Pipeline v0.1 – Demo

charley

10/01/2021

This work is using species distribution modeling to map the spatial distribution of eight benthic taxa. This short report is looking at three modeling aspects:

- predictors selection
- the influence of spatial variation when explicitly accounted in the modeling
- prediction and uncertainty mapping

Modelling implementation

We analyse the data with Hierarchical Modelling of Species Communities (HMSC) (Tikhonov et al. 2020). HMSC is a joint species distribution model (Warton et al. 2015), using a Bayesian inference framework. For the spatial model, we use spatially structured latent variables recently proposed by (Tikhonov et al. 2020).

The scripts used in this report are all available on a publicly-available Github repository. The followings have been generated using the release v0.1.

Dataset

Loading data…

```
csv_path = "C:/Users/cgros/code/qms517/SXY.csv"
SXY = read.csv(csv_path, stringsAsFactors=TRUE)
S = SXY[, 1:4]  # Study design data
X = SXY[, 5:9] # Predictors
Y = SXY[, 10:17] # Response variables
```

The dataset includes records of 200 observed sites from eight photographic surveys. Labeling has been done using CoralNet. Percentage cover data from eight benthic taxa are used in this report.

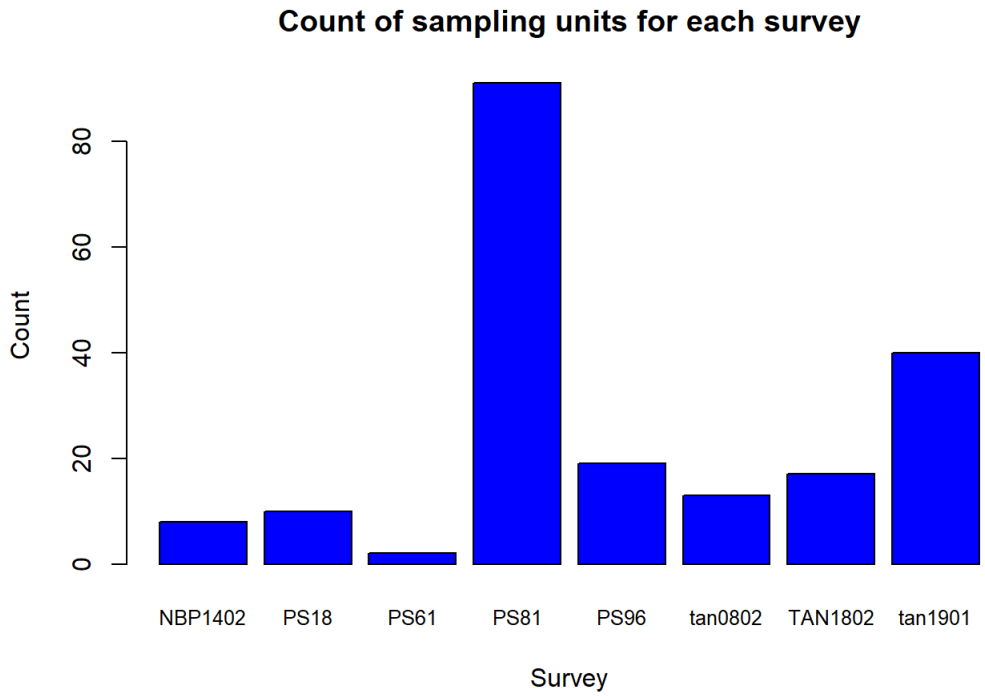
```
dim(Y)
```

```
## [1] 200   8
```

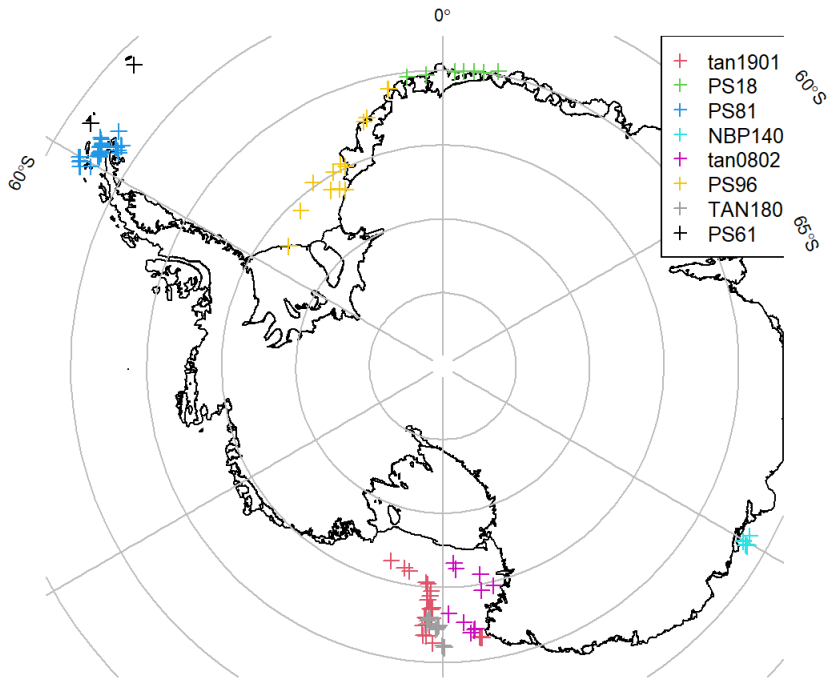
```
colnames(Y)
```

```
## [1] "Anemones"    "Ascidians"   "Bryozoa"     "HydroCorals" "Hydroids"
## [6] "Octocorals"  "Sponges"     "Worms"
```

```
barplot(table(S$Survey), main="Count of sampling units for each survey", xlab="Survey", ylab="Count", col="b
lue", cex.names=0.8)
```



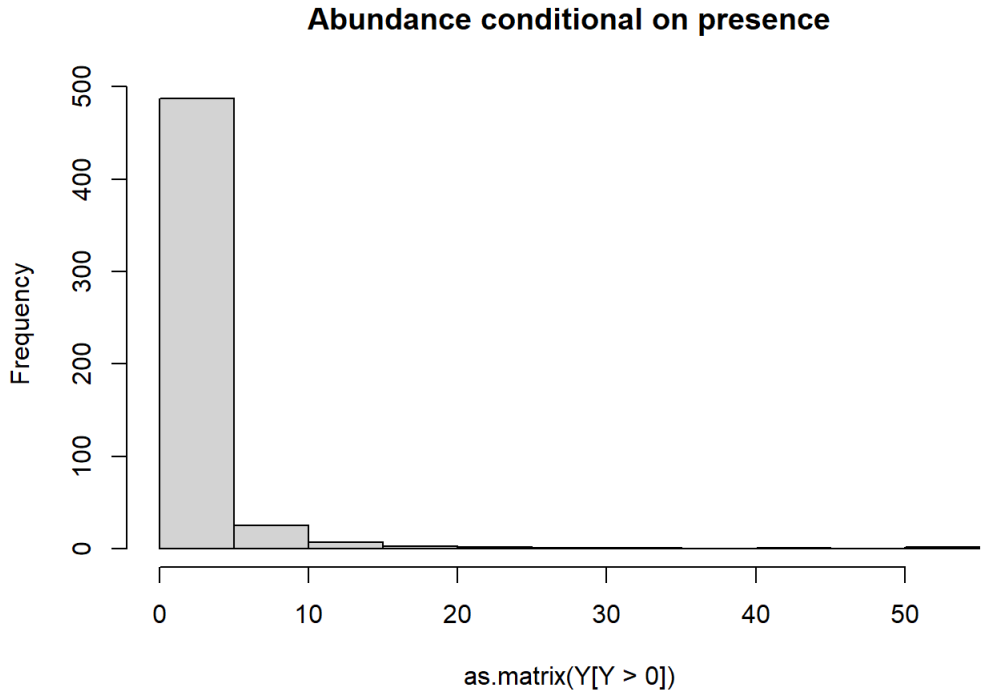
Surveys are spread over the Southern Ocean continental shelf:



Looking at the prevalence for each taxa, and at the log abundance conditional on presence:

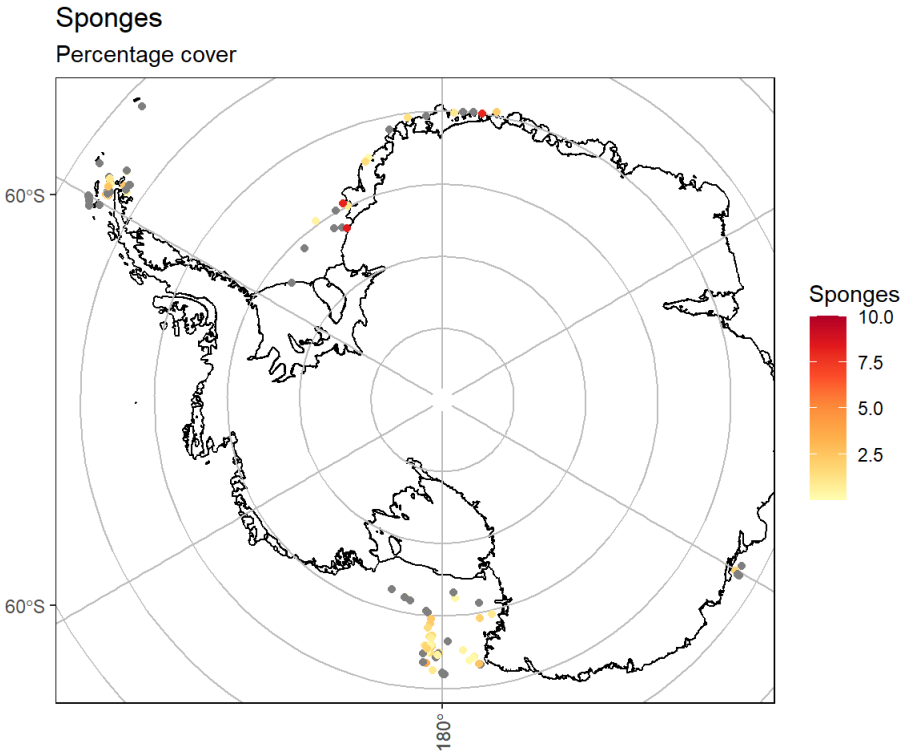
```
## [1] "Prevalence"
```

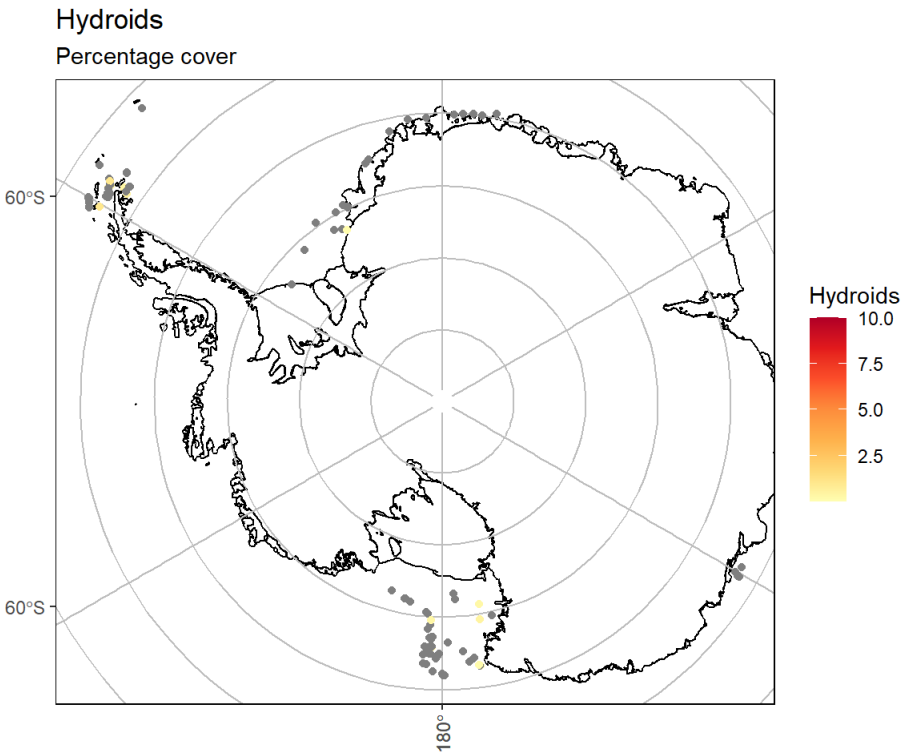
```
##    Anemones   Ascidians     Bryozoa HydroCorals    Hydroids  Octocorals
##       0.170       0.520       0.485       0.155       0.165       0.370
##     Sponges       Worms
##       0.550       0.230
```



… we observe that:

- the prevalence varies between taxa: Sponges are present in 55.0% of the cells while Hydroids are only present in 16.5% of the sampling units (see maps below: grey dots = absent).
- the abundance is highly zero-inflated.





Response variables

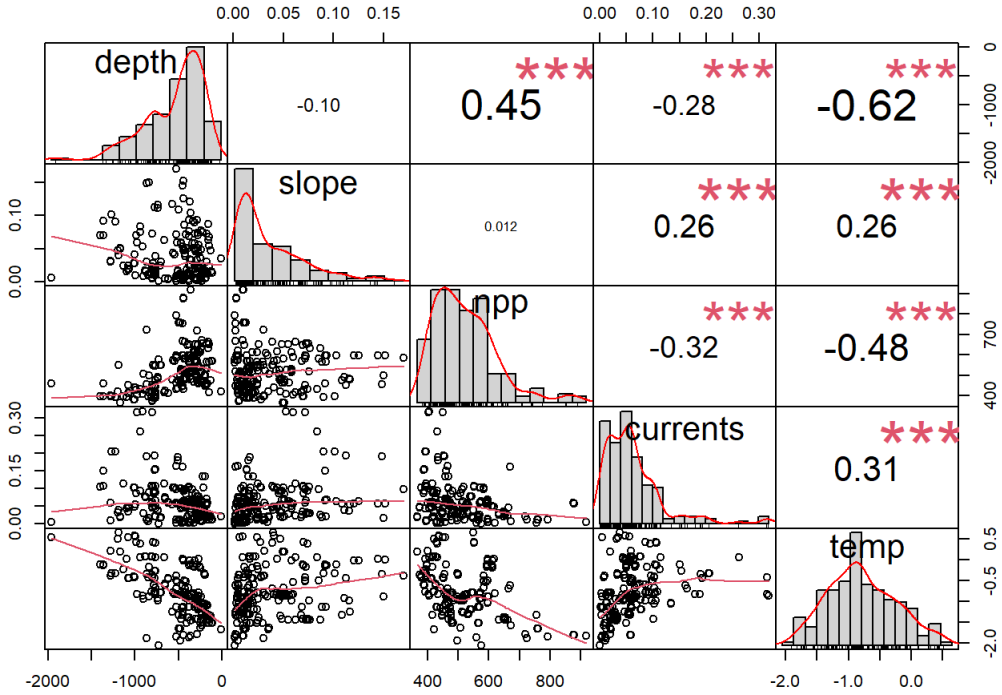
The percentage cover of each taxon is used as response variable. Due to the zero-inflated nature of the data, we resort to a Hurdle model, i.e. one model for presence-absence and another one for abundance conditional on presence (henceforth abundance COP model). We apply probit regression in the presence-absence model, and linear regression for transformed percentage cover data in the abundance COP model. The percentage cover data is transformed by declaring zeros as missing data, log-transforming, and then scaling the data to zero mean and unit variance within each taxon.

```
# Presence-absence model
Ypa = 1*(Y>0)
m1 = Hmsc(Y=Ypa, XData = X,  XFormula = XFormula,
          distr="probit")

# Log-normal abundance data conditional on presence
Yabu = Y
Yabu[Yabu==0] = NA
Yabu=log(Yabu)
# Log-normal model for abundance conditional on presence
m2 = Hmsc(Y=Yabu, YScale = TRUE,
          XData = X,  XFormula = XFormula,
          distr="normal")

# Hurdle model
models = list(m1,m2)
```

Predictors

We include five continuous variables as fixed effects: depth, temperature, slope, NPP, and current speed.

```
##      depth              slope               npp           currents
##  Min.   :-1959.45   Min.   :0.001077   Min.   :366.6   Min.   :0.0003197
##  1st Qu.: -768.79   1st Qu.:0.010559   1st Qu.:434.3   1st Qu.:0.0235201
##  Median : -440.38   Median :0.020567   Median :500.4   Median :0.0519949
##  Mean   : -535.57   Mean   :0.036221   Mean   :519.6   Mean   :0.0612806
##  3rd Qu.: -282.25   3rd Qu.:0.052864   3rd Qu.:582.9   3rd Qu.:0.0768935
##  Max.   :  -11.58   Max.   :0.170071   Max.   :918.1   Max.   :0.3180830
##       temp
##  Min.   :-2.0470
##  1st Qu.:-1.2493
##  Median :-0.8419
##  Mean   :-0.7948
##  3rd Qu.:-0.3908
##  Max.   : 0.6451
```



Model fitting

We fit the HMSC model with the R-package Hmsc (Tikhonov et al. 2020) assuming the default prior distributions. We sampled the posterior distribution with four Markov Chain Monte Carlo (MCMC) chains, each of which was run for 375 iterations, of which the first 125 were removed as burn-in. The chains were thinned by 1 to yield 250 posterior samples per chain and so 1000 posterior samples in total.

```
# Hyperparamters
samples = c(250)
#thin_list = c(1, 10, 100, 1000) # thin is the thinning parameter of the MCMC chain.
thin = c(1)
nChains = 4
```

Note: `thin=1` is here used for prototyping. Thinning is applied to avoid storing model objects of very large size. Higher thinning will be used in future works.

Predictor selection

Comparison between:

- **No predictor selection**, i.e. using all available predictor variables
- Predictor selection using **spike and slab prior, jointly for all taxa**
- Predictor selection using **spike and slab prior, separately for each taxon**
- Predictor selection using **PCA**
- Predictor selection using **Reduced Rank Regression**

Note: See code here for implementation details.

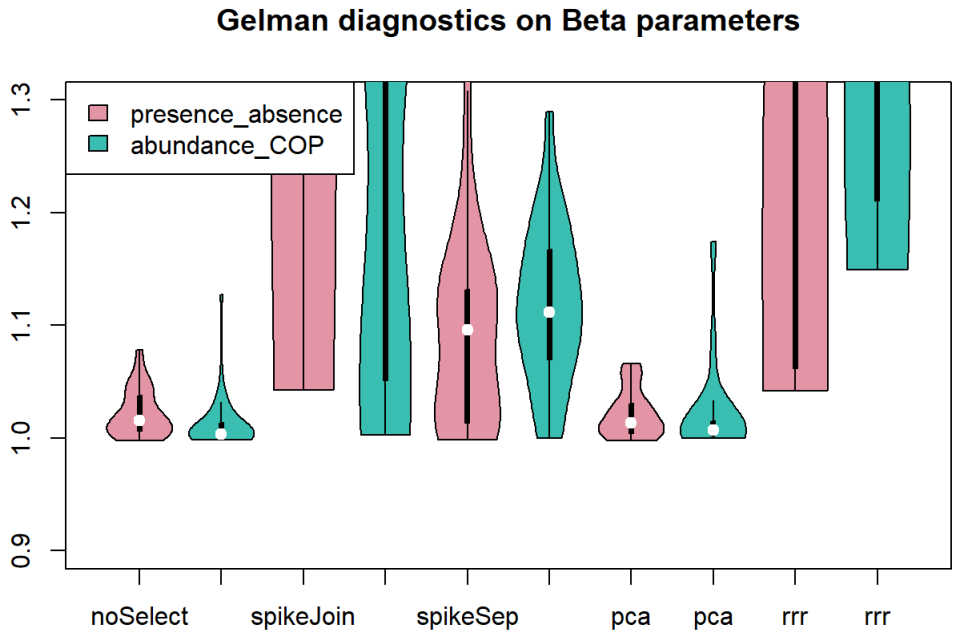
We define one model for each of the above:

```
XFormula = ~depth + slope + temp + npp + currents
# No selection
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expPred", spatial=TRUE, predictor_selection=FALSE)
# Spike and Slab prior, jointly
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expPred", spatial=TRUE, predictor_selection="spike_
slab_jointly")
# Spike and Slab prior, separately
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expPred", spatial=TRUE, predictor_selection="spike_
slab_separately")
# PCA
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expPred", spatial=TRUE, predictor_selection="pca")
# RRR
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expPred", spatial=TRUE, predictor_selection="rrr")
```

Evaluate convergence

We extract the posterior distribution from the model object and convert it into a coda object. We examine MCMC convergence by examining the potential scale reduction factors (Gelman, Rubin, and others 1992) of each model parameters. This diagnostic compares if different chains give consistent results.

```
## [1] "presence_absence"
## [1] "abundance_COP"
## [1] "presence_absence"
## [1] "abundance_COP"
## [1] "presence_absence"
## [1] "abundance_COP"
## [1] "presence_absence"
## [1] "abundance_COP"
## [1] "presence_absence"
## [1] "abundance_COP"
```



We observe that only the models `noSelect` (i.e. no predictor selection) and `pca` (i.e. predictor selection using PCA) reached convergence. In the next subsection, we evaluate the model fit of these two approaches only.

Evaluate model fit

We examine the explanatory and predictive powers of the probit models through taxon-specific AUC (Pearce and Ferrier 2000) and Tjur's R2 (Tjur 2009) values. The explanatory and predictive powers of the abundance COP models are measured by R2.

To compute explanatory power, we make model predictions based on models fitted to all data. To compute predictive power, we performe

5-fold cross-validation, in which the sampling units were assigned randomly to five folds, and predictions for each fold is based on model fitted to data on the remaining four folds.

```
# Explanatory power - Presence Absence model
print(ep.pa.mat)
```

```
##        noSelect            pca
## AUC    "0.857076020757913" "0.837368693494784"
## TjurR2 "0.282925587881566" "0.236575291701225"
```

```
# Explanatory power - Abundance COP model
print(ep.cop.mat)
```

```
##    noSelect            pca
## R2 "0.416930672586489" "0.502748497175392"
```

```
# Predictive power - Presence Absence model
print(pp.pa.mat)
```

```
##        noSelect            pca
## AUC    "0.666567767990675" "0.592440999917068"
## TjurR2 "0.107062457406027" "0.0369037952688906"
```

```
# Predictive power - Abundance COP model
print(pp.cop.mat)
```

```
##    noSelect              pca
## R2 "0.00747350653537623" "-0.0632657130565071"
```

Conclusion

Both approaches have reasonably good explanatory power (AUC > 80% for the presence absence models) but have very very low predictive power (AUC < 70% for the presence absence models and abs(R2) < 0.1 for the abundance COP models). Overall both approaches perform similarly. In the remaining of this work, we do not perform predictor selection.

Benefit of spatial model

In this section, we assess whether a spatial model (i.e. which accounts explicitly for spatial variation) has a better explanatory and predictive power than a non-spatial model. The spatial model accounts for the spatial nature of the study design by including a spatially explicit random effect based on the sample site coordinates. The random effect is implemented through the predictive Gaussian process for big spatial data (see Tikhonov et al. 2020).

Let's define the two models, both without predictor selection:

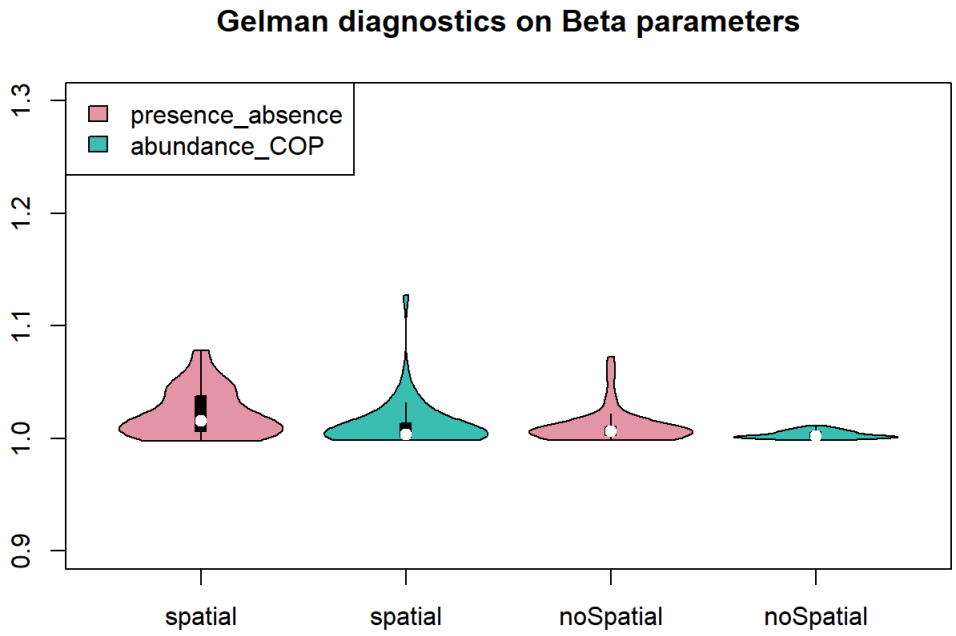
```
XFormula = ~depth + slope + temp + npp + currents
# Spatial model
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expSpatial", spatial=TRUE, predictor_selection=FALS
E)
# Non spatial model
define_hurdle_model(S, X, Y, XFormula, "../qms517/models/expSpatial", spatial=FALSE, predictor_selection=FAL
SE)
```

Note: See here for implementation details.

Evaluate convergence

As previously, we evaluate the models' convergence by looking at the Gelman diagnostic.

```
## [1] "presence_absence"
## [1] "abundance_COP"
## [1] "presence_absence"
## [1] "abundance_COP"
```



Both approaches converged (i.e. spatial and non-spatial models).

Evaluate model fit

We evaluate the model fit in the same way than in the previous section, by looking at its explanatory and predictive powers.

```
# Explanatory power - Presence Absence model
print(ep.pa.mat)
```

```
##         spatial             noSpatial
## AUC    "0.857076020757913" "0.728531230529402"
## TjurR2 "0.282925587881566" "0.139446482303469"
```

```
# Explanatory power - Abundance COP model
print(ep.cop.mat)
```

```
##    spatial             noSpatial
## R2 "0.416930672586489" "0.123977062696847"
```

```
# Predictive power - Presence Absence model
print(pp.pa.mat)
```

```
##         spatial             noSpatial
## AUC    "0.666567767990675" "0.675103442103995"
## TjurR2 "0.107062457406027" "0.110316236451613"
```

```
# Predictive power - Abundance COP model
print(pp.cop.mat)
```

```
##    spatial              noSpatial
## R2 "0.00747350653537623" "0.00107956683471399"
```

Conclusion

The spatial approach leads to a better explanatory power: AUC of 87% vs. 73% for the presence-absence models and R2 of 0.42 vs. 0.12 for the COP abundance models. However, both approaches have very low predictive power. In the last section of this work, we use the spatial model.

Spatial prediction and uncertainty mapping

In this last section, we make some spatial predictions (extrapolation) of the spatial distribution of one taxon, `Sponges`, using the spatial model with no predictor selection. We import a grid of spatial coordinates and environmental predictors for 10,000 locations.

```
source("../hmsc_pipeline/make_spatial_predictions.R")

grid_path <- "C:/Users/cgros/code/qms517/grid.csv"
grid = read.csv(grid_path, stringsAsFactors=TRUE)
S = grid[, 2:3]  # Study design data
X = grid[, 4:8] # Predictors

model_path <- "../qms517/models/expSpatial/models_spatial_thin_1_samples_250_chains_4.Rdata"

fname_out <- "grid_predictions.Rdata"

# Run predictions
make_spatial_predictions(S, X, model_path, fname_out)
```

Looking at the results for the presence-absence model:

```
fname_preds <- "grid_predictions.Rdata"
load(fname_preds)
predY_pa <- results$presence_absence

print(length(predY_pa)) # samples x nChains
```

```
## [1] 1000
```

```
print(dim(predY_pa[[1]])) #  nLocations x nSpecies
```

```
## [1] 10000     8
```

`predY_pa` is a list of samples x nChains matrices. Each matrix has dim: nLocations x nSpecies, filled in with occurrence probabilities.

Looking at the posterior mean prediction:

```
EpredY_pa = apply(simplify2array(predY_pa), 1:2, mean)
EpredY_pa = as.data.frame(EpredY_pa)

print(head(EpredY_pa))
```

```
##    Anemones Ascidians   Bryozoa HydroCorals  Hydroids Octocorals   Sponges
## 1 0.3144769 0.6826737 0.6856444  0.02358385 0.2575399  0.4970546 0.7272056
## 2 0.3197489 0.7074154 0.7085730  0.02713046 0.2719252  0.5074000 0.7525488
## 3 0.3254537 0.7048774 0.6930848  0.02665288 0.2560496  0.5038434 0.7523335
## 4 0.3271514 0.7193351 0.7109836  0.02814259 0.2647902  0.5065228 0.7679802
## 5 0.3278367 0.7168786 0.7130896  0.02853801 0.2601335  0.5110931 0.7639925
## 6 0.3225934 0.7194596 0.7136013  0.02815193 0.2685267  0.5093004 0.7661796
##       Worms
## 1 0.1776526
## 2 0.1823183
## 3 0.1804827
## 4 0.1850760
## 5 0.1840698
## 6 0.1856507
```

where `EpredY_pa` (nLocations x nSpecies) is filled with posterior mean occurrence probabilities.

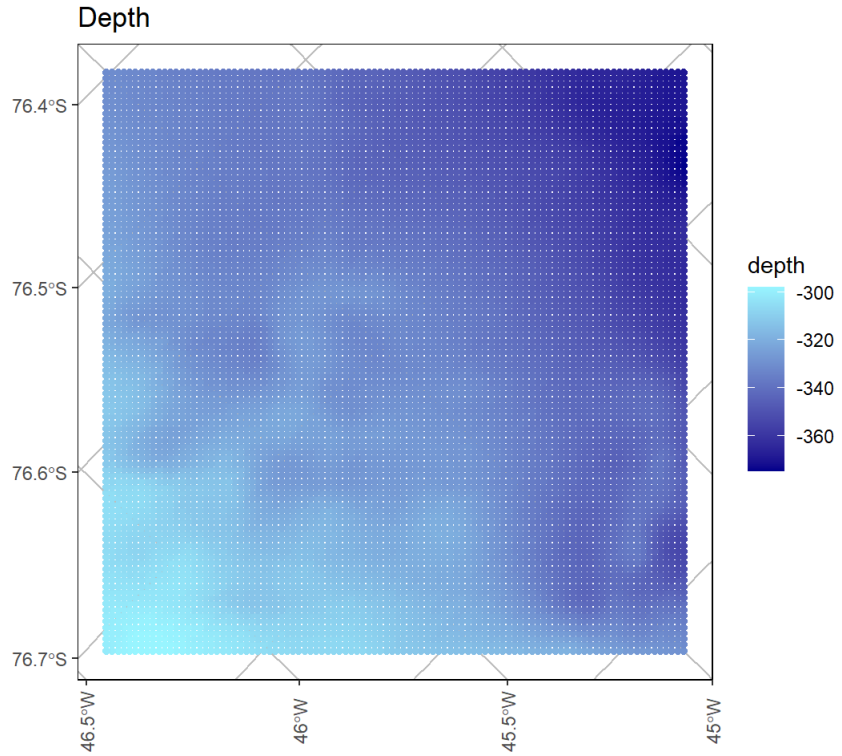
To estimate the parameter uncertainty, we can compute the posterior width (95% confidence interval):

```
ci <- function(x, z=1.96) {
  mean(x) + z * sd(x) / sqrt(length(x)) - (mean(x) - z * sd(x) / sqrt(length(x)))
}
CI95predY_pa = apply(simplify2array(predY_pa), 1:2, ci)
CI95predY_pa = as.data.frame(CI95predY_pa)

print(head(CI95predY_pa))
```

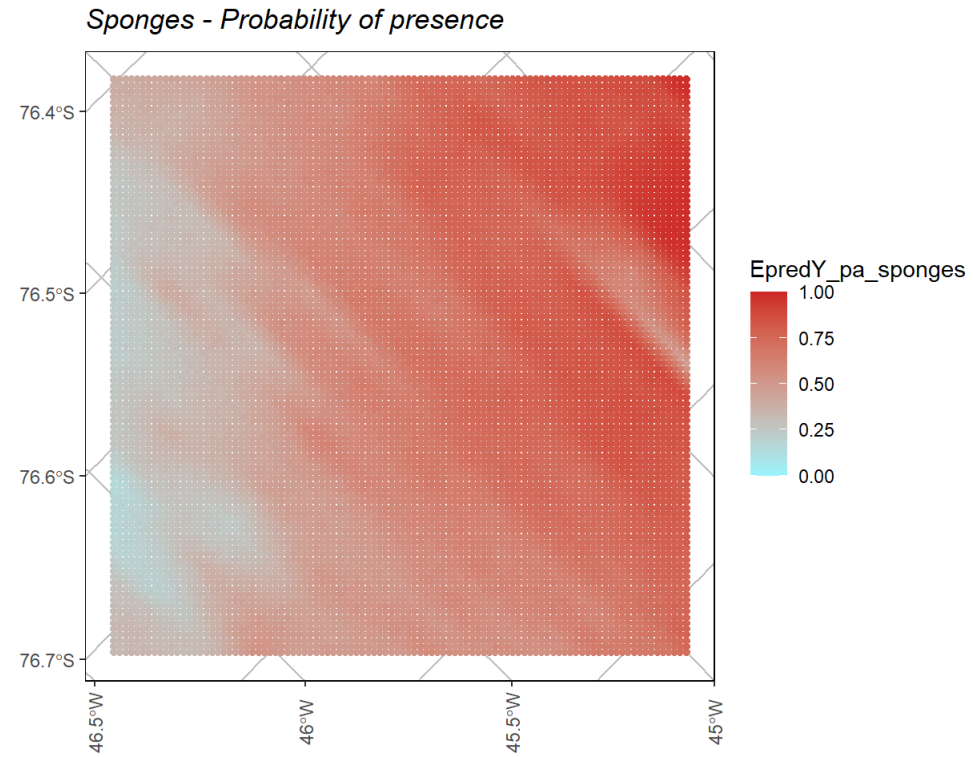
```
##      Anemones  Ascidians    Bryozoa HydroCorals   Hydroids Octocorals    Sponges
## 1 0.01217195 0.03323830 0.03654252 0.003737593 0.02503540 0.01946801 0.02976310
## 2 0.01204834 0.03289601 0.03620538 0.004633308 0.02615388 0.02022212 0.02957694
## 3 0.01261282 0.03221514 0.03621689 0.004673724 0.02530651 0.02017928 0.02899641
## 4 0.01250842 0.03183004 0.03457606 0.004643190 0.02573384 0.01987372 0.02776559
## 5 0.01276230 0.03230967 0.03568878 0.005037035 0.02411133 0.02002990 0.02902482
## 6 0.01256855 0.03234198 0.03516051 0.004931511 0.02562679 0.01989932 0.02851960
##         Worms
## 1 0.01113929
## 2 0.01180671
## 3 0.01185566
## 4 0.01217816
## 5 0.01164988
## 6 0.01198141
```

Let's look at predictions the occurrence probability of `Sponges` and species richness.

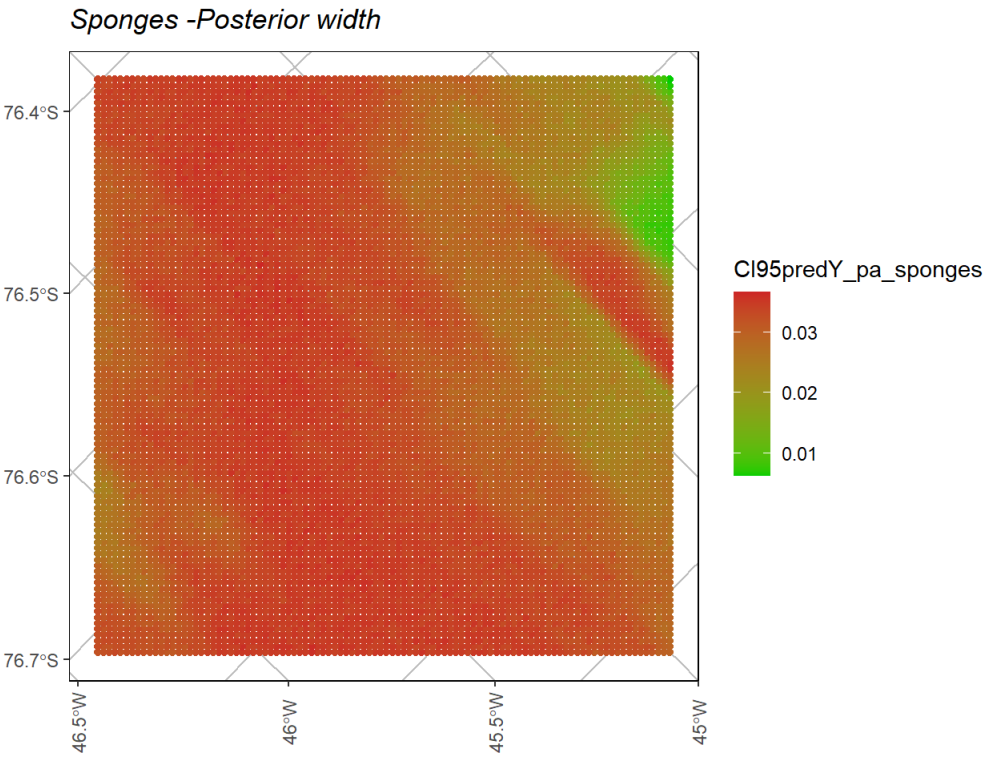
We first plot variation in depth, one of the predictors on which the predictions are based on.



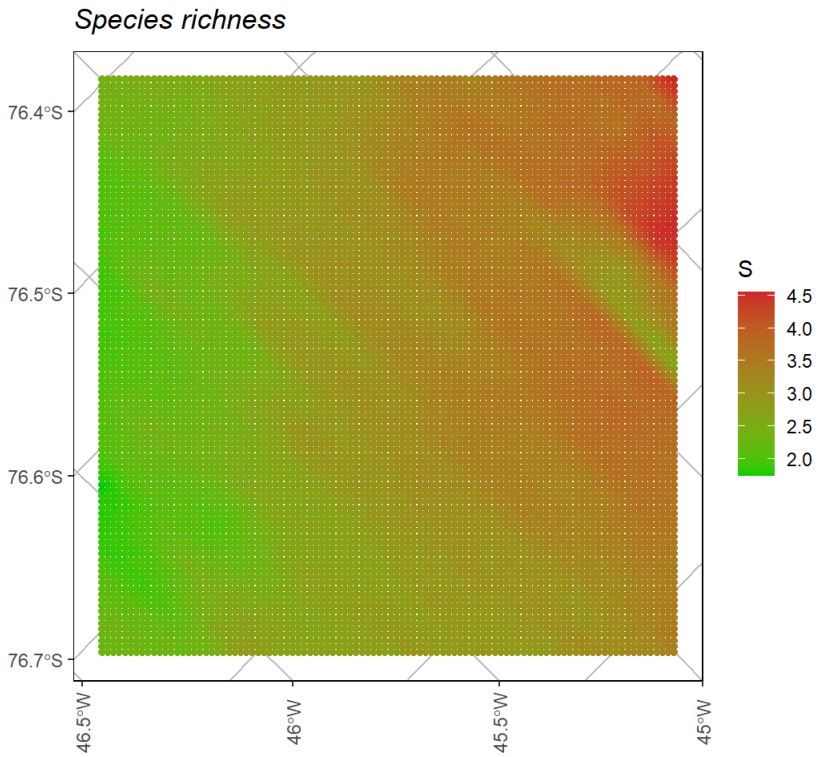
We then exemplify prediction for one focal taxon, `Sponges`:



… and the associated parameter uncertainty:



We finally plot predicted species richness.



References

Gelman, Andrew, Donald B Rubin, and others. 1992. “Inference from Iterative Simulation Using Multiple Sequences.” *Statistical Science* 7 (4): 457–72.

Pearce, Jennie, and Simon Ferrier. 2000. “Evaluating the Predictive Performance of Habitat Models Developed Using Logistic Regression.” *Ecological Modelling* 133 (3): 225–45.

Tikhonov, Gleb, Øystein H Opedal, Nerea Abrego, Aleksi Lehikoinen, Melinda MJ de Jonge, Jari Oksanen, and Otso Ovaskainen. 2020. “Joint Species Distribution Modelling with the R-Package Hmsc.” *Methods in Ecology and Evolution* 11 (3): 442–47.

Tjur, Tue. 2009. “Coefficients of Determination in Logistic Regression Models-a New Proposal: The Coefficient of Discrimination.” *The American Statistician* 63 (4): 366–72.

Warton, David I, F Guillaume Blanchet, Robert B O’Hara, Otso Ovaskainen, Sara Taskinen, Steven C Walker, and Francis KC Hui. 2015. “So Many Variables: Joint Modeling in Community Ecology.” *Trends in Ecology & Evolution* 30 (12): 766–79.