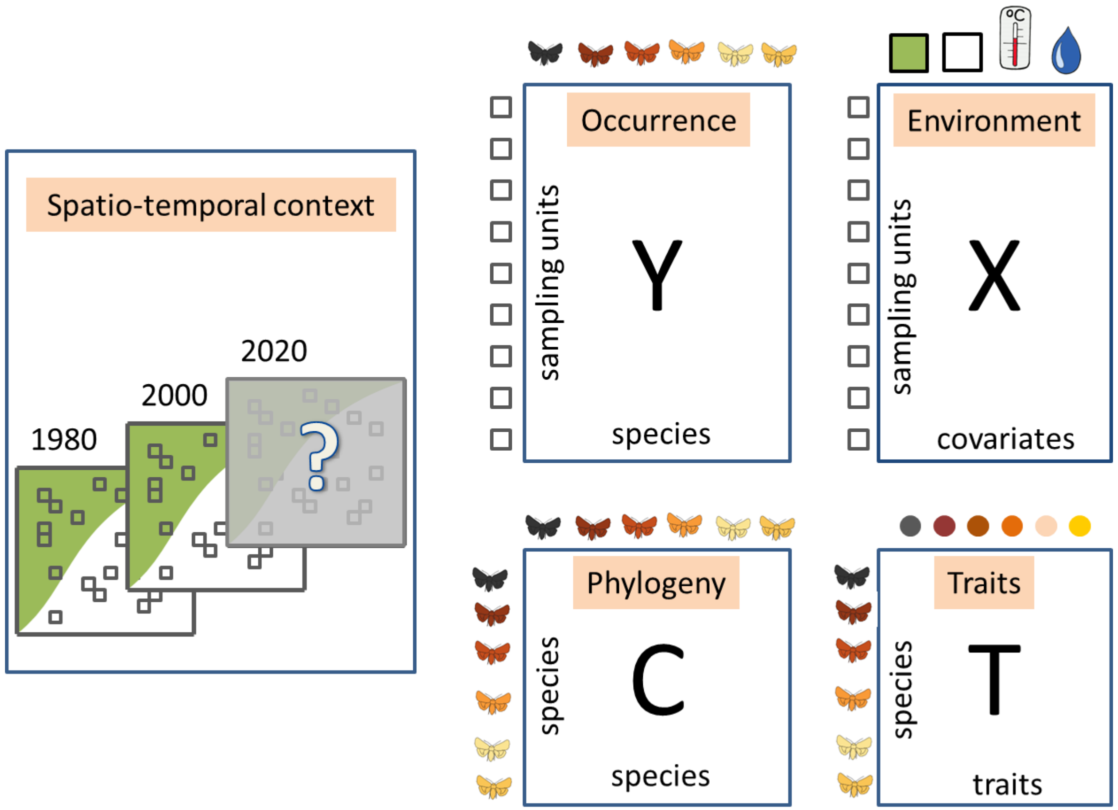
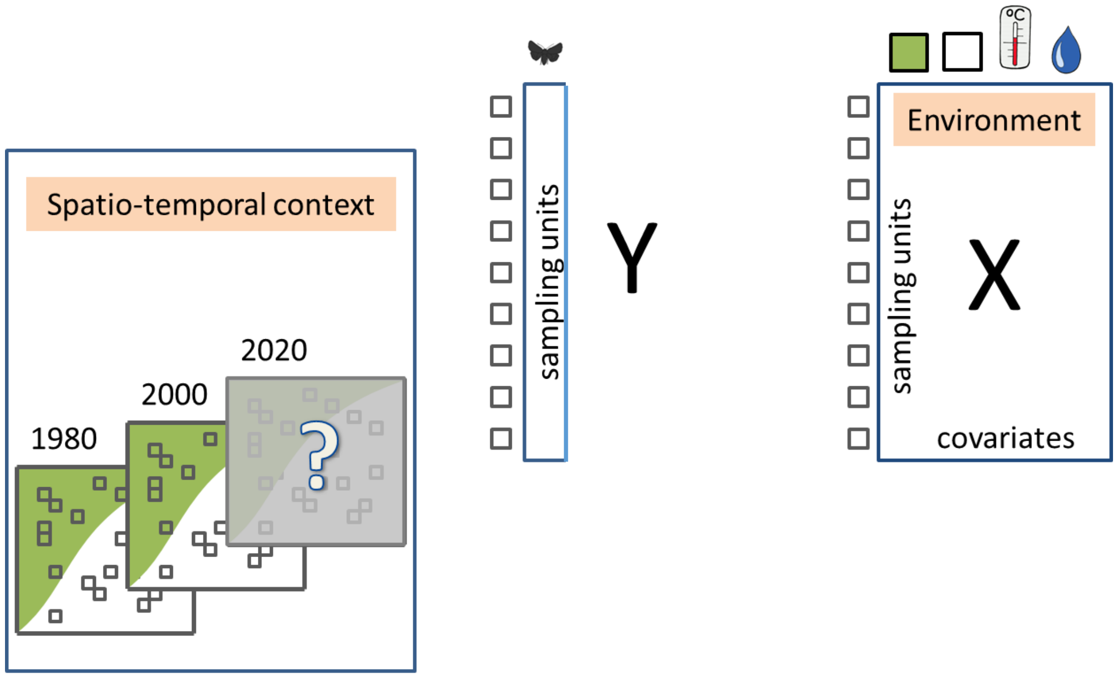
Building HMSC step by step: single-species distribution modelling

Full HMSC



Single-species HMSC



Back to basics: the linear model

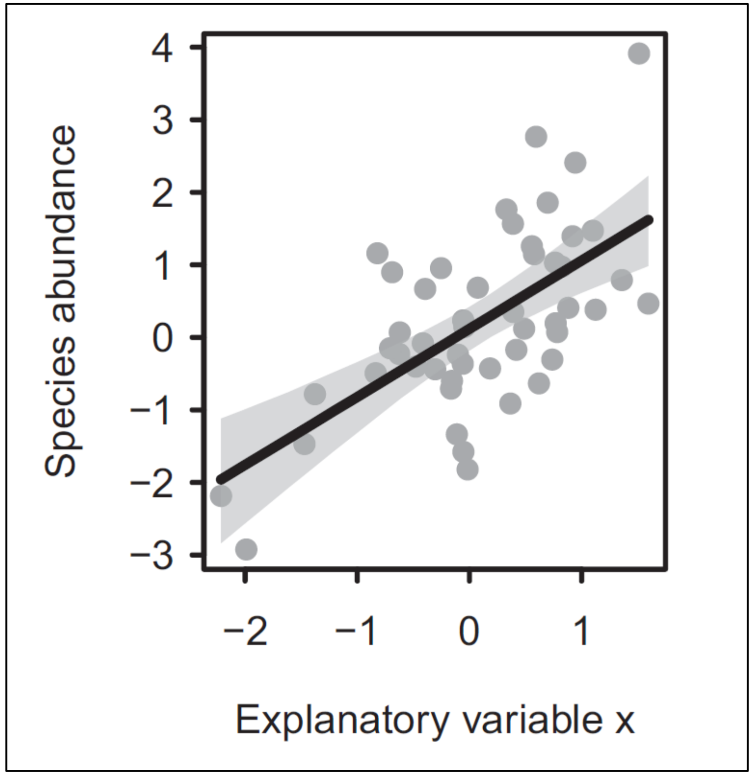
The linear model: $y_i = \alpha + \beta x_i + \varepsilon_i$

Index for data points: $i = 1,2,3,\dots,n$

Response (or dependent) variable: y

Explanatory (or independent or predictor) variable: x

Residual: $\varepsilon_i \sim N(0, \sigma^2)$



Intercept: α

Slope: β

Several explanatory variables and the linear predictor

The linear model with two variables: $y_i = \alpha + \beta_1 x_{i1} + \beta_1 x_{i2} + \varepsilon_i$

Can also be parameterized as: $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

where $x_{i1} = 1$ for all sampling units i

Can be written more compactly as $y_i = L_i + \varepsilon_i$ where

$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$ is the linear predictor and n_c is the number of covariates (including the intercept)

Continuous versus categorical predictors

In the basic linear model $y_i = \alpha + \beta x_i + \varepsilon_i$

x is a continuous explanatory variable (covariate)

Often x is a categorical explanatory variable (factor), e.g. habitat type classified as coniferous forest, broadleaved forest, or mixed forest.

This can be incorporated as:

$$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$$

$x_{i1} = 1$ for all sampling units

$x_{i2} = 1$ if i is in broadleaved forest, otherwise $x_{i2} = 0$

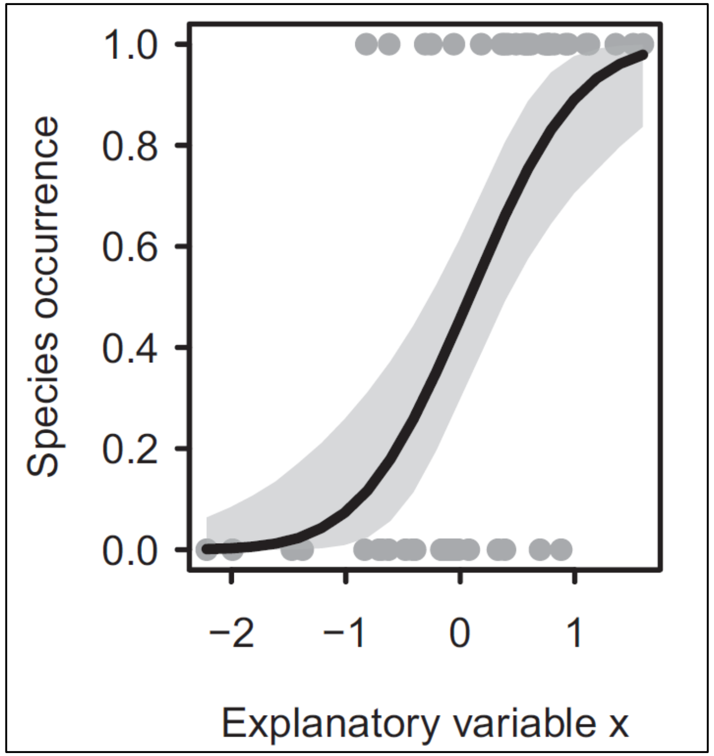
$x_{i3} = 1$ if i is in mixed forest, otherwise $x_{i3} = 0$

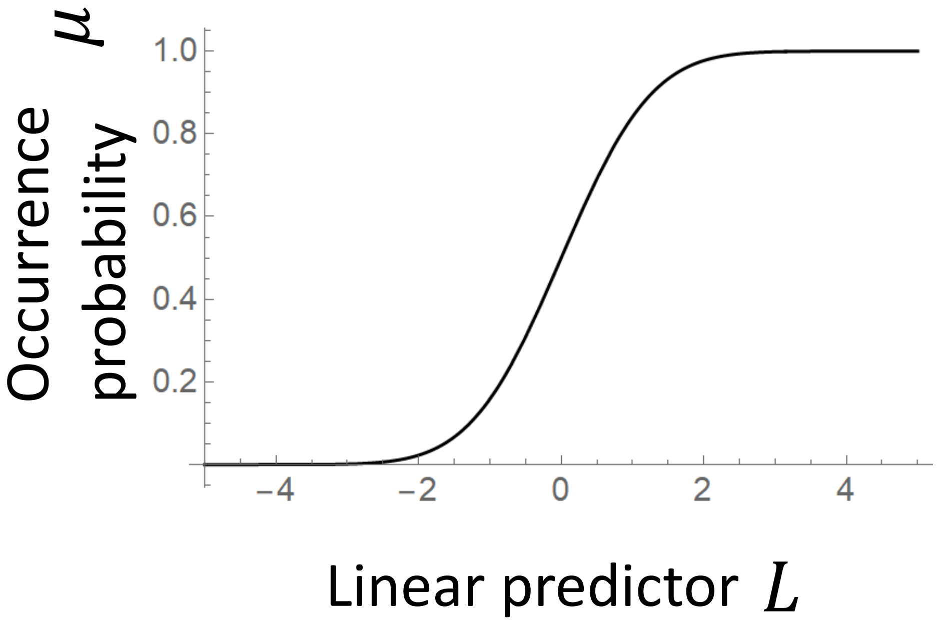
Generalized linear models

Presence-absence data: $y_i \sim \mathrm{Bernoulli}(\mu_i)$

Logistic regression: $\mu_i = \mathrm{logit}^{-1}(L_i)$

Probit regression: $\mu_i = \Phi(L_i)$





Generalized linear models

Count data: $y_i \sim \text{Poisson}(\mu_i)$

Poisson model

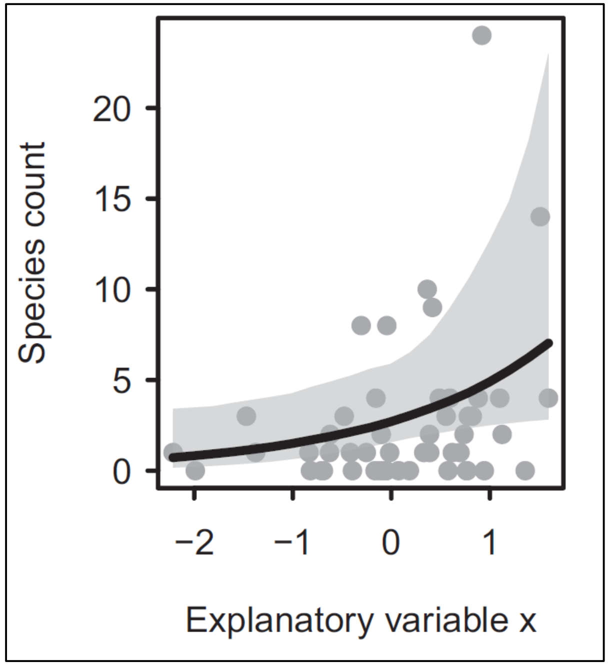
$y = 3,0,0,2,1$

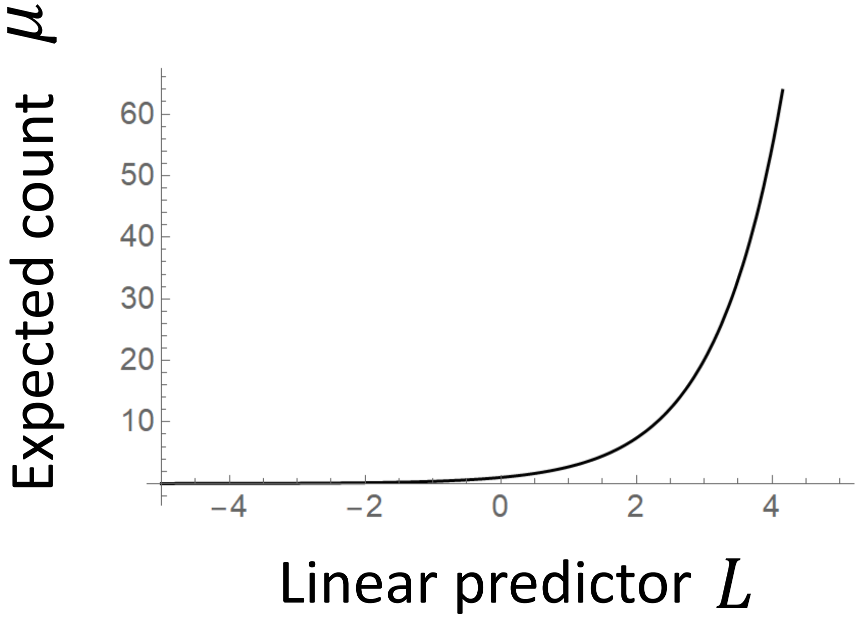
$$\mu_i = \exp(L_i)$$

Lognormal Poisson model

$y = 3,0,0,2,30,\ldots$

$$\mu_i = \exp(L_i + \varepsilon_i)$$





Hurdle models for zero inflated data

Assume that the data looks like $y = 0, 42, 39, 0, 43$

We can model separately presence-absence $y = 0, 1, 1, 0, 1$

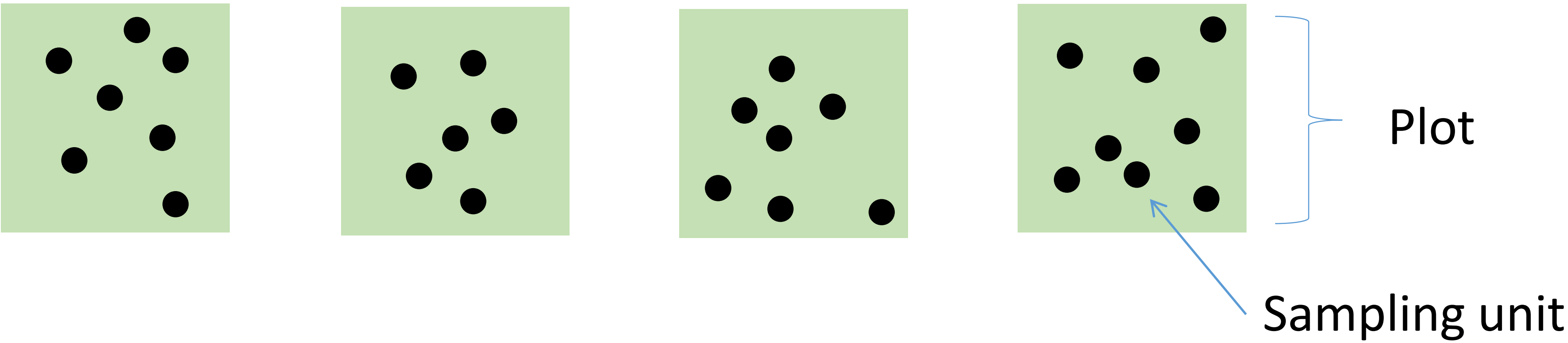
and abundance conditional on presence $y = \mathrm{NA}, 42, 39, \mathrm{NA}, 43$

These two parts of the hurdle-model can be fitted independently of each other, but they can be thought to jointly form one model.

The presence-absence part predicts if the species is present or not. If it is present, then the abundance (conditional on presence) part predicts how abundant the species is.

Mixed models: fixed effects and random effects

Hierarchical study design:



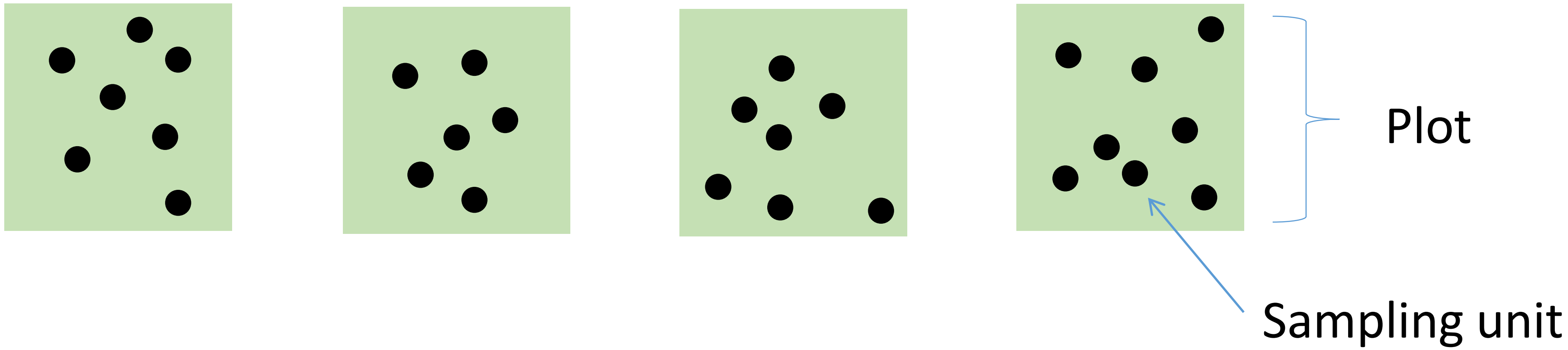
Linear model with fixed effects only: $y_i = L_i + \varepsilon_i$

$$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$$

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Mixed models: fixed effects and random effects

Hierarchical study design:



Linear model with fixed and random effects: $y_i = L_i + a_{p(i)} + \varepsilon_i$

$$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$$

$$a_p \overset{\text{iid}}{\sim} N(0, \sigma_P^2)$$

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Mixed models: fixed effects and random effects

Spatial study design:

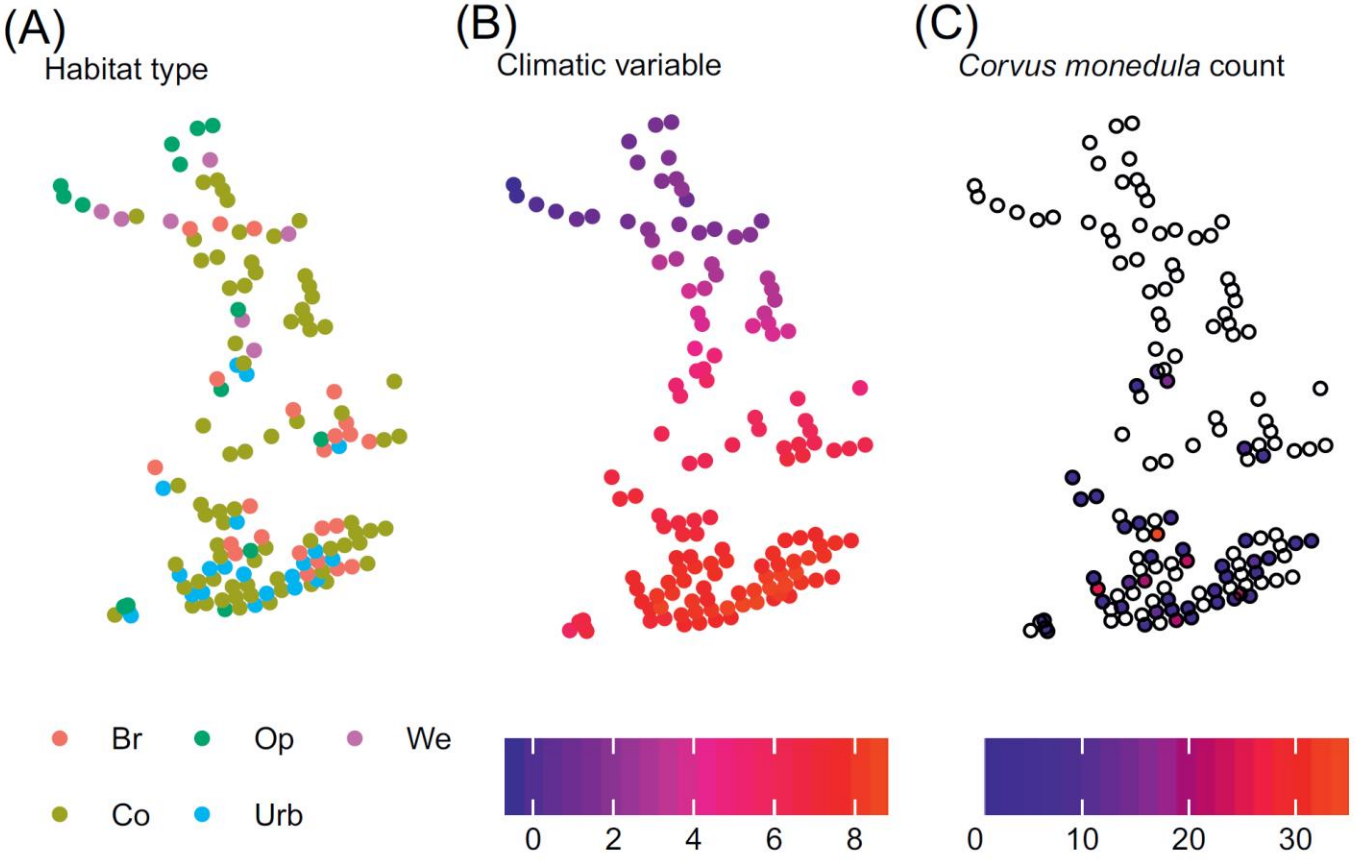
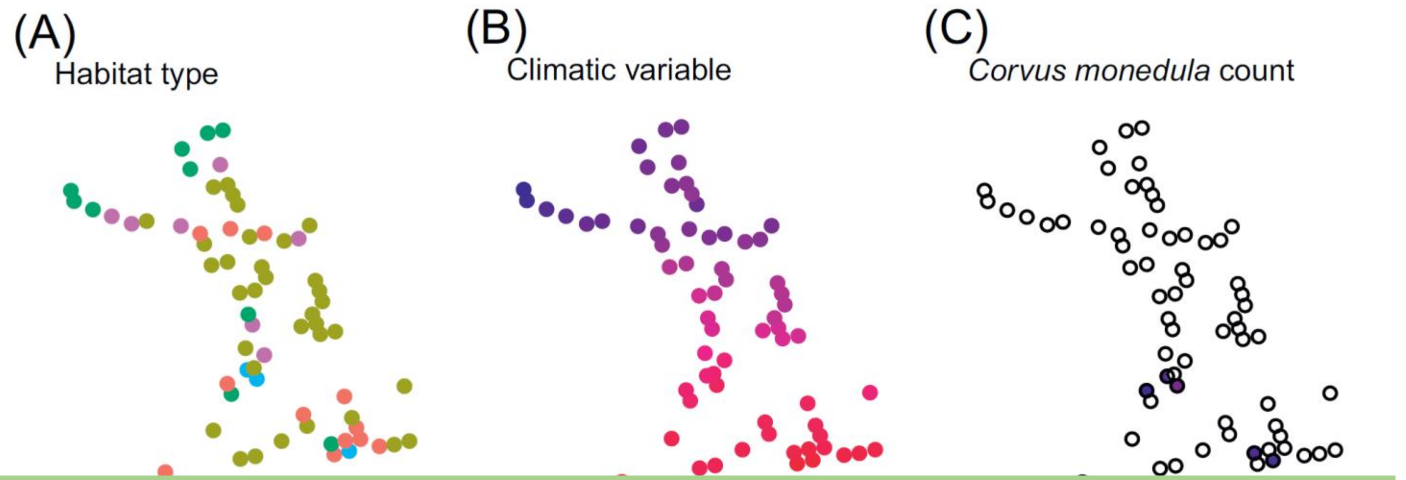


Figure 5.9 An illustration of environmental and species data used in this example. The panels show spatial variation in habitat type (A), climatic conditions (B), and the counts of the target species across Finland (C).

Mixed models: fixed effects and random effects

Spatial study design:



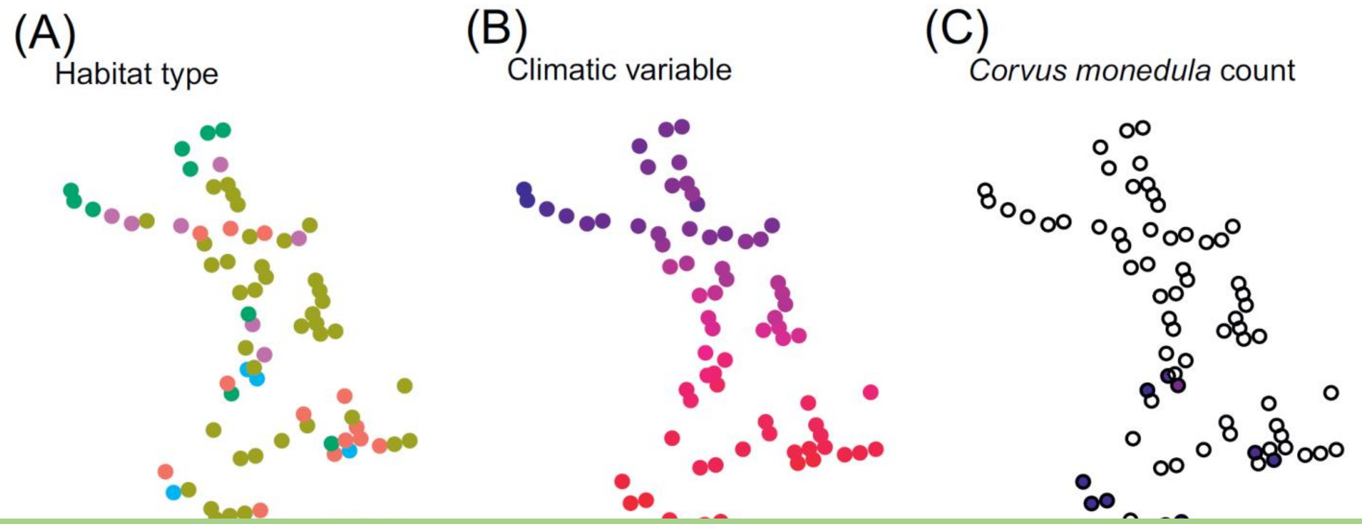
Linear model without spatial structure: $y_i = L_i + \varepsilon_i$

$$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$$

$$\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Mixed models: fixed effects and random effects

Spatial study design:



Linear model with spatial structure: $y_i = L_i + a_i + \varepsilon_i$

$$L_i = \sum_{k=1}^{n_c} \beta_k x_{ik}$$

$$a_i \sim N(0, \sigma_S^2)$$

$$\mathrm{Cov}(a_i, a_j) = \sigma_S^2 \exp(-d_{ij}/\alpha)$$

iid

$$\varepsilon_i \sim N(0, \sigma^2)$$

Fitting a spatial model enables using spatial information when generating predictions