The data for the case study originate from Purhonen et al. (2020):

Purhonen, J., Ovaskainen, O., Halme, P., Komonen, A., Huhtinen, S., Kotiranta, H., Læssøe, T. and Abrego, N. 2020. Morphological traits predict host-tree specialization in wood-inhabiting fungal communities. *Fungal Ecology* **46**, 100863.

**Brief description of the study design:** The study was carried out in central Finland. All 12 study sites were semi-natural spruce dominated forests that varied relatively little in their age and management history. From each forest, 16 large (base diameter > 15 cm) naturally fallen logs were selected: four birch logs, four spruce logs, four pine logs and four aspen logs. The study thus consists of a total of 12 x 16 = 192 logs. These logs were thoroughly surveyed for fungal sexual fruiting bodies, which were observed in total for 657 species. For these species, information about their traits and taxonomy were compiled.

**Environmental data are provided in the file environment.csv.** These data contain information about the 192 logs (column id) included in the study: their tree species (column tree), their volume in cubic meters (column volume), their decay stage classified as 2, 3 or 4 (column decay), the name of the forest site (column site), and an index characterizing how natural the forest is (column index; the higher, the more natural).

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | id | tree | volume | decay | site | index |
| 2 | 1 | Spruce | 0.6678345 | 2 | Kuu | 25 |
| 3 | 2 | Spruce | 0.8387816 | 3 | Kuu | 25 |
| 4 | 3 | Spruce | 0.7244591 | 3 | Kuu | 25 |
| 5 | 4 | Spruce | 0.8868480 | 2 | Kuu | 25 |
| 6 | 5 | Pine | 1.5393804 | 3 | Kuu | 25 |
| 7 | 6 | Pine | 1.6986827 | 3 | Kuu | 25 |
| 8 | 7 | Pine | 0.8867145 | 3 | Kuu | 25 |
| 9 | 8 | Pine | 0.5150143 | 2 | Kuu | 25 |
| 10 | 9 | Birch | 0.2567990 | 2 | Kuu | 25 |
| 11 | 10 | Birch | 0.8398419 | 2 | Kuu | 25 |
| 12 | 11 | Birch | 1.4109625 | 2 | Kuu | 25 |
| 13 | 12 | Birch | 1.1833620 | 2 | Kuu | 25 |
| 14 | 13 | Aspen | 0.6268524 | 2 | Kuu | 25 |
| 15 | 14 | Aspen | 2.2058823 | 2 | Kuu | 25 |
| 16 | 15 | Aspen | 1.8035333 | 2 | Kuu | 25 |
| 17 | 16 | Aspen | 2.1555723 | 2 | Kuu | 25 |
| 18 | 17 | Spruce | 0.6367170 | 2 | Tik | 6 |
| 19 | 18 | Spruce | 1.1338911 | 3 | Tik | 6 |
| 20 | 19 | Spruce | 0.5834251 | 3 | Tik | 6 |

*Shown is the head of the file environment.csv.*

**Species occurrence data are provided in the file species.csv.** These data contain information about the presence-absences of the 657 species (the columns with species names) on the 192 logs (column id) included in the study.

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | id | Acanthosti | Acrogenos | Actidium.h | Alutaceod | Amphinem | Amphispha | Amphispha | Amylocort | Amylocort | Amylocort | Amylocor |
| 2 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 6 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 11 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 13 | 12 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 13 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

*Shown is the head of the file species.csv.*

**Species trait and taxonomical data are provided in the file traits.csv.** These data contain the following information about the presence-absences of the 657 species (column species) included in the study: fruit body type (column fb), thickness of spore cell wall (column wall), whether the spores have ornamentation (column orn), the shape of the spore (column shape; ratio of length to width), spore volume (column volume), and taxonomical classification (columns phylum, class, order, family, genus).

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | species | fb | wall | orn | shape | volume | genus | family | order | class | phylum | |
| 2 | Acanthostigma.sp1. | Pyrenoid | Thin | No | 2.2368421 | 96.399770 | Acanthosti | Tubeufiace | Tubeufiale | Dothideon | Ascomycota | |
| 3 | Acrogenospora.carmichaeliana | Pyrenoid | Thick | No | 2.1428571 | 1948.2783 | Acrogenos | Hysteriace | Hysteriale | Dothideon | Ascomycota | |
| 4 | Actidium.hysterioides | Pyrenoid | Thin | No | 6.4444444 | 57.653133 | Actidium | Mytilinidia | Mytilinidia | Dothideon | Ascomycota | |
| 5 | Alutaceodontia.alutacea | Resupinate | Thin | No | 4.2424242 | 14.967725 | Alutaceod | Schizopora | Hymenoch | Agaricomy | Basidiomycota | |

*Shown is the head of the file traits.csv.*

**OVERVIEW OF EXERCISES 1-4.** All four exercises are based on the same case study. Your starting point are the three data files described above, and the following pipeline for Hmsc scripts:

- S2_fit_models.R (you must define a model before running this one)
- S3_evaluate_convergence.R (you must run S2 before running this one)
- S4_compute_model_fit.R (you must run S2 before running this one)
- S5_show_model_fit.R (you must run S4 before running this one)
- S6_show_parameter_estimates.R (you must run S2 before running this one)
- S7_make_predictions.R (you must run S2 before running this one)

What is NOT provided is the beginning of the pipeline, where you define the model. So your main challenge will be to read in the data, and write the script “S1_define_models.R” that reads in the datafiles, and produces as output the file “unfitted_models.RData” which contains the model(s) that you have defined. After defining the model, your main challenge is to apply the Hmsc pipeline (scripts S2-S7) to fit the model and make some inference and predictions. To help you with defining

the model, you can use as a template the scripts of the plant case study used in the first days R-demonstration. Note that this file does not apply directly, as it concerns another case study, it just works as an example to show how Hmsc models are defined.

**EXERCISE 1.**

**Tasks.**

- Write an R-script named "S1_define_models.R" that reads in these data (except traits.csv), defines a highly simplified Hmsc model (include only species data and environmental covariates, but not yet traits, taxonomy, nor the random effect of the site), and saves the model to the file "unfitted_models.RData".
- Run the following parts of the Hmsc pipeline for this model to generate some standard output: "S2_fit_models.R" (samples=250, thin=1 is sufficient), and "S6_show_parameter_estimates" (we consider the other parts of the pipeline in the later exercises).
- Explore the parameter estimates and try to interpret them!

**Hints/suggestions.**

- Select only those species that occur at least 20 times (or at least 10 times) in the data, to avoid modelling species with very little information in the data, and to minimize computational times (hint: use colSums() function).
- Try to first construct a very simple model that makes some sense and that you can fit without error messages; you can then make the model more complex/complete as needed.
- Make some selection of environmental covariates based on your intuition and exploration of the variables (we think more about model selection in later exercises).
- Make sure that the data are consistent and that your model makes sense.
    - Are the sampling units in the same order for the environmental data and species occurrence data?
    - Are the covariates continuous or factors?
    - Should you apply some transformations to the covariates?

**EXERCISE 2.**

**Tasks.** Repeat Exercise 1 so that you now include also traits, taxonomy and the random effect of site in the model. After you are done with Exercise 2, if possible, leave the model fitting with at least until thin=10 (this will take some time), as that will be helpful for Exercise 3.

**Hints/suggestions.**

- Make some selection of the traits based on your intuition and exploration of the variables.
- Include a taxonomical tree as a proxy for phylogeny.
- Account for the hierarchical study design in the random effect structure. You can optionally include a sampling unit level random effect as well in case you are interested in residual species associations.
- Make sure that the data are consistent and that your model makes sense.
    - Are the species in the same order in the species matrix and the trait/taxonomy matrix, and do all the names match?

- Should you apply some transformations to the covariates?

**EXERCISE 3.**

**Tasks.** Continue from Exercise 2 by applying the other parts of the Hmsc pipeline to check MCMC convergence (S3; ideally, fit the models at least until thin=10), examine model fit (S4 and S5), and make predictions over environmental gradients (S7).

**EXERCISE 4.**

**Tasks.** Continue from Exercises 1-3 by defining some alternative models and selecting among them.

**Hints/suggestions.**

- Let's assume that you are interested in how much knowing the naturalness index of the site helps in making predictions to new sites. Then you may define alternative models that either include or not include the naturalness index as covariate.
    - To study the predictive powers of the models, you may apply site-level cross validation to ask specifically how well the models would predict to new sites. You can achieve this by partition = createPartition(m, nfolds = nfolds, column="site"). If you wish to apply leave-out-out cross-validation over the 12 sites, set nfolds=12. [but this may be computationally intensive].
- If you wish to study also the rare species, you can construct a species richness model for the selected and unselected species (a bivariate model where both are included simultaneously) and compare their outputs to see if rare species seem to have systematically different responses than the selected ones.
    - Hint: sum the presences of each category of relevant species (e.g. 'rare' and 'common' species) for this species richness model. These new summed values then become the counts of species at each site, and the categories of species become the 'species' columns in Hmsc. What distribution would you use on this type of model (i.e., *not* probit)?