Incorporating species' correlations

with Joint Species Distribution Models

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- JSDM
- GLLVMs background
- `gllvm` R-package

Questions so far?

Distribution Modelling

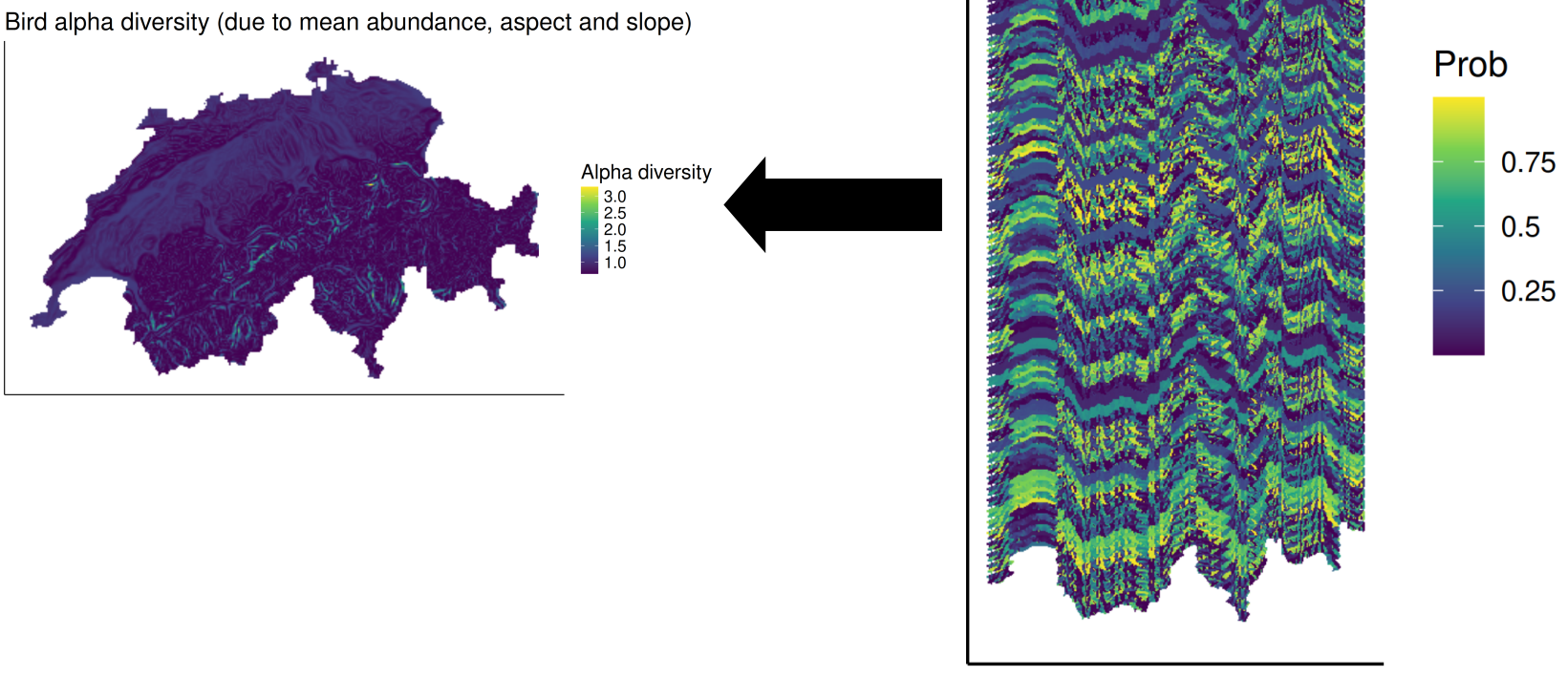
If you have presence-absence data of a species, you fit a Species Distribution Model of the form:

$$y_{ij} = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta} \tag{1}$$

- $\mathbf{x}_i$ is usually a bioclimatic variable
- Then you want to predict where a species may occur
- Potentially based on future climate scenarios
- Similar to the model from yesterday; it requires

Co-occurrence of Swiss birds

Can the distribution of one species inform us of anothers'?



Leveraging shared information

In our mixed-effects model yesterday, we had some components already shared for all species:

The random effect $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

- had mean $\boldsymbol{\mu}$ the same for all species
- had covariance $\boldsymbol{\Sigma}$ the same for all species

Sharing information across species helps; on some species we have more information than others, which we can use to inform ourselves on the occurrence of less frequent species.

Leveraging shared information

In our mixed-effects model yesterday, we had some components already shared for all species:

The random effect $\boldsymbol{\beta}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

- had mean $\boldsymbol{\mu}$ the same for all species
- had covariance $\boldsymbol{\Sigma}$ the same for all species

Sharing information across species helps; on some species we have more information than others, which we can use to inform ourselves on the occurrence of less frequent species.

This is also the idea of JSDM: use co-occurrence information to improve the model's knowledge of the community

Co-occurrence patterns

- if one species occurs somewhere, we know another does too, for various reasons

Species correlation

If we fit a GLM to data of multiple species, we assume **independence** (so we need much more information for accurate estimates)

But, observations of the same species form groups. Co-occurring species have more similar observations than for other species

In GLMM language: **observations of species exhibit correlation**

1) Part of this can be explained by shared environmental responses
2) The other part remains

We never know how much of the "remainder" is explainable by the environment.

Independence

Correlation means non-independence. Violation of the independence assumption cannot be ignored:

- Fixed effect parameter estimates are biased (estimated environmental preferences are wrong)
- Standard errors are too small (inflated type I error; too optimistic)
  - Consequently, p-values and CIs are too small
- Predictions may be poor
- Equivalently, means we have pseudoreplication
- Random effect estimates can be inaccurate

So, there is also a statistical need to adjust the model

Assessing the independence assumption

This is done via:

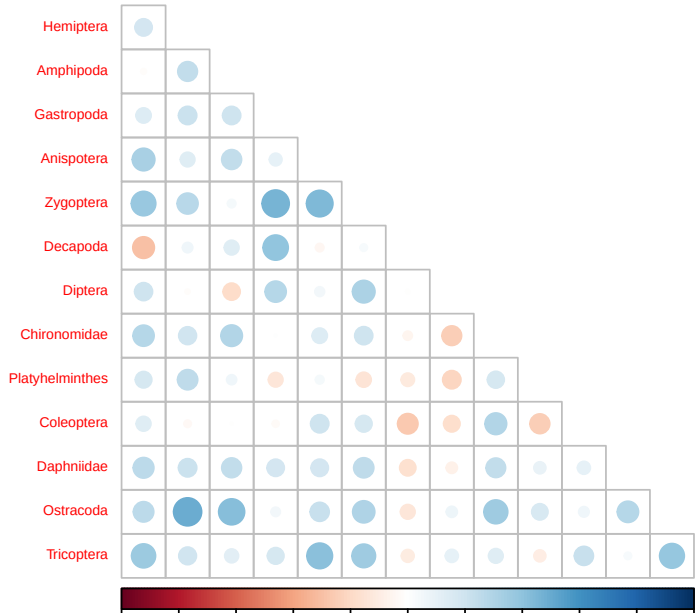
- Nature of the data (clustering)
- Visually via residual diagnostics (see yesterday)

The previous model

```
model4 <- gllvm::gllvm(y, X = X, formula = ~NO3, num.lv = 0,
                       family = "negative.binomial")
```

We can see this in the correlations of the residuals from our model:

```
corrplot::corrplot(cor(residuals(model4)$resi), type = "lower", diag = FALSE, tl.pos = "l", tl.cex = 0.5)
```



Joint Species Distribution Modeling

A decade ago, Joint Species Distribution Models (JSDM) were introduced to model binary data of multiple species

- Pollock et al. (2015): co-occurrence of frogs and trees
- Clark et al. (2015): co-occurrence of trees

The goal: to incorporate covariation of species for better predictions

Species associations

What induces covariation between species?

- Shared environmental responses (abiotic conditions)
- Biotic interactions

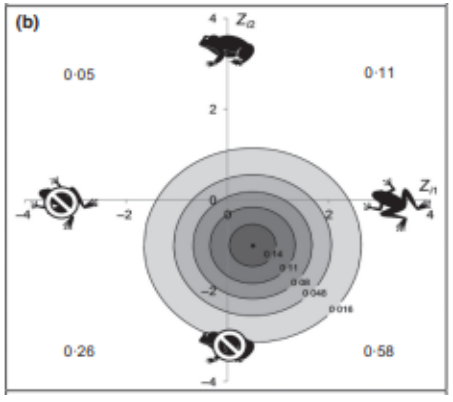
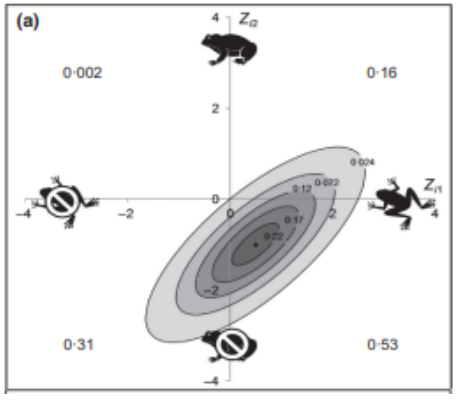
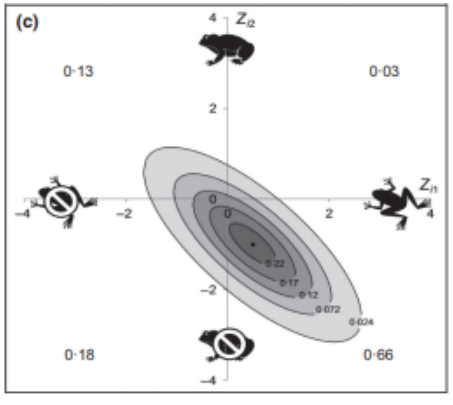


Figure 1: Pollock et al. (2015, fig. 1)



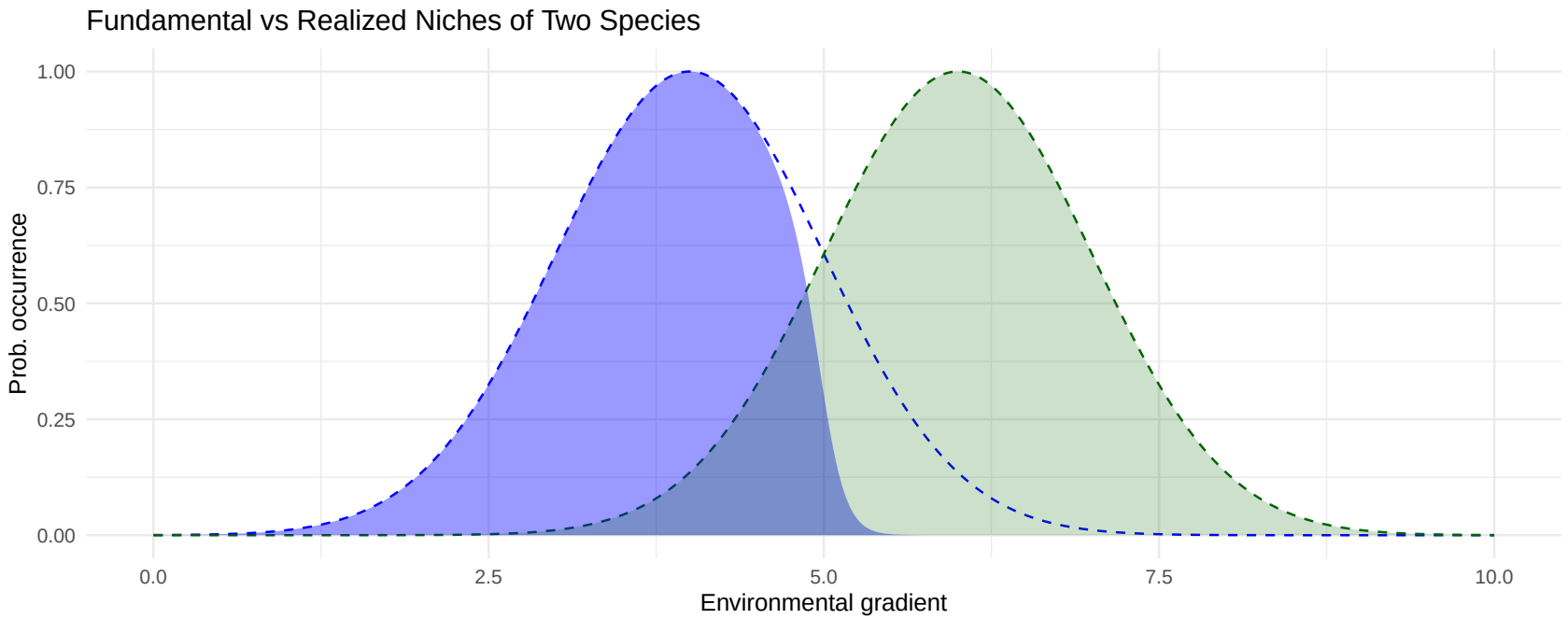


Interactions and co-occurrence

Interactions induce correlation, but we cannot infer interactions from (non-temporal) co-occurrence data.

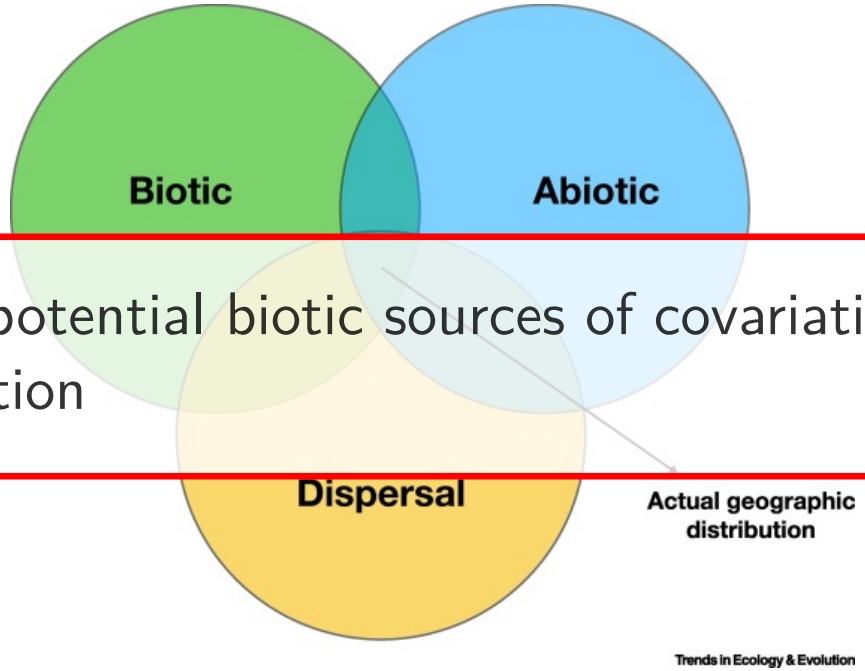
The niche concept

We can also take a different angle; correlation is introduced to improve our estimates for species' responses.



Fundamental niche: total occupiable space without other species' intervention
Realized niche: the space occupied due to other species

When reality kicks in



Incorporating potential biotic sources of covariation is a step in the right direction

Figure 2: Poggiato et al. (2021)

Niches are shaped more than by environment and interactions; historical limitations, dispersal, and other processess prevent us from estimating the fundamental niche.

JSDM vs. classical multivariate analysis

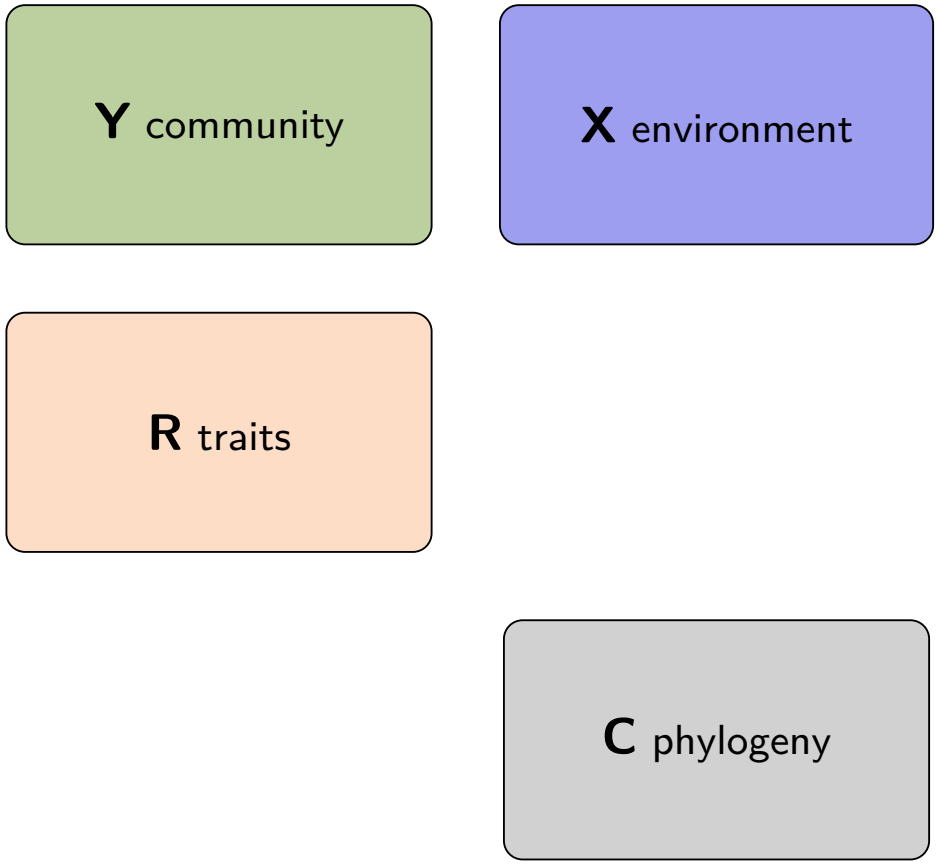
	Classic	JSDM
Focus	Low-dimensional space	Distributions
Goal	Inference	Prediction
Data type	Usually quantitative	Binary
Scale	Local	Regional
Covariates	Environmental	Bioclimatic
Presentation	Ordination diagram	Correlation plot/map
Audience	Community ecologists	Macro ecologists

JSDM vs. classical multivariate analysis

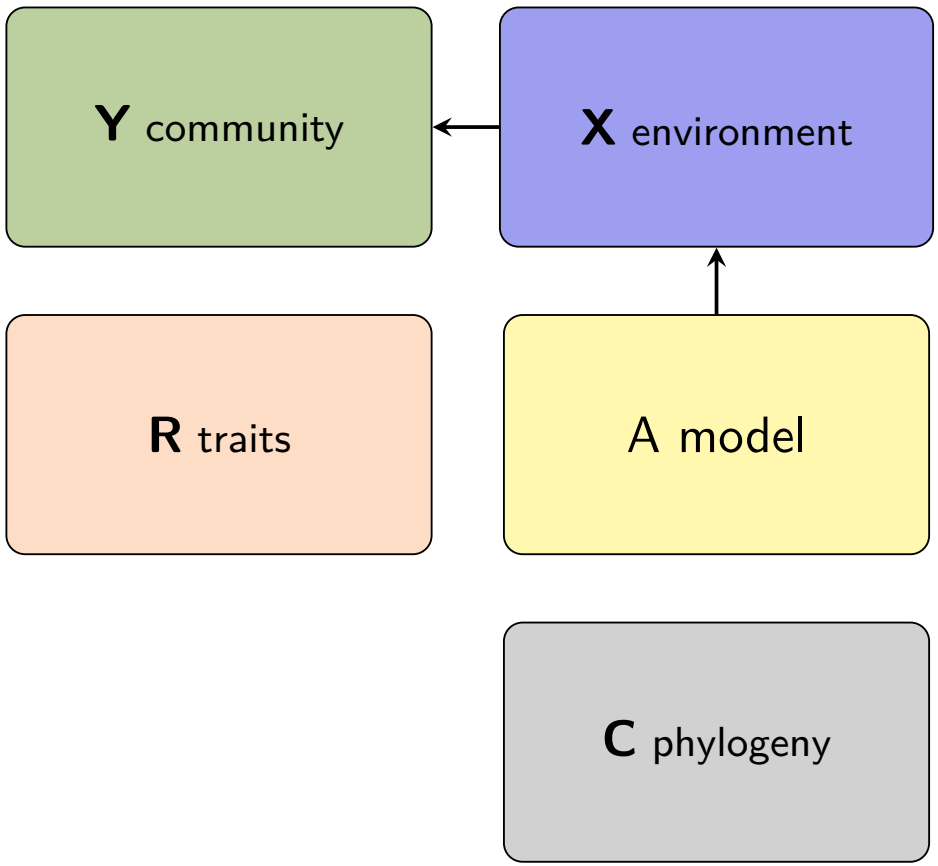
	Classic	JSDM
Focus	Low-dimensional space	Distributions
Goal	Inference	Prediction
Data type	Usually quantitative	Binary
Scale	Local	Regional
Covariates	Environmental	Bioclimatic
Presentation	Ordination diagram	Correlation plot/map
Audience	Community ecologists	Macro ecologists

That is not to say JSDMs cannot be used for non-binary data, for inference, or for local scales

Typical questions in the framework



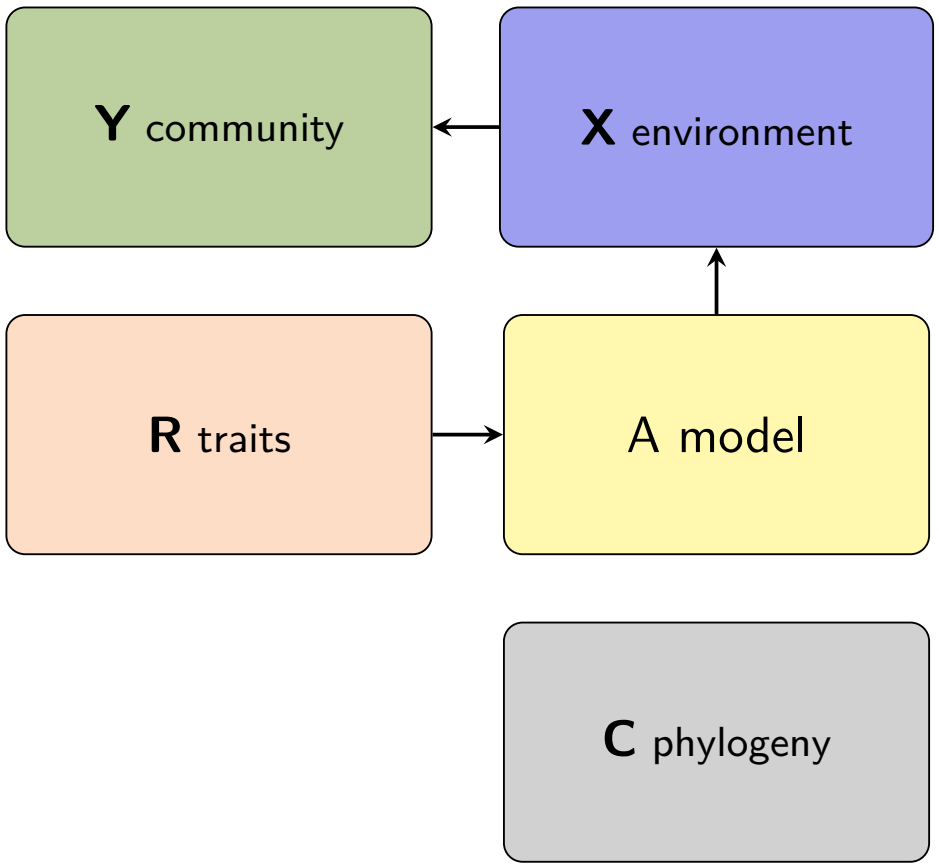
Typical questions in the framework



Q: How does the environment structure the community?

environmental filtering

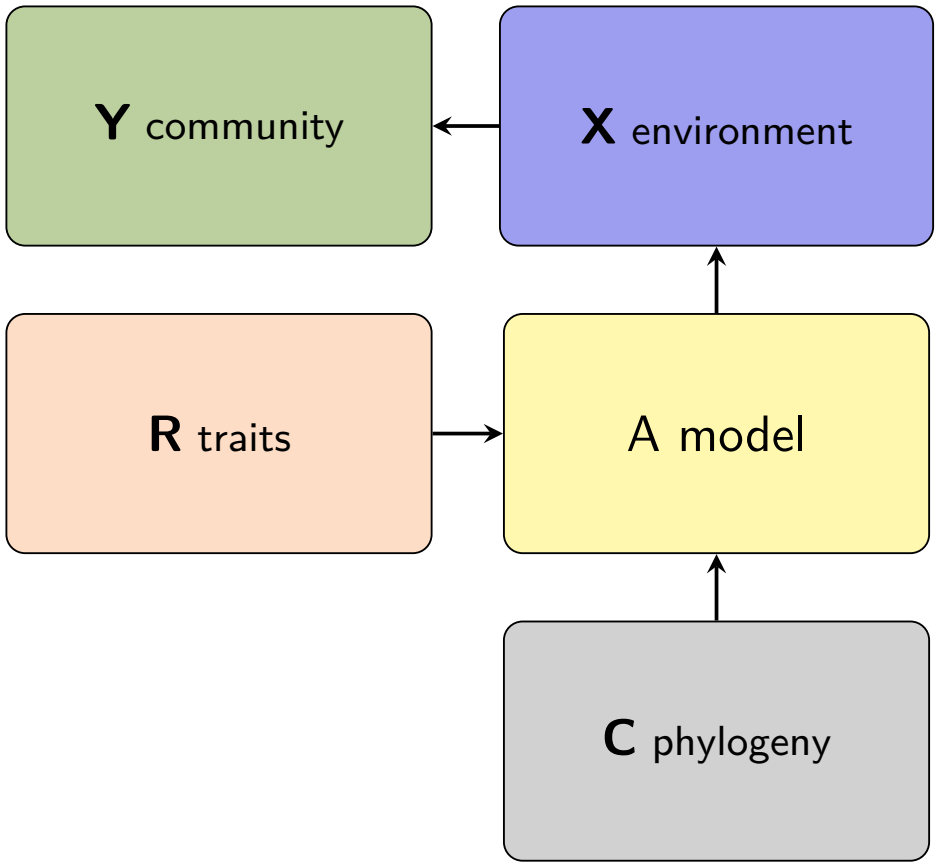
Typical questions in the framework



Q: How do traits affect species' responses to the environment?

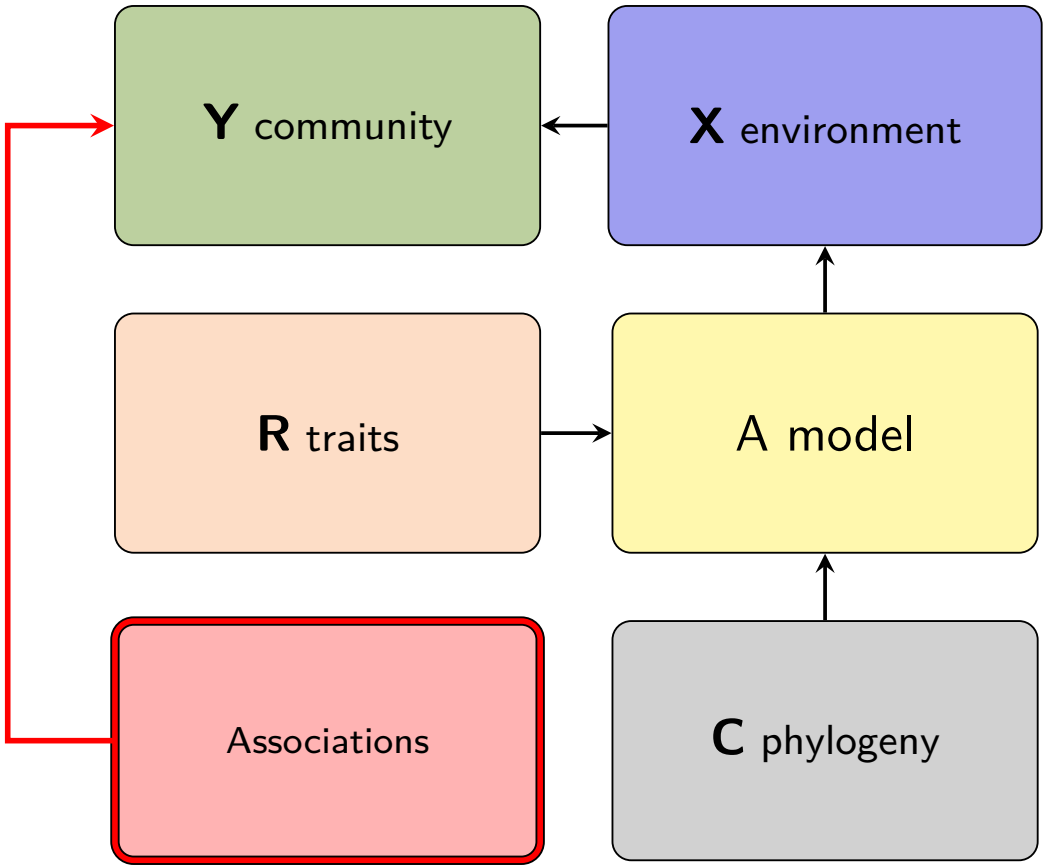
environmental filtering (more later)

Typical questions in the framework



Q: Do species with shared evolutionary history co-occur?
(phylogenetic structuring) environmental filtering

Typical questions in the framework



Q: Do species co-occur **after** the environment has been considered? biotic filtering

Joint Species Distribution Modeling

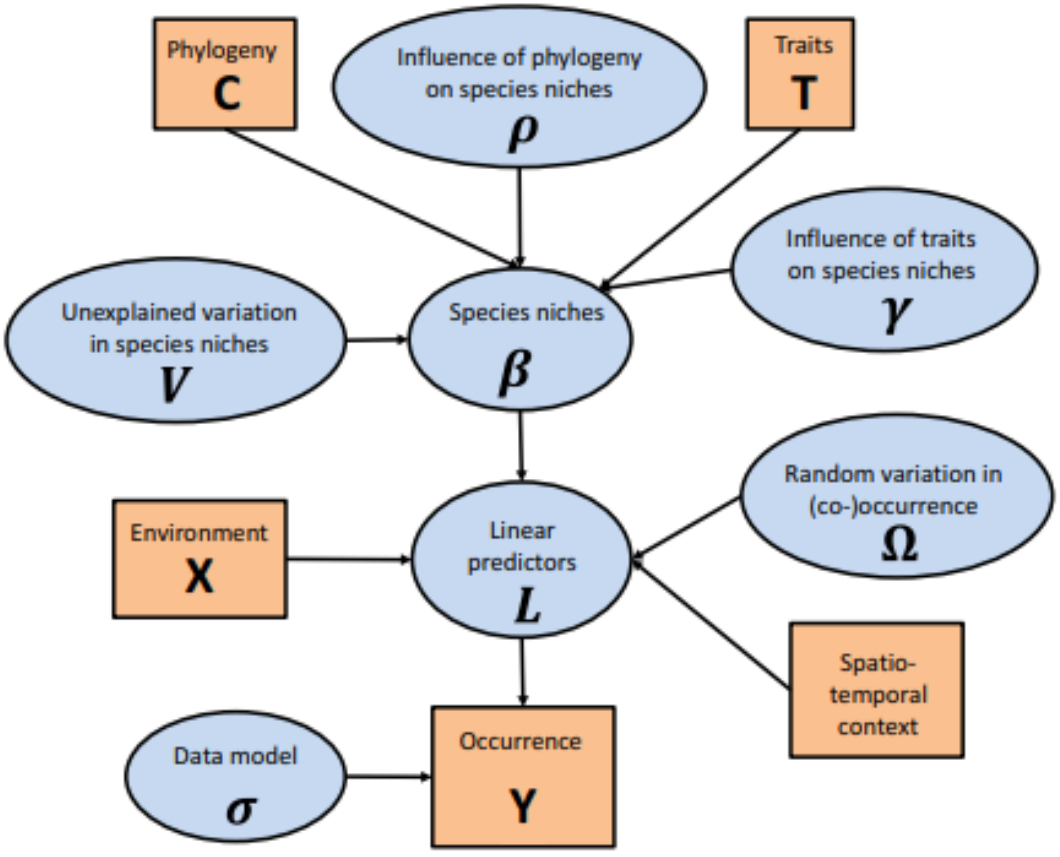


Figure 3: Figure from Ovaskainen et al. (2017)

Joint Species Distribution Modeling

The aim of JSDMs is to incorporate *species associations*

- Species may co-occur due to biotic interactions
- Due to similar environmental preferences
- Or because they have a similar history

Either how, it results in correlations between responses

Joint Species Distribution Model (JSDM)

- For community data, we want to incorporate correlation of species
- We have Multivariate data (in contrast to multivariable)

$$g\{\mathbb{E}(\mathbf{y}_i|\boldsymbol{\epsilon}_i)\} = \boldsymbol{\beta}_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{2}$$

- we add $\boldsymbol{\epsilon}_i$ relative to the VGLM(M)
- This random effect takes care of the left-over (co)variation of species
- so we assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ is the matrix of **species associations**

JSDM: the model

$$\eta_{ij} = \beta_{0j} + \ldots + \epsilon_{ij} \quad (3)$$

- The stuff from yesterday
- $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ is the matrix of *species associations*
- So we expect a positive values of species co-occur, and negative if they do not

Species associations

- Difficult to estimate: there are usually too many parameters
- The number of pairwise associations grows quadratically
  - 2 with 2 species, 6 for 4 species, 45 for 10 species, 4950 for 100

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & sp_{12} & \cdots & sp_{1j} \\ sp_{21} & 1 & \cdots & sp_{2j} \\ \vdots & & \ddots & \vdots \\ sp_{j1} & sp_{j2} & \cdots & 1 \end{bmatrix} \tag{4}$$

This very quickly becomes an issue for fitting models

JSDM: it is just a mixed-effects model

The JSDM is "just" a mixed-effects model. So we can fit it with available software:

In `lme4`:

```
glmer(abundance ~ species + x:species + (0+species|sites), data = data)
```

- There are $p(p+1)/2$ correlations between species
- This model becomes (very) large very quickly do not try this at home
- Will usually not fit
- So we need to do something smart!

JSDM software implementations

There are many!

JSDM software implementations

There are many!

- Boral (Bayesian, slow and somewhat outdated)
- sJSDM (Bayesian, relatively slow, but faster than Boral)
- Hmsc (Bayesian, generally slow, loads of functionality)
- ecoCopla (Frequentist, very fast but limited functionality)
- CBFM (Frequentist, geared towards spatio-temporal analysis)
- sjSDM (Frequentist, very fast but limited functionality, requires python)
- glmmTMB (Frequentist, fast and very versatile, not purpose-coded)
- gllvm (Frequentist, fast and very versatile, purpose-coded)

JSDM software implementations

There are many!

- Boral (Bayesian, slow and somewhat outdated)
- sJSDM (Bayesian, relatively slow, but faster than Boral)
- Hmsc (Bayesian, generally slow, loads of functionality)
- ecoCopla (Frequentist, very fast but limited functionality)
- CBFM (Frequentist, geared towards spatio-temporal analysis)
- sjSDM (Frequentist, very fast but limited functionality, requires python)
- glmmTMB (Frequentist, fast and very versatile, not purpose-coded)
- gllvm (Frequentist, fast and very versatile, purpose-coded)

Which software is most suitable on your aim, data type, and model.

Example: alpine plants in France

- Data by Choler 2005
- Occurrence of 92 species at 75 5 by 5 plots
- 6 environmental variables: aspect, slope, microscale landform, disturbance level (physical and trampling/burrowing), and mean Julian snowmelt date
- In the jSDM package

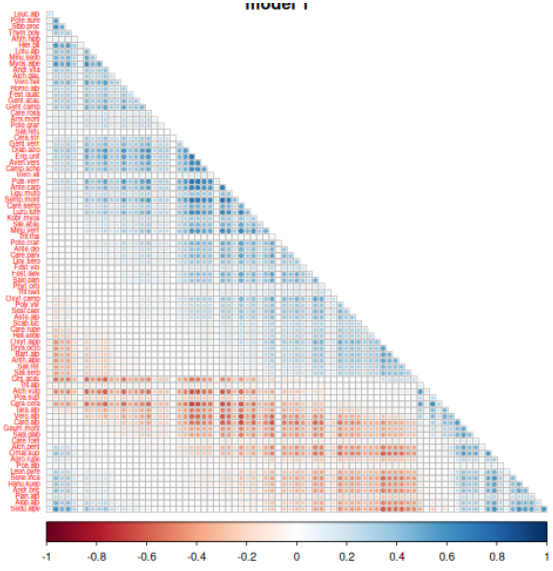
Example: fitting a JSDM

```
model1 <- gllvm(Y, family = "binomial")
```

Pretty straightforward!

Example: visualizing associations

```
cors <- getResidualCor(model1)
corrplot(cors, type = "lower", diag = FALSE, tl.pos = "l", order = "AOE",
```



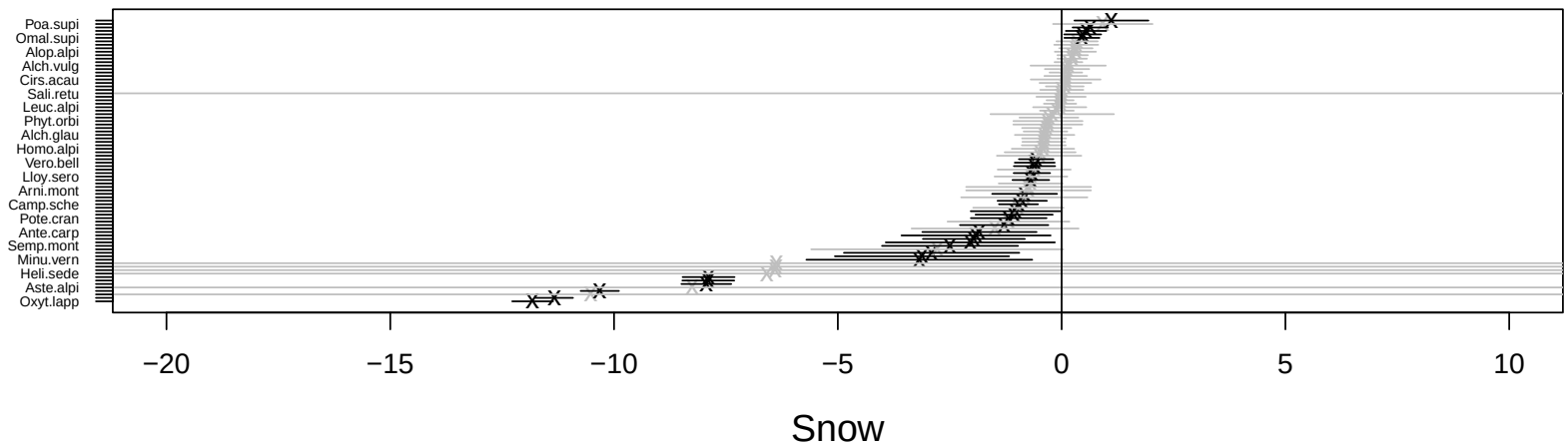
Blue: species that are predicted to co-occur

Red: species that are predicted to avoid each other

Example: adding a environmental variable

Now that we have incorporated associations, we can add in environmental variables as yesterday (fixed or random):

```
model2 <- gllvm(Y, X = X, formula = ~Snow, family = "binomial")
```



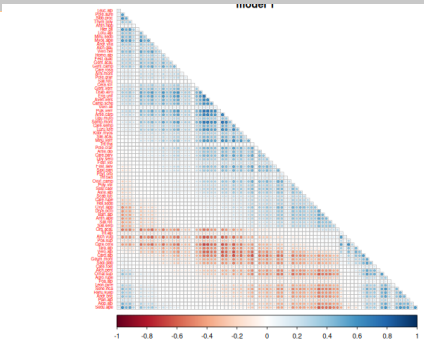
Adding environmental covariates tends to improve the model, but reduce the signal in the associations.

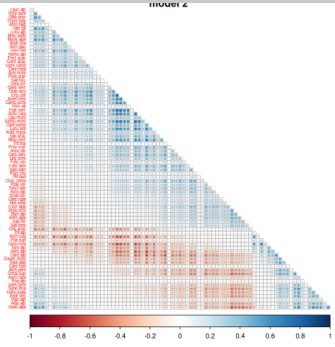
Example: comparing the models

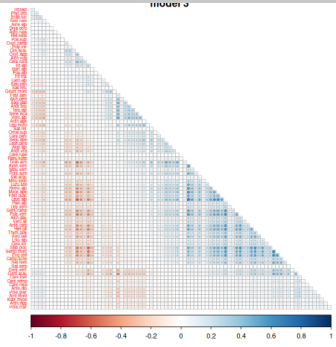
```
model3 <- gllvm(Y, X = X, formula = ~Snow+Form+Slope+Aspect, family = "binomial")
```

```
AIC(model1, model2, model3)
```

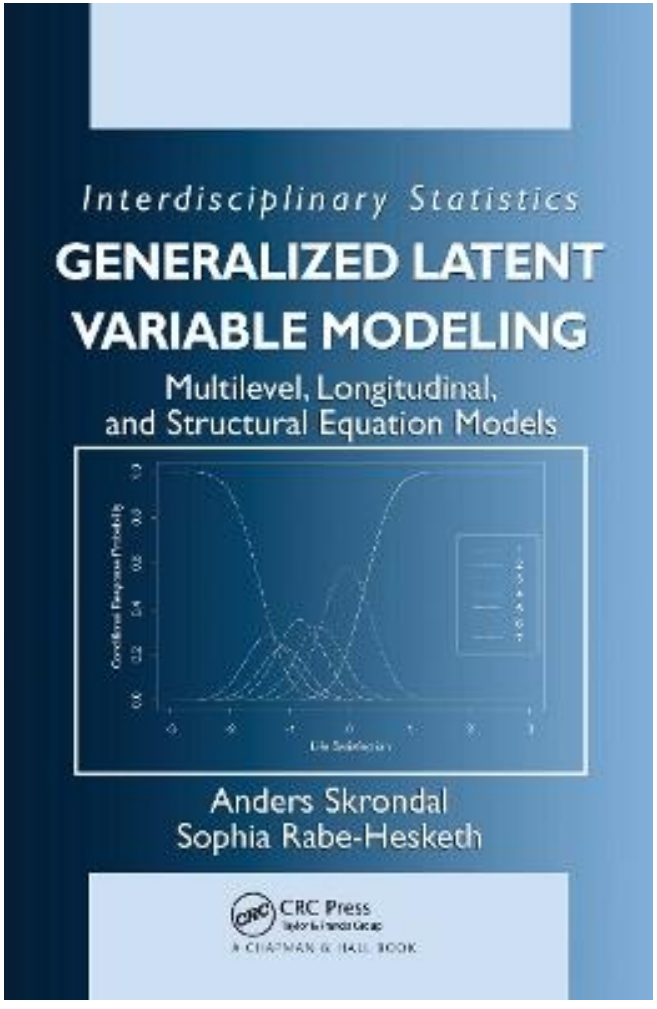
```
##          df      AIC
## model1 245 4148.531
## model2 327 4140.254
## model3 573 4171.055
```







Generalized Linear Latent Variable Models (GLLVMs)



Generalized Linear Latent Variable Models (GLLVMs)

- A framework for model-based multivariate analysis
- That does dimension reduction
- Similarly as in VGLMM, you need to specify:
  1. A distribution
  2. A link function
  3. The model its structure
- But now also the number of dimensions for the associations

Factor analysis to the rescue

- GLLVMs were introduced to ecology as a technical solution to this problem
- We represent the covariance matrix with fewer **dimensions**: $\boldsymbol{\Sigma} \approx \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$

"The factor analytic solution" because factor analysis (Spearman, 1904) is the precursor of GLLVMs

GLLVM: the model

$$\eta_{ij} = \beta_{0j} + \ldots + \epsilon_{ij} \tag{5}$$

- The stuff from yesterday
- $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma})$
- $\boldsymbol{\Sigma}$ is the matrix of *species associations*
- So we expect a positive values of species co-occur, and negative if they do not

GLLVM: the model

$$\eta_{ij} = \beta_{0j} + \ldots + \epsilon_{ij} \tag{5}$$

- The stuff from yesterday
- $\boldsymbol{\epsilon}_i = \mathbf{u}_i^\top \boldsymbol{\Gamma}^\top \sim \mathcal{N}(0, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top)$
- $\boldsymbol{\Sigma}$ is the matrix of *species associations*
- So we expect a positive values of species co-occur, and negative if they do not

Prediction

So, we represent:

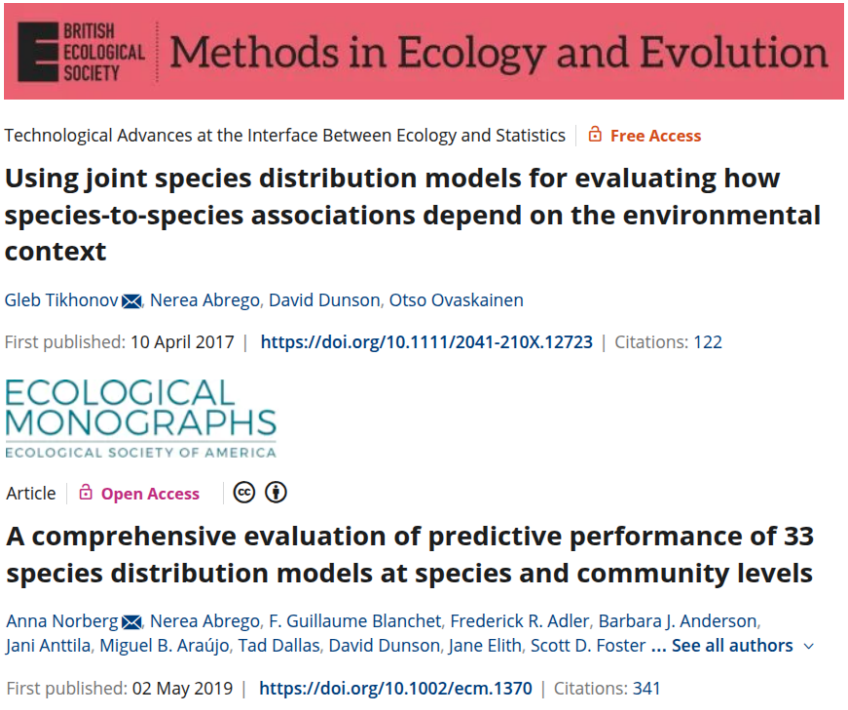
$$\boldsymbol{\Sigma} \approx \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \tag{6}$$

The number of columns in $\boldsymbol{\Gamma}$ is equal to the number of latent variables.

- The more latent variables we use, the better we represent the associations
- But, more latent variables slows down the model!
- So; it is a trade-off that we need to measure (somehow)
  - Can use information criteria or hypothesis tests
  - Variation explained
  - Cross-validation
  - Or some measure of predictive performance

Does it improve predictions?

Yes



BRITISH ECOLOGICAL SOCIETY | Methods in Ecology and Evolution

Technological Advances at the Interface Between Ecology and Statistics | Free Access

Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context

Gleb Tikhonov, Nerea Abrego, David Dunson, Otso Ovaskainen

First published: 10 April 2017 | https://doi.org/10.1111/2041-210X.12723 | Citations: 122

ECOLOGICAL MONOGRAPHS
ECOLOGICAL SOCIETY OF AMERICA

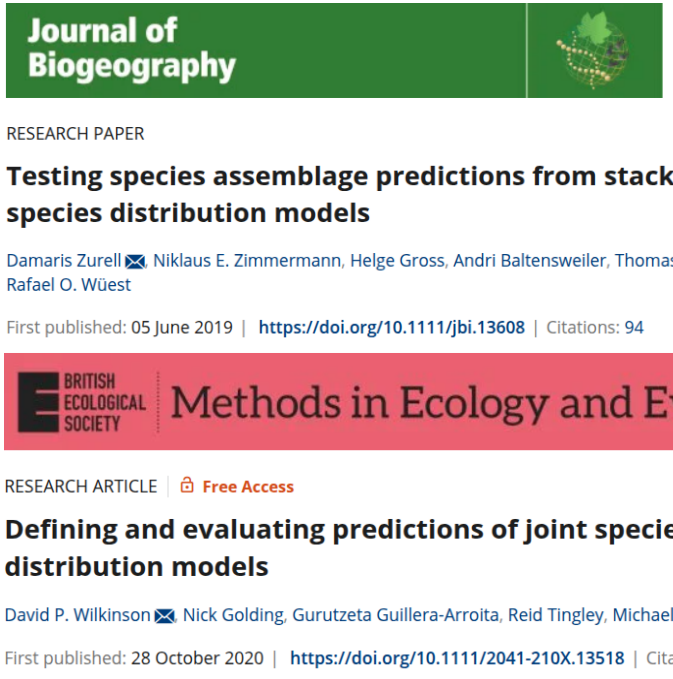
Article | Open Access

A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels

Anna Norberg, Nerea Abrego, F. Guillaume Blanchet, Frederick R. Adler, Barbara J. Anderson, Jani Anttila, Miguel B. Araújo, Tad Dallas, David Dunson, Jane Elith, Scott D. Foster ... See all authors

First published: 02 May 2019 | https://doi.org/10.1002/ecm.1370 | Citations: 341

No



Journal of Biogeography

RESEARCH PAPER

Testing species assemblage predictions from stacked and joint species distribution models

Damaris Zurell, Niklaus E. Zimmermann, Helge Gross, Andri Baltensweiler, Thomas Sattler, Rafael O. Wüest

First published: 05 June 2019 | https://doi.org/10.1111/jbi.13608 | Citations: 94

BRITISH ECOLOGICAL SOCIETY | Methods in Ecology and Evolution

RESEARCH ARTICLE | Free Access

Defining and evaluating predictions of joint species distribution models

David P. Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, Michael A. McCarthy

First published: 28 October 2020 | https://doi.org/10.1111/2041-210X.13518 | Citations: 41

It depends how you evaluate it. Statistically, we need to incorporate non-independence.

Prediction of focal species

For a subset of species we are particularly interested in **A**, we can do **conditional** prediction: utilize information from other species **B** to improve its prediction. We define our residual covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_B \end{bmatrix} \tag{7}$$

we set $\boldsymbol{\epsilon}_{iA} = \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\epsilon}_{iB}$

Now, even if species **A** is absent somewhere, we are also using its known relation with species **B**. But, this does not work if we don't have response data at a site.

Prediction of focal species

For a subset of species we are particularly interested in **A**, we can do **conditional** prediction: utilize information from other species **B** to improve its prediction. We define our residual covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_A & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_B \end{bmatrix} \tag{7}$$

we set $\boldsymbol{\epsilon}_{iA} = \boldsymbol{\Sigma}_{AB}\boldsymbol{\Sigma}_B^{-1}\boldsymbol{\epsilon}_{iB}$

Now, even if species **A** is absent somewhere, we are also using its known relation with species **B**. But, this does not work if we don't have response data at a site.

Note, this requires $\boldsymbol{\Sigma}$ to be full rank.

The residual variance

$\boldsymbol{\Sigma}$ is of full rank, as long as we also have residual variance.

We write our model with a residual:

$$\eta_{ij} = \beta_{0j} + \epsilon_{ij} + e_{ij} \tag{8}$$

- where ϵ_{ij} is our term for covariation as before
- $e_{ij} \sim f(0, \phi_j)$ is an independent residual

e_{ij} takes different forms depending on the response distribution. In some cases (Poisson) it is hard to define.

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^{\star}$ for thresholding.

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^{\star}$ for thresholding.

$$y_{ij} = \begin{cases} 1, & \text{if } \eta_{ij}^{\star} > 0 \\ 0, & \text{otherwise} \end{cases}$$

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^{\star}$ for thresholding.

$$y_{ij} = \begin{cases} 1, & \text{if } \eta_{ij}^{\star} > 0 \\ 0, & \text{otherwise} \end{cases}$$

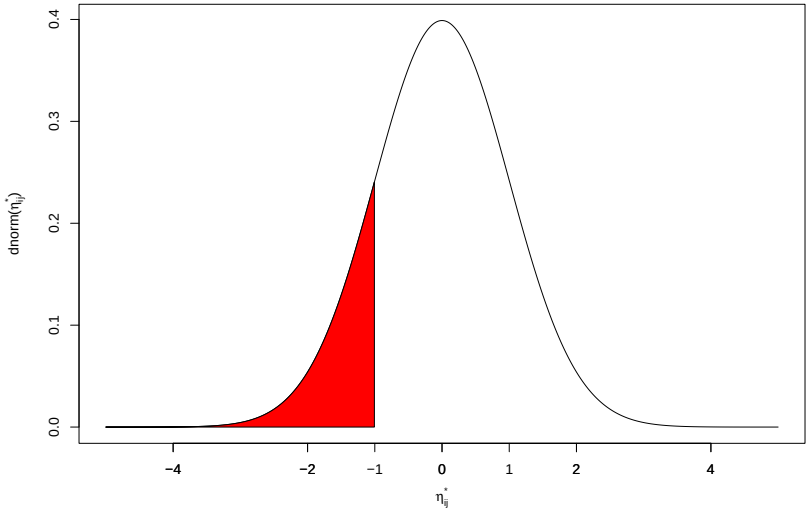
and we model this auxiliary variable:

$$\eta_{ij}^{\star} = \eta_{ij} + E_{ij} \sim \mathcal{N}(0, 1)$$

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^{\star}$ for thresholding.

$$y_{ij} = \begin{cases} 1, & \text{if } \eta_{ij}^{\star} > 0 \\ 0, & \text{otherwise} \end{cases}$$



and we model this auxiliary variable:

$$\eta_{ij}^{\star} = \eta_{ij} + E_{ij} \sim \mathcal{N}(0, 1)$$

Which is the same as:

$$p_i = \Phi(\eta_{ij})$$

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^{\star}$ for thresholding.

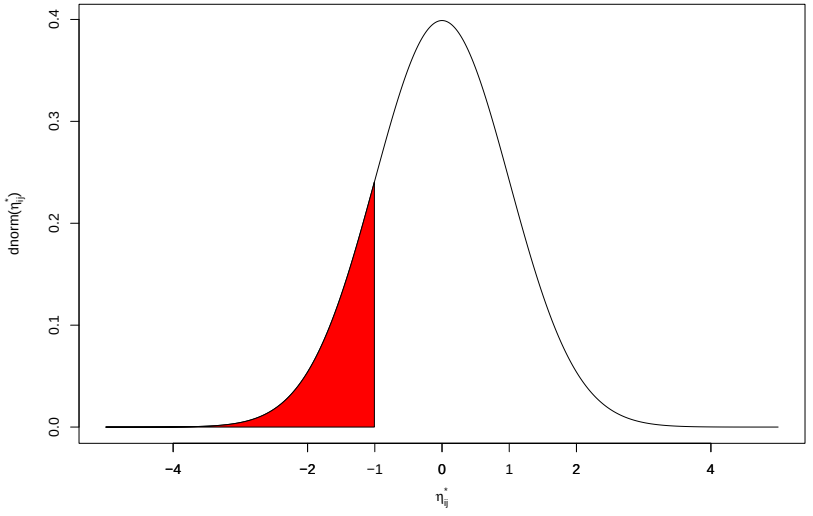
$$y_{ij} = \begin{cases} 1, & \text{if } \eta_{ij}^{\star} > 0 \\ 0, & \text{otherwise} \end{cases}$$

and we model this auxiliary variable:

$$\eta_{ij}^{\star} = \eta_{ij} + E_{ij} \sim \mathcal{N}(0, 1)$$

Which is the same as:

$$p_i = \Phi(\eta_{ij})$$



if $\eta_{ij}^{\star}$ is positive, we have 1 and 0 if it is negative

Link functions: probit

In probit regression, we use a latent variable $\eta_{ij}^\star$ for thresholding.

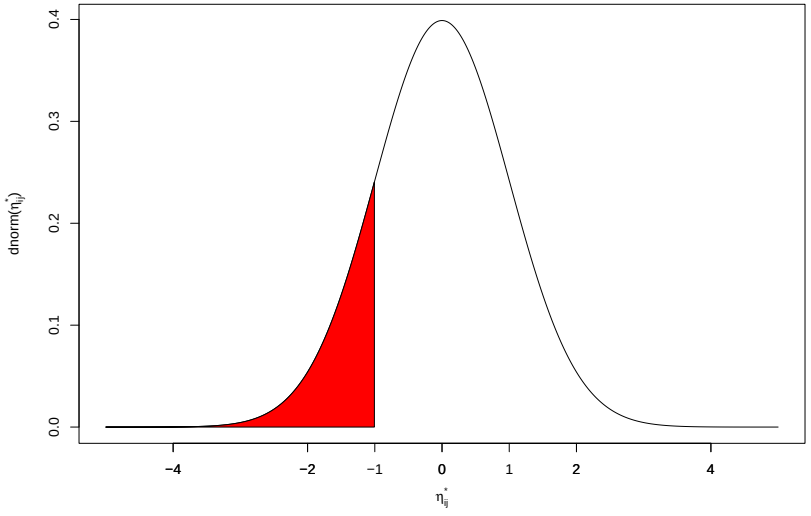
$$y_{ij} = \begin{cases} 1, & \text{if } \eta_{ij}^\star > 0 \\ 0, & \text{otherwise} \end{cases}$$

and we model this auxiliary variable:

$$\eta_{ij}^\star = \eta_{ij} + E_{ij} \sim \mathcal{N}(0, 1)$$

Which is the same as:

$$p_i = \Phi(\eta_{ij})$$

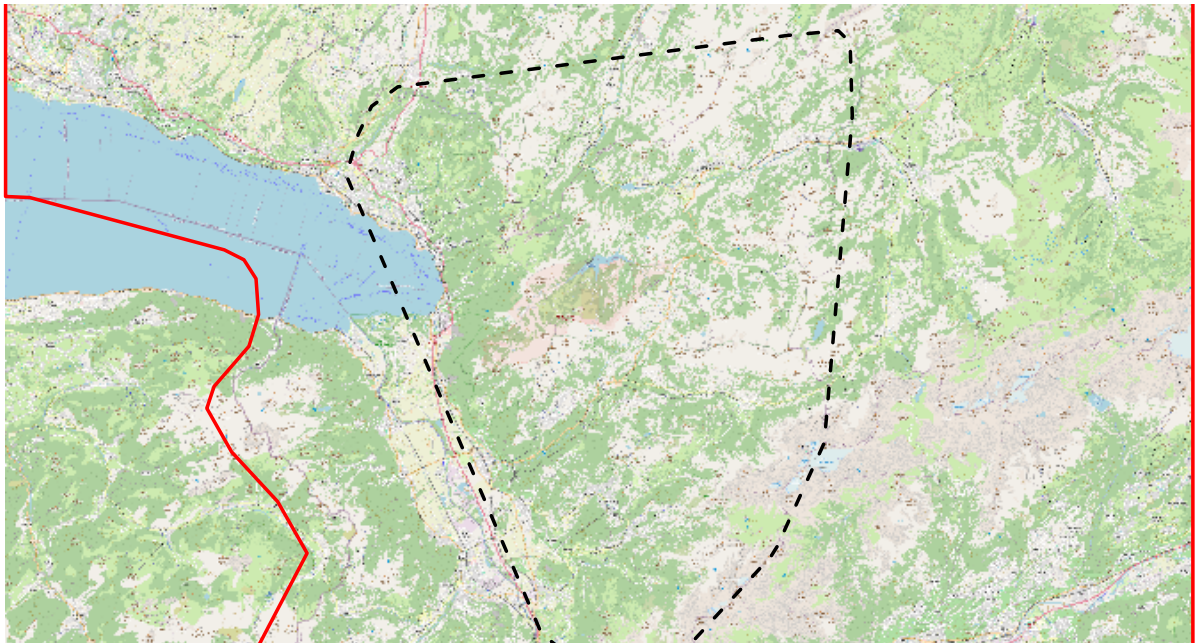


if $\eta_{ij}^\star$ is positive, we have 1 and 0 if it is negative

So, really, $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^\top + \mathbf{I}$

Example: alpine plants in Switzerland

- Data by D'amen et al. (2017)
- Occurrence of 175 species at 840 $4m^2$ plots
- Environmental variables: Degree days above zero, slope, moisture, solar radiation, topography (and coordinates)



Example: fit JSDMs

```
model5 <- gllvm(Y, num.lv = 2, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
model6 <- gllvm(Y, num.lv = 3, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
model7 <- gllvm(Y, num.lv = 4, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
```

Calculate predictive performance

```
goodnessOfFit(Y, object = model5, measure = "RMSE")$RMSE
```

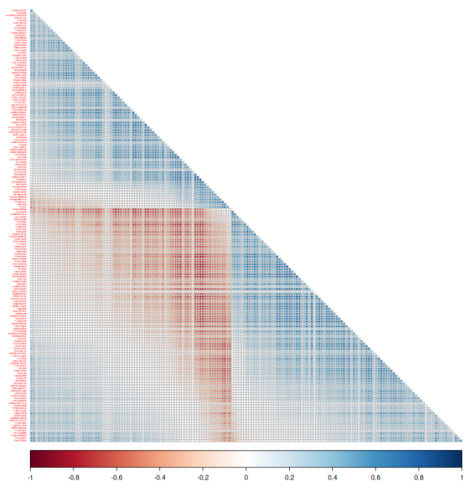
```
## [1] 0.2625214
```

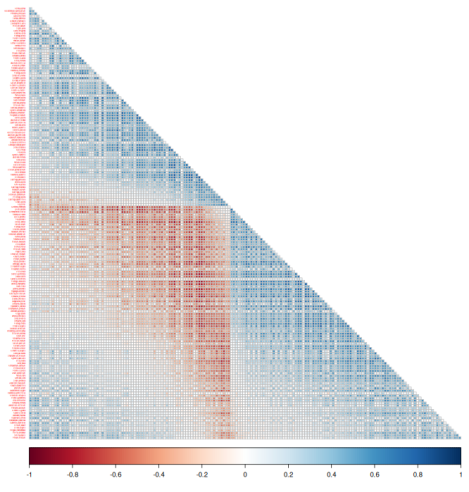
```
goodnessOfFit(Y, object = model6, measure = "RMSE")$RMSE
```

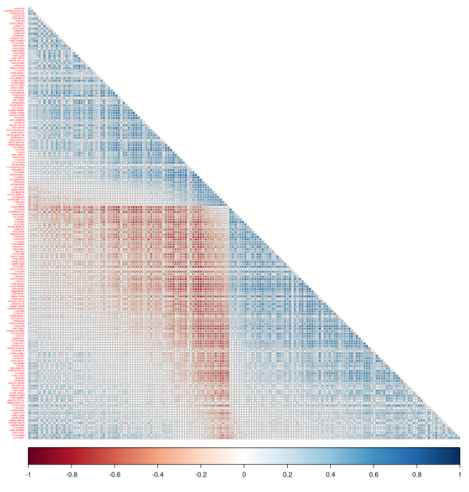
```
## [1] 0.2454212
```

```
goodnessOfFit(Y, object = model7, measure = "RMSE")$RMSE
```

Example: resulting associations





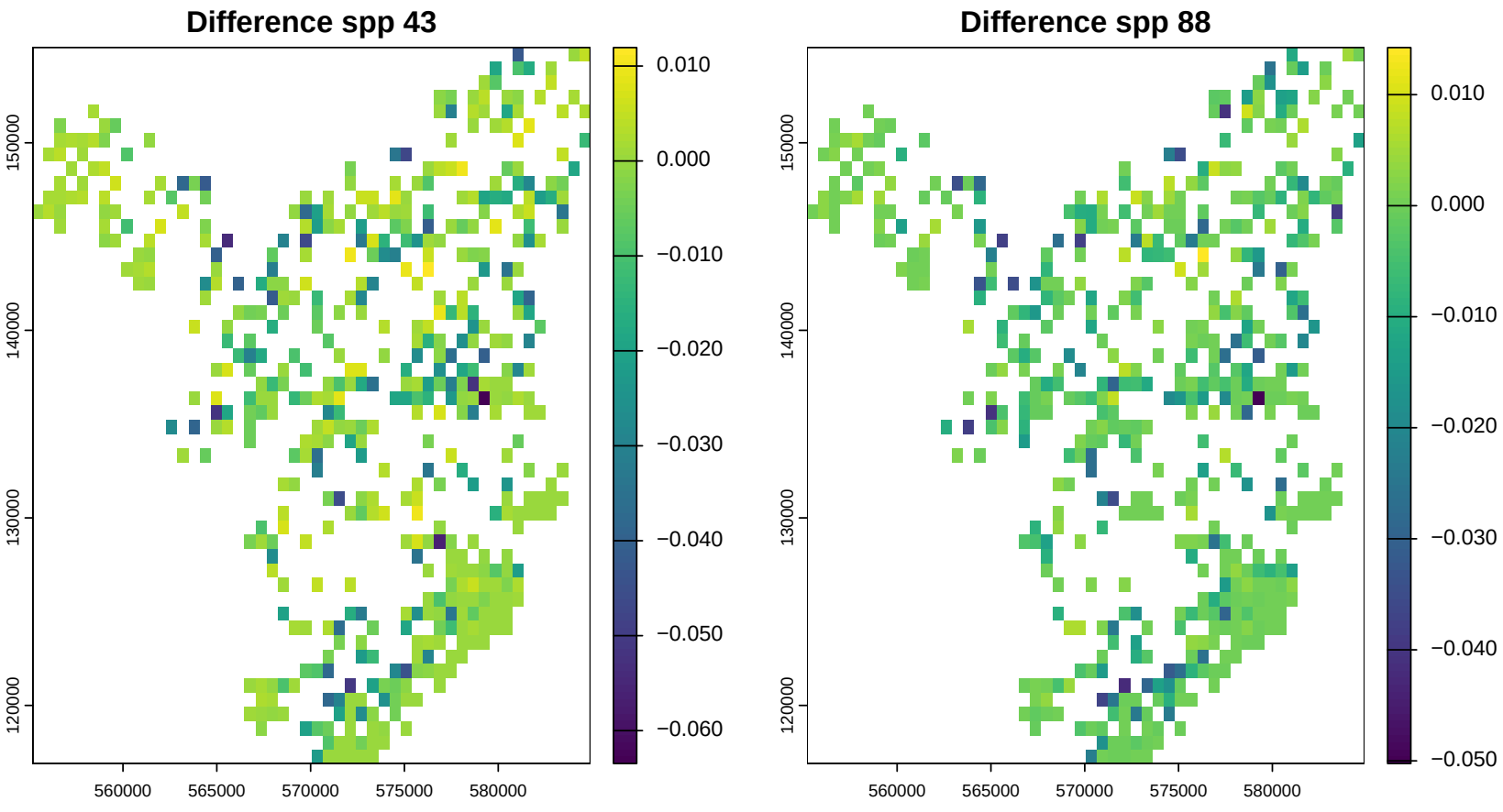


Example: conditional prediction

Let's take one of the models, and do a conditional prediction.

```
condPred <- pnorm(conditionalPredict(c(43,88),model7))
pred <- pnorm(predict(model7)[,c(43,88)])
colnames(pred) = NULL
pts = cbind(pts, condPred = condPred)
pts = cbind(pts, pred = pred)
pts2 <- terra::vect(pts)
```

Example: result of conditional prediction



Defining predictions

BRITISH ECOLOGICAL SOCIETY Methods in Ecology and Evolution

RESEARCH ARTICLE | Free Access

Defining and evaluating predictions of joint species distribution models

David P. Wilkinson, Nick Golding, Gurutzeta Guillera-Arroita, Reid Tingley, Michael A. McCarthy

First published: 28 October 2020 | https://doi.org/10.1111/2041-210X.13518 | Citations: 41

Wilkinson et al. defined multiple types of predictions:

- marginal (from `predict`)
- conditional marginal (aforementioned)
- joint: occurrence of multiple species simultaneously
- conditional joint: occurrence of two or more species together, given the information from another

Summary

- JSDMs were introduced: models to incorporate species' correlation
- JSDM mostly focuses on predicting (e.g., on a map)
- Usually, JSDMs are implemented using latent variables
  - The number used affects predictive performance
- Conditional predictions can facilitate a focus on focal species
- There are different types of predictions possible, to target particular ecological questions