



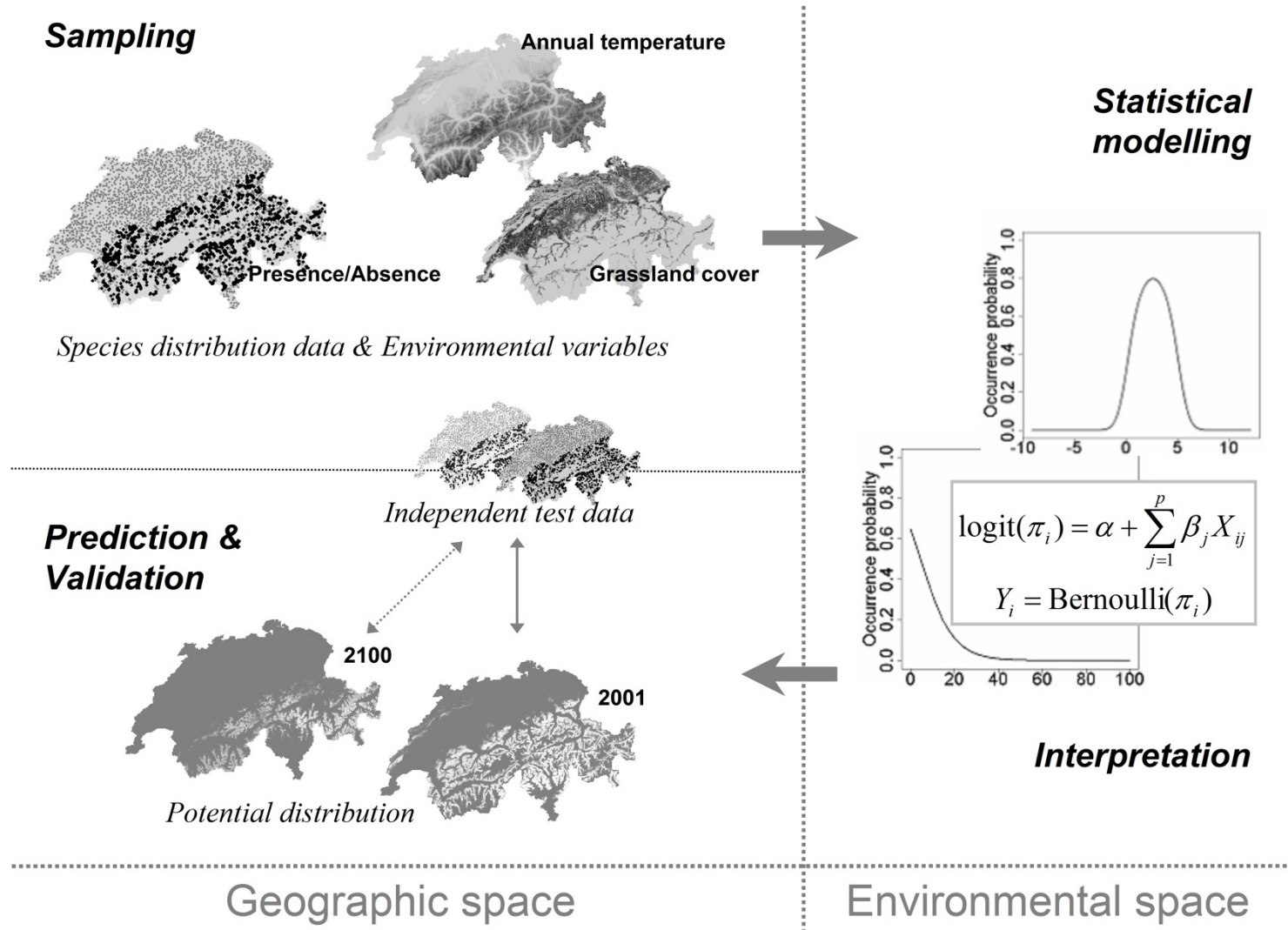
# Fitting and evaluating species distribution models

**Damaris Zurell**

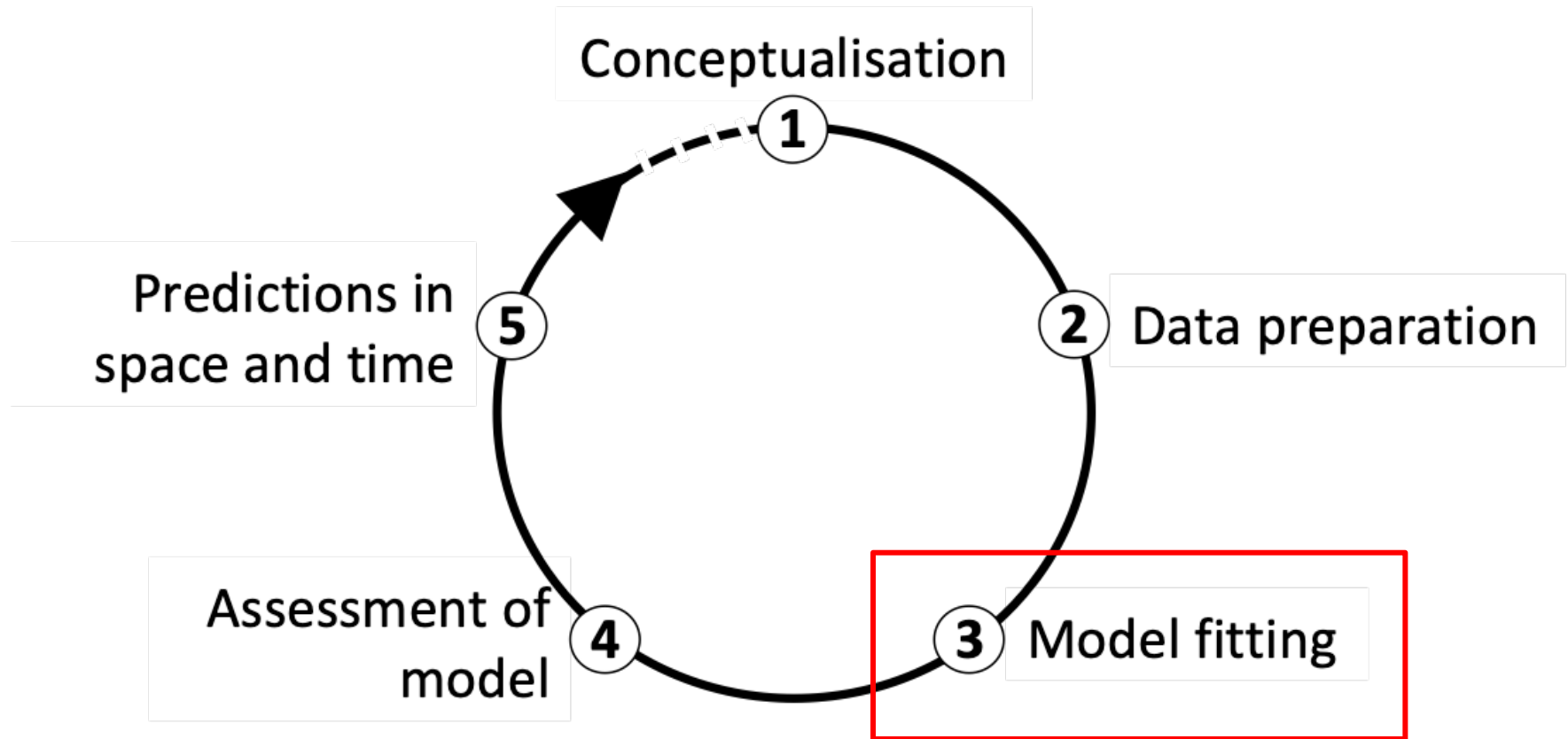
<https://damarisurell.github.io>

 @ZurellLab

# Species distribution models



# SDMs – model building steps

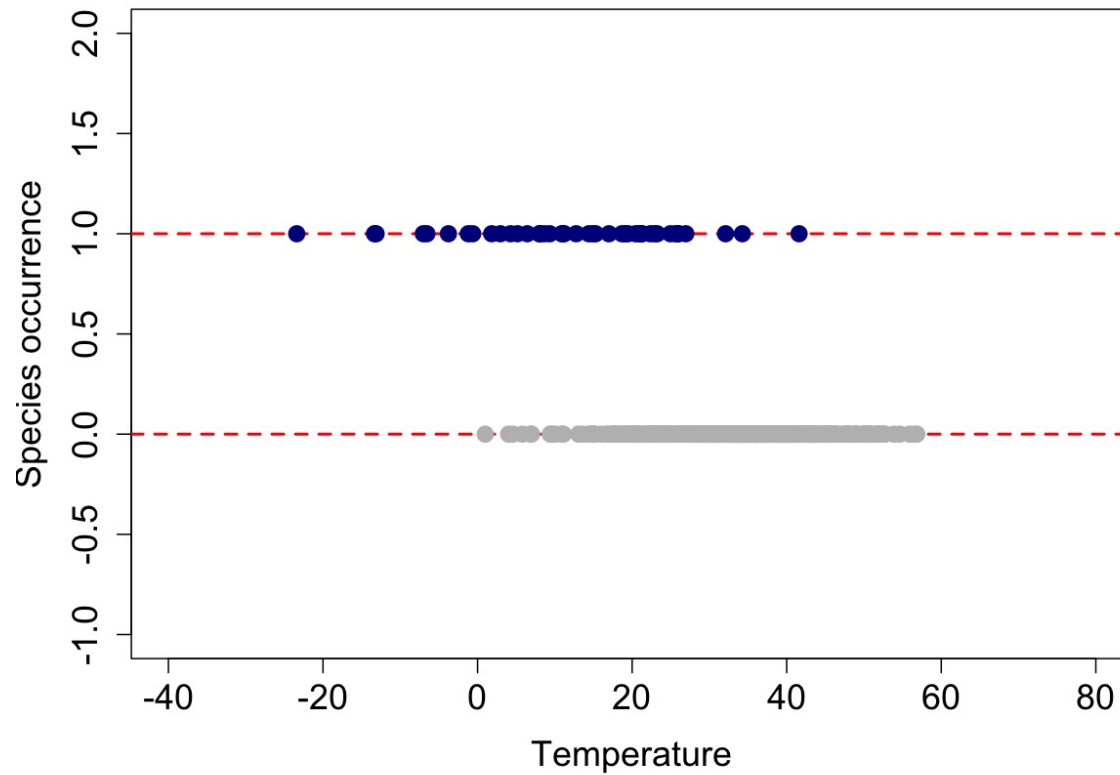


# Generalised linear models (GLMs)

- Parametric regression method based on maximum likelihood estimation
- Allow error distributions different from Normal distribution
- The linear predictor is related to the response variable by a link function
- Logistic regression uses the logit link

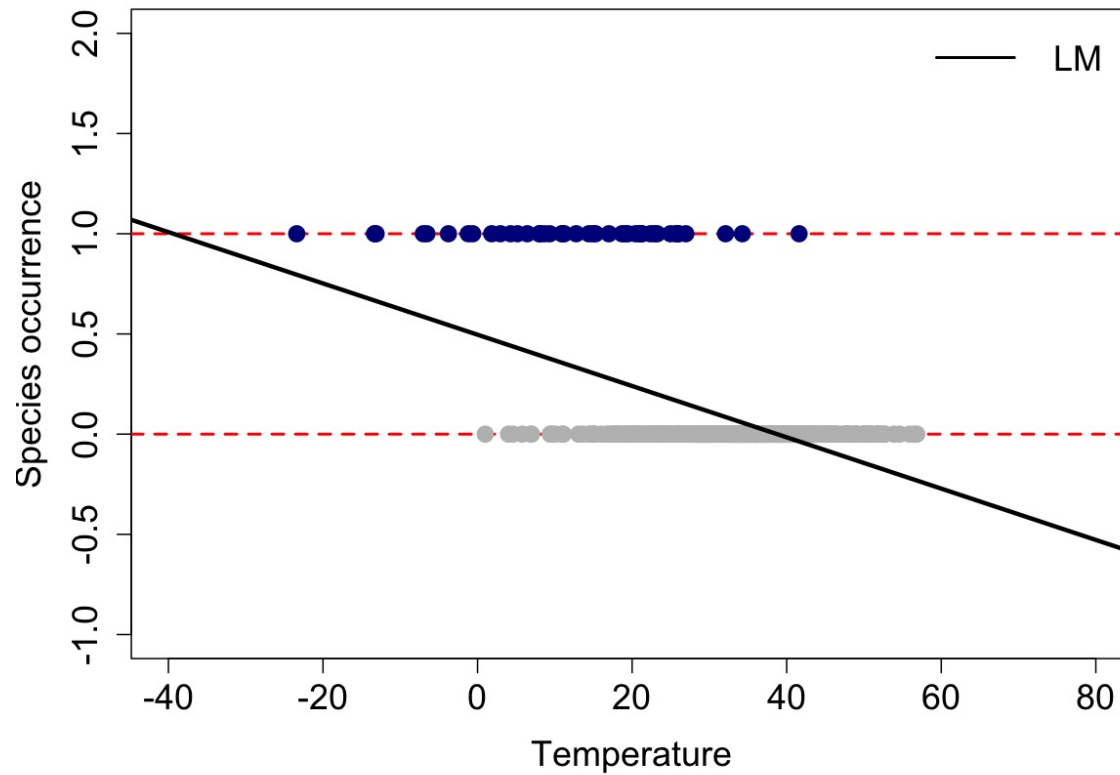
# Generalised linear models (GLMs)

Our data are bounded (0,1)



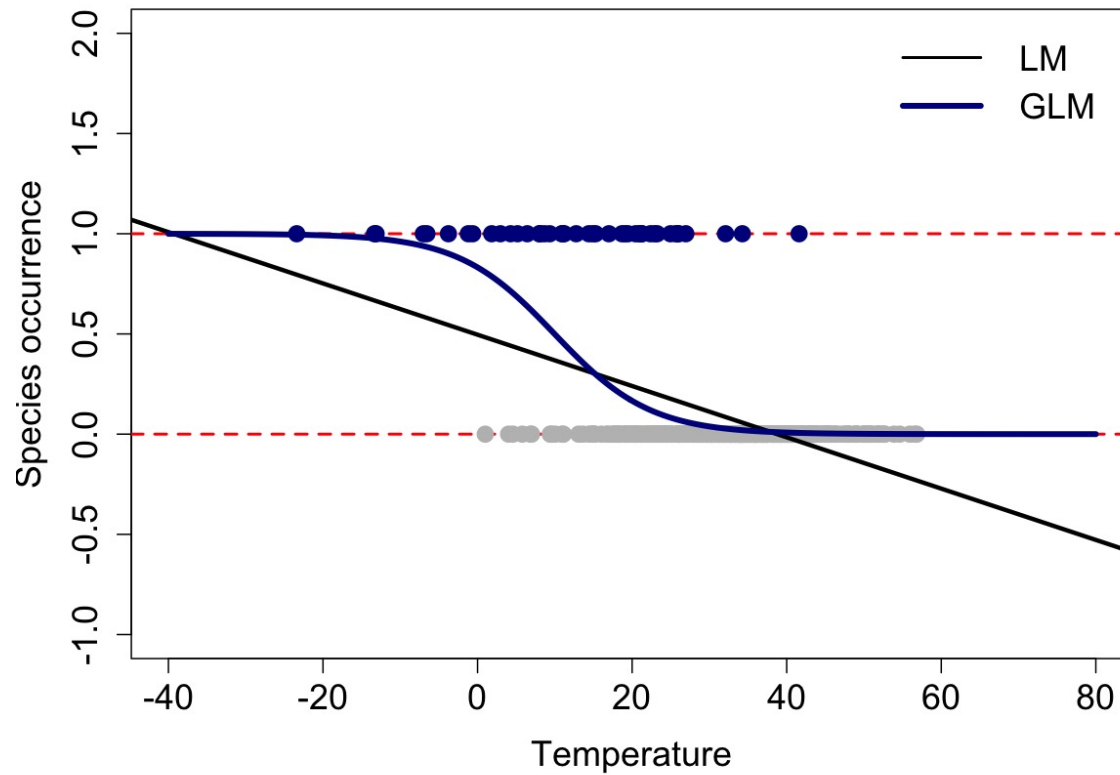
# Generalised linear models (GLMs)

The response OLS regression is unbounded  $(-\infty, \infty)$



# Generalised linear models (GLMs)

In GLMs, the response is bounded through link function



# Generalised linear models (GLMs)

- In OLS regression the normally distributed error ranges  $(-\infty, \infty)$

$$E(Y|X) = \beta X + \epsilon$$

- Presence/Absence data are bounded  $(0,1)$
- The link function is used to transform the response to normality

$$E(Y|X) = \pi(X) = \frac{e^{\beta X + \epsilon}}{1 + e^{\beta X + \epsilon}}$$

- The logit  $g(X)$  is linear in its parameters

$$g(X) = \ln \left( \frac{\pi(X)}{1 - \pi(X)} \right) = \beta X + \epsilon$$



# Fitting GLMs in R

- Example with Ring Ouzel in UK



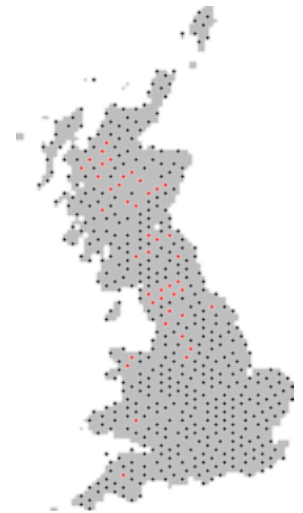
```
m1 <- glm(Turdus_torquatus ~ bio11, family="binomial", data= sp_dat)
```

Response variable

Predictor variable

Link function

Data



# Fitting GLMs in R

- Example with Ring Ouzel in UK

```
summary(m1)
```

Call:

```
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",  
     data = sp_dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8183	-0.2813	-0.1803	-0.1157	3.1911

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.60045	0.44867	3.567	0.000361 ***
bio11	-0.16071	0.01978	-8.126	4.44e-16 ***

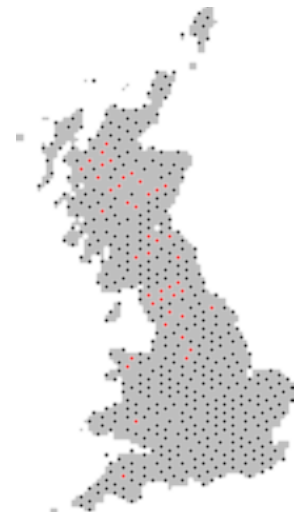
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 294.43 on 506 degrees of freedom  
Residual deviance: 176.39 on 505 degrees of freedom  
AIC: 180.39

Number of Fisher Scoring iterations: 7



# Fitting GLMs in R

- Example with Ring Ouzel in UK

```
summary(m1)
```

Call:

```
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",  
     data = sp_dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8183	-0.2813	-0.1803	-0.1157	3.1911

Coefficients:

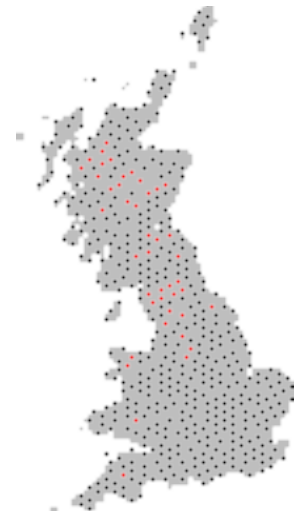
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.60045	0.44867	3.567	0.000361 ***
bio11	-0.16071	0.01978	-8.126	4.44e-16 ***

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 294.43 on 506 degrees of freedom  
Residual deviance: 176.39 on 505 degrees of freedom  
AIC: 180.39

Number of Fisher Scoring iterations: 7



# Fitting GLMs in R

- Example with Ring Ouzel in UK

```
summary(m1)
```

Call:

```
glm(formula = Turdus_torquatus ~ bio11, family = "binomial",  
     data = sp_dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8183	-0.2813	-0.1803	-0.1157	3.1911

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.60045	0.44867	3.567	0.000361 ***
bio11	-0.16071	0.01978	-8.126	4.44e-16 ***

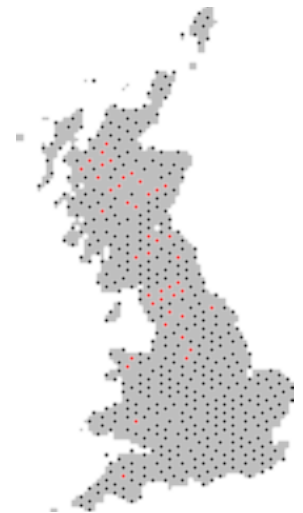
---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 294.43	on 506	degrees of freedom
Residual deviance: 176.39	on 505	degrees of freedom
AIC: 180.39		

Number of Fisher Scoring iterations: 7



# Likelihood, deviance and AIC

- Likelihood compares fitted against observed values

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n (y_i \times \ln[\pi(x_i)] + (1 - y_i) \times \ln[1 - \pi(x_i)])$$

- Deviance is defined as:

$$D = -2 \times L$$

- Explained deviance is the amount of variation explained by the model compared to the null model

$$D^2 = 1 - \frac{D(model)}{D(Null.model)}$$

- AIC is the Akaike Information criterion that penalizes model complexity (number of parameters  $p$ )

$$AIC = -2 \times L + 2 \times (p + 1) = D + 2 \times (p + 1)$$

# Consideration for model fitting

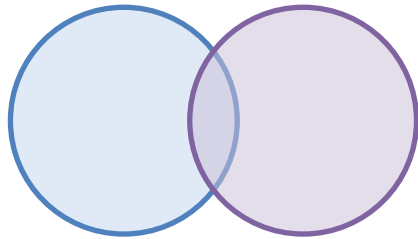
- How to deal with multicollinearity in the environmental data?
- How many variables should be included in the model (without overfitting) and how should we select these?
- Which model settings should be used?
- When multiple model algorithms or candidate models are fitted, how to select the final model or average the models?
- Do we want to threshold the predictions, and which threshold should be used?
- Do we need to test or correct for non-independence in the data (spatial or temporal autocorrelation, nested data)?

# Consideration for model fitting

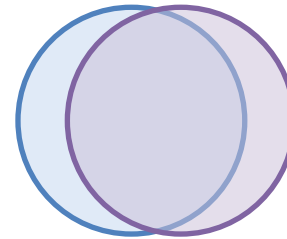
- How to deal with multicollinearity in the environmental data?
- How many variables should be included in the model (without overfitting) and how should we select these?
- Which model settings should be used?
- When multiple model algorithms or candidate models are fitted, how to select the final model or average the models?
- Do we want to threshold the predictions, and which threshold should be used?
- Do we need to test or correct for non-independence in the data (spatial or temporal autocorrelation, nested data)?

# What is multicollinearity

- Predictor variables are not independent from each other



Weak correlation

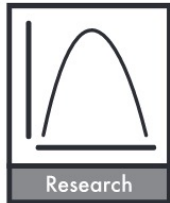


Strong correlation

- Makes it difficult to identify the more meaningful predictor
- Lead to inflated errors
- Most problematic when extrapolating



# How to deal with multicollinearity



EDITOR'S  
CHOICE

**Ecography 36: 027–046, 2013**

doi: 10.1111/j.1600-0587.2012.07348.x

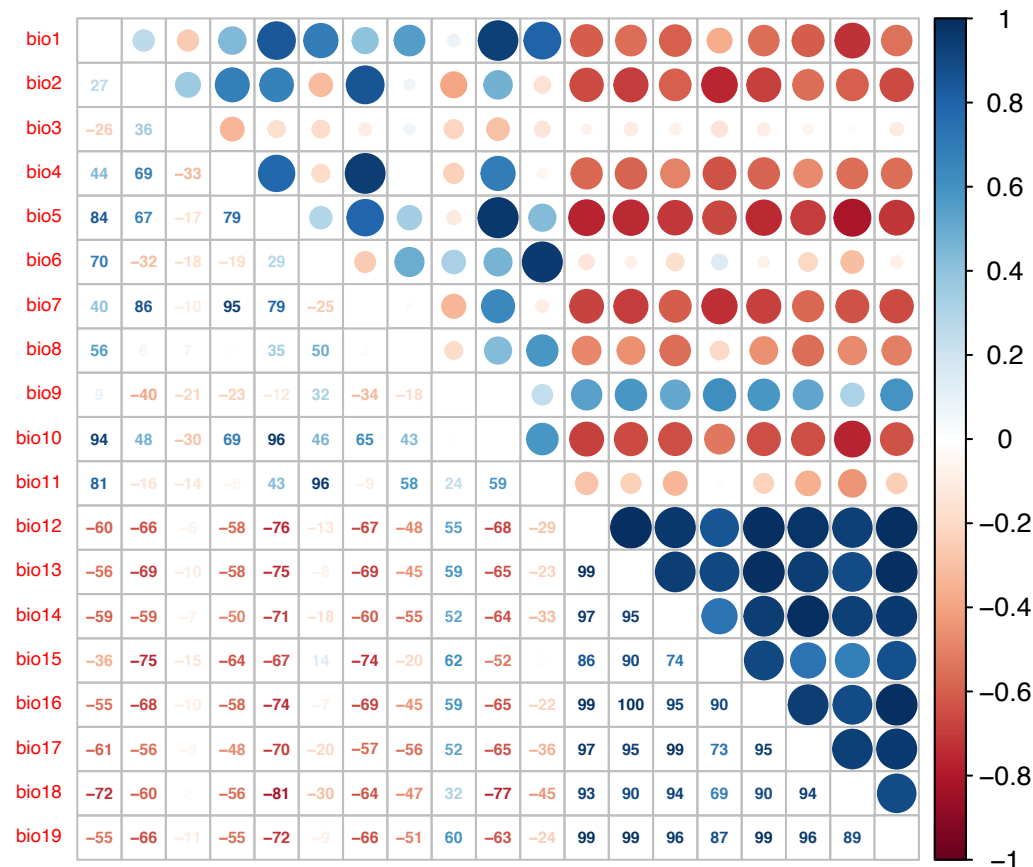
© 2012 The Authors. Ecography © 2012 Nordic Society Oikos  
Subject Editor: Marti Jane Anderson. Accepted 24 February 2012

## **Collinearity: a review of methods to deal with it and a simulation study evaluating their performance**

**Carsten F. Dormann, Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Márquez, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell and Sven Lautenbach**

# How to deal with multicollinearity

- We inspect the pairwise correlations between the 19 bioclimatic variables :



# How to deal with multicollinearity

- Then, we determine the univariate importance of all predictors

```
> AIC(m1)
[1] 180.3946
```

- From pairs of highly correlated variables, we select the variable with lower AIC

```
> cor(sp_dat$bio11, sp_dat$bio1, method='spearman')
[1] 0.8091144
```

```
> m2 <- glm(Turdus_torquatus ~ bio1, family="binomial", data= sp_dat)
> AIC(m2)
[1] 189.9521
```

- In the Script, we use the function select07()

# Variable selection and model selection

- How many parameters to include in the model?
  - Rule of thumb: 10 presences per parameter
- What is the most parsimonious model?
  - Step-wise variable selection using AIC

```
m_full <- glm( Turdus_torquatus ~ bio11 + I(bio11^2) + bio8 + I(bio8^2),  
              family='binomial', data=sp_dat)
```

```
m_step <- step(m_full)
```

```
> m_step
```

```
Call:  glm(formula = Turdus_torquatus ~ bio11 + bio8, family = "binomial",  
          data = sp_dat)
```

```
Coefficients:
```

(Intercept)	bio11	bio8
1.86289	-0.11329	-0.03293

```
Degrees of Freedom: 506 Total (i.e. Null); 504 Residual
```


```
Null Deviance: 294.4
```

```
Residual Deviance: 172.2            AIC: 178.2
```

# Thank you for your interest

**Contact:**  
**Damaris Zurell**  
Ecology & Macroecology  
University of Potsdam



<https://damarisurell.github.io>  
Email: [damaris.zurell@uni-potsdam.de](mailto:damaris.zurell@uni-potsdam.de)  
 @ZurellLab