Generalised Linear Models

for data of multiple species

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- Data collection and common data types
- Generalised Linear Models background
  - Assumption checking
- "Vector" models
- GLMMs
  - Estimating variation in species responses
  - Estimating correlation in species responses
- Model diagnostics and selection
- Building on some material from the GLM workshop

Statistical modeling

Instead of a focus on data, we consider the *data generating process*

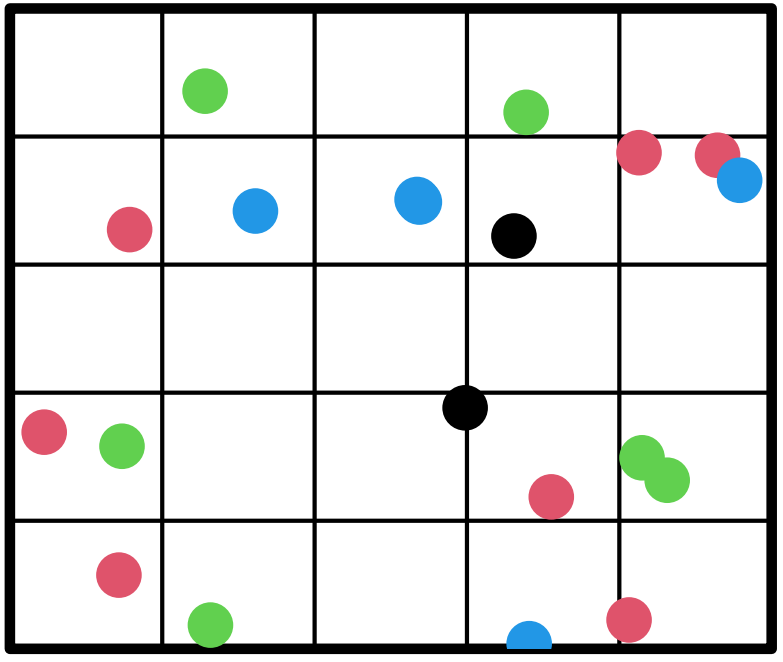
- We collect data
- Decide on a research question for *the population*
- Learn about the variation in the data
  - Which requires formulating a model
- Work out distribution of the estimates
  - And find the "best" estimate
- Conclude if our answer is robust for the population

Sampling data

Figure 1: dw.com

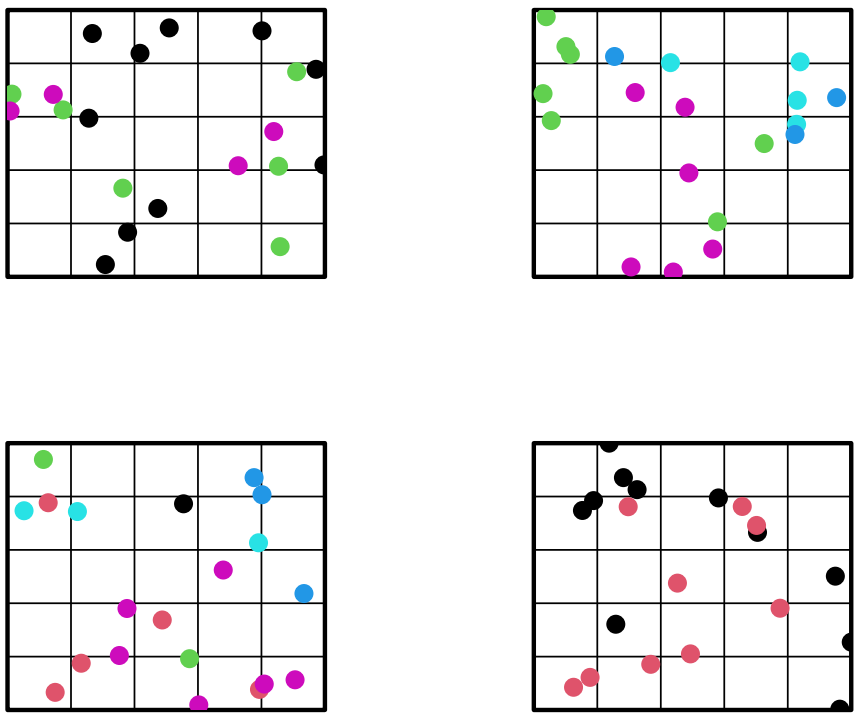
A meadow in the Dutch dunes.

How many plant species are there in this meadow?



We go into the meadow and count plants in quadrats. We find 4 species.

Resampling the dunes

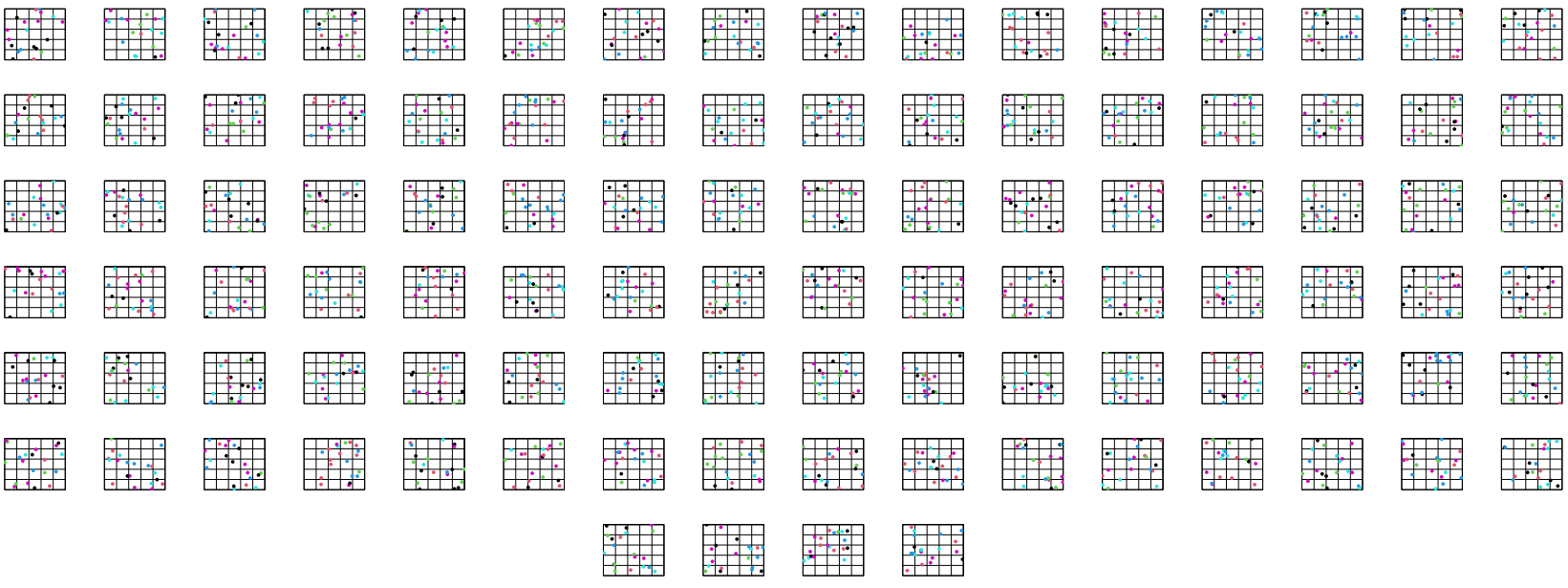


We resample the dune meadow, and find different numbers of species: 3, 3, 6, 2. And different compositions. On average we have found 3.5 species per quadrat.

Sampling variation

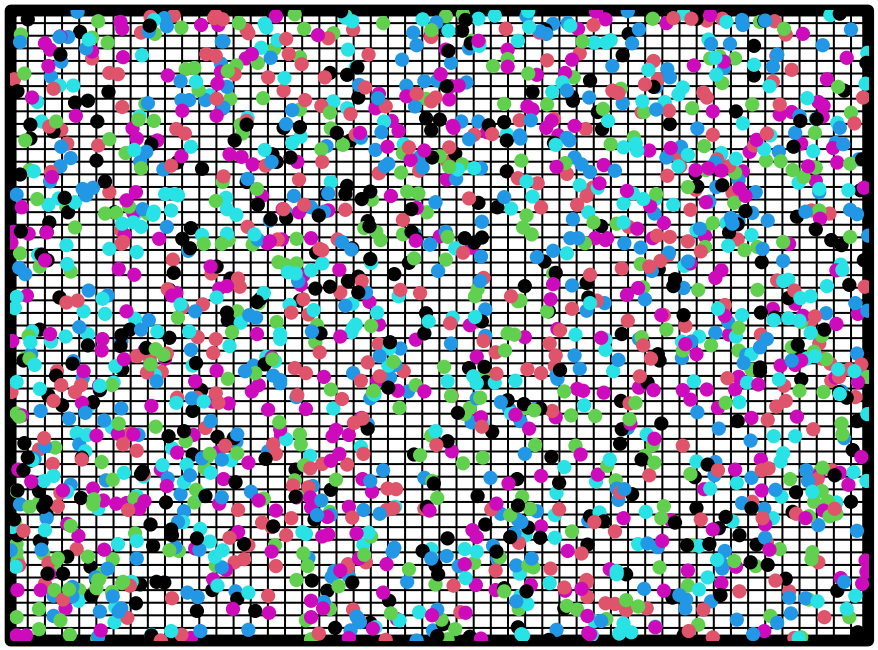
1) Each time we sample, we have a slightly different sample
2) Each time we estimate a parameter, it might be slightly different due to this sampling variation
3) The more data we sample, the better we can represent this variability of our estimate
4) And the better we can represent the true richness/cover in the meadow

Sampling variation



If we sample many times, we have an accurate picture of the whole meadow (and variability in the number of species in a quadrat that we might find).

Sampling variation



Statistical modeling

Instead of a focus on data, we consider the *data generating process*

- We collect data
- Decide on a research question for *the population*
- Learn about the variation in the data
  - Which requires formulating a model
- Work out distribution of the estimates
  - And find the "best" estimate
- Conclude if our answer is robust for the population (e.g., fields like this have more than 6 species)

The ecological process

What do we know of the processes that generate these data?

- Meta-community theory
- Assembly processes (filtering)
- Ecological gradient theory

The ecological process

What do we know of the processes that generate these data?

- Meta-community theory
- Assembly processes (filtering)
- Ecological gradient theory

Multispecies models provide a statistical connection to these ecological frameworks. We do not just use a fancy tool, we use a fancy tool because we believe it aligns well with our understanding of the ecological process.

The ecological process (2)

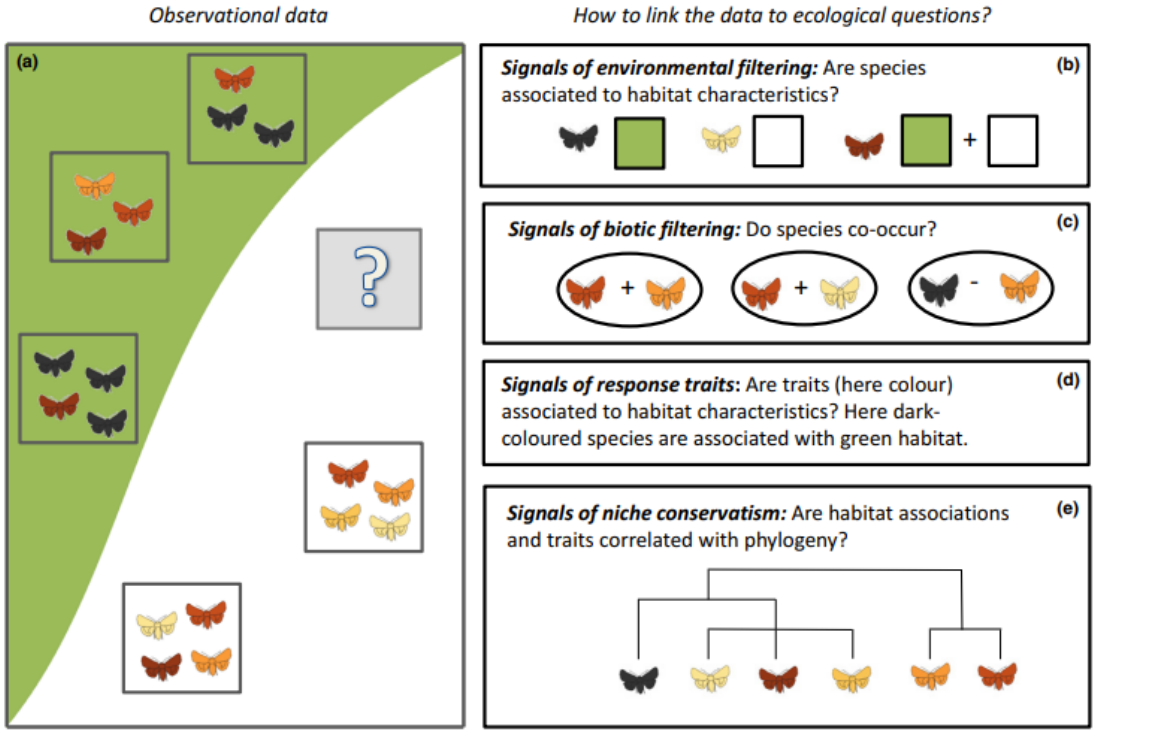


Figure 2 A conceptual illustration of some key questions in community ecology. The green and white colours represent differences in the environmental

Figure 2: Figure 2 from Ovaskainen et al. (2017)

On ecological communities

The concept of an ecological community is of limited use. By definition:

An ecological community is a group or association of two or more species occupying the same geographical area at the same time

- We often think of ecological communities as groups
- We can also think of a community as a continuum that changes along a gradient (Austun 1985)
- We can also think of them as the species in our data

On ecological communities

The concept of an ecological community is of limited use. By definition:

An ecological community is a group or association of two or more species occupying the same geographical area at the same time

- We often think of ecological communities as groups
- We can also think of a community as a continuum that changes along a gradient (Austun 1985)
- We can also think of them as the species in our data

Connecting model outputs to ecological concepts requires some deep thoughts

An ecological gradient

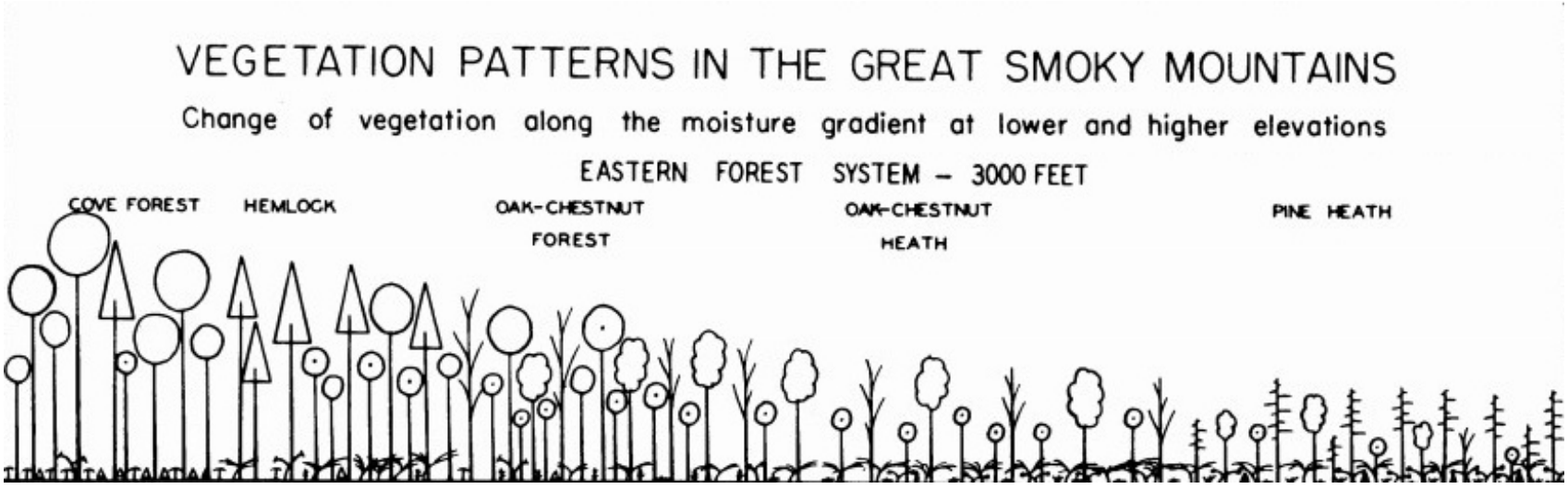


Figure 3: Whittaker (1956)

Response curves

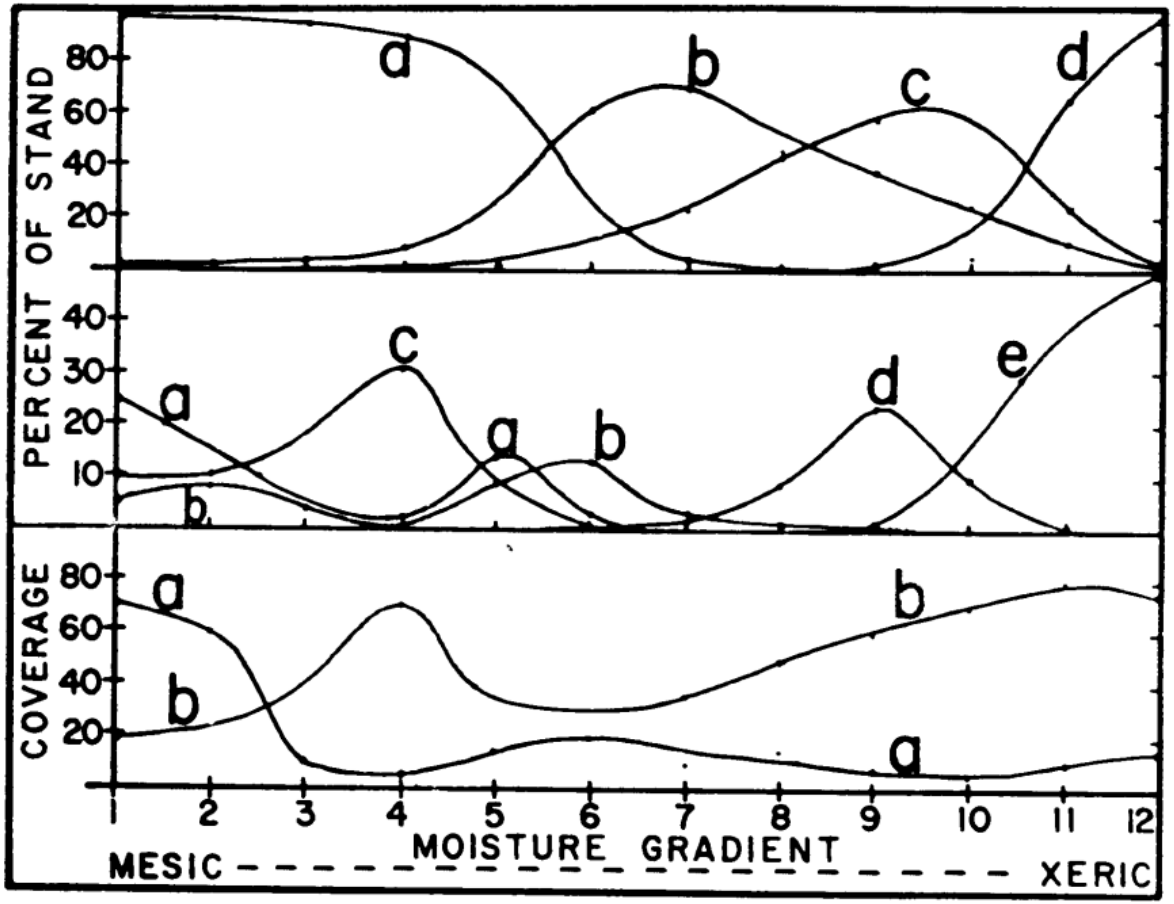


FIG. 4. Transect of the moisture gradient, 3500-4500 ft. Top—curves for tree classes; a, mesic; b, submesic; c, subxeric; d, xeric. Note expansion of mesic stands, compared with Figs. 2 and 3. Middle—curves for tree species: a, *Tilia heterophylla*; b, *Halesia monticola* (both

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Or ecologically:

- Species distribution models
- Joint Species Distribution Models
- Ordination

Multispecies models

There are multiple statistical frameworks for studying the processes:

- Generalised Linear Models
- Generalised Linear Mixed-effects Models
- Generalised Additive Models (not covered here)
- Generalised Linear Latent Variable Models

Or ecologically:

- Species distribution models
- Joint Species Distribution Models
- Ordination

and more. Each method has its limitations (assumptions). It is up to us to assess which are appropriate.

Generalised linear models (GLMs)

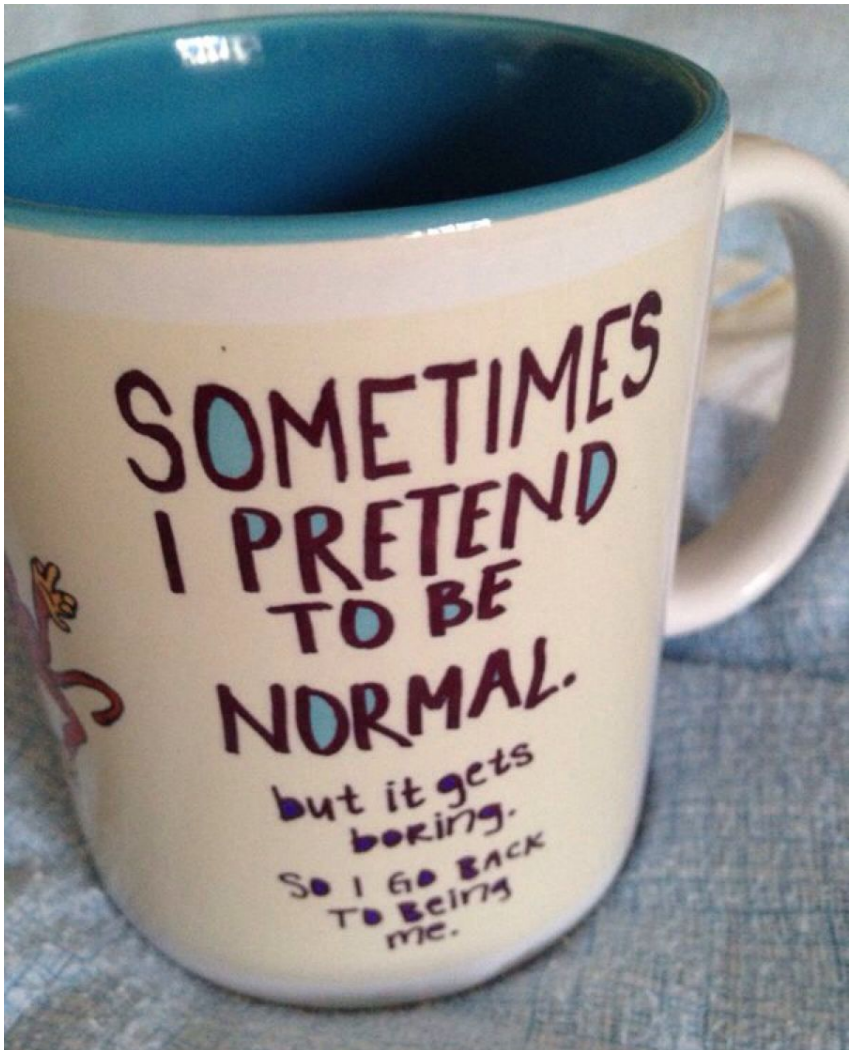
GLMs as a framework were introduced by Nelder and Wedderburn (1972) uniting many different models. With a special focus on teaching statistics.

- Linear regression
- Logistic regression
- Probit regression
- Complementary log-log regression
- Log-linear regression
- Gamma regression

Generalised Linear Models

For when the assumptions of linear regression fail.

- Linearity (straight line)
- Independence of errors
- Homoscedasticity (same variance for all errors)
- Normality (distribution of errors)



Generalised linear models (2)

GLMs extend the linear model framework to address:

- Variance changes with the mean
- Range of **y** is bounded

The basis of many statistical models in Biology

Components of a GLM

- Systematic component: η
- Random component: data/distribution)
- The link function: connects these components
  - This is not a data transformation
- The variance function

But no explicit error term

GLM Likelihood

- We use MLE for estimation
- With a distribution in the "exponential family" (for fixed ϕ)

All GLMs have the likelihood:

$$\mathcal{L}(y_i; \Theta) = \exp\{\frac{y_i \eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\} \quad (1)$$

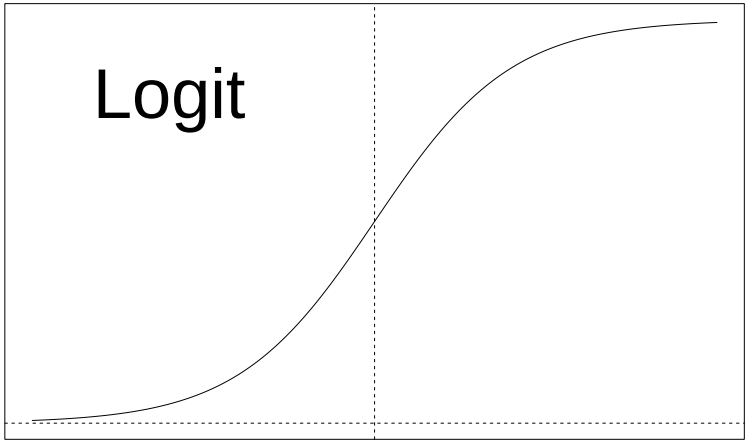
Generalised linear model

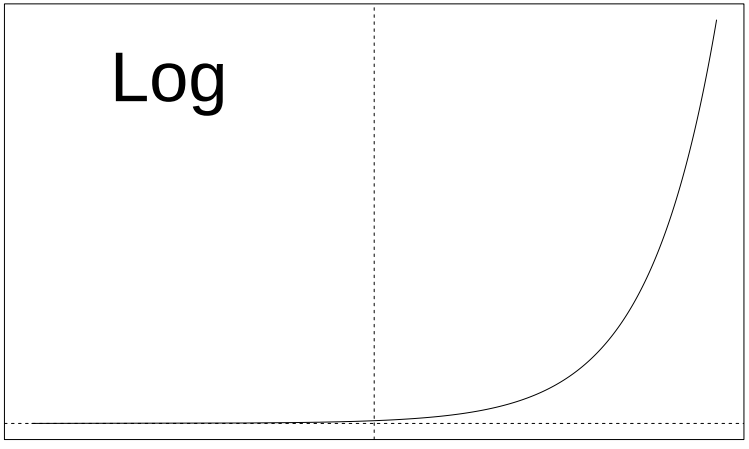
$$
\begin{aligned}
g\{\mathbb{E}(y_i|x_i)\} &= \eta_i = \alpha + x_i\beta \\
\mathbb{E}(y_i|x_i) &= g^{-1}(\eta_i) = g^{-1}(\alpha + x_i\beta)
\end{aligned}
\tag{2}
$$

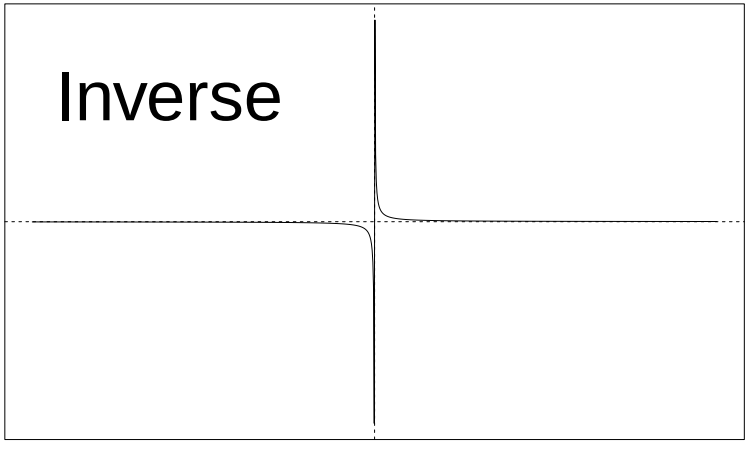
g$(\cdot)$ is the **link function**

The link function

- Is a smooth/monotone function
- Has an inverse $g^{-1}(\cdot)$
- Restricts the scale
- g$(\cdot)$ can be e.g.







Variance function

Perhaps most critically, variance changes with the mean:

$$\text{var}(y_i; \mu_i, \phi) = \frac{\partial^2 g(\eta_i)}{\partial \eta_i^2} a(\phi)$$

- ϕ: the dispersion parameter, constant over observations
  - Fixed for some response distributions
- $a(\phi)$ is a function of the form ϕ / w_i (McCullagh and Nelder 1989)

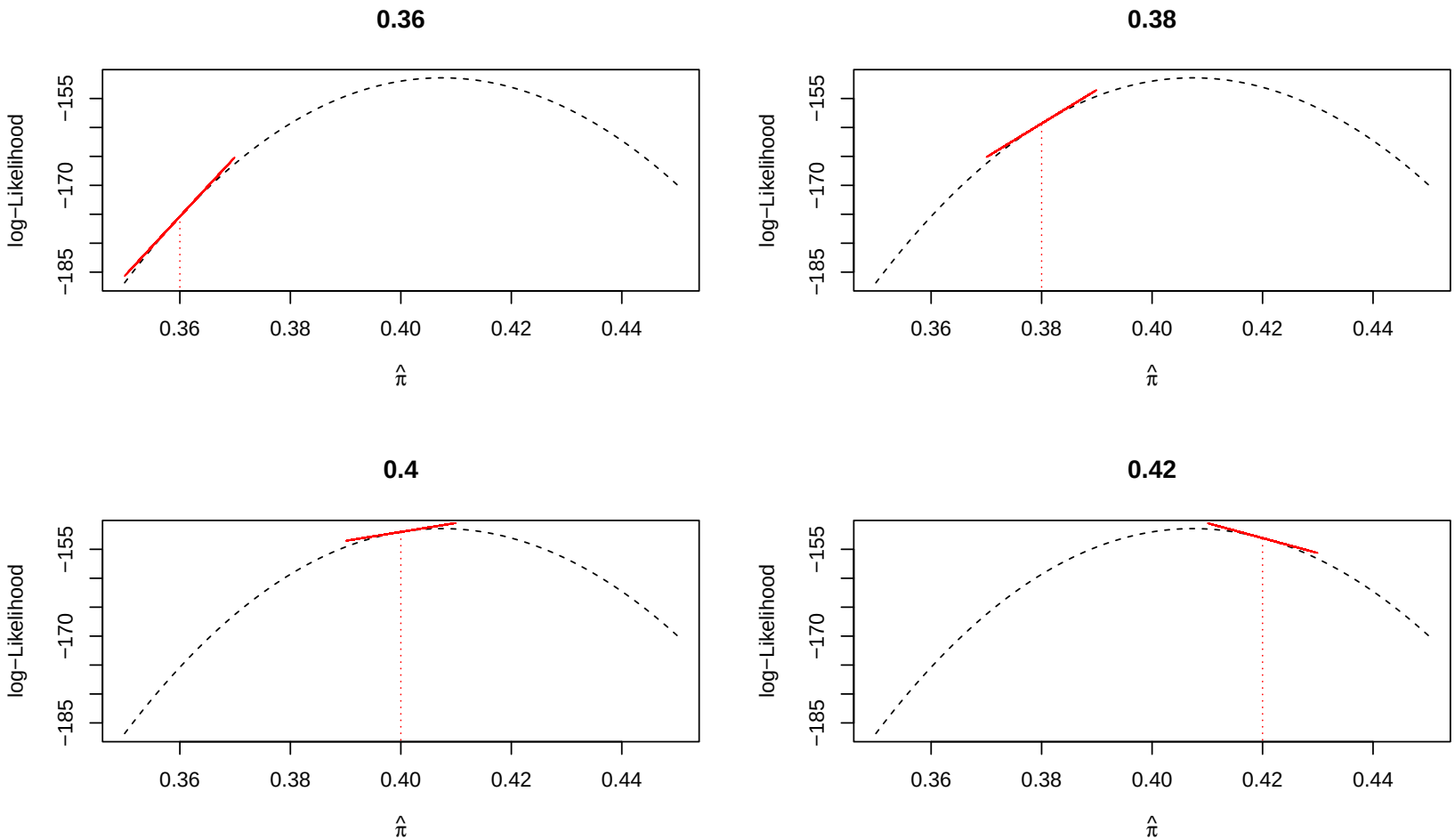
Fitting GLMs

Parameters in GLMs need to be estimated **iteratively**.

- More difficult to fit
- Requires numerical optimisation
- Susceptible to local convergence

Holds for GLLVMs too

Estimating GLMs



We need a good algorithm to find the maximum!

Why is this important?

1) A basic (mathematical) understanding helps apply methods correctly.
2) GLMMs/GLLVMs may not always converge to the best solution immediately.
3) This can help to diagnose your model.

Often used distributions in ecology

- Binomial: occurrence/counts. Presence of species, number of germinated seeds out of a total
- Poisson: counts. Abundance
- Negative binomial (fixed dispersion): counts. Number of species or abundance
- Gamma: (positive) continuous. Body size or biomass
- Ordinal (cumulative link). Cover classes
- Beta (logit link). Cover (note: not a GLM)

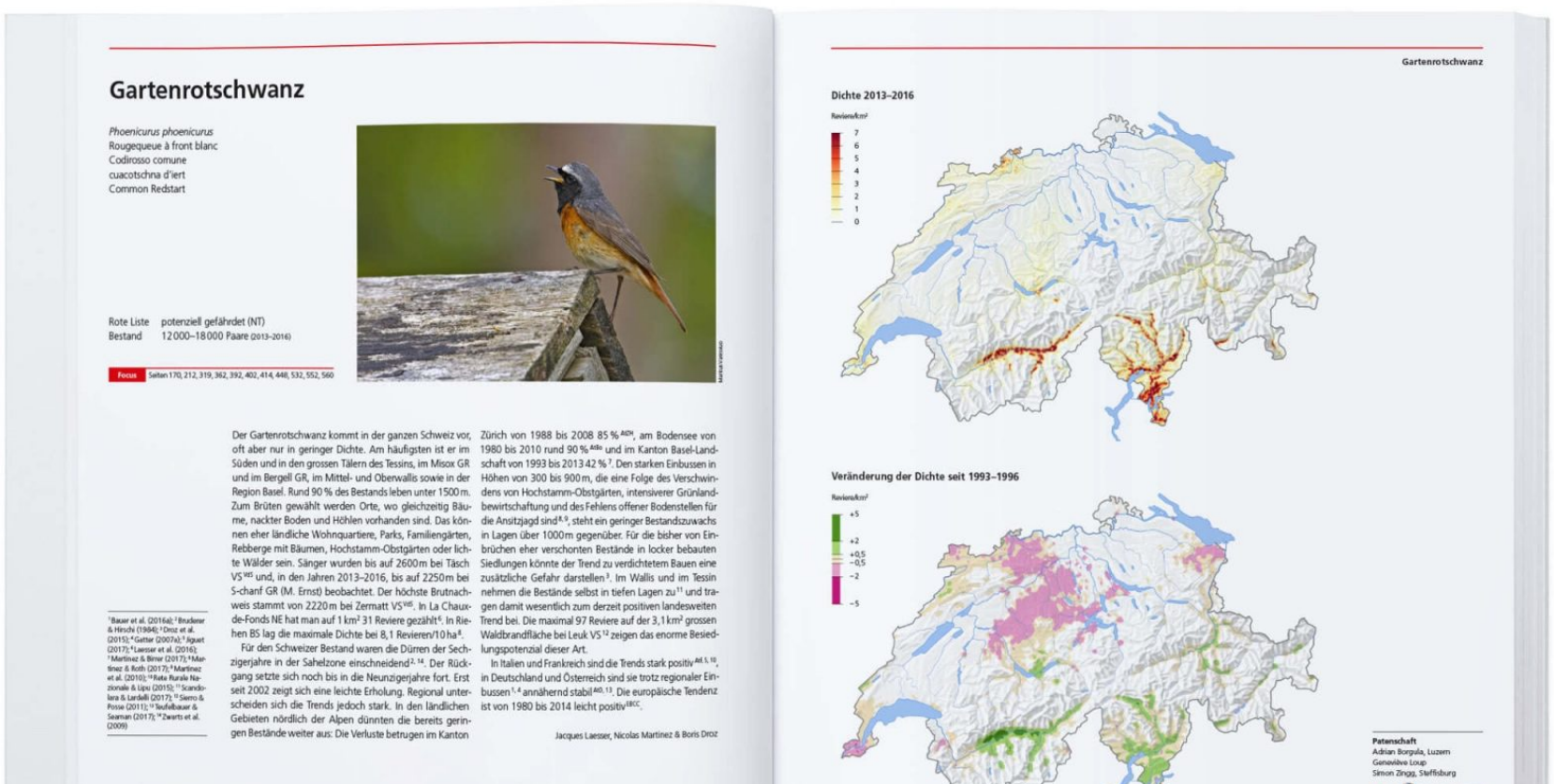
Example: Swiss bird occurrence

Observation process: see if a bird is present (or we might hear it)
Alternatively: The proportion of a species in a place
Alternatively: Count of birds in the forest

There are often many ways to observe the same ecological process.
We need to **disentangle** this from the ecological process.

Example: Swiss birds

- Data by Schmid et al. (1998): the Swiss breeding bird atlas
- Occurrence of 56 species at 2524 locations recorded over a 4-year period



Gartenrotschwanz

Phoenicurus phoenicurus
Rougequeue à front blanc
Codirosso comune
cuacotschna d'iert
Common Redstart

Rote Liste potenziell gefährdet (NT)
Bestand 12000–18000 Paare (2013–2016)

Focus Seiten 170, 212, 319, 362, 392, 402, 414, 448, 532, 552, 560

Der Gartenrotschwanz kommt in der ganzen Schweiz vor, oft aber nur in geringer Dichte. Am häufigsten ist er im Süden und in den grossen Tälern des Tessins, im Misox GR und im Bergell GR, im Mittel- und Oberwallis sowie in der Region Basel. Rund 90 % des Bestands leben unter 1500 m. Zum Brüten gewählt werden Orte, wo gleichzeitig Bäume, nackter Boden und Höhlen vorhanden sind. Das können eher ländliche Wohnquartiere, Parks, Familiengärten, Rebberge mit Bäumen, Hochstamm-Obstgärten oder lichte Wälder sein. Sänger wurden bis auf 2600 m bei Täsch VS[WS] und, in den Jahren 2013–2016, bis auf 2250 m bei S-chanf GR (M. Ernst) beobachtet. Der höchste Brutnachweis stammt von 2220 m bei Zermatt VS[WS]. In La Chaux-de-Fonds NE hat man auf 1 km² 31 Reviere gezählt[6]. In Riehen BS lag die maximale Dichte bei 8,1 Revieren/10 ha[8].

Für den Schweizer Bestand waren die Dürren der Sechzigerjahre in der Sahelzone einschneidend[2, 14]. Der Rückgang setzte sich noch bis in die Neunzigerjahre fort. Erst seit 2002 zeigt sich eine leichte Erholung. Regional unterscheiden sich die Trends jedoch stark. In den ländlichen Gebieten nördlich der Alpen dünnten die bereits geringen Bestände weiter aus: Die Verluste betrugen im Kanton Zürich von 1988 bis 2008 85 %[AEH], am Bodensee von 1980 bis 2010 rund 90 %[Atlas] und im Kanton Basel-Landschaft von 1993 bis 2013 42 %[7]. Den starken Einbussen in Höhen von 300 bis 900 m, die eine Folge des Verschwindens von Hochstamm-Obstgärten, intensiverer Grünlandbewirtschaftung und des Fehlens offener Bodenstellen für die Ansitzjagd sind[8, 9], steht ein geringer Bestandszuwachs in Lagen über 1000 m gegenüber. Für die bisher von Einbrüchen eher verschonten Bestände in locker bebauten Siedlungen könnte der Trend zu verdichtetem Bauen eine zusätzliche Gefahr darstellen[9]. Im Wallis und im Tessin nehmen die Bestände selbst in tiefen Lagen zu[11] und tragen damit wesentlich zum derzeit positiven landesweiten Trend bei. Die maximal 97 Reviere auf der 3,1 km² grossen Waldbrandfläche bei Leuk VS[12] zeigen das enorme Besiedlungspotenzial dieser Art.

In Italien und Frankreich sind die Trends stark positiv[Atl. 5, 10], in Deutschland und Österreich sind sie trotz regionaler Einbussen[1, 4] annähernd stabil[AO, 13]. Die europäische Tendenz ist von 1980 bis 2014 leicht positiv[EBCC].

Jacques Laesser, Nicolas Martinez & Boris Droz

[1] Bauer et al. (2016a); [2] Bruderer & Hirschi (1984); [3] Droz et al. (2015); [4] Gatter (2007a); [5] Iguet (2017); [6] Laesser et al. (2016); [7] Martinez & Binz (2017); [8] Martinez & Roth (2017); [9] Martinez et al. (2010); [10] Rete Rurale Nazionale & Lipu (2015); [11] Scandolara & Lardelli (2017); [12] Sierro & Posse (2011); [13] Teufelbauer & Seaman (2017); [14] Zwarts et al. (2009)

Gartenrotschwanz

Dichte 2013–2016

Revier/km²
7
6
5
4
3
2
1
0

Veränderung der Dichte seit 1993–1996

Revier/km²
+5
+2
+0,5
−0,5
−2
−5

Patenschaft
Adrian Borgula, Luzern
Geneviève Loup
Simon Zingg, Steffisburg

The data

Falco_subbuteo	Anthus_trivialis	Phylloscopus_bonelli	Tetrao_tetrix	Parus_caeruleus	Dendrocopos_major
0	1	0	1	0	1
1	0	0	0	1	1
0	0	0	0	1	1
0	1	0	0	1	1
0	0	1	1	0	0
1	0	0	0	1	1

The environmental variables

avg	cov	cv	dns	fhd	p10	p25	p95	rt_p2595	std	uhd
13.549999	34.1	0.6287823	34.1	1.8422778	2.68	5.950000	28.14	0.2114428	8.52	2.381401
13.790000	32.0	0.5547498	32.0	1.8031981	3.96	7.090000	28.20	0.2514184	7.65	2.276368
19.340000	14.3	0.5351603	14.3	2.0021141	4.79	9.139999	34.03	0.2685865	10.35	2.364149
15.460000	34.5	0.4618370	34.5	1.7484871	5.19	9.790000	26.36	0.3713961	7.14	2.369797
2.290000	2.0	0.7292576	2.0	0.3038808	1.06	1.200000	5.74	0.2090592	1.67	1.270930
8.929999	61.3	0.5890258	61.3	1.4213525	2.21	4.570000	18.05	0.2531856	5.26	2.392563

- Bioclimatic variables (bioclim)
- Topography (slope, aspect, TPI, TWI) from a DEM
- Potential evapotranspiration (PET) from solar radiation
- Moisture index, degree days above zero
- Vegetation structure from LiDAR

The binomial GLM

$$p(y_{ij} = 1) = p_{ij} = g^{-1}(\eta_{ij}) \tag{3}$$

The binomial GLM

Link functions:

- Logit: $\log(\frac{\pi_i}{1-\pi_i})$ and inverse $\frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ - *the canonical link*
- Probit: $\Phi^{-1}(\pi_i)$ and inverse $\Phi(\eta_i)$
- Complementary log-log: $\log(-\log(1-\pi_i)$ and inverse $1-\exp(-\exp(\eta_i))$
- ~~Log-log~~
- Logit is canonical and easier to interpret
- Probit is sometimes easier mathematically than Logit
- Complementary log-log for counts

Data format

There are two ways to format these data:
Wide format: Species as columns (as presented)
Long format: Species is one column, and "Site" is another column

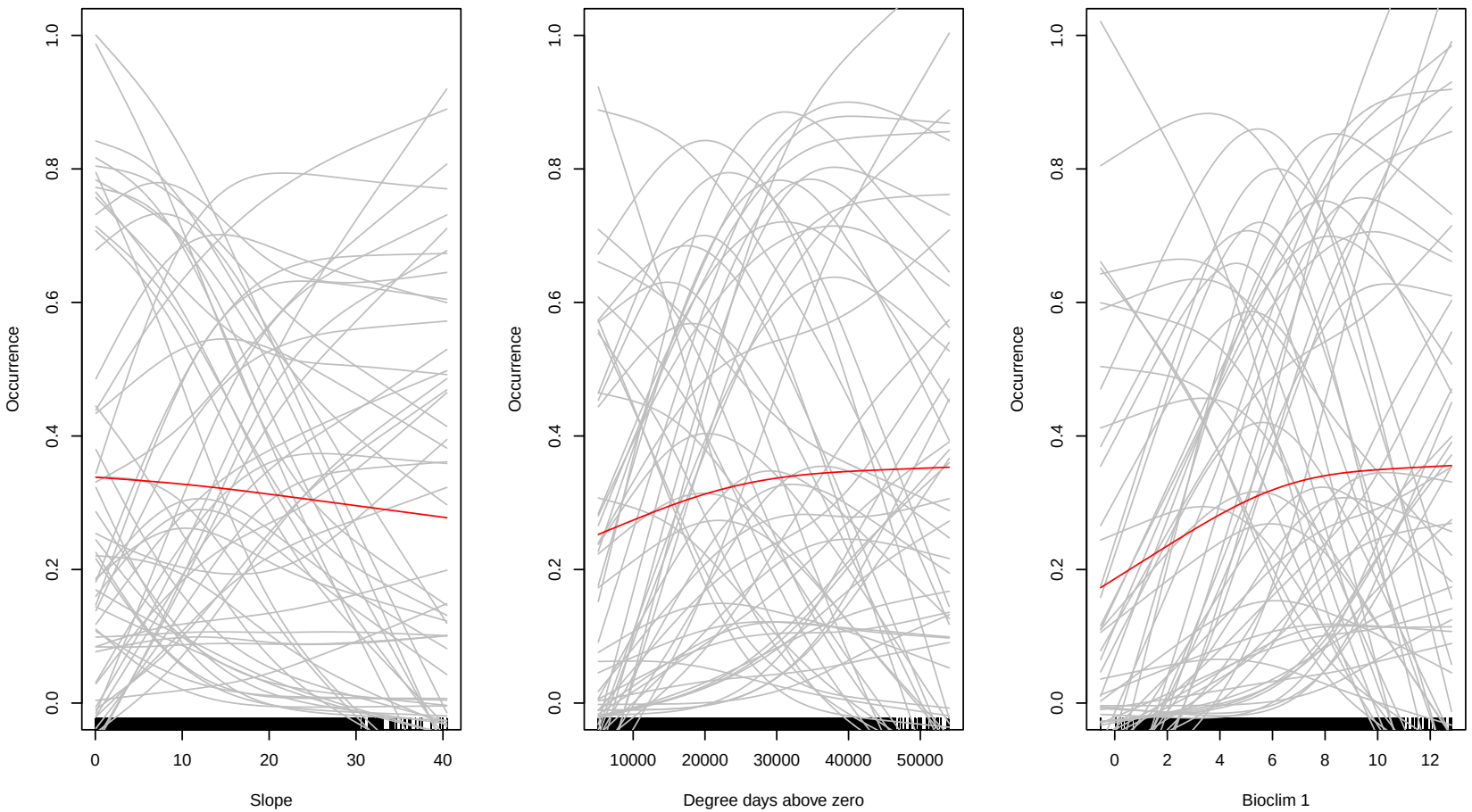
The format of the data does not affect the model. Some functions accept long format, other wide format, but the formulation of the model is up to us.

Swiss birds: to long format

```
data <- data.frame(y, X)
datalong <- reshape(data,
                    varying = colnames(y),
                    v.names = "occ",
                    idvar = "Site",
                    timevar = "Species",
                    direction = "long")

datalong$Species <- factor(datalong$Species,
                           labels = colnames(y))
```

Swiss birds: visually inspect the data



Is there a community-level trend?

Swiss birds: fit a model

```
model1 <- glm(occ~slp,
              data = datalong, family="binomial")
coef(model1)
```

```
## (Intercept)         slp
##  -0.6589784  -0.0065449
```

Here we assume that the intercept and slp effect are the same for all species

Multispecies modeling

1) Is the same effect for all species realistic?
2) Is the same (average) probabiliity of occurrence for all species realistic?

Multispecies modeling

1) Is the same effect for all species realistic?
2) Is the same (average) probabiliity of occurrence for all species realistic?

3) We usually assume that species have their own preferred environmental conditions
4) Some species might still like similar conditions; there is a common component
5) We can separate this out with GLMMs or with a "sum-to-zero" contrast

Swiss birds: species-specific effects

```
model2 <- glm(occ~slp*Species,
              data = datalong, family="binomial")
```

- One intercept per species
- One slp effect per species
- But all are relative to the first species

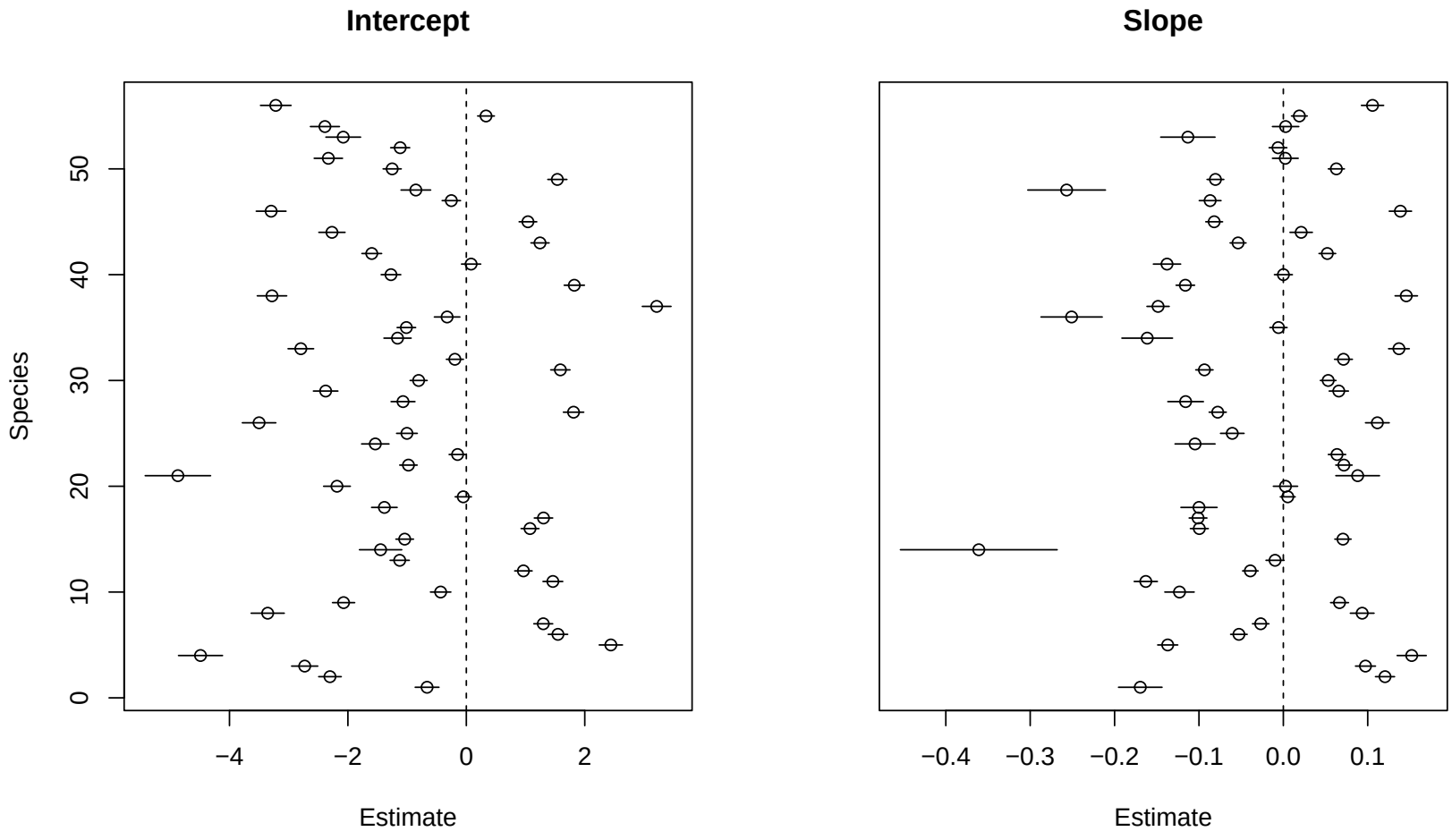
Swiss birds: species-specific effects

```
model3 <- glm(occ ~ 0 + Species + slp:Species,
              data = datalong, family="binomial")
```

The same model, but a bit easier to interpret

- One intercept per species
- One slp effect per species
- Not relative to each other (prevents post-hoc processing of tests and CI)

Swiss birds: results



Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)
  - At 0.5 there is equal probability to observe and not

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)
  - At 0.5 there is equal probability to observe and not
- Odds ratio for the first species at slp 0 is `exp(-0.663) = 0.52:1`

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)
  - At 0.5 there is equal probability to observe and not
- Odds ratio for the first species at slp 0 is `exp(-0.663) = 0.52:1`
  - We are two times more likely to not observe the species on flat ground, than to observe it

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)
  - At 0.5 there is equal probability to observe and not
- Odds ratio for the first species at slp 0 is `exp(-0.663) = 0.52:1`
  - We are two times more likely to not observe the species on flat ground, than to observe it
- This decreases by `exp(-0.17)` for every unit of slp `0.52*exp(-0.17) = 0.52*0.84 = 0.44`

Interpreting Binomial GLM coefficients

- Below one we are more likely to not observe the species
- The likelihood of observing the species 1 at flat ground is `exp(-0.663)/(1+exp(-0.663)) = 0.34` (given by the intercept)
  - At 0.5 there is equal probability to observe and not
- Odds ratio for the first species at slp 0 is `exp(-0.663) = 0.52:1`
  - We are two times more likely to not observe the species on flat ground, than to observe it
- This decreases by `exp(-0.17)` for every unit of slp `0.52*exp(-0.17) = 0.52*0.84 = 0.44`

Contrasts

There are other "contrast" treatments in R than "dummy"

- We can instead use "sum-to-zero" contrasts
  - If the sum is zero, the mean must be too
  - The coefficient of the last species is set to the negative sum

```
(contr <- contr.sum(levels(datalong$Species)))
```

```
##                          [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## Falco_subbuteo              1    0    0    0    0    0    0    0    0
## Anthus_trivialis            0    1    0    0    0    0    0    0    0
## Phylloscopus_bonelli        0    0    1    0    0    0    0    0    0
## Tetrao_tetrix               0    0    0    1    0    0    0    0    0
## Parus_caeruleus             0    0    0    0    1    0    0    0    0
## Dendrocopos_major           0    0    0    0    0    1    0    0    0
## Garrulus_glandarius         0    0    0    0    0    0    1    0    0
## Carduelis_spinus            0    0    0    0    0    0    0    1    0
## Loxia_curvirostra           0    0    0    0    0    0    0    0    1
## Phylloscopus_trochilus      0    0    0    0    0    0    0    0    0
## Certhia_brachydactyla       0    0    0    0    0    0    0    0    0
## Sylvia_borin                0    0    0    0    0    0    0    0    0
## Phoenicurus_phoenicurus     0    0    0    0    0    0    0    0    0
## Hippolais_icterina          0    0    0    0    0    0    0    0    0
```

Swiss birds: species-specific responses with common effect

```
model4 <- glm(occ~0+slp+Species + slp:Species, data = datalong, family = "binomial",
              contrasts = list(Species = contr))
coef(model4)[1]
```

```
##         slp
## -0.02703675
```

- The intercept is the same for all species
- One slp effect that is the same for all species (the mean of effects)
- One slp effect per species, relative to the common effect

The benefit: the average effect gets a statistical test.

By design corresponds to the result from our previous model:

`mean(coef(model3)[-c(1:ncol(y))])` = -0.0270367

Swiss birds

The three models have the same number of parameters, but are just differently parameterized. So, their log-likelihoods and AIC are the same:

```
AIC(model2, model3, model4)
```

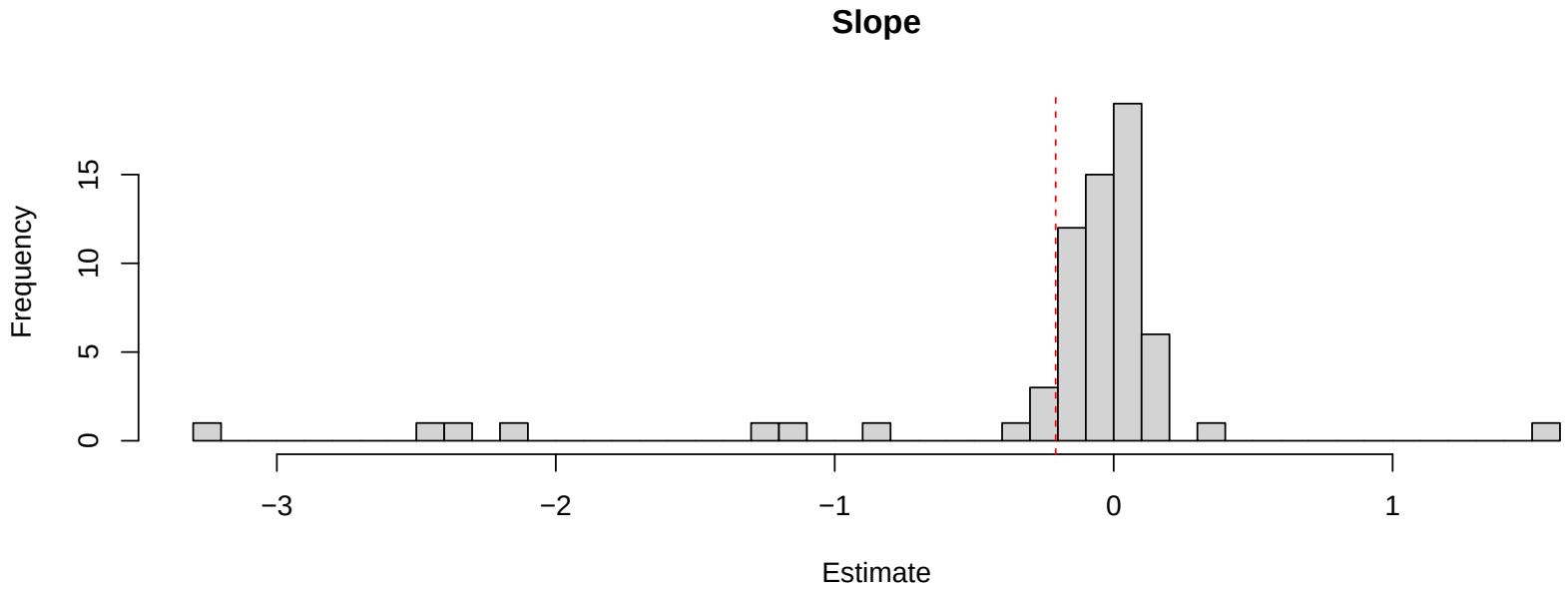
```
##          df      AIC
## model2 112 132240.8
## model3 112 132240.8
## model4 112 132240.8
```

Swiss birds: species-specific responses with common effect

```
##
## Call:
## glm(formula = occ ~ 0 + slp + Species + slp:Species, family = "binomial",
##     data = datalong, contrasts = list(Species = contr))
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## slp                              -0.0270367  0.0013931 -19.408  < 2e-16 ***
## SpeciesFalco_subbuteo            -0.6629549  0.1024622  -6.470 9.78e-11 ***
## SpeciesAnthus_trivialis          -2.3028109  0.0972711 -23.674  < 2e-16 ***
## SpeciesPhylloscopus_bonelli      -2.7295986  0.1127350 -24.213  < 2e-16 ***
## SpeciesTetrao_tetrix             -4.4890112  0.1896457 -23.671  < 2e-16 ***
## SpeciesParus_caeruleus            2.4406078  0.1003095  24.331  < 2e-16 ***
## SpeciesDendrocopos_major          1.5471247  0.0838875  18.443  < 2e-16 ***
## SpeciesGarrulus_glandarius        1.2992201  0.0813667  15.967  < 2e-16 ***
## SpeciesCarduelis_spinus          -3.3544392  0.1426728 -23.511  < 2e-16 ***
## SpeciesLoxia_curvirostra         -2.0729215  0.0955422 -21.696  < 2e-16 ***
## SpeciesPhylloscopus_trochilus    -0.4337227  0.0880349  -4.927 8.36e-07 ***
## SpeciesCerthia_brachydactyla      1.4614529  0.0852668  17.140  < 2e-16 ***
## SpeciesSylvia_borin               0.9639614  0.0750798  12.839  < 2e-16 ***
## SpeciesPhoenicurus_phoenicurus   -1.1252670  0.0827956 -13.591  < 2e-16 ***
## SpeciesHippolais_icterina        -1.4481047  0.1822145  -7.947 1.91e-15 ***
## SpeciesPyrrhula_pyrrhula         -1.0409062  0.0757068 -13.749  < 2e-16 ***
## SpeciesEmberiza_citrinella        1.0753753  0.0771653  13.936  < 2e-16 ***
## SpeciesMuscicapa_striata          1.3028822  0.0793269  16.424  < 2e-16 ***
## SpeciesPicus_canus               -1.3840458  0.1104130 -12.535  < 2e-16 ***
## SpeciesPicus_viridis             -0.0517754  0.0699471  -0.740 0.459174
## SpeciesAccipiter_gentilis        -2.1845813  0.1151459 -18.972  < 2e-16 ***
## SpeciesBonasa_bonasia            -4.8713731  0.2817025 -17.293  < 2e-16 ***
```

Swiss birds: conclusions

We can conclude that fewer bird species occur in steep places
Most species are more negatively affected than the average
Some species are positively affected by slope, but most negatively



Interpreting the coefficients

Or with `predict`:

```
predict(model3, newdata =
          data.frame(Species = factor("Falco_subbuteo", levels = colnames(y)), slp = 1),
        type = "response")
```

```
##         1
## 0.3031163
```

Example: macroinvertebrate counts in USA desert

Observation process: count of macroinvertebrates in three "dips"
Alternatively: The proportion of a species in a dip
Alternatively: Was this species found in the dip

Example: macroinvertebrate counts in USA desert

- Data by Pina and Lougheed 2022
- Counts of 14 species, in 2018 and 2019, in 14 wetlands
- Main goal: assess impacts of water quality on macroinvertebrates

The abundance data

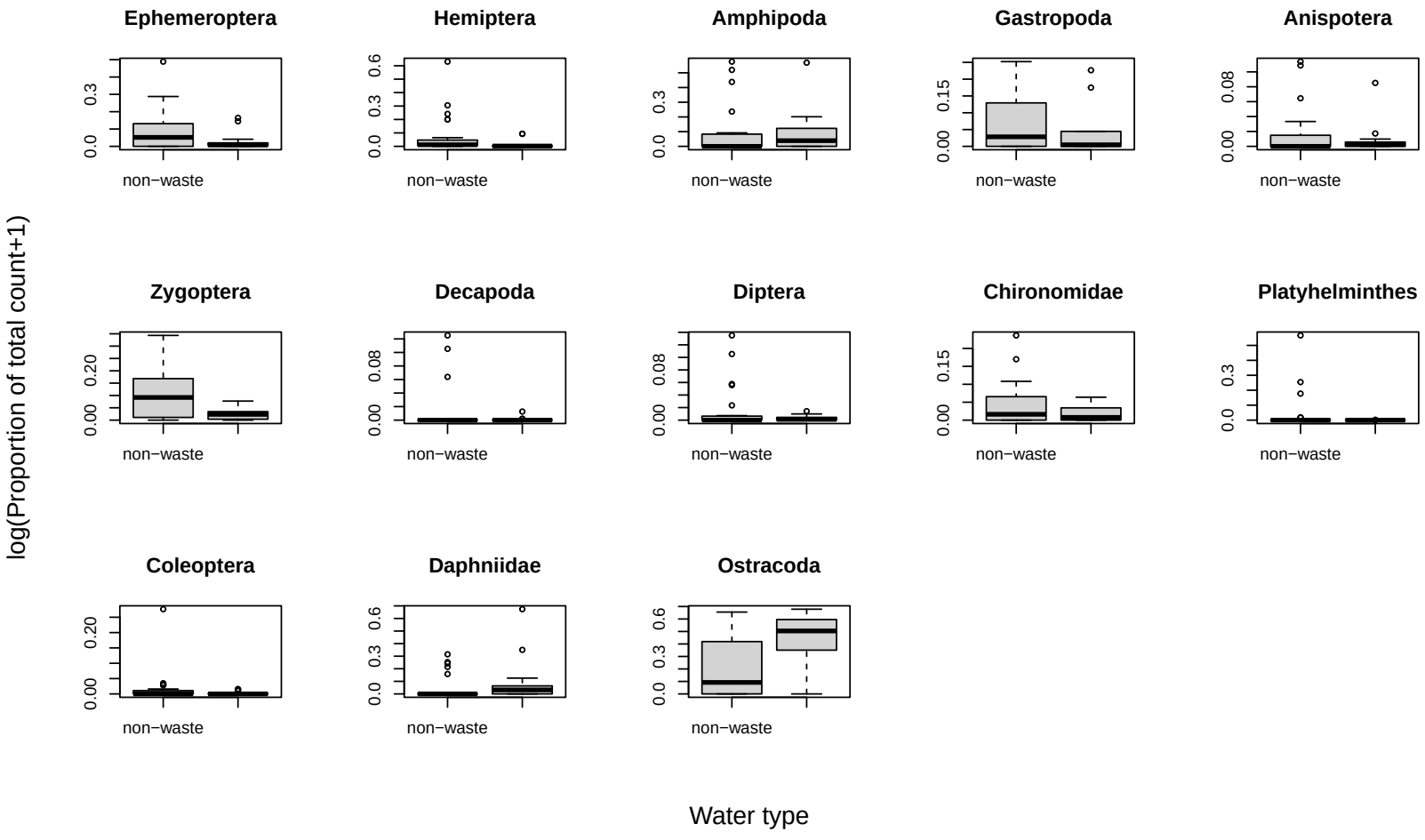
Ephemeroptera	Hemiptera	Amphipoda	Gastropoda	Anispotera	Zygoptera	Decapoda	Diptera	Chiron
0	1	100	8	0	1	12	0	
21	0	0	5	3	10	0	1	
11	3	5	0	4	32	0	0	
0	1	11	0	0	0	0	0	
80	1	60	25	0	7	0	2	
9	15	0	6	0	10	0	1	
30	0	0	10	7	25	0	1	
10	1	60	190	4	60	0	0	
6	0	0	5	0	3	2	2	
5	41	0	5	0	2	0	0	
32	2	0	0	1	35	0	0	
6	25	0	31	2	13	0	0	
28	15	400	0	21	14	0	0	
28	15	400	0	21	14	0	0	
6	150	200	70	2	48	3	4	
6	150	200	70	2	48	3	4	
9	1	0	35	14	50	0	7	
26	1	0	1	10	21	0	12	
0	1	0	2	2	7	0	0	
1	0	60	6	1	9	0	0	
13	0	15	26	0	8	0	1	
87	4	0	0	0	3	0	0	
2	1	11	4	2	1	0	1	

The environment data

Year	Hydro	Water_Type	Conductivity	DOC	TDN	Turbidity	Alkalinity	Total_CHL	Correc
2018	permanent	non-waste	4.060	2.846	0.306	4.40	63.050	2.231	
2018	permanent	waste	2.582	23.160	3.544	60.73	412.400	95.211	
2018	permanent	non-waste	8.563	28.120	2.450	17.40	363.708	16.915	
2018	permanent	non-waste	15.710	75.040	7.160	38.50	457.625	26.160	
2018	ephemeral	non-waste	1.029	4.012	0.386	3.78	198.042	12.657	
2018	ephemeral	non-waste	1.204	5.356	0.491	24.70	168.042	30.353	

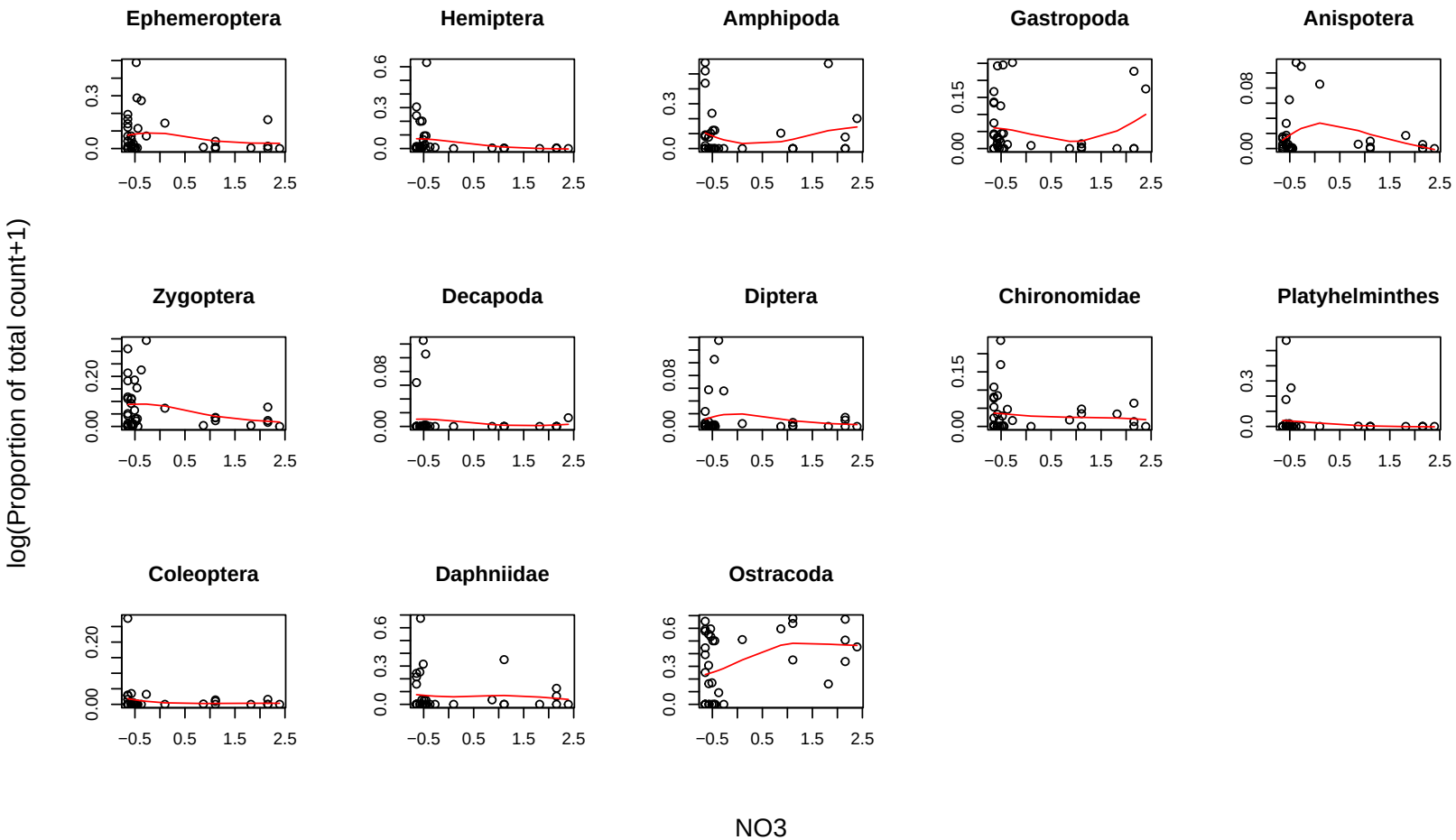
- 11 environmental variables
  - Water chemistry
  - Water type
  - Presence of hydro power
  - Permanent or temporary wetland

Visually inspect the data: categorical covariate



Is there a common effect?

Visually inspect the data: continuous covariate



Is there a common effect?

Wetlands: species-specific responses with common effect

```
model5 <- glm(Count~0+Species+NO3+NO3:Species, data = long, family = "poisson", contrasts = list(Species = con
```

```
##   SpeciesEphemeroptera       SpeciesHemiptera       SpeciesAmphipoda
##             2.67111728             2.36082751             3.86298300
##      SpeciesGastropoda      SpeciesAnispotera       SpeciesZygoptera
##             2.55157209             1.21032834             2.70445223
##        SpeciesDecapoda        SpeciesDiptera    SpeciesChironomidae
##            -0.51635148             0.44859004             2.18230972
## SpeciesPlatyhelminthes      SpeciesColeoptera      SpeciesDaphniidae
##             0.93797899             0.46774680             4.02318285
##       SpeciesOstracoda                    NO3          Species1:NO3
##             5.76826016            -0.02504095             0.04674984
##           Species2:NO3           Species3:NO3           Species4:NO3
##            -0.74506167             0.21656771             0.51506145
##           Species5:NO3           Species6:NO3           Species7:NO3
##             0.14244009             0.12796859            -0.43765004
##           Species8:NO3           Species9:NO3          Species10:NO3
##             0.06845114             0.34248963            -0.95659058
##          Species11:NO3          Species12:NO3
##             0.48870124            -0.05675712
```

Wetlands: species-specific responses with common effect

Count data is usually overdispersed; we might want to switch to a NB.

```
coef(model6)
```

```
##   SpeciesEphemeroptera      SpeciesHemiptera           Specie
##             2.67109283            2.24282790
##      SpeciesGastropoda     SpeciesAnispotera           Specie
##             2.56892440            1.20450595
##        SpeciesDecapoda       SpeciesDiptera     SpeciesCh
##            -0.49779235            0.44836747
## SpeciesPlatyhelminthes     SpeciesColeoptera        Species
##             0.80901413            0.43073882
##       SpeciesOstracoda                   NO3             Sp
```

Wetlands: conclusions

We can conclude that NO3 has, **on average**, a negative effect on our species pool (but this is not statistically significant)
Some species are more negatively affected than the average, some more positive
Some species are positively affected by NO3, but most negatively

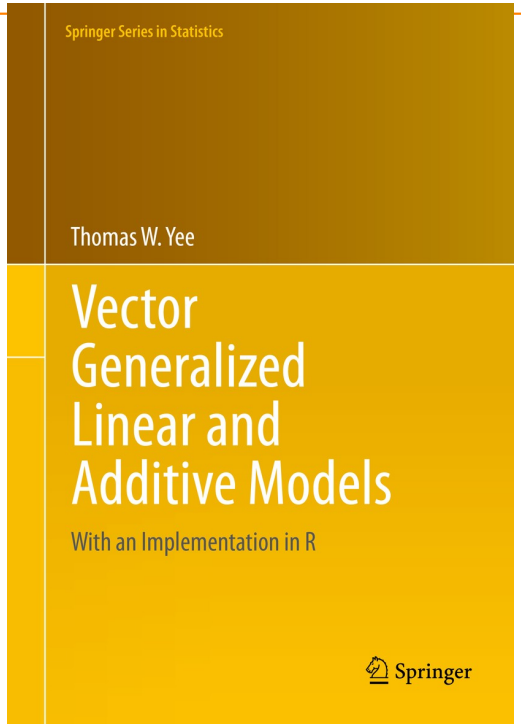
Interpreting the coefficients

```
##   SpeciesEphemeroptera       SpeciesHemiptera       SpeciesAmphipoda
##             2.67109283             2.24282790             3.85799261
##      SpeciesGastropoda      SpeciesAnispotera       SpeciesZygoptera
##             2.56892440             1.20450595             2.70360747
##        SpeciesDecapoda        SpeciesDiptera   SpeciesChironomidae
##            -0.49779235             0.44836747             2.15746463
## SpeciesPlatyhelminthes      SpeciesColeoptera     SpeciesDaphniidae
##             0.80901413             0.43073882             4.02257593
##       SpeciesOstracoda                    NO3         Species1:NO3
##             5.75226893            -0.03750298             0.06151128
##           Species2:NO3           Species3:NO3         Species4:NO3
##            -1.10550200             0.27812794             0.46946826
##           Species5:NO3           Species6:NO3         Species7:NO3
##             0.25306729             0.15672356            -0.32428547
##           Species8:NO3           Species9:NO3        Species10:NO3
##             0.09136078             0.48958482            -1.33007265
##          Species11:NO3          Species12:NO3
##             0.62804597            -0.06027316
```

- Negative means a decrease in the response and positive increase
- More specifically here: the coefficient is multiplicative decrease in `exp(intercept)` for a unit change in `NO3`
- E.g., for "Ephemeroptera": `exp(2.243)*exp(-0.038+ 0.062)` = `9.4*1.025`

Vector GLMs

- One GLM per species
- Each gets their own dispersion parameter
- Slightly more flexible than what we have done so far



Methods in Ecology and Evolution
British Ecological Society
Methods in Ecology and Evolution 2012, **3**, 471–474 doi: 10.1111/j.2041-210X.2012.00190.x

mvabund – an R package for model-based analysis of multivariate abundance data

Yi Wang[1,2], Ulrike Naumann[1], Stephen T. Wright[1], and David I. Warton[1,3]*

[1]School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052, Australia; [2]School of

Fitting vector GLMs

A few software implementations exist:

- The `VGAM` R-package
- The `glmmTMB` R-package
- The `gllvm` R-package

Clearly, we will use the last one.

VGLM Likelihood

- We use MLE for estimation
- With a distribution in the "exponential family" (for fixed ϕ)

All GLMs have the likelihood:

$$\mathcal{L}(y_{ij};\Theta) = \exp\{\frac{y_{ij}\eta_{ij} - b(\eta_{ij})}{a(\phi_j)} + c(y_{ij},\phi_j)\} \tag{4}$$

So, now we have ϕ_j instead of ϕ

gllvm

Received: 7 May 2019 | Accepted: 5 September 2019

DOI: 10.1111/2041-210X.13303

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

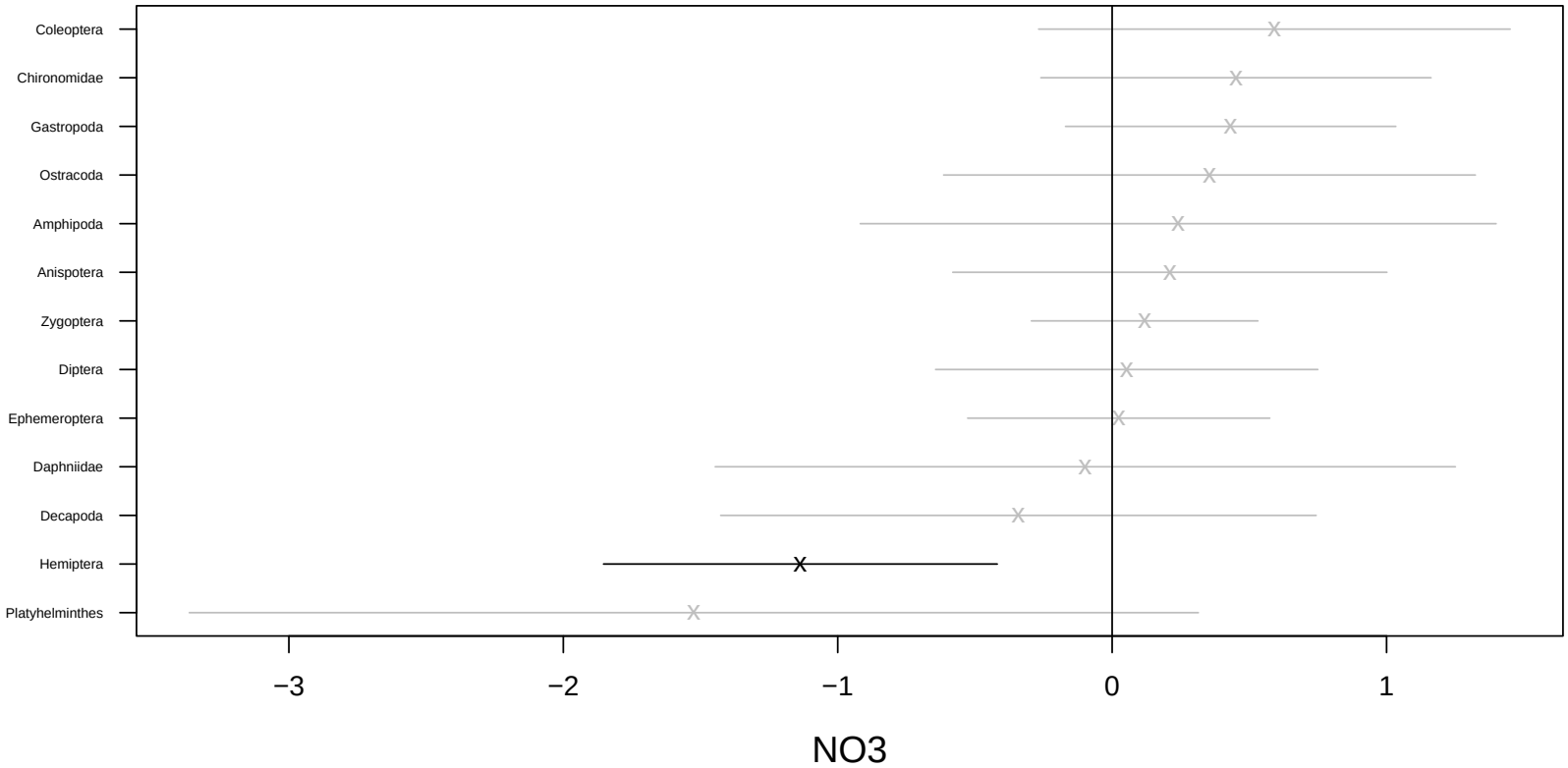
gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R

Jenni Niku[1] | Francis K. C. Hui[2] | Sara Taskinen[1] | David I. Warton[3]

- Originally published in 2019 by Niku et al. I "joined in" shortly after
- For model-based multivariate analysis of community ecological data
- Models are fitted in C++ (Kristensen et al. 2015)
- Can fit many different models: VGLM(M), JSDM, and ordination

VGLM in gllvm

```
model7 <- gllvm::gllvm(y, X = X, formula = ~NO3, family = "negative.binomial", num.lv = 0)
gllvm::coefplot(model7)
```



(the "species" covariate is implicit)

Does NO3 improve the model?

```
model8 <- gllvm::gllvm(y, formula = ~1, family = "negative.binomial", num.lv = 0)
anova(model7, model8)
```

```
## Model  1 :  y ~ NULL
## Model  2 :  ~ NO3
```

```
##   Resid.Df        D Df.diff  P.value
## 1      455  0.00000       0
## 2      442 15.75468      13 0.262631
```

NO3 does not improve the model.

Does NO3 improve the model?

```
AIC(model7, model8)
```

```
##         df      AIC
## model7 39 2663.824
## model8 26 2653.578
```

There are 14 more parameters in model7 than model8, so AIC needs to be 14*2 points lower.

Comparison to an "ordinary" NB GLM

```
model9 <- gllvm::gllvm(y, X = X, formula = ~NO3, family = "negative.binomi
                       disp.formula = rep(1, ncol(y)), num.lv = 0)

data.frame(VGLM = c(coef(model7, "Xcoef")), GLM = c(coef(model9, "Xcoef"))
```

```
##             VGLM         GLM
## 1   0.02397540  0.02400115
## 2  -1.13605814 -1.14307638
## 3   0.24075786  0.24067679
## 4   0.43199257  0.43191304
## 5   0.21027053  0.21574989
## 6   0.11847738  0.11916327
## 7  -0.34157725 -0.36176403
## 8   0.05311675  0.05386589
## 9   0.45118842  0.45205838
## 10 -1.52436675 -1.36752294
## 11  0.59123059  0.59062790
```

Comparison NB GLM: standard errors

```
data.frame("VGLM" = c(model7$sd$Xcoef), "GLM" = c(model9$sd$Xcoef))
```

```
##          VGLM       GLM
## 1   0.2808008 0.3835172
## 2   0.3659933 0.4283106
## 3   0.5910672 0.4062612
## 4   0.3070433 0.3435331
## 5   0.4035991 0.5086298
## 6   0.2106435 0.3928682
## 7   0.5535556 0.4188498
## 8   0.3553199 0.4292097
## 9   0.3627100 0.4282210
## 10  0.9379919 0.5400630
## 11  0.4383469 0.4185581
## 12  0.6880387 0.3973900
## 13  0.4942781 0.4538778
```

Downsides

- VGLM defaults to 1 dispersion parameter per species
- VGLM assumes 1 parameter per species per covariate
- This does not tend to work very well for real (sparse) community data
- VGLM assumes independence of species
- Does not include random effects (pseudoreplication, autocorrelation)

Summary

- GLMs are fun, but not usually suitable for multispecies data
- VGLMs; fitting one model per species gives more flexibility
- This facilitates adding components that are shared across species
- Which is especially helpful when working with random effects

So far we have assumed that species do not influence each other