***Notes:***

This document is a draft of how the Hmsc-based analyses could be described in the subsection Statistical analysesthat could be part of the section Materials and Methods.Note that this document does not describe the data collection process, we assume that it is described in earlier sections of Materials and Methods.

This document has several sub-subheadings to clarify its structure. For an actual manuscript, only few sub-subheadings or none of them will be needed. The document is currently a minimalistic version, slightly longer explanations could be added e.g. to motivate the choices.

This document describes the model that could eventually be fitted to the data. The pilot version of the model that has actually been fitted deviates from the description of this document as follows: the pilot model has been fitted using less MCMC iterations than needed for the actual model. Further, cross-validation is 2-fold instead of 5-fold.

***Statistical analyses***

*Overview*

We analyzed the data with a Hierarchical Model of Species Communities (HMSC; Ovaskainen et al., 2017, Ovaskainen and Abrego 2020), a joint species distribution model (Warton et al., 2015).

*Study design*

As described above, the data consist of the estimated biomass of 518 Malaise-trapped forest invertebrate taxa (quasi-probability species) with a DNA spike-in standard used to estimate how the species' biomass changes from sample to sample. The samples were taken during two time periods (sessions) using 1-2 traps at each sampled site.

*Response variable*

The response variable was a vector of the estimated biomass of each of the 518 taxa across the trap samples. Due to the zero-inflated nature of the data, we fitted a hurdle model, i.e. one model for presence-absence data (probit regression), and another model for abundance conditional on presence (a linear regression of the log-transformed biomass data).

*Data selection*

Species-level analyses are not informative for species with very sparse data, so we included in these analyses only those species with at least five occurrences across samples.

*Environmental covariates*

We included as fixed effects the mean elevation and forest canopy height (both continuous variables) at the sample sites and the minimum number of years since logging.

*Phylogeny*

To examine whether phylogenetic relatedness influences the similarity of species’ environmental responses, we assumed a phylogenetically structured residual. As phylogenetic data, we included a quantitative phylogenetic tree.

*Random effects*

To account for the nature of the study design, we included a spatially explicit random effect, implemented through the predictive Gaussian process for big spatial data (Tikhonov et al. 2020b) for the sample sites. The temporal sampling session (1 vs 2) was also included as a random effect, because smoke from a fire in the broader region may have altered species distributions and co-occurrences between the two sessions.

*Model fitting*

We fitted the model with the R-package Hmsc (Tikhonov et al., 2020a) assuming the default prior distributions. We sampled the posterior distribution with four MCMC chains, each of which was run for 375,000 iterations, of which the first 125,000 were removed as burn-in. The iterations were thinned by 1000 to yield 250 posterior samples per chain, and thus 1000 posterior samples in total. To explore the rate of MCMC convergence we first fitted otherwise identical models but with 375 iterations (burn-in 125, thin 1), 3,750 iterations (burn-in 1,250, thin 10) and 37,700 iterations (burn-in 12,500, thin 100).

We examined the explanatory and predictive powers of the probit models through taxon-specific AUC and Tjur R2, which measure how well the model discriminates those cells at which the taxon occurs from those in which it does not occur. The explanatory and predictive powers of the normal models were measured by R2. To compute explanatory power, we computed model predictions based on models fitted to all the data. To compute predictive power, we performed a five-fold cross validation, in which the sampling units were assigned randomly to five folds, and predictions for each fold were based on a model fitted to data on the remaining four folds.

*Addressing the study questions with the fitted models*

[TO BE WRITTEN]

**References**

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. and Abrego, N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* **20,** 561-576

Ovaskainen, O. and Abrego, N. 2020. Joint Species Distribution Modelling – With Applications in R. *Cambridge University Press*.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. and Ovaskainen, O. 2020a. Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution* **11**, 442-447.

Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D. and Ovaskainen, O. 2020b. Computationally efficient joint species distribution modelling of big spatial data. *Ecology*, e02929.

Warton, D., Blanchet, G., O’Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. and Hui, F. K. C. 2015. So many variables: joint modelling in community ecology. *Trends in Ecology and Evolution* **30**, 766-779