# Generalized biodiversity assessment by Bayesian nested random effects models with spyke-and-slab priors

Giovanna Jona Lasinio[1]

*Department of Statistical Sciences, Sapienza University, Rome Italy*

Alessio Pollice

*Department of Economics and Finance, "Aldo Moro" University, Bari, Italy*

Elisa Anna Fano

*Department of Life Sciences and Biotechnology, University of Ferrara Italy*

**Abstract**

We analyze variations in $\alpha$-diversity of benthic macroinvertebrate communities in an Italian lagoon system using Bayesian hierarchical models with nested random effects. Our aim is to understand how spatial scales influence microhabitat definition. Tsallis entropy measures diversity and spike-and-slab regression selects predictors.

*Keywords:* Tsallis entropy, biodiversity modeling, Bayesian random effects models, spike and slab, climate change monitoring, Po River Delta lagoon (Italy).

## 1. Introduction

The analysis of benthic assemblages is a valuable tool to describe the ecological status of transitional water ecosystems. As some species are extremely sensitive and respond to both microhabitat and seasonal differences, changes in the composition of the macrobenthic community can be used as an "early warning" for environmental changes possibly due to climate variations and affecting the economic and ecological importance of lagoons, through their provision of

Ecosystem Services [1]. Lagoons are fragile ecosystems, highly sensitive to climate change. Their ecological status can then be considered a powerful tool for climate monitoring. In this context, the appropriate definition of the spatial scale of microhabitats is of crucial importance from a conservational point of view. The objective of this work is to understand and describe the influence that different microhabitat spatial scales have on the variation of the biodiversity of lagoons. In [2] the same data were analyzed: the variation of three biodiversity indices was described using mixed effects models, including fixed effects of biotic and abiotic factors and random effects ruled by nested sources of variability, corresponding to alternative definitions of microhabitats. Some of the questions addressed by [2] are worth some further investigation. A deeper understanding of the spatial structure of the data at hand is required, but made difficult by the small number of samples at the finest spatial scale. Building upon [2], where a mixed effects model was estimated in a REML framework [3], we turn to the hierarchical Bayesian modeling paradigm in order to achieve a more complex specification of the spatial structure. In this framework, the spike-and-slab approach [4] is considered for model selection. The latter returns posterior inclusion probabilities of fixed effects, allowing variable selection in a fully Bayesian framework. Inclusion probabilities are also used to produce a ranking of the fixed effects according to their influence on the biodiversity variation. The composition of the macrobenthic community is measured by three biodiversity indices corresponding to as many versions of the Tsallis entropy [5, 2] and giving decreasing relevance to rare species (or singletons[1]).

[1]Species found at only one location regardless their abundace.

## 2. The Data

Benthic macroinvertebrates were collected monthly during the period 1997-2000 in the Po River Delta observation field. Samples were taken with 3 replicates at each of 23 monitoring stations, divided in 10 areas belonging to 3 lagoons (nested grouping structure) and abundance of each species was evaluated as *number of individuals per square meter*. In the following, the mean abundances of the 3 replicates are considered throughout. The selected sites present from one to three dominant habitat types defined by a factorial classification of sediment granulometry (sand, mud) and vegetation cover/type (without vegetation, submerged macrophytes, emerged macrophytes and macroalgae, here recoded as *with and without macroalgae*) and for details see [2] and references therein. A total of 47 taxa were identified. The total dataset size is 272 records.

## 3. The Tsallis Entropy

Information theory and entropy measures have extensively been applied to ecological processes in areas as diverse as biodiversity assessment, evolution, species interactions and landscape analysis. Within the common focus of measuring ecosystem structural and functional complexity, the Tsallis measure has considerable statistical relevance when used as biodiversity index; properties of this measure were extensively studied in [6]. Given a discrete set of species probabilities $p = \{p_i\}$ and any real number $q$, the Tsallis entropy of order $q$ is defined as

$$H_q(p) = \frac{1}{q-1}\left(1 - \sum_i p_i^q\right) \tag{1}$$

Many known biodiversity measures, including the *number of species*, *Shannon* and *Simpson indices*, are obtained by the *deformed exponential transformation* of the Tsallis entropy with known values of $q$: $D_q(p) = e_q(H_q(p))$. When

$x < \frac{1}{q-1}$ holds, the deformed exponential transformation of order $q$ is defined as $e_q(x) = [1 + (1-q)x]^{\frac{1}{1-q}}$ that converges to the standard exponential when $q \to 1$. In the following, we consider the deformed exponential transformations of the Tsallis entropy of order $q = 0, 1, 2$ as biodiversity measures. When $q = 0$ we obtain the number of species, $q = 1$ corresponds to the Shannon biodiversity index and $q = 2$ to the Simpson index. Hence when $q$ increases, decreasing relevance is given to rare species. A consequence of this property is that with increasing $q$ the heterogeneity of the biodiversity measure decreases. Indeed, with the data at hand and $q = 0, 1, 2$ we obtain the values of the inter-quartile range reported in table 1. Notice that also the central position of the biodiversity indices decreases when $q$ increases, as is clearly shown in table 1 where the medians by lagoon are also reported.

| | L1 | L2 | L3 |
|---|---|---|---|
| $q = 0$ | 6.000 (6.25) | 16.015 (10.00) | 8.000 (5.99) |
| $q = 1$ | 1.157 (0.77) | 1.769 (0.57) | 1.276 (0.64) |
| $q = 2$ | 0.570 (0.21) | 0.768 (0.15) | 0.634 (0.24) |

Table 1: Median and inter quartile range (in parenthesis) of the Tsallis diversity for $q = 0, 1, 2$ by lagoon (L1=Comacchio, L2=Fattibello, L3=Goro).

## 4. Modeling biodiversity

In order to investigate the influence of alternative definitions of the micro-habitats on the biodiversity of the Po River Delta, the deformed exponential transformations of the Tsallis entropy of order $q = 0, 1, 2$ at 23 monitoring stations and 12 time points are considered as responses within Gaussian linear mixed effects models, where the fixed effects part depends on habitat and seasonal descriptors and the random part accounts for the nested spatial effects of lagoons, areas and monitoring stations. Let $Y_{iljs}$ denote one of the three biodiversity measures at time $i$ ($i = 1, \ldots, 12$), lagoon $l$ ($l = 1, 2, 3$), area $j$

$(j = 1, \ldots, 10)$, station $s$ $(s = 1, \ldots, 23)$:

$$Y_{iljs} | \mu_{Yljs}, \tau_Y \sim N(\mu_{Yljs}, \tau_Y) \tag{2}$$

with

$$\mu_{Yljs} = \beta_1 X_{1ljs} + \beta_2 X_{2ljs} + \beta_3 X_{3ljs} + \beta_4 X_{4ljs} + \beta_5 X_{5ljs} + \beta_6 X_{6ljs} + \mu_{ljs} \tag{3}$$

$X_1, \ldots, X_6$ are indicator variables for the winter, spring, summer and autumn seasons, for the presence of macro algae and for muddy sediments. As for the nested random effects $\mu_{ljs}$ in equation (3), they are specified as follows: $\mu_{ljs} \sim N(\mu_{jl}, \tau_S)$, $\mu_{jl} \sim N(\mu_l, \tau_A)$, $\mu_l \sim N(0, \tau_L)$; where $l = 1, 2, 3, j = 1, \ldots, n_l, s = 1, \ldots, n_{js}$ and $\sum_l n_l = 10$ and $\sum_{j,s} n_{js} = 23$. For the fixed effects part of the predictor, in order to better appreciate the role of available covariates, we adopt the spike-and-slab regression approach, described in [4] by Kuo and Mallick for variable selection. Among the advantages of this approach is the selective shrinkage property that allows the posterior mean of the coefficients to shrink toward zero for truly zero coefficients, while for non-zero coefficients, posterior estimates are similar to the ordinary least squares (OLS) estimates [7]. Let $\beta_k = \tilde{\beta}_k \times \gamma_k$ for $k = 1, \ldots, 6$, with $\gamma_k$'s being latent indicator variables. Independent priors are then assumed for $\tilde{\beta}_k$, $\gamma_k$ and the error precision $\tau_Y$. $\tilde{\beta}_k \sim N(\mu_\beta, \tau_\beta)$ , $\gamma_k \sim$ Bern$(p_k)$, $p_k \sim Beta(a_p, b_p)$, and $\tau_Y \sim$ Gamma$(a, b)$, where $k = 1, \ldots, 6$, $\mu_\beta$ and $\tau_\beta$ reflect prior beliefs about the distribution of $\tilde{\beta}_k$'s, $p_k$ is the preference to include the $k$-th predictor in the model, assumed to be Beta distributed, $a_p$, $b_p$, $a$ and $b$ are user-defined tuning parameters. Notice that when $\gamma_k = 0$, the updated value of $\tilde{\beta}_k$ is sampled from the full conditional distribution, which is its prior distribution. Mixing will be poor if this is too vague and the sampler will only rarely flip from $\gamma_k = 0$ to $\gamma_k = 1$. Furthermore, some tuning of

the prior distributions is necessary to control for false positives (see for example [8]). For these reasons, our final prior setting includes independent $N(0, 10)$ priors for all $\tilde{\beta}_k$'s and $a = 2$, $b = 0.001$ for the error precision. In addition, we consider $p_k \sim \text{Beta}(5, 5)$, $k = 1, \ldots, 6$ that favors equal probabilities for all terms. Only with $q = 2$ a more informative $\text{Beta}(50, 5)$ prior was chosen for $p_5$, to ensure 90% credible intervals for the corresponding $\tilde{\beta}_5$ excluding the zero value. These choices were carefully checked by sensitivity analysis with alternative prior settings; the final selection of explanatory variables was highly robust and largely independent on the priors. Prior knowledge on the precision parameters is modeled in the same way for all $q$'s as $\tau_L, \tau_S \sim \text{Gamma}(2, 0.1)$ and $\tau_A \sim \text{Gamma}(2, 0.01)$. Notice that the prior for the precision of the random effects of areas $\tau_A$ is more concentrated on larger values than those for lagoons and monitoring stations. This is motivated by areas being not well defined in terms of physical boundaries, actually intersecting each other, and plausibly characterized by smaller internal variability.

## 5. Results

Posterior estimates of model parameters and predictions of random effects were obtained as MCMC simulation summaries by a JAGS [9] implementation, running 2 parallel chains of 250000 iterations with a burnin phase of 50000 and thinning by 10 (133.106 seconds computation time on a latest generation computer with 16Gb Ram). To better comply with the Gaussian assumption, the biodiversity index of order $q = 0$ (the number of species) is modeled on the log scale. Posterior estimates of inclusion probabilities $\hat{p}_k = \frac{1}{ns} \sum_{i=1}^{ns} \gamma_{ki}$, where $ns$ is the number of MCMC samples, highlight the fundamental role of the presence/absence of macroalgae, regardless of the biodiversity measure chosen (see table 2). When the role of rare species is relevant ($q = 0$), both the

presence of macroalgae and the sediment type have large posterior probabilities of inclusion. Sediment is not included in models for $q = 1, 2$, with all alternative priors used for sensitivity analysis suggesting that most of the variation of the indices is explained by the spatial random effects (Figure 1). Again regardless the choice of prior distributions and the index order, the posterior inclusion probabilities indicate that the presence of macroalgae is the most informative descriptor among those considered ($\hat{p}_5 > 0.95$). In Figure 1 the larger spatial

| | $q = 0$ | | | | $q = 1$ | | | |
|---|---|---|---|---|---|---|---|---|
| $k$ | $\tilde{\beta}_k$ | $\tilde{\beta}_{k,0.025}$ | $\tilde{\beta}_{k,0.975}$ | $p$ | $\tilde{\beta}_k$ | $\tilde{\beta}_{k,0.025}$ | $\tilde{\beta}_{k,0.975}$ | $p$ |
| 1 | 0.030 | -0.455 | 0.453 | 0.324 | 0.011 | -0.484 | 0.492 | 0.204 |
| 2 | 0.016 | -0.471 | 0.461 | 0.274 | 0.002 | -0.484 | 0.488 | 0.193 |
| 3 | -0.064 | -0.398 | 0.387 | 0.526 | -0.052 | -0.413 | 0.413 | 0.476 |
| 4 | -0.002 | -0.470 | 0.477 | 0.224 | 0.033 | -0.451 | 0.455 | 0.340 |
| 5 | 0.496 | 0.211 | 0.749 | 0.977 | 0.375 | 0.155 | 0.584 | 0.970 |
| 6 | 0.520 | 0.097 | 0.860 | 0.946 | 0.209 | -0.287 | 0.556 | 0.718 |

| | $q = 2$ | | | |
|---|---|---|---|---|
| $k$ | $\tilde{\beta}_k$ | $\tilde{\beta}_{k,0.025}$ | $\tilde{\beta}_{k,0.975}$ | $p$ |
| 1 | 0.000 | -0.507 | 0.504 | 0.082 |
| 2 | 0.003 | -0.501 | 0.505 | 0.077 |
| 3 | -0.009 | -0.482 | 0.471 | 0.220 |
| 4 | 0.010 | -0.475 | 0.478 | 0.233 |
| 5 | 0.150 | 0.009 | 0.290 | 0.951 |
| 6 | 0.059 | -0.423 | 0.427 | 0.438 |

Table 2: Posterior estimates of fixed effects with 90% credible intervals and the posterior inclusion probabilities $p$ for three order of the Tsallis entropy.

random effects are always associated to lagoon 2, namely the Fattibello lagoon. Indeed the overall biodiversity of this lagoon is larger than the other two, no matter the order of the Tsallis biodiversity, i.e. the importance given to rare species (see Table 1). Figure 2 shows that, regardless the order of the Tsallis entropy, a larger portion of variance is always explained by the higher level of spatial aggregation (lagoons), followed by the most detailed one (stations). The intermediate spatial aggregation level (areas) is the least relevant in terms of

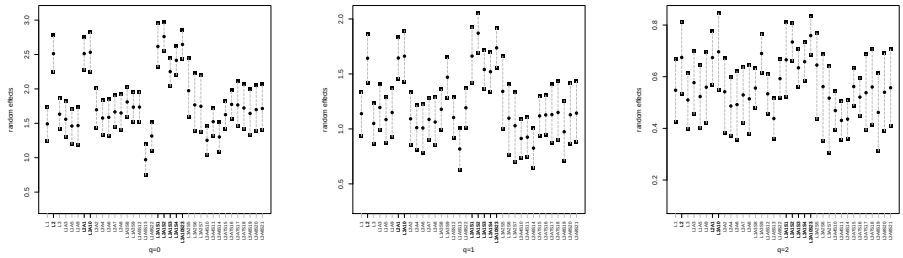
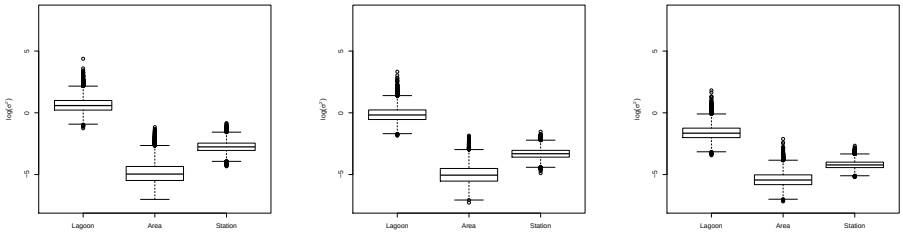



Figure 1: Posterior predicted means (bullet) of latent spatial random effects (dots) with 90% (squares) credible intervals for three orders of the Tsallis entropy. Bold thicks and labels highlight the Fattibello lagoon effects.

spatial random effects variability.




$q = 0$ $q = 1$ $q = 2$

Figure 2: Boxplots of logs of posterior estimates of random effects variances ($\tau_L^{-1}$, $\tau_A^{-1}$, $\tau_S^{-1}$) for three orders of the Tsallis entropy.

## 6. Concluding remarks

The main advantage of the proposed approach is the ability to account for the relation across the spatial aggregation levels: stations within areas, areas within lagoons. For this purpose, we adopt a hierarchical specification of spatial random effects, where each spatial aggregation level influences the finer one. This allows us to define spatial random effects at the finer station level, representing the intrinsic spatial variation within areas varying within lagoons. This

complex random spatial structure provides an alternative to the proper consideration of a spatial correlation model accounting for site proximities. Indeed, a spatially structured correlation model could not be used with the data at hand, given that lagoons are not spatially connected. Hence we would need to model the spatial effects of areas and monitoring stations within lagoons, with a very small number of sampling points per lagoon: 2, 3, 5 areas, 5, 5, 13 monitoring stations. In this paper, the spike-and-slab approach to posterior variable selection proved to provide useful information on the relevance of seasonal and habitat features. Posterior inclusion probabilities allow to rank the effects of explanatory factors and analyze their relative importance in the explanation of the response variation. Parameter estimation is readily obtained at the same time, thus simplifying the entire process. A possible drawback of the spike-and-slab variable selection is its computational complexity. However, the accuracy and interpretability of the results further justify the choice. Possible alternative applications of the proposed methodology include understanding and quantifying the geographic patterns of biodiversity in water body and landscape monitoring for climate change assessment, in those cases where the number of spatial samples is relatively small, e.g. in the analysis of water bodies status trough isotope signature [10].

**Acknowledgements**

Giovanna Jona Lasinio and Alessio Pollice were partially supported by the PRIN2015 project "Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)" funded by MIUR - Italian Ministry of University and Research.

## References


[1] M. Scott, R. Smith, J. Dick, Quantitative approaches to ecosystem services assessment, Environmetrics 22 (5) (2011) 597–597. doi:10.1002/env.1114.

[2] G. Jona Lasinio, A. Pollice, E. Marcon, E. A. Fano, Assessing the role of the spatial scale in the analysis of lagoon biodiversity. a case-study on the macrobenthic fauna of the po river delta, Ecological Indicators 80 (Supplement C) (2017) 303 – 315. doi:https://doi.org/10.1016/j.ecolind.2017.05.037.

[3] A. F. Zuur, E. N. Ieno, A. A. Walker, Neil Saveliev, G. M. Smith, Mixed Effects Models and Extensions in Ecology with R, Springer, 2009.

[4] L. Kuo, B. Mallick, Variable selection for regression models, Sankya: The Indian Journal of Statistics, Series B 60 (1998) 65–81.

[5] E. Màrcon, I. Scotti, B. Hérault, V. Rossi, G. Lang, Generalization of the partitioning of shannon diversity., PLoS ONE 9 (3) (2014) e90289.

[6] N. Gupta, R. K. Bajaj, On partial monotonic behaviour of some entropy measures, Statistics & Probability Letters 83 (5) (2013) 1330 – 1338. doi:https://doi.org/10.1016/j.spl.2013.02.001.

[7] H. Ishwaran, J. S. Rao, Consistency of spike and slab regression, Statistics & Probability Letters 81 (12) (2011) 1920 – 1928. doi:https://doi.org/10.1016/j.spl.2011.08.005.

[8] R. B. O'Hara, M. J. Sillanp a a, A review of bayesian variable selection methods: what, how and which, Bayesian Analysis 4 (1) (2009) 85–117.

[9] M. Plummer, JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling, in: Proceedings of the 3rd International Workshop on Distributed Statistical Computing, 2003.

[10] F. Fiorentino, D. Cicala, G. Careddu, E. Calizza, G. Jona-Lasinio, L. Rossi, M. L. Costantini, Epilithon $\delta^{15}$n signatures indicate the origins of nitrogen loading and its seasonal dynamics in a volcanic lake, Ecological Indicators 79 (Supplement C) (2017) 19 – 27. doi:https://doi.org/10.1016/j.ecolind.2017.04.007.