Joint species distribution modelling of the species of the estuary and gulf of St. Lawrence

David Beauchesne, Kévin Cazelles, Guillaume Blanchet, Philippe Archambault, Dominique Gravel

08 April 2017

To do

Next steps:

- Incorporate categorical environmental covariables (*e.g.* sediment type, habitat type)
- Integrate R code to text for parameters (*e.g.* number of taxa in analysis)
- Read Ovaskainen 2017 Ecology
- Format and incorporate other datasets
- Get better environmental data
- Contact people for zooplankton and phytoplankton data
- Get marine mammals data (not easy)
- Find how multiple different datasets could be grouped together in a single JSDM analysis?

Questions:

- Is it possible to use categorical variables with JSDMs? Oui mais non, il faut décoder les variables?

```
x<-factor(rep(letters[1:3],4))
model.matrix(~x)
model.matrix(~x-1)
```

Contacts

- Denis Bernier (data)
- Guillaume Blanchet (method)
- Claude Nozères (data)

Overview

The goal of this document is to present the process to evaluate the distribution of species in the estuary and gulf of St. Lawrence (EGSL). This corresponds to the second half of my second thesis objective, which aims at predicting the spatial distribution of EGSL species. I still have in mind to establish a similar process to the one I did to predict species interactions, but with species distribution and . I think at some point I should spend some time working on this. I should be able to code something much more quickly than I did for species interactions, by using a few species as examples to evaluate whether it could cork or not. For now, I am using real occurrence data for St. Lawrence species in species distribution models. Here is the data I have so far (make a table with this, with data information and reference):

Objectives

Evaluate the spatial structure of biotic interactions networks in the estuary and gulf of St. Lawrence

1. Predict the spatial distribution of species in the estuary and gulf of St. Lawrence (This document)
2. Predict biotic interactions among co-occurring species in the estuary and gulf of St. Lawrence (iEat algorithm)
3. Characterize the spatial structure of interaction networks (*e.g.* richness, connectance, number of links, etc) in the estuary and gulf of St. Lawrence (Full 2nd chapter)

Methodology

Data

Occurrence data

The data used to predict the spatial structure of estuary and northern gulf of St. Lawrence is from DFO's annual plurispecific survey of the Northern Gulf pf St. Lawrence between 2010 and 2015. The raw data was formatted to retain, to the extent possible, only taxa at the species level. Groups that were too coarse taxonomically were generally removed from the dataset, except certain widely distributed and abundant groups (*e.g.* Porifera). Taxa that were closely related taxonomically, shared similar functional roles and were hard to distinguish in the field during identification were aggregated to avoid biases. Taxa removal and aggregation is documented on my personal wiki and the code is available on github. Furthermore, only taxa with a minimum number of **50** records were retained for further analyses to allow for the use of JSDMs (@ref).

The dataset used for the analyses contains 27088 observations for 124 taxa (218.5 ± 183 observations per taxon). The survey data spans 5 years, for a total of 878 stations (176 ± 16 stations per year).

Environmental covariables

More details is available here

Les données environnementales utilisées proviendront principalement d'un exercice de modélisation des habitats benthiques (Dutil et al., 2011) et épipélagiques (c.-à-d. les 30 premiers mètres de la colonne d'eau; Dutil et al., 2012). Ces bases de données caractérisent une multitude de données environnementales géographiques, physiques et sur la colonne d'eau pour l'ensemble de l'EGSL. Les variables utilisées pour prédire la distribution des organismes benthiques seront la profondeur, la salinité, la température de fond et de surface, et le niveau de saturation en oxygène (Moritz et al., 2013, 2015; Albouy et al., 2014).

Joint species distribution models

We performed joint species distribution models (@ref) to predict the potential distribution of all taxa for which we had occurrence data. Survey number (*i.e.* corresponds to years the survey was performed) and station (trawl sessions) were used as a random factor. *I don't think that stations need to be mentionned, as there is no station replication. Eventually I could use the cells from my study grid to evaluate which stations are spatially dependent. Ultimately some formal analysis of station independence should also be performed*

Describe methodology!

MCMC diagnostics: trace and density plots

Mixing of the chains has to be checked when performing MCMC-based Bayesian inference, *i.e.* verify if the chains become stationary and sufficiently long to provide a representative sample from the posterior. Parameter estimates should also be reviewed to make sure that they are properly estimated, *i.e.* the distribution of the density of

parameter values should be normally distributed. Visual inspection of trace and density plots (@figure1) shows that chains seem to be converging properly and that parameter are properly estimated.

Parameter summaries: posterior probabilities or credible intervals

We use the 95% credible intervals of parameters in order to assess the level of statistical support for the influence of environmental covariables on species occurrence (@figure2). Parameters with intervals that overlap with 0 are deemed uninformative in predicting species spatial occurrence.

Variance partitioning

We partition the proportion of the variance explained by groups of parameters (*i.e.* Depth, Temperature, Oxygen and Spatial) for each taxa. The sum of variance explained by the groups thus amounts to 100% for all taxa, although the model does not explain all the variation in species distribution.

Predictive power of the model

Add text

Monte Carlo cross-validation - AUC under ROC curves

Cross-validation of HMSC models was performed using Monte Carlo cross-validation, which is a repeated random sub-sampling validation process using 20% of observations (trawl sessions) as validation data and the remaining 80% as training data for each iteration (n = 20). Rather than sampling at the species scale, we sampled at the scale of trawling activities since HMSC are community level models. Random sampling was also stratified within years. Validation was evaluated using the area under the curve (AUC) of ROC curves (@ref), which ranges between 0 (reverse predictions) and 1 (perfect predictions), with models with AUC values close to 0.5 performing no better than random expectation.

Association networks

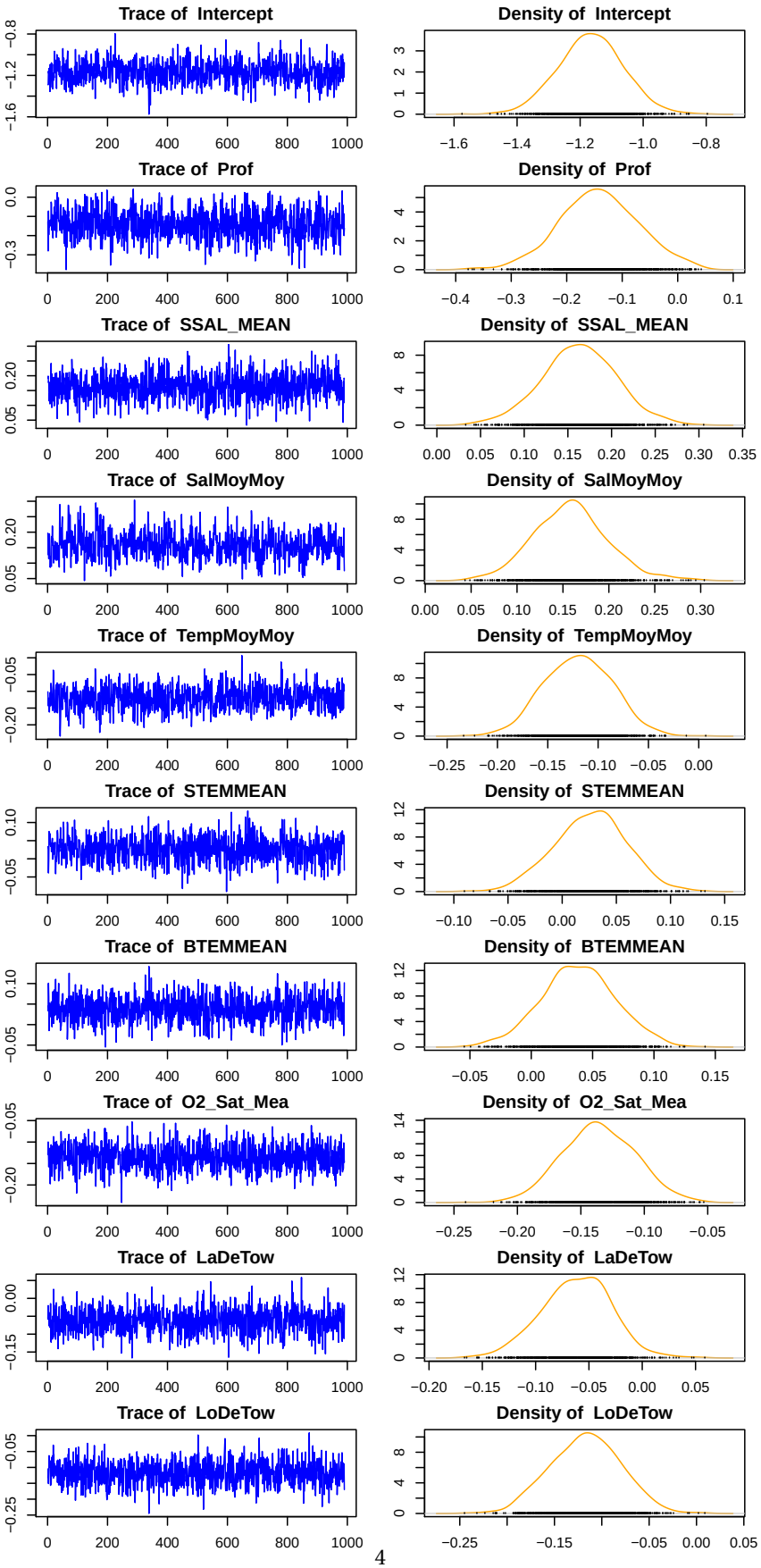

Figure 1. MCMC trace (blue) and density (orange) plots showing the mixing chains for environmental parameters: depth (Prof), annual mean surface salinity (SSAL-MEAN), annual bottom mean salinity (SalMoyMoy), annual bottom mean temperature (TempMoyMoy), annual mean surface temperature (STEMMEAN), annual mean temperature at 30 meters depth (BTEMMEAN), mean bottom oxygen saturation (O2-Sat-Mea). MCMC iterations are shown on the x-axis. Trace and density plot y-axes show parameter values at each iteration and density for parameter values, respectively. The black dots on the density plots represent the MCMC iterations. The MCMC chain was run for 10 000 iterations with 1000 burning iterations and a thin value set to 10. Trace and density plots were respectively produced using traceplot and densplot from the coda package.

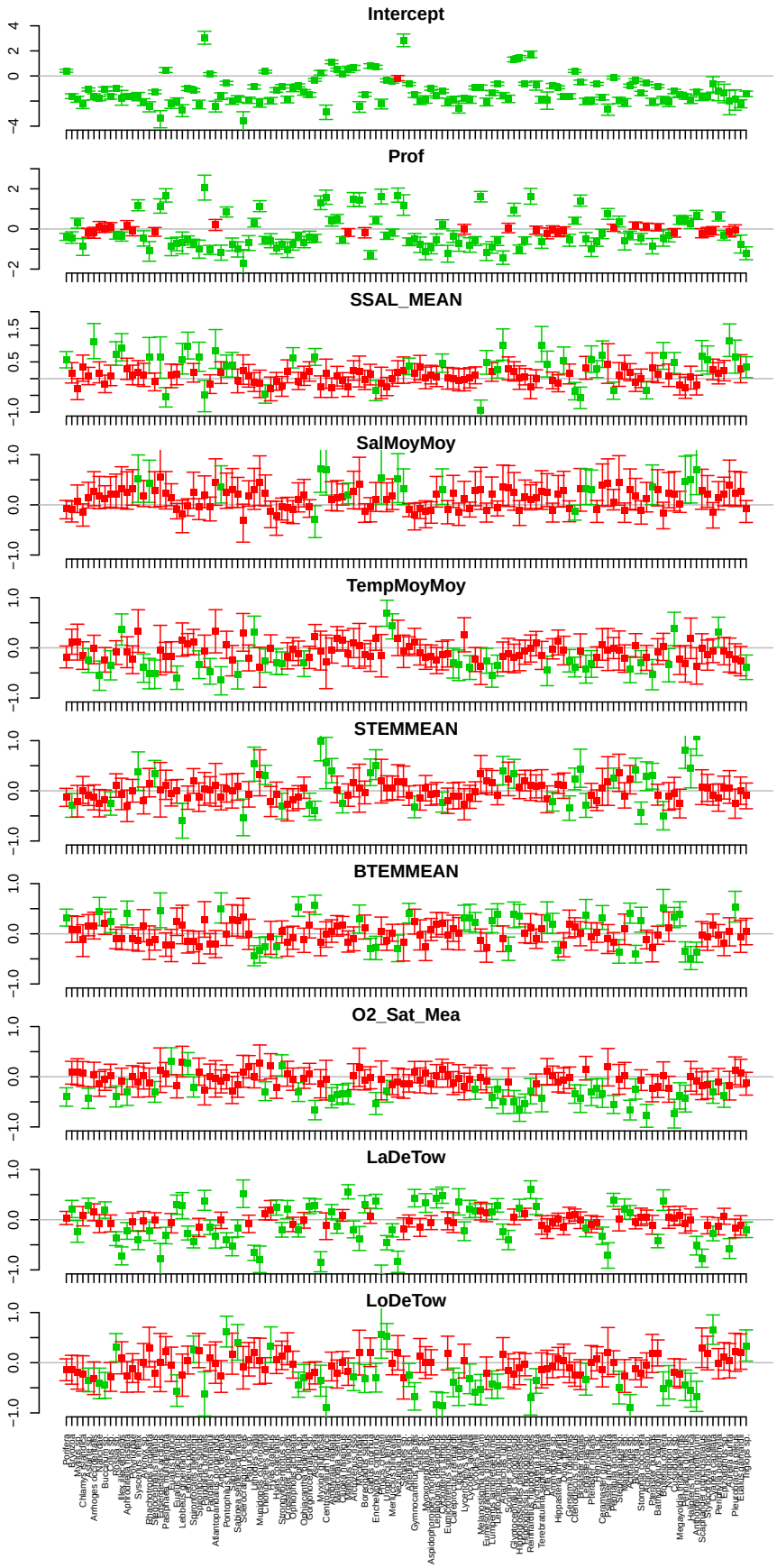
Intercept
Prof
SSAL_MEAN
SalMoyMoy
TempMoyMoy
STEMMEAN
BTEMMEAN
O2_Sat_Mea
LaDeTow
LoDeTow

Figure 2. 95% Credible intervals for environmental parameters: depth (Prof), annual mean surface salinity (SSAL_MEAN), annual bottom mean salinity (SalMoyMoy), annual bottom mean temperature (TempMoyMoy), annual mean surface temperature (STEMMEAN), annual mean temperature at 30 meters depth (BTEMMEAN), mean bottom oxygen saturation (O2_Sat_Mea). Informative parameters per taxa per environmental covariables are those whose credible interval does not overlap with 0. Informative and uninformative parameters are identified in green and red, respectively. The MCMC chain was run for 10 000 iterations with 1000 burning iterations and a thin value set to 10.

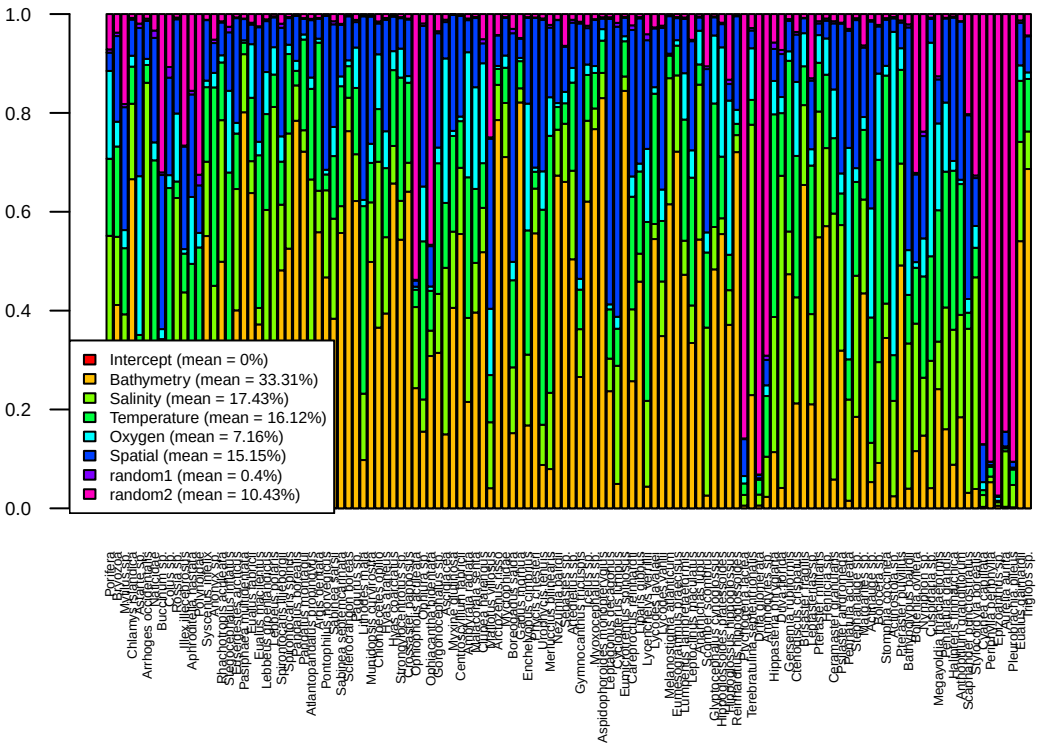

Figure 3. Plot of the variance partitioning between, *i.e.* the proportion of the variance explained by environmental covariables. All environmental covariables (*i.e.* depth (Prof), annual mean surface salinity (SSAL_MEAN), annual bottom mean salinity (SalMoyMoy), annual bottom mean temperature (TempMoyMoy), annual mean surface temperature (STEMMEAN), annual mean temperature at 30 meters depth (BTEMMEAN), mean bottom oxygen saturation (O2_Sat_Mea)) were grouped in order to visually represent the proportion of variation in spatial distribution they explain compared to the Intercept and the random variables. The legend shows mean values for all taxa.

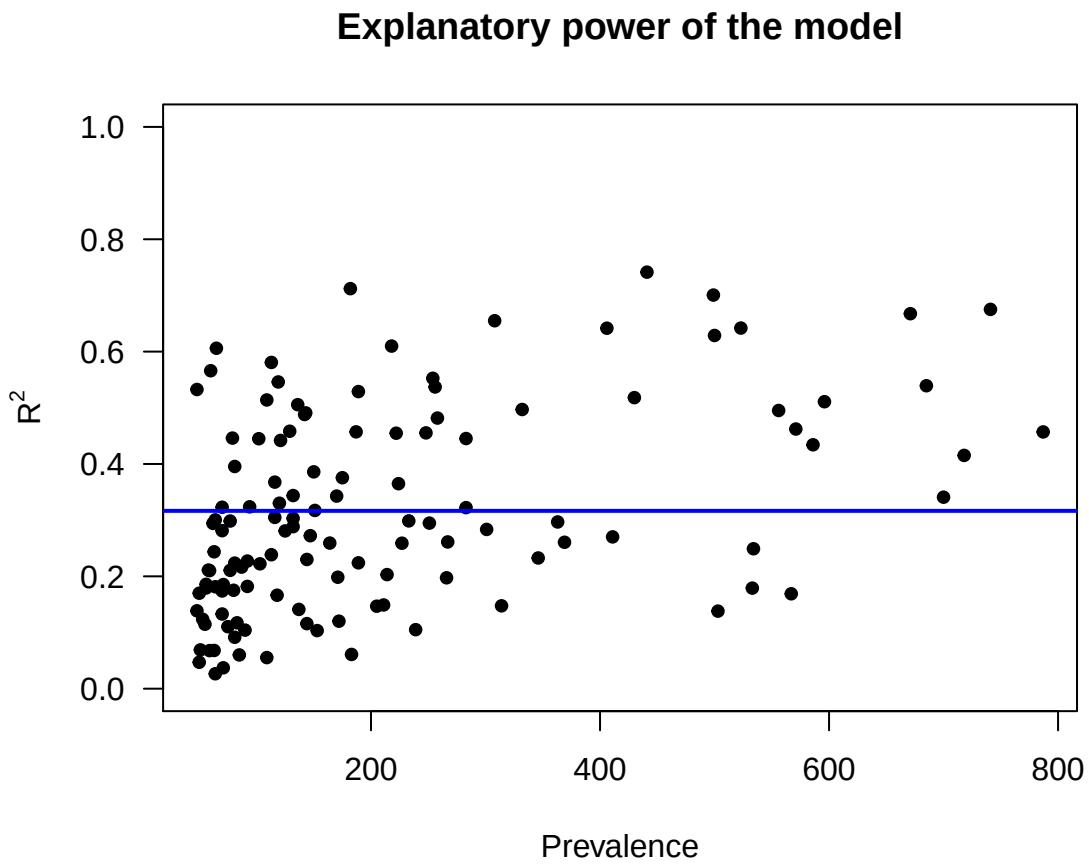

Figure 4. Predictive power of the model evaluated by comparing coefficients of determination (Rˆ2) as a function of the prevalence, *i.e.* the number of trawl sessions in which taxa were captured. Each dot represent an individual taxa. The evaluated Rˆ2 is the Tjur's Rˆ2, which is the mean model prediction for sampling units where a taxa occurs minus the mean model prediction for sampling units where the species does not occur (@Tjur2009). The current version of predictive power assessment is based on the same data that was used to generate the model. It is therefore likely that Rˆ2 values are overestimated due to overfitting. We will ultimately do cross-validation by using only 80% of the original data points for each taxa, keeping the remaining 20% for predictive power assessment.

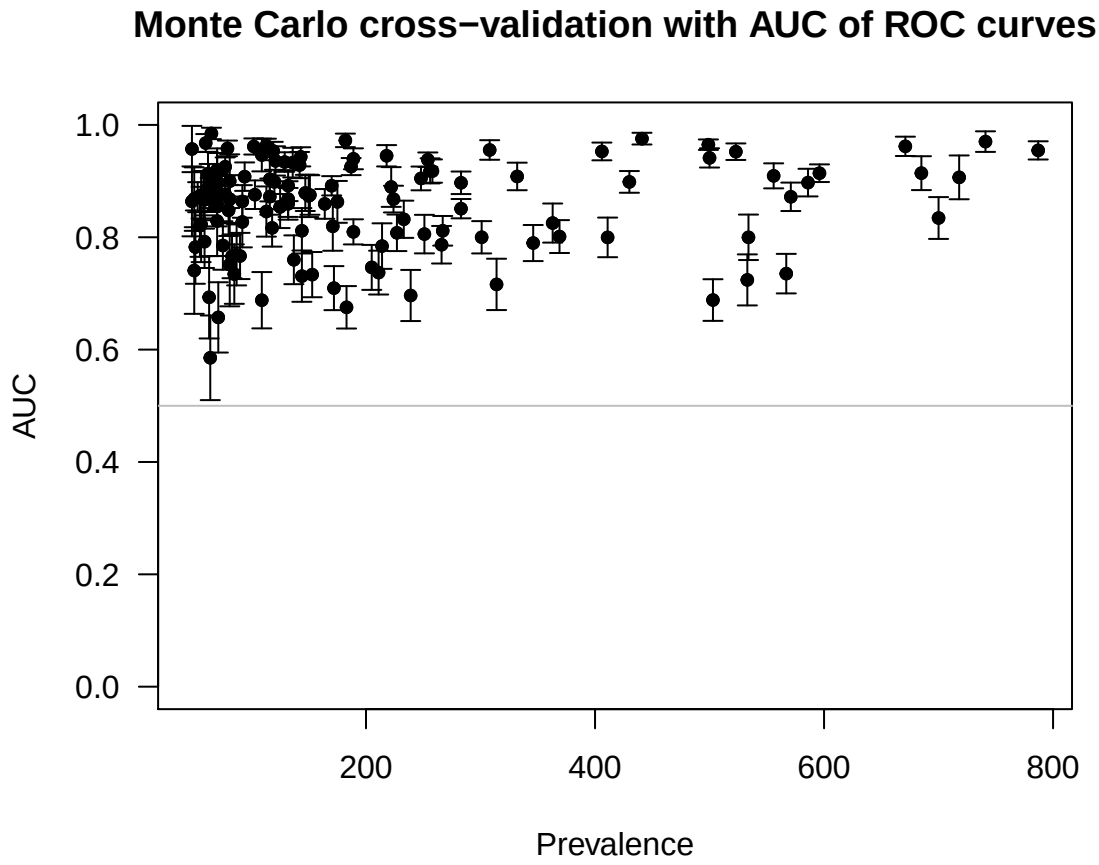

Figure 5. Monte Carlo Cross-validation using AUC (mean +- sd) as a function of prevalence *i.e.* the number of trawl sessions in which taxa were captured. Each dot represent an individual taxa. Random sampling was stratified within years in order to set aside 20% of observations (*i.e.* trawl sessions) used as validation data and the remaining 80% used as training data for the model for each iteration (n = 20). Environmental covariables used to build the model were depth (Prof), annual mean surface salinity (SSAL_MEAN), annual bottom mean salinity (SalMoyMoy), annual bottom mean temperature (TempMoyMoy), annual mean surface temperature (STEMMEAN), annual mean temperature at 30 meters depth (BTEMMEAN), mean bottom oxygen saturation (O2_Sat_Mea), latitude and longitude of trawling activities. The MCMC chain was run for 10 000 iterations with 1000 burning iterations and a thin value set to 10.

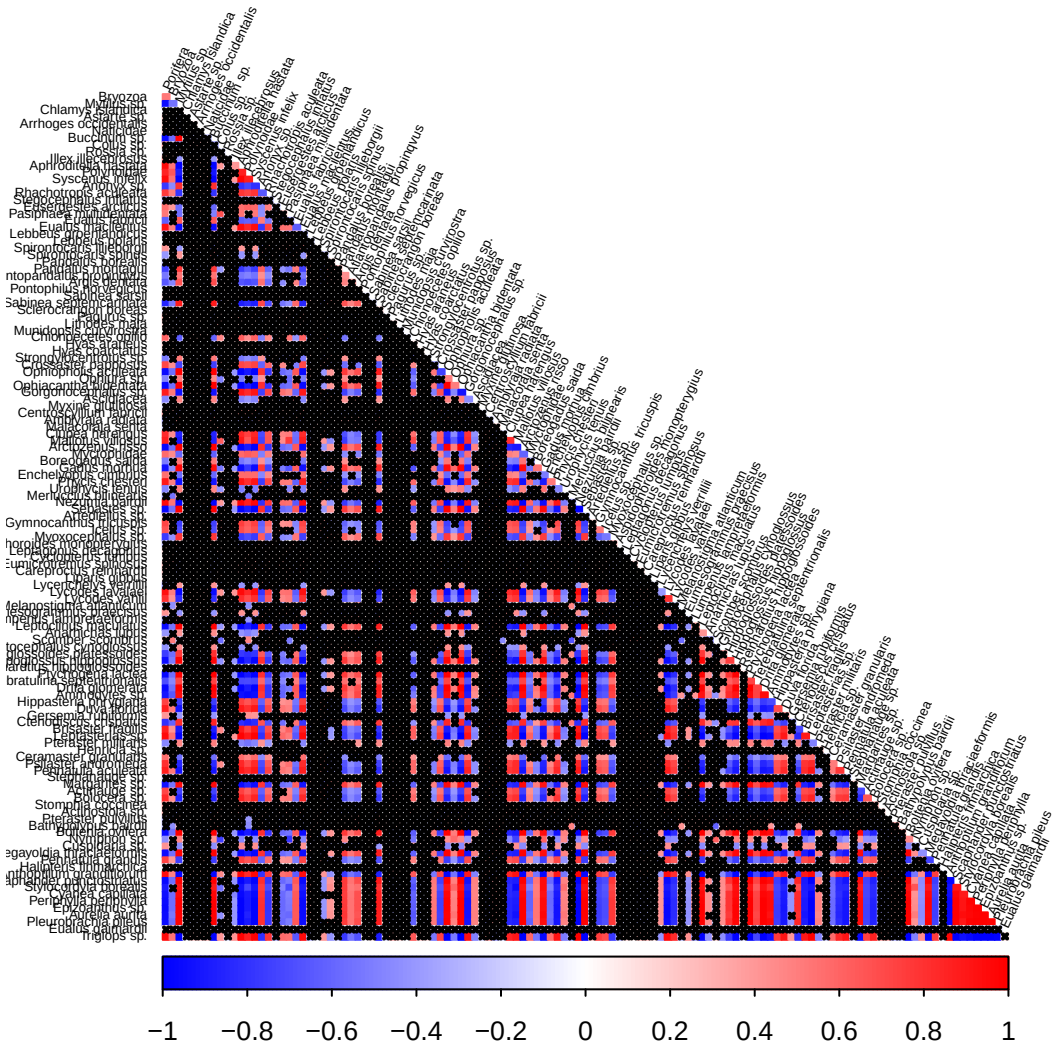

Figure 4. The HMSC framework allows for the evaluation of taxa spatial association networks. This figure is a matrix plot depicting the spatial associations between 124 St. Lawrence taxa. Red and blue cells depict taxa that are respectively positive or negative associations, *i.e.* associations that are more or less likely to be associated spatially compared to random expectations. Cells marked with a 'X' have absolute correlation values smaller than 0.4.