estimated entropy and estimation error with different data options: one trial per site (left panels), 1–20 trials (right panels)

Table 3 Spatial effect - entropy estimation

	ML	MM	Zh	BLapl	BJeff	$\overline{\text{BMB}}$ (sd)
Opt 1: one trial/site	0.6773	0.6777	0.6777	0.6784	0.6778	0.5483 (0.1657)
Opt 2: 1-20 trials/site	0.6780	0.6779	0.6780	0.6781	0.6781	0.5483 (0.1056)

- The system's latent biodiversity has a spatially structured behaviour, with a cold-spot at the centre and increasing heterogeneity when moving to the corners: one can be confident about what species is found in the middle, while at the corners, subjects may belong to any species;
- If desired, a "baseline" biodiversity level can be quantified over the surface, using only the estimates of the intercepts in the model;
- The uncertainty plot shows where local estimates are more or less reliable;
- Model comparison helps finding the best fit for the data;
- The global $\overline{\text{BMB}}$ value is different from the results of the available estimators (Table 3) and proves to be more accurate (MSEs in Appendix B);
- The available estimators detect a low level of heterogeneity (around 62% of the maximum entropy, i.e. log(3)), but cannot explain what this is due to; on the contrary, we can show it is due to a spatial structure;
- Local estimates with the existing methods are not possible for option 1, as the number of trials per site is too small; they can be computed for option 2, but are very noisy, less accurate (Appendix B) and lack motivation.

4 A study on rainforest tree data

The motivating dataset regards the presence of lowland tropical rainforest tree species over Barro Colorado Island, Panama (https://stri.si.edu). We have a rectangular observation window of size 1000×500 metres, with four tree species; information is available about the soil elevation and slope (Fig. 4).

The dataset is a marked point pattern with $n = 5639$ trees, where the point categorical mark X is the tree species, with $I = 4$: $x_1 = Inga\ sapindoides$, $x_2 = Heisteria\ concinna$, $x_3 = Beilschmiedia\ pendula$, $x_4 = Astronium\ graveolens$, with $n_1 = 487$, $n_2 = 1141$, $n_3 = 3887$, $n_4 = 124$. *Elevation* (Z_E) is a continuous variable representing the soil altitude in meters, and ranges from 119 to 159 with a mean value of 144. *Slope* (Z_S) is also continuous, displaying values between 0.001 and 0.328

Fig. 4 Rainforest tree data (upper panel), soil elevation (left) and slope (right)
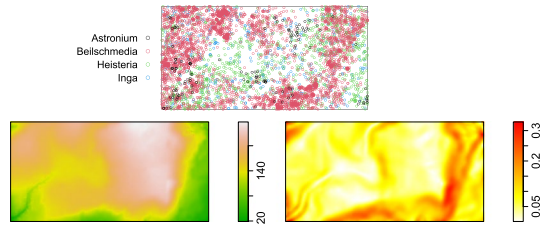
(mean 0.082) to measure the slope of the terrain. It is plausible that the two covariates somehow affect the growth of trees over the area as regards both richness and abundance, and thus influence the biodiversity of the system. By visually comparing the patterns in Fig. 4, one can detect similarities. The BMB approach is able to check this possibility.

We partition the area into 50×100 cells, that contain 0 to 40 trees and mimic the spatial resolution of the environmental covariates. This produces a very fine grid that the human eye perceives as a nearly continuous surface; we tried different grid options that return analogous entropy behaviours, except for the quality of the image. A discussion on the spatial resolution is left to Sect. 5. Each cell is one observation site u, with its spatial location represented by the centroid's coordinates. For each cell, we know the values of the covariates and the counts of all species: the multinomial response variable is a 5000×4 matrix of species counts.

The most general model for the BMB estimation approach has a linear predictor of the form $g_{ui} = \beta_{0i} + \beta_{1i}z_{Eu} + \beta_{2i}z_{Su} + I(i = i^*)\phi_u$: all fixed effect coefficients are allowed to be category-specific; the covariates and the random effect are observation-specific. The spatial effect may be applied to any of the four tree species; we consider two CAR models with different structures for the precision matrix: the ICAR model with nearest neighbours and the RW2d model with 12 nearest neighbours (Sect. 2). We fit the general model and all possible sub-models: with/without intercept, with covariate elevation, slope, both or no covariate, with a ICAR or RW2d spatial effect; results take only a few seconds for each option. We take the WAIC, returned by INLA, to select the best one for entropy estimation. Other selection methods may be used, but this point is not the main core of the present paper (see Kadane and Lazar (2004) for a review). The chosen model is fitted to the data, and parameter estimates are used to estimate the tree species probabilities p_{ui}. Following Equation (8), the entropy surface is estimated.

4.1 Biodiversity evaluation

According to WAIC, the best fitting model includes covariate *slope* and a RW2d spatial effect applied to species 2, i.e. *Beilschmedia Pendula*. The estimated entropy surface is displayed in the left panel of Fig. 5 in relative terms, i.e. ranging in [0, 1]. The biodiversity of the rainforest tree system depends on the soil slope, whose effect is significant on three of the four species: the coefficient is estimated equal to −0.71 (sd=0.112) for species *Astronium graveolens*, 0.21 (sd=0.116) for *Beilschmiedia*

Table 4 Global estimators for rainforest tree data		ML	MM	Zh	BLapl	BJeff	$\overline{\text{BMB}}$ (sd)
	$\widehat{H}$	0.875	0.875	0.875	0.876	0.876	0.912(0.076)
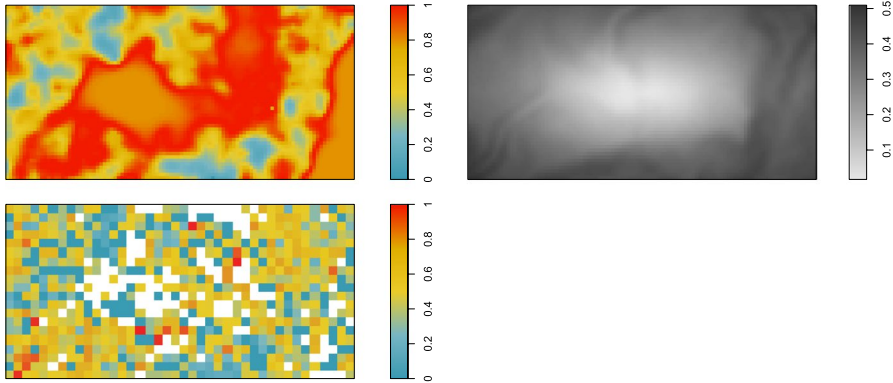
Fig. 5 BMB estimate (upper left); standard error (upper right); local ML estimate (lower)

pendula and 0.67 (sd=0.112) for species *Heisteria concinna*. Indeed, by comparing the data behaviour of Fig. 4 to the slope values, it can be seen that *Astronium* trees tend to grow over flat areas, while *Beilschmiedia* and *Heisteria* trees tend to follow the pattern of the steepest areas. Moreover, the latent biodiversity shows a spatial structure, whose effect particularly affects species *Beilschmiedia pendula*. The estimated entropy surface ranges from 12% to 99% of the maximum possible heterogeneity, with the central half of the distribution concentrated between 60% and 90%; by looking at Fig. 5, one can detect the cold- and hot-spots as regards the latent biodiversity level. The estimation procedure has no issues with areas containing few or zero trees, thanks to the exploitation of a model.

Literature entropy estimators start from the estimated probability distribution based on the relative frequencies $\widehat{p}_X^{ML} = (0.087, 0.202, 0.689, 0.022)$. Results for all global estimators are shown in Table 4, and their difference is negligible, except for the $\overline{\text{BMB}}$.

In relative terms, the available estimates correspond to 63% of the maximum possible entropy ($\log 4$), hinting at a low biodiversity level, but no explanation or interpretation can be drawn. Conversely, the global $\overline{\text{BMB}}$ value can be motivated with the above comments, since it is computed using the whole model information and after careful checking of the model fitting to the data. The estimation error (Fig. 5, right panel) is affected by the number of trials, but also depends on the covariate and on the neighbourhood structure: borrowing information from surrounding sites improves the stability of the estimates.

The computation of a local ML estimator is unreliable with 50×100 cells, since only 26% of them have more than 1 tree. A rougher grid must be chosen to have a satisfactory amount data in most observation sites. With a partition into 20×40 cells

of size 25×25 ms, 78 cells contain one single tree (entropy equal to zero), and 47 have no trees (no entropy); in most cells at least one species is absent and influences the results, but there is no way to control *which* species are absent for each site. The local ML estimate is in the right panel of Fig. 5, in relative terms: many cells are white (zero/non-computable values) and the remaining ones look like white noise. Some resemblance to the BMB estimated pattern can be detected over the area; for example, the cold spots (light blue areas) are placed at similar locations and indicate low biodiversity, but the BMB approach is the only one giving motivation and associating uncertainty to the result. The high entropy areas identified by the model-based approach sometimes correspond to hot (orange/red) spots in the ML panel, sometimes correspond to 0- or 1-individual sites (white cells in the ML estimate): the BMB approach identifies the areas with a high *latent* biodiversity according to the structure of the environmental covariate and of the spatial effect; conversely, ML values only rely on local observations.

5 Discussion and conclusion

The present work proposes a novel approach for entropy estimation suitable for ecological and environmental data, where independence is an unrealistic assumption. In the methodological Sect. 2, we highlight some key novel contributions of such proposal to the field of biodiversity studies. First, entropy is related to covariates and/or data dependence structures, exploiting all the available information. Second, it allows to include information about local abundance and about which species are present/absent at each observation point and at neighbouring points, by using multinomial data instead of presence/absence. Third, it uses INLA instead of MCMC (Markov Chain Monte Carlo) for estimation, overcoming many computational issues, and allows for extensions to continuous space if wished. The empirical illustration of Sect. 3 underlines the ability of the BMB estimation approach to draw conclusions on the latent biodiversity of the system over a variety of situations, to check assumptions, select the most suitable model and evaluate the estimation uncertainty. We also show how to easily deal with empty, or nearly empty, observation points, thanks to the exploitation of a model structure. When wished, results can be further synthesized in a global entropy value for the whole dataset, that is well-grounded as it is built out of consideration of the data structures and not only of relative frequencies. In the application Sect. 4, we analyse the biodiversity of a system of four rainforest tree species. We describe the relationship between biodiversity and the terrain slope; a smooth spatial effect is included in the model, and the BMB estimation result is an entropy surface that captures the latent biodiversity of the system, where cold- and hot-spots can be identified, sided by uncertainty evaluation.

We conclude with a few points for discussion and further work.

First, results may be extended to other diversity indices, such as Hill's numbers, and to species distribution models, where interpretation should focus on the latent process, rather than on observations and description. In addition, our model captures the spatial variation in biodiversity often referred to as β diversity; the possibility to use Shannon's

entropy to quantify simultaneously α and β diversity at a local scale might stimulate this study area, normally more focused on description.

Second, we grid the observation area for a simpler presentation to applied studies, with advantages for computations and for covariate matching. Thanks to the computational fastness of INLA, in many cases we do not have limits to the grid resolution and results are very accurate (Moller and Waagepetersen 2004; Illian 2012). Computational difficulties may arise when the model includes a large number of random effects, or when the dataset is really large. For such situations, we suggest an approach presented in Altieri et al. (2016) that exploits advanced computational techniques in R and the use of Bayesian P-splines to substantially speed up computations. Alternatively, more complex approaches for continuous space are possible, such as Gaussian Random Fields based on the Matérn family of covariances, or the SPDE approach (Lindgren et al. 2011) implemented in INLA.

Last, it is worth mentioning the choice of priors for the precision hyperparameters in the model. We show how the degree of smoothness may be tuned by the random effect, and different types of models may be compared to choose the most suitable one. However, in cases where data are not informative enough about the underlying entropy pattern, the prior choice on the precision parameter of the random effect may play a role in determining the degree of smoothness and this may need further investigation.

Appendix A: A State of the art in entropy estimation

A.1 Frequentist approach

Given X, suppose we have a number of observations n, where each observation is indexed by $u = 1, \ldots, n$ and takes values in the I categories: $x_u \in \{1, \ldots, I\}$. The total number of observations for each category is n_i, with $\sum_{i=1}^{I} n_i = n$. The maximum likelihood estimate for $H(X)$ is:

$$\widehat{H}^{ML}(X) = \sum_{i=1}^{I} \widehat{p}^{ML}(x_i) \log \frac{1}{\widehat{p}^{ML}(x_i)}, \tag{9}$$

where $\widehat{p}^{ML}(x_i) = n_i/n$ for $i = 1, \ldots I$. This is also the non-parametric estimator for entropy, known as plug-in or naive estimator (Paninski 2003). Following the Central Limit Theorem, for $n \to \infty$, then $\widehat{\boldsymbol{p}}_X^{ML}$ is asymptotically normal with mean $\boldsymbol{p}_X$ and variance decreasing with $1/\sqrt{n}$. As a consequence, $\widehat{H}^{ML}(X)$ is also asymptotically normal with bias and variance decreasing with $1/n$, assuming $I < \infty$ (Zhang 2012). When $I = \infty$ the upper limit for the variance is $V(\widehat{H}^{ML}(X)) \leq \frac{\log n^2}{n}$, which goes to zero for $n \to \infty$. A standard correction proposed in the literature to improve the estimator bias is known as the Miller-Madow correction (Miller 1955):

$$\widehat{H}^{MM}(X) = \widehat{H}^{ML}(X) + \frac{I-1}{2n}. \tag{10}$$

Note that I may be an unknown quantity, whose estimation is not trivial. See Miller (1955) for some work in this direction.

It is proved (Antos and Kontoyiannis 2001) that the estimation error does not have a unique decay speed with any of the above estimators. Moreover, when $I < \infty$, the estimator variance $V(\widehat{H}^{ML}(X)) \leq \frac{1}{n}$ decays much faster, and the difference in performance across estimators may be negligible. Furthermore, the ML estimator is negatively biased: $E(\widehat{H}^{ML}(X)) \leq H(X)$, with $E(\widehat{H}^{ML}(X)) = H(X)$ iff $H(X) = 0$, i.e. only one category is present. The bias $B(\widehat{H}^{ML}(X))$ goes to zero for $n/I \to \infty$, i.e. when the number of observations is much larger than the number of categories; it increases with increasing I and small n.

A.2 Non-parametric approach

Zhang (2012) proposes an alternative non-parametric estimator:

$$\widehat{H}^Z(X) = \sum_{v=1}^{n-1} \frac{1}{v} \left[\frac{n^{v+1}[n-(v+1)]!}{n!} \times \sum_{i=1}^{I} \left(\widehat{p}^{ML}(x_i) \prod_{j=0}^{v-1} \left(1 - \widehat{p}^{ML}(x_i) - \frac{j}{n} \right) \right) \right].$$

(11)

This approach starts from a rewriting of Simpson's diversity index (Frosini 2004), combined with Taylor's series expansion and Fubini's Lemma. It can be proved that for $H(X) < \infty$ the estimator is consistent and asymptotically normal, but with a faster decaying variance wrt the ML estimator for $I = \infty$, and a faster decaying bias for $I < \infty$.

A.3 Bayesian approach

For a Bayesian approach, a hierarchical model is needed: we have n observations $\boldsymbol{x}^{obs} = (x_1^{obs}, \ldots, x_n^{obs})$ from an unknown probability distribution $\boldsymbol{p}_X$ depending on parameters $\boldsymbol{\theta}$. For instance, for a multinomial distribution $\boldsymbol{\theta}$ contains the category-specific probabilities of success. The factorization of the joint distribution is $\pi(H, \boldsymbol{x}^{obs}, \boldsymbol{p}_X, \boldsymbol{\theta}) = \pi(H|\boldsymbol{p}_X)\pi(\boldsymbol{x}^{obs}|\boldsymbol{p}_X)\pi(\boldsymbol{p}_X|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

In this case, the entropy H is a variable that takes different values in $[0, \log I]$ according to $\boldsymbol{p}_X$, therefore it has its own probability distribution (Paninski 2003):

$$\pi(H|\boldsymbol{p}_X) = \delta(H - H(\boldsymbol{p}_X)) = \delta\left(H - \sum_{i=1}^{I} p(x_i) \log \frac{1}{p(x_i)} \right)$$

(12)

where H is the unknown entropy and δ is a function measuring the similarity between H and all possible values of $H(X)$, so that $\pi(H|\boldsymbol{p}_X)$ increases as we approach the true value.

Let us take a uniform proper prior for $\boldsymbol{p}_X$, called $\pi(\boldsymbol{p}_X)$. The likelihood is $\pi(\boldsymbol{x}^{obs}|\boldsymbol{p}_X)$; consequently, the posterior is $\pi(\boldsymbol{p}_X|\boldsymbol{x}^{obs}) \propto \pi(\boldsymbol{x}^{obs}|\boldsymbol{p}_X)\pi(\boldsymbol{p}_X)$ and the Bayesian point estimator can be its expected value:

$$\widehat{H}^B = E(H|\boldsymbol{x}^{obs}) = \int H(X)\pi(H|\boldsymbol{p}_X)\pi(\boldsymbol{p}_X|\boldsymbol{x}^{obs})d\boldsymbol{p}_X \qquad (13)$$

with credibility intervals coming from the posterior distribution.

An intuitive choice for the prior distribution $\pi(\boldsymbol{p}_X)$ is Dirichlet's prior (Nemenman et al. 2002), as it is conjugate to the multinomial distribution for $I < \infty$: $\pi_D(\boldsymbol{p}_X) \propto \prod_{i=1}^{I} p(x_i)^{a-1}$, where a tunes how much a uniform distribution is favoured (large values of a) or discouraged (small values of a). For specific choices of a different estimators may be derived: the most common options are $a = 0$, which produces the ML estimator, $a = 1/2$ defining Jeffrey's prior and $a = 1$ for Laplace's prior (Paninski 2003).

Appendix B: Empirical study—additional details

We refer to Sect. 3 for a suitable introduction.

B.1 Independent data and dependence on covariates

The simplest case of independence coincides with constant probabilities of occurrence for each of the three species $p_{uA} = p_A$, $p_{uB} = p_B$ and $p_{uC} = p_C$, so that entropy takes a single value for the whole dataset. The linear predictor is $g_{ui} = \beta_i$, where $i = A, B, C$ and we fix the three values for β as (0.5, 4, 3), tuning how close probabilities are to a uniform distribution. Other options for β do not affect the scope of this work and are not reported here. Table 5 shows the generated probabilities and the consequent entropy.

The distribution for $\widehat{H}^{BMB}$ is approximately Gaussian. Table 6 shows the results across estimators, where the point estimate is reported for the BMB approach, together with a standard deviation computed over the posterior samples, which evaluates the estimation uncertainty.

A few highlights for the independence scenario are:

- All estimates are very similar: the BMB approach is a safe option;

Table 5 Independence—true probabilities and entropy

p_A	p_B	p_C	H
0.0351	0.7054	0.2595	0.7139

Table 6 Independence - entropy estimation

	ML	MM	Zh	BLapl	BJeff	BMB(sd)
$\widehat{H}$	0.7133	0.7137	0.7137	0.7145	0.7139	0.7135(1.768e−04)

Table 7 Binary covariate—true probabilities, entropy and BMB entropy estimates

	p_A	p_B	p_C	H	$\hat{H}^{BMB}$ (sd)
$Z = 0.1$	0.2700	0.3832	0.3467	1.0884	1.0870 (9.169e−06)
$Z = 2$	0.0008	0.8801	0.1191	0.3716	0.3715 (3.564e−04)

Table 8 Binary covariate - entropy estimation

	ML	MM	Zh	BLapl	BJeff	$\overline{\text{BMB}}$ (sd)
$\hat{H}$	0.9001	0.9005	0.9005	0.9006	0.9004	0.7292(0.1285)

- Despite the complexity of the procedure, results are achieved in a few seconds with R and INLA: the computational burden is not an issue;
- The BMB approach is the only one where uncertainty can be assessed, since it allows to compute the estimate's standard deviation. The Mean Square Errors across replicates would be useless in real cases, where only one dataset is available;
- The independence model can be tested against more complex models, to check whether data are actually independent; instead, the existing methods simply assume independence with no possibility to check its validity.

Advantages increase substantially when moving away from independence.

A simple departure from independence consists on data depending on a binary covariate. The linear predictor is $g_{ui} = z_u \beta_i$, where covariate Z is the same for the three categories, but varies across observations taking one of two values: $z_u \in \{0.1, 2\}$ for $u = 1, \ldots, n$, while, again, $\beta = (0.5, 4, 3)$ for the three categories A, B and C. The dataset is split into two subsets according to the covariate value; each subset is made of $n/2 = 1250$ observations. The two data subsets conditional on each covariate value have different entropy levels, as shown in Table 7: one set has an entropy close to its maximum $\log(3)$, due to evenly spread probabilities, while the second has a very low entropy, due to a predominance of p_B. Other options have been tried and do not change the substance of the results.

The BMB estimation procedure differentiates between the two behaviours, as can be seen in Table 7. All existing estimators ignore the covariate and estimate entropy based on the estimated ML probabilities, i.e. the overall data relative frequencies $\hat{p}_A^{ML} = 0.126$, $\hat{p}_B^{ML} = 0.643$ and $\hat{p}_C^{ML} = 0.231$. The resulting estimates are in Table 8, where the global entropy value $\overline{\text{BMB}}$ for the BMB approach is computed as a weighted mean of the two estimates in Table 7.

A few key points for this scenario are:

- BMB is able to include the covariate and check how it affects the presence of species and the latent biodiversity of the system. A larger value for Z induces the predominance of species B and decreases heterogeneity;

- If a global entropy value is needed, the $\overline{\text{BMB}}$ estimate takes the covariate into account, therefore it differs and is more accurate than the available estimates (see Section B.2) and allows uncertainty evaluation;
- The standard deviation may be computed to measure local uncertainty and check whether the accuracy changes for different covariate values;
- Standard tools for model comparison can help checking the model;
- It is not recommendable to compute separate estimates for entropy conditional on the two covariate values. In Section B.2, we show that such local estimates are less accurate, and we refer to Sect. 5 for a discussion of the major limits of this procedure.

The same conclusion holds for categorical or discrete covariates with a limited number of values. The case of a continuous covariate is analogous to the one of temporal dependence: the model takes a different value for each observation.

B.2 Comparative evaluation of the estimators

The BMB approach is compared to the existing estimators at a global level for all scenarios with a departure from independence; in addition, a local entropy comparison is run for the binary covariate scenario and for the spatial scenario, option 2 (1–20 trials per observation site), where the amount of data is sufficient for computing local relative frequencies. The comparison is run in terms of mean square error (MSE); the dataset-specific MSE for $s = 1, \ldots, S$ is:

$$MSE_s = \frac{1}{n} \sum_{u=1}^{n} (\widehat{H} - H_u)^2, \tag{14}$$

where $\widehat{H}$ is the global entropy estimate for any of the considered methods (shown in Tables 2, 3, 8), while H_u is the observation-specific true entropy, that may vary according to the scenario. The dataset-specific MSE for the local estimates is computed as in (14), where the local $\widehat{H}_u$ substitutes the global $\widehat{H}$. A distribution of $S = 1000$ values is obtained for each scenario and for each estimation method; the MSE boxplots over the 1000 datasets are in the log scale. Figures 6, 7 and 8 show that the MSE of the BMB approach is the smallest for all scenarios.

For data with a binary covariate, Fig. 6 displays the (log)MSEs for the global estimate (Table 8) in the left panel, and the (log)MSEs for the local estimates (Table 7) in the right panel. The local existing estimates perform quite well, since a large number of observations ($n/2 = 1250$) contributes to the estimation of each conditional probability given the covariate value. Nevertheless, Fig. 6 highlights the improvement in the estimation accuracy with the BMB approach with both global and local estimates,

For data with temporal dependence, MSEs are shown in Fig. 7, referring to the estimates in Table 2: despite the great loss of information linked to a global value, and despite the fitting model being different than the generating model, the BMB approach leads again to the smallest (log)MSE.
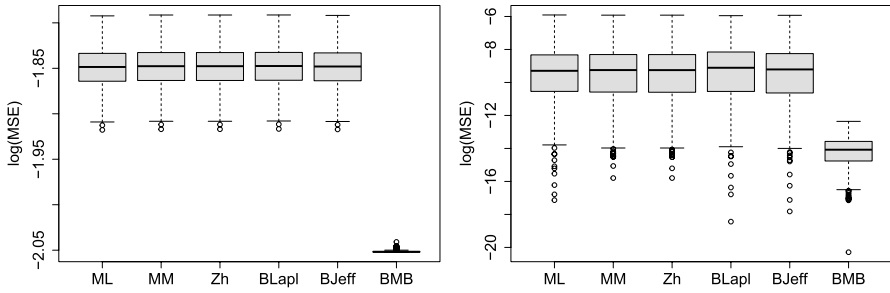
Fig. 6 Binary covariate—MSE boxplots, log scale, for global (left) and local (right) estimate

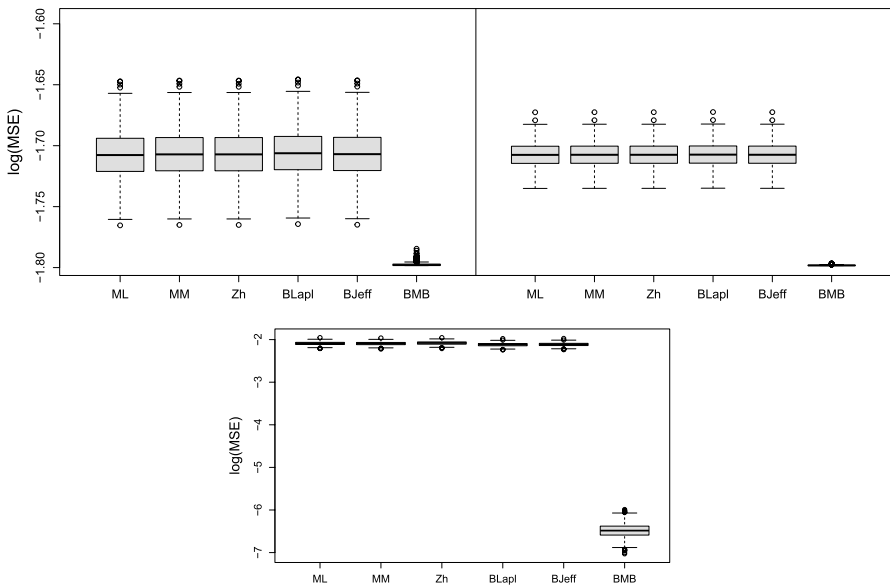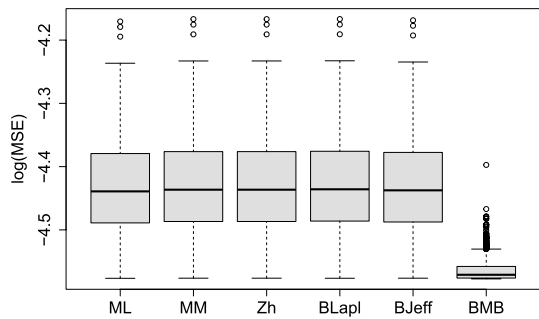**Fig. 7** Temporal effect—MSE boxplots, log scale, for global estimate

Fig. 8 Spatial effect—MSE boxplots, log scale, for global estimate: option 1-1 trial per site (top-left panel), option 2-1-20 trials (top-right panel). MSE boxplots, log scale, for local estimate with 1-20 trials (bottom panel)

For data with spatial dependence, MSEs for the global estimates of Table 3 are shown in Fig. 8, top panels: again, even if the fitting model is different than the generating one, the (log)MSE is the smallest; we can appreciate the difference in the uncertainty wrt to the existing approaches, especially when the amount of available data is small. For spatial option 2, the local comparison is in Fig. 8, bottom panel, and shows the improvement of the BMB proposal.

Data Availibility The presented dataset is available as Supplementary Material, with a sample `R` code. The original dataset of tropical rainforest trees over Barro Colorado Island is available at https://stri.si.edu/research-computing/databases.

Declarations

Competing interest The authors have no competing interests to declare that are relevant to the content of this article.No funds, grants, or other support was received, and there are no other financial or non-financial competing interests.

References

Altieri L, Cocchi D (2021) Entropy estimation for binary data with dependence structures. In: Perna C, Salvati N, Spagnolo FS (eds) Book of short papers SIS 2021, Pisa, Pearson, pp 1088–1093. https://it.pearson.com/content/dam/region-core/italy/pearson-italy/pdf/Docenti/Universit%C3%A0/pearson-sis-book-2021-parte-1.pdf

Altieri L, Cocchi D, Greco F et al (2016) Bayesian P-splines and advanced computing in R for a change-point analysis on spatio-temporal point processes. J Stat Comput Simulation 86(13):2531–2545

Altieri L, Cocchi D, Roli G (2018) A new approach to spatial entropy measures. Environ Ecol Stat 25(1):95–110

Altieri L, Cocchi D, Roli G (2019) Advances in spatial entropy measures. Stoch Environ Res Risk Assess 33:1223–1240

Anderson MJ, Crist TO, Chase JM et al (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. Ecol Lett 14:19–28

Antos A, Kontoyiannis I (2001) Convergence properties of functional estimates for discrete distributions. Random Struct Algorithms 19:163–193

Baddeley A, Rubak E, Turner R (2015) Spatial point patterns: methodology and applications with R. Chapman and Hall, London

Baker GS (1994) The multinomial-poisson transformation. Statistician 43:495–504

Barmoudeh L, Baghishani H, Martino S (2022) Bayesian spatial analysis of crash severity data with the INLA approach: assessment of different identification constraints. Accid Anal Prev 167(106):570

Batty M (1976) Entropy in spatial aggregation. Geogr Anal 8:1–21

Cressie NAC (2015) Statistics for spatial data, rev. Wiley, New York

Frosini BV (2004) Descriptive measures of ecological diversity. In Environmetrics. Edt J. Jureckova, A. H. El-Shaarawi in encyclopedia of life support systems (EOLSS), revised edn 2006, Paris

Gelfand AE (2022) Spatial modeling for the distribution of species in plant communities. Spatial Stat 50:100582

Gelfand AE, Shirota S (2019) Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological monographs 89:e01372

Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432

Illian JB (2012) A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). Ann Appl Stat 6:1499–1530

Kadane JB, Lazar NA (2004) Methods and criteria for model selection. J Am Stat Assoc 99:279–290

Krainski ET, Gómez-Rubio V, Bakka H et al (2019) Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC Press, Boca Raton

Leibovici DG, Claramunt C, LeGuyader D et al (2014) Local and global spatio-temporal entropy indices based on distance ratios and co-occurrences distributions. Int J Geogr Information Sci 28(5):1061–1084

Lindgren F, Rue H, Lindstrom J (2011) An explicit link between Gaussian fields and Gaussian markov random fields: the stochastic partial differential equation approach. J Royal Stat Soc B 73:423–498

Magurran AE (2004) Meas Biol Divers. Blackwell Publishing, Oxford

Martinez-Minaya J, Cameletti M, Conesa D et al (2018) Species distribution modeling: a statistical review with focus in spatio-temporal issues. Stoch Environ Res Risk Assess 32:3227–3244

Miller G (1955) Note on the bias of information estimates. In: Quastler H (ed) Information theory in psychology II-B. Glencoe, IL free press, pp 95–100

Moller J, Waagepetersen R (2004) Statistical inference and simulation for spatial point processes. Chapman and Hall, Boca Raton

Moller J, Waagepetersen R (2007) Modern statistics for spatial point processes. Scand J Stat 34:643–684

Nemenman I, Shafee F, Bialek W (2002) Entropy and inference, revisited. In: Dietterich TG, Becker S, Ghahramani Z (eds) Advances in neural information processing, 14, Cambridge. MIT Press, MA

Paninski L (2003) Estimation of entropy and mutual information. J Neural Comput 15:1191–1253

Rue H, Held L (2005) Gaussian Markov random fields. Chapman and Hall, Boca Raton

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J Royal Stat Soc 71(2):319–392

Serafini F (2019) Multinomial logit models with INLA. https://inla.r-inla-download.org/r-inla.org/doc/vignettes/multinomial.pdf, R-INLA tutorial

Shannon C (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Ventrucci M, Cocchi D, Burgazzi G et al (2020) PC priors for residual correlation parameters in one-factor mixed models. Stat Methods Appl 29:745–765

Watanabe S (2013) A widely applicable Bayesian information criterion. J Mach Learn Res 14:867–897

Zhang Z (2012) Entropy estimation in Turing's perspective. J Neural Comput 24:1368–1389