HMSC Paper - Temporary title

# Materiel and methods

## HMSC

Since the rise of Joint Species Distribution Models using latent variables mid-2010 (Warton *et al.* [2015](#ref-Warton_2015)), several implementations have been proposed : *HMSC* (Ovaskainen *et al.* [2017b](#ref-Ovaskainen_2017a)), *Boral* (Hui [2016](#ref-Hui_2016)), *gllvm* (Niku *et al.* [2019](#ref-Niku_2019)). Among these different implementations, we choose for this study *HMSC* because this implementation performs well on occurrence data against other classical *SDM* and *JSDM* (Norberg *et al.* [2019](#ref-Norberg_2020)). Moreover, *HMSC* allow the user to use different types of ecological meaningful data like traits and/or phylogenetic data (Ovaskainen *et al.* [2017b](#ref-Ovaskainen_2017a) ; Tikhonov *et al.* [2019](#ref-Tikhonov_2019b)) by solving the fourth corner problem (Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)).

“*HMSC* is a mutilvariate hierarchical generalized linear mixed model adjusted with Bayesian inference” rooted in assembly theory (Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). Therefore, a *HMSC* model is composed of two parts, one fix and one random. The fix part models the realised niche of each species ( matrix), where each dimension of the niche is a covariate included in the model (Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). Without any traits, the *HMSC* framework assume that all species share the same expected niche value for each covariate, however, the width of each niche is assumed to be species-independent. If the modeler includes some trait data, the information contained in it enables *HMSC* to estimate a specie-specific expected niche value (Ovaskainen *et al.* [2017b](#ref-Ovaskainen_2017a) ; Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). In addition, traits are phylogenetically inherited, therefore, closely related phylogenetically species tend to share more traits. Thus, using phylogenetic information, *HMSC* can capture residual traits information not include in the trait data (Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). Since *HMSC* can include a random part, it can constrain the estimated residuals with random effects. These random effects are estimated using latent variables: for each random effect, two matrices of latent variables are estimated (Ovaskainen\_2017a ; Tikhonov *et al.* [2019](#ref-Tikhonov_2019b) ; Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). One of them, thereby called *site loadings* ( matrix) contains the values of all the covariates not included in the model at the level of this random effect (Ovaskainen *et al.* [2017b](#ref-Ovaskainen_2017a) ; Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). The other matrix, thereafter called *species loadings* ( matrix) correspond to the response of the species to each of these missing predictors at the level of this random effect (Ovaskainen *et al.* [2017b](#ref-Ovaskainen_2017a) ; Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)). These missing covariate can be missing environmental features, but, if the model models well the environmental niche of each species, the remaining information are more likely to contain some biotic interaction signals (Ovaskainen *et al.* [2017a](#ref-Ovaskainen_2017b) , [2017b](#ref-Ovaskainen_2017a) ; Ovaskainen & Abrego [2020](#ref-Ovaskainen_2020)).

## Dataset

Faunistic data were provided by the *REBENT* monitoring program. This program aims to collect data about benthic habitats and their associated fauna. From this survey program, we focused on two monitored habitats: the intertidal bare sediments and intertidal seagrass (*Zostera marina*) beds of Brittany. We have selected 21 sites samples all over Brittany (est-ce que l’on met une carte dans le corps du texte ou en annexe ?). The sampling methodology was described in length in Boyé *et al.* ([2017](#ref-Boye_2017)). Each site was sampled at least 6 times between 2006 and 2014 (est-ce que l’on met en annexe le tableau contenant l’année, le site et l’habitat échantillonné ?) and one sample consisted of a set of three replicates of three sediment cores. This dataset, hereafter called *training dataset* had 180 observational units defined as the unique combination of years, site and habitat. Additionally, we have selected two more sites, which included the two habitats and 35 observational units constituting the *test dataset* in order to test our models against unseen data.

The training dataset originally included 519 species, later reduce to 278 species after removing species with fewer than 4 occurrences across the 180 observational units. Of these 278 species, 99 of them were Polychaeta. The taxonomic classification of these 99 Polychaeta species was retrieved on Worms to be used as a phylogeny proxy. Hence, phylogenetic correlation has been estimated using ape R package (Paradis & Schliep [2019](#ref-Paradis_2019)). In addition, Polychaeta traits data matrix was obtained from a previous study of these benthic community (Boyé *et al.* [2019](#ref-Boye_2019a)). The trait matrix contained 11 fuzzy coded traits for a total of 41 categories. Prior to using the trait matrix, we reduced its dimensionality by applying fuzzy-PCA algorithm. We have kept the first three axes, accounting for 59% of the total variance of the trait matrix to be used as synthetic traits data. (Est-ce que l’on met la Fuzzy ACP et les statistiques afférentes en annexe ?)

Based on the work of Boyé ([2018](#ref-Boye_2018_test)), we selected 7 environmental variables which allowed to characterise the realised niche of each species of the sampled communities. These variables describe the fetch, a proxy for the degree of hydrodynamic exposure of the sites; hydrological properties as sea water temperature and salinity and current velocity; and sedimentological properties as mud and organic matter content, and granulometric properties. To capture the non-linear nature of the ecological niches, we then calculated orthogonal polynomials of degree 1 and degree 2 for each of these variables. These polynomials transformed variables have been used as explanatory covariates in the models presented below.

## Fitted models & Analysis

We fitted four alternative candidate models of increasing complexity with the R-package Hmsc (Tikhonov *et al.* [2019](#ref-Tikhonov_2019b)) assuming the default priors. The first candidate model uses only Polychaeta community data and the environmental covariates, the second one add to the previous model the phylogenetic data, the third one add to the previous trait data. The last model uses the whole community data and the environmental covariates, but it did not include any trait nor phylogenetic data. All these four models were fitted twice, using occurrence and abundance data. Moreover, all these models and their occurrence/abundance variants shared the same random effect: one temporal random effect accounting for the sampling year, one random factor accounting for the sampling site and the last one accounting for the sampling habitat. We sampled the posterior distributions of each model with same parameters: we used 15 Markov Chain Monte Carlo chains, each one ran for 30 000 iterations. The first 10 000 iterations were discarded as burn-in and the remaining were thinned by 20 to yield 1000 posterior samples per chain. In total, we sampled 15,000 posterior samples per parameter. We then examined MCMC convergence by examining the potential scale reduction factors (Gelman & Rubin [1992](#ref-Gelman_1992)) of each model parameters.

Goodness of fit and predictive power of all models were assessed using AUC (for abundance data, AUC was computed using abundance predictions truncated to occurrence). Additionally, the root mean squared errors (RMSE) was also computed for abundance based models. We compared the and matrices of the different models two by two using co-inertia analysis. Co-inertia analysis allows to search for common structures between two or more matrices (Legendre & Legendre [2012](#ref-Legendre_2012)). Besides, co-inertia analysis allows the calculation of the RV coefficient, a multivariate generalisation of the Pearson coefficient (Legendre & Legendre [2012](#ref-Legendre_2012)). The RV coefficient helps to identify whether one or more of the models differs completely from the others in terms of the estimated coefficients. The species loading matrix was further used to infer networks of associations between species by extracting the residual correlations between species (Ovaskainen *et al.* [2017a](#ref-Ovaskainen_2017b)) and process these correlations to create probabilistic species association networks (Momal *et al.* [2020](#ref-Momal_2020)). These networks were compared between models by measuring several of their structuring properties.

Boyé, A. (2018). Diversité taxinomique et fonctionnelle des habitats benthiques dans l’espace et dans le temps : une perspective régionale et décennale. PhD thesis.

Boyé, A., Legendre, P., Grall, J. & Gauthier, O. (2017). Constancy despite variability: Local and regional macrofaunal diversity in intertidal seagrass beds. *Journal of Sea Research*, 130, 107–122.

Boyé, A., Thiébaut, Grall, J., Legendre, P., Broudin, C. & Houbin, C. *et al.* (2019). Trait-based approach to monitoring marine benthic data along 500 km of coastline. *Diversity and Distributions*, 25, 1879–1896.

Gelman, A. & Rubin, D.B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7, 457–472.

Hui, F.K.C. (2016). boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, 7, 744–750.

Legendre, P. & Legendre, L. (2012). *Numerical Ecology*. Developments in Environmental Modelling. 3rd edn. Elsevier.

Momal, R., Robin, S. & Ambroise, C. (2020). Tree-based inference of species interaction networks from abundance data. *Methods in Ecology and Evolution*, 11, 621–632.

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J. & Anttila, J. *et al.* (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, e01370.

Ovaskainen, O. & Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R*. Ecology, Biodiversity and Conservation. Cambridge University Press.

Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S. & Sæther, B.-E. *et al.* (2017a). How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20170768.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L. & Dunson, D. *et al.* (2017b). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20, 561–576.

Paradis, E. & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35, 526–528.

Tikhonov, G., Opedal, O., Abrego, N., Lehikoinen, A. & Ovaskainen, O. (2019). Joint species distribution modelling with HMSC-R. *bioRxiv*.

Warton, D.I., Blanchet, F.G., O’Hara, R.B., Ovaskainen, O., Taskinen, S. & Walker, S.C. *et al.* (2015). So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution*, 30, 766–779.