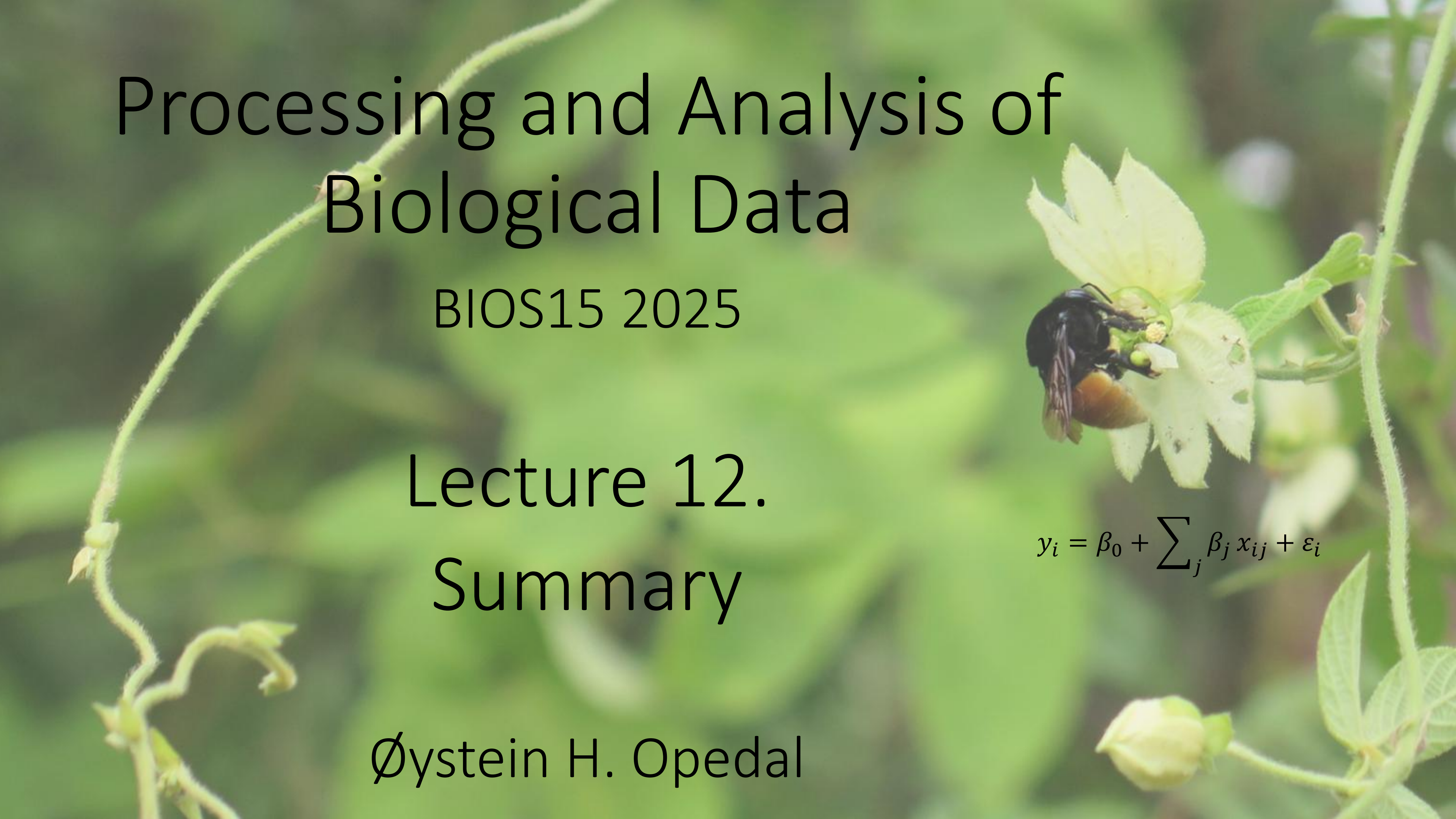


Processing and Analysis of Biological Data

BIOS15 2025

Lecture 12. Summary

$$y_i = \beta_0 + \sum_j \beta_j x_{ij} + \varepsilon_i$$

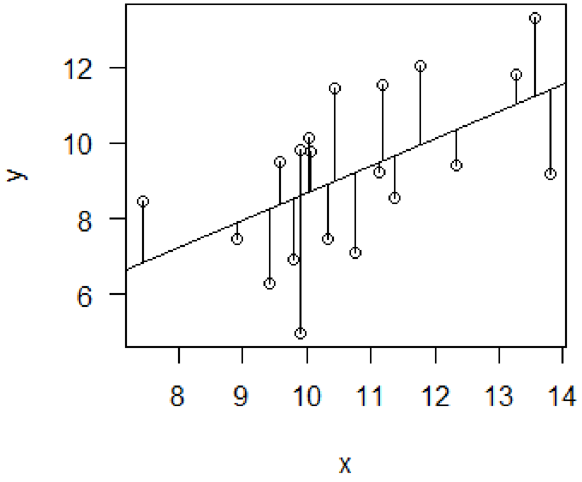
Øystein H. Opedal

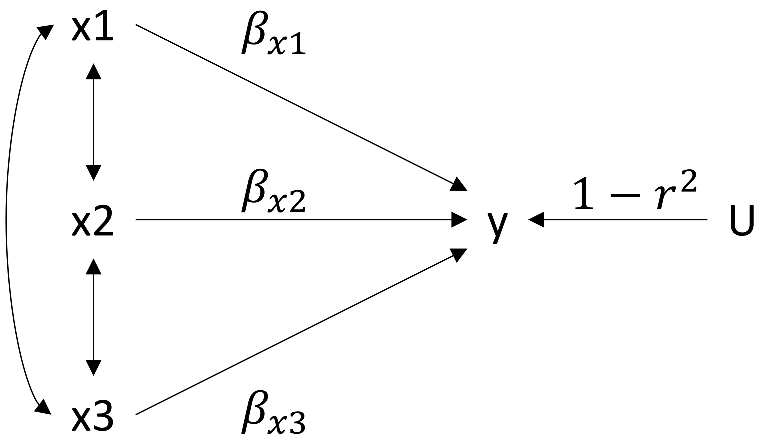
A data analysis ABC?

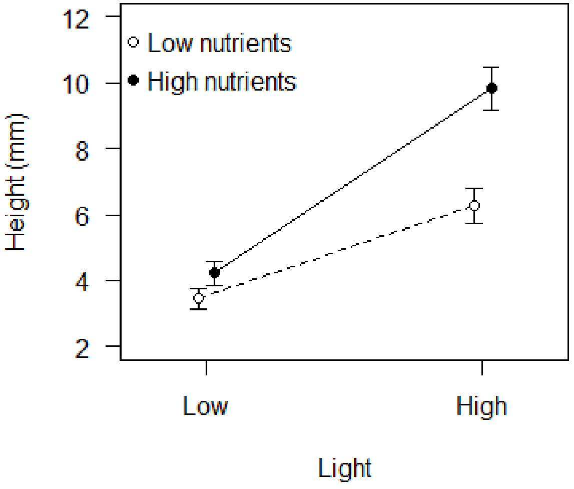
- 1. Think about research question, hypotheses and predictions
- 2. Explore data (graphically, summary stats)
- 3. Sketch the main figures
- 4. Decide on analytical approach:
  - Which kind of response variable? GLM(M)?
  - Which kinds of predictors? (categorical, continuous, “nuisance variables”)
  - Random structure (what do you know about the data-generating process?)
- 5. Fit models and do model selection/evaluate support
- 6. Compute relevant metrics, interpret quantitatively, add e.g. regression lines to figures
- 7. Write results by using output of 6 to answer 1

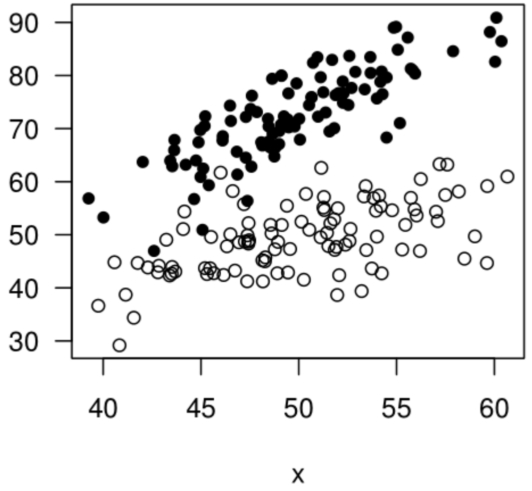
Overview of
(generalized) linear (mixed) models

- Continuous covariates: (multiple) regression
- Categorical covariates: N-way ANOVA
- Continuous and categorical covariates: ANCOVA



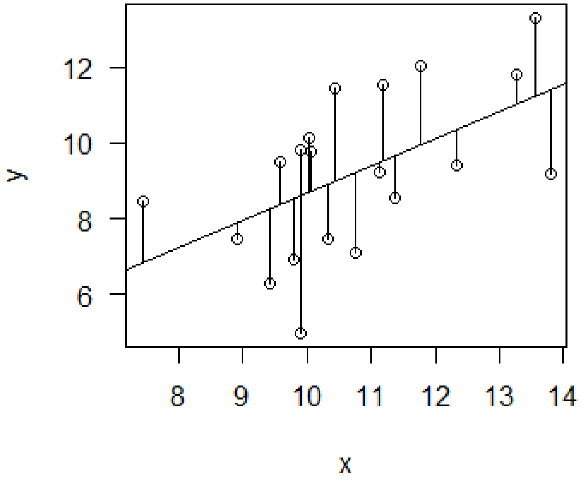


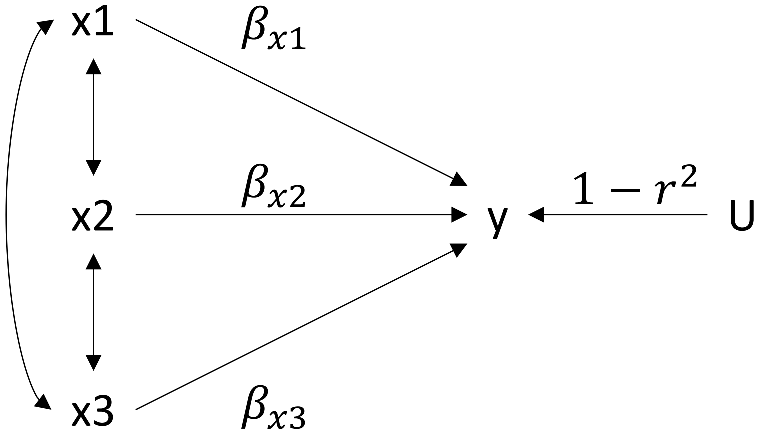


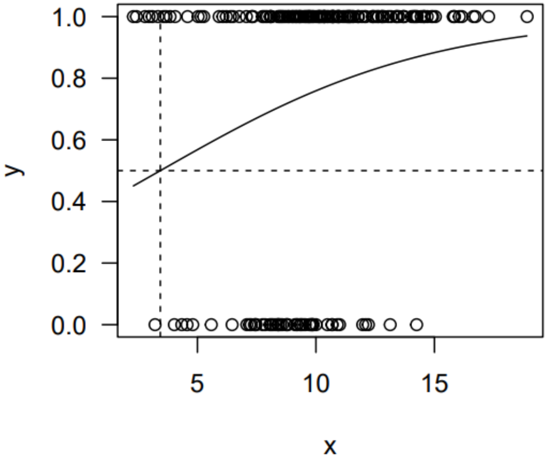


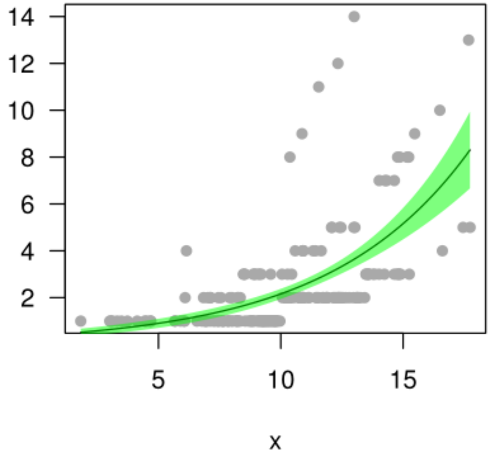
Overview of (generalized) linear models

- Binary/proportional data: Logistic regression
- Count data: Poisson GLM
- Overdispersed count data: Negative binomial GLM



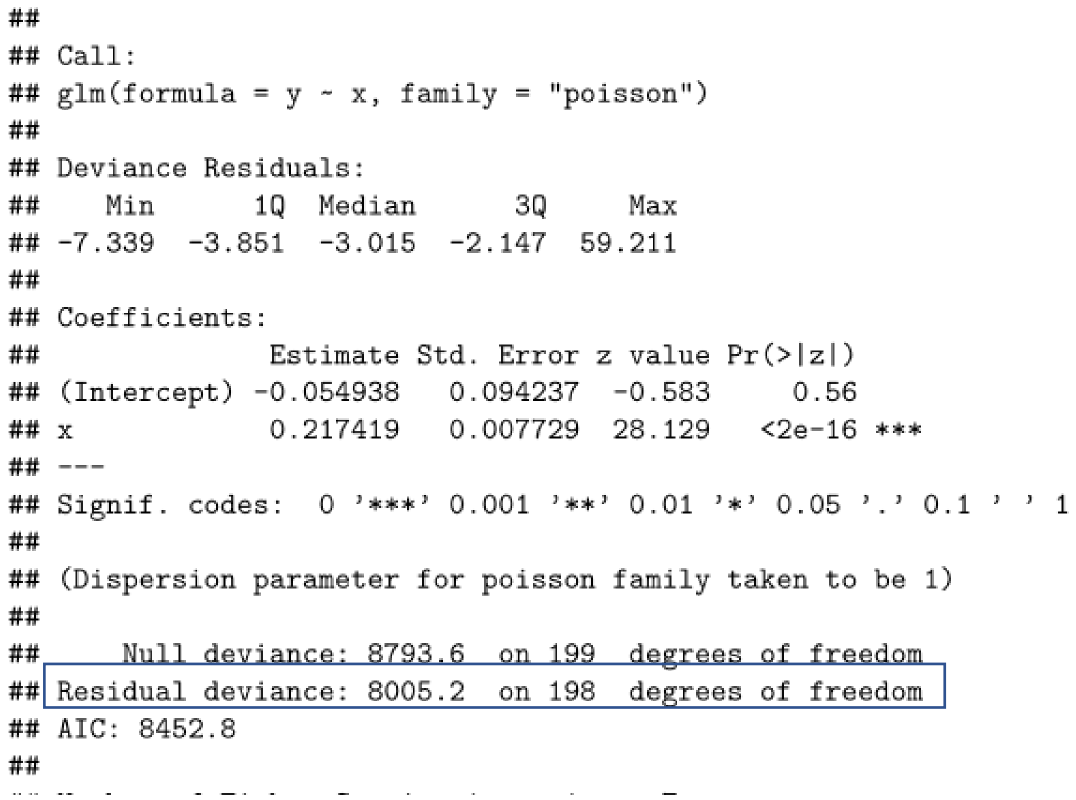






Poisson regression model in R

- If the variance increase disproportionally with the mean, there is **overdispersion** in the data
- Overdispersion is a problem if the residual deviance is much higher than the residual degrees of freedom



```
##
## Call:
## glm(formula = y ~ x, family = "poisson")
##
## Deviance Residuals:
##    Min       1Q   Median       3Q      Max
## -7.339   -3.851   -3.015   -2.147   59.211
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.054938   0.094237  -0.583     0.56
## x            0.217419   0.007729  28.129   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8793.6  on 199  degrees of freedom
## Residual deviance: 8005.2  on 198  degrees of freedom
## AIC: 8452.8
##
```

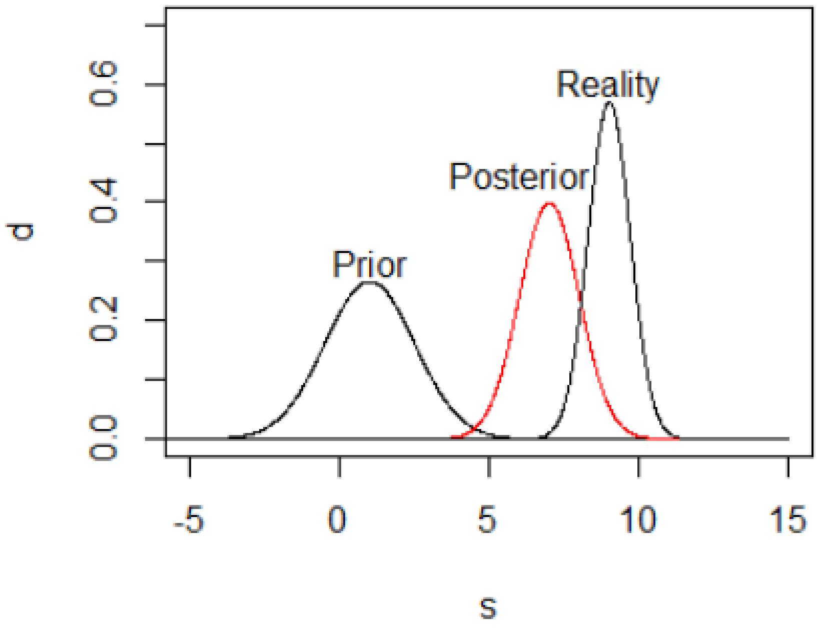
Frequentist vs. Bayesian statistics

- There are two main schools or approaches to statistical analysis: **frequentist** approaches and **Bayesian** approaches
- So far we have fitted models with least-square or maximum likelihood methods. These are frequentist methods yielding a test statistic we can compare to a known distribution
- In practise, biologists often use a mix of approaches, depending on the task or availability of software/R packages

Rev. T. Bayes

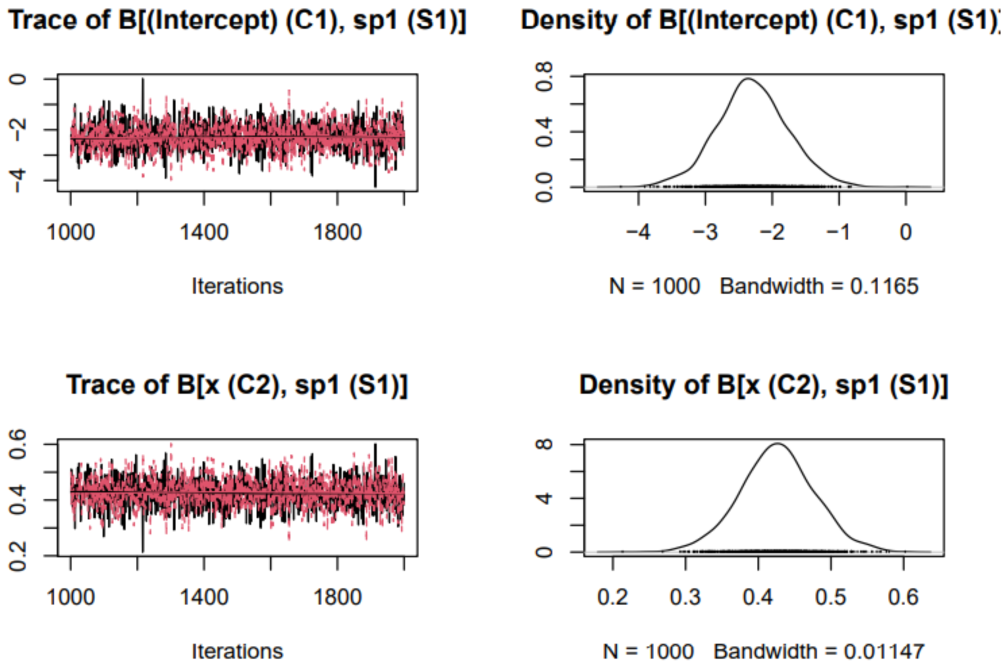
Prior and posterior distributions

- The prior distribution reflects our belief in different values of a given parameter
- After model fitting (posterior sampling), the posterior distribution will (hopefully!) be closed to reality than is the prior



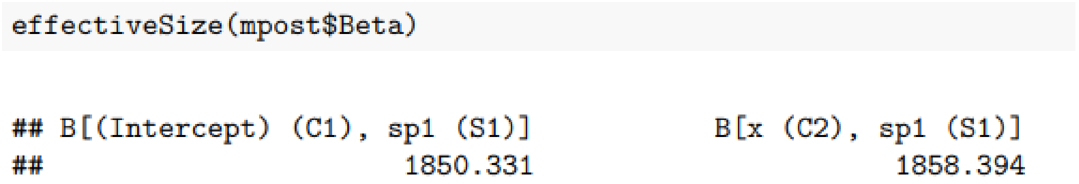
Evaluating model convergence

- After MCMC posterior sampling, we have to evaluate that the sampler has reached a stable posterior distribution



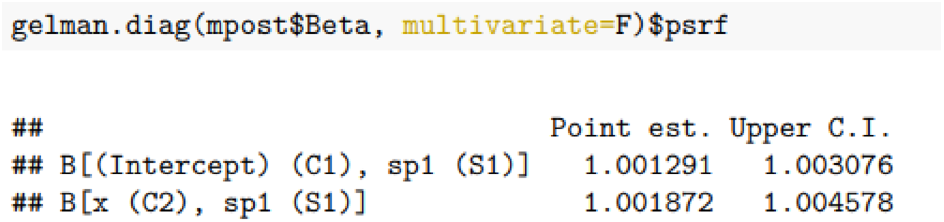
Evaluating model convergence

- We can evaluate mixing and convergence quantitatively by computing **effective sample sizes** (posterior simples accounting for autocorrelation)
- With multiple replicate chains, we can also assess how well the results correspond by **potential scale reduction factors** (psrf)



```
effectiveSize(mpost$Beta)
```

```
## B[(Intercept) (C1), sp1 (S1)]          B[x (C2), sp1 (S1)]
##                      1850.331                     1858.394
```



```
gelman.diag(mpost$Beta, multivariate=F)$psrf
```

```
##                               Point est. Upper C.I.
## B[(Intercept) (C1), sp1 (S1)]   1.001291   1.003076
## B[x (C2), sp1 (S1)]             1.001872   1.004578
```

What is «effect size»

- Wikipedia: an **effect size** is a value measuring the strength of the relationship between two variables in a population, or a sample-based estimate of that quantity
- Which measure is relevant depends on the (theoretical) context of the study, and the kind of comparison (e.g. regression vs. ANOVA-style analyses)
- Example: In a study of natural selection, the relevant effect size is the selection gradient, which gives the % change in fitness per % change in a trait
- Example 2: In a study of pesticide effects on bees, the relevant effect size is the LD_{50}, the dose of the toxin that kills half of the bees on average

Effect size in standard deviation units

- Effects are quite often measured in units of standard deviations
- We have seen this in path analysis, where path coefficients measure the number of SDs change in the response per SD change in the predictor
- For categorical data, measured as the difference scaled by the (pooled) SD,

  Cohen's d, **Hedges' g**, etc.

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Effect sizes in percent

- In regression: elasticities (log-log slope)
- For categorical data: log ratios: log(a/b) = log(a) – log(b)
- Percent difference (≈ log-ratio)

What to report from a fitted model?

- In a table
  - All parameter estimates (including intercept). **With units!**
  - Standard errors (or CI if Bayesian model)
  - Sample size
  - r^2, if mixed model then marginal and conditional
  - If used, test statistic and p-value
  - If reporting several candidate models, delta AIC, (AIC weight)
- In the text
  - Quantification of effects for those variables relevant to the research question
  - Reference to the table/figure

No evidence that the orchid bee *Eulaema nigrita* is more abundant in more heterogeneous landscapes

Øystein H. Opedal

Introduction

Euglossine bees are important pollinators in the neotropics (Dressler 1982). Due to their diverse resource needs, including pollen, nectar, fragrances (males) and nesting resources (females), these bees have been proposed to have large home ranges incorporating multiple habitat types (Janzen 1982). This hypothesis yields the testable prediction that euglossine bees should be more frequently encountered in heterogeneous landscapes comprising variable habitat types. To test this prediction, we analyse data from baiting surveys of male *Eulaema nigrita* from the Brazilian Atlantic forest, a highly fragmented region. Specifically, we ask whether *El. nigrita* is more frequent in more heterogeneous landscapes after accounting for climatic factors.

Methods

The data comprise bee abundances and sampling effort for 178 baiting surveys conducted at 72 sampling sites. We also obtained climatic and land-use data from each sampling site. To test the hypothesis that *El. nigrita* are more frequent in more heterogeneous landscapes, we fitted a generalized linear model with negative binomial errors (account for overdispersion). To account for variation in sampling effort, we included sampling effort as a covariate. To account for climatic drivers unrelated to local land-use, we included annual precipitation and annual temperature as additional covariates. Thus, in R syntax, the model took the following form:

```
y ~ land-use heterogeneity + sampling effort + climatic variables
```

Results

The model linking bee abundance to land-use heterogeneity, sampling effort and climatic covariates (annual temperature, annual precipitation) explained 29.9% of the variance in bee abundances across the landscape (pseudo $r^2 = 0.299$, Table 1). Bee abundance did not depend detectably on local land-use heterogeneity (slope = 0.012 ± 0.26 log bees/*S*, Table 1, Figure 1) or annual temperature (slope = 0.005 ± 0.39 log bees/°C) but increased in drier parts of the study region (slope = -0.002 ± 0.0002 log bees/mm, Table 1, Figure 1). For example, the expected number of bees at a site with no land-use heterogeneity and an average annual precipitation of 1457 mm (SD = 409 mm) is 59 and would drop to 32 at a site with one standard deviation greater annual precipitation (1457 + 409 = 1866 mm).

Table 1. Parameter estimates from a generalized linear model with negative binomial errors. Land-use heterogeneity (*S*) is the Shannon diversity of land-use proportions surrounding each sampling site. Model is supported over a null model with only effort as a fixed effect (ΔAIC = 30.97). Pseudo r^2 = 29.9%.

Parameter	Estimate ± SE
Intercept (log bees)	4.08 ± 0.24
Land-use heterogeneity (*S*)	0.012 ± 0.26
Effort (log hours)	0.38 ± 0.06
Mean annual temperature (°C)	0.005 ± 0.039
Mean annual precipitation (mm)	-0.002 ± 0.0002

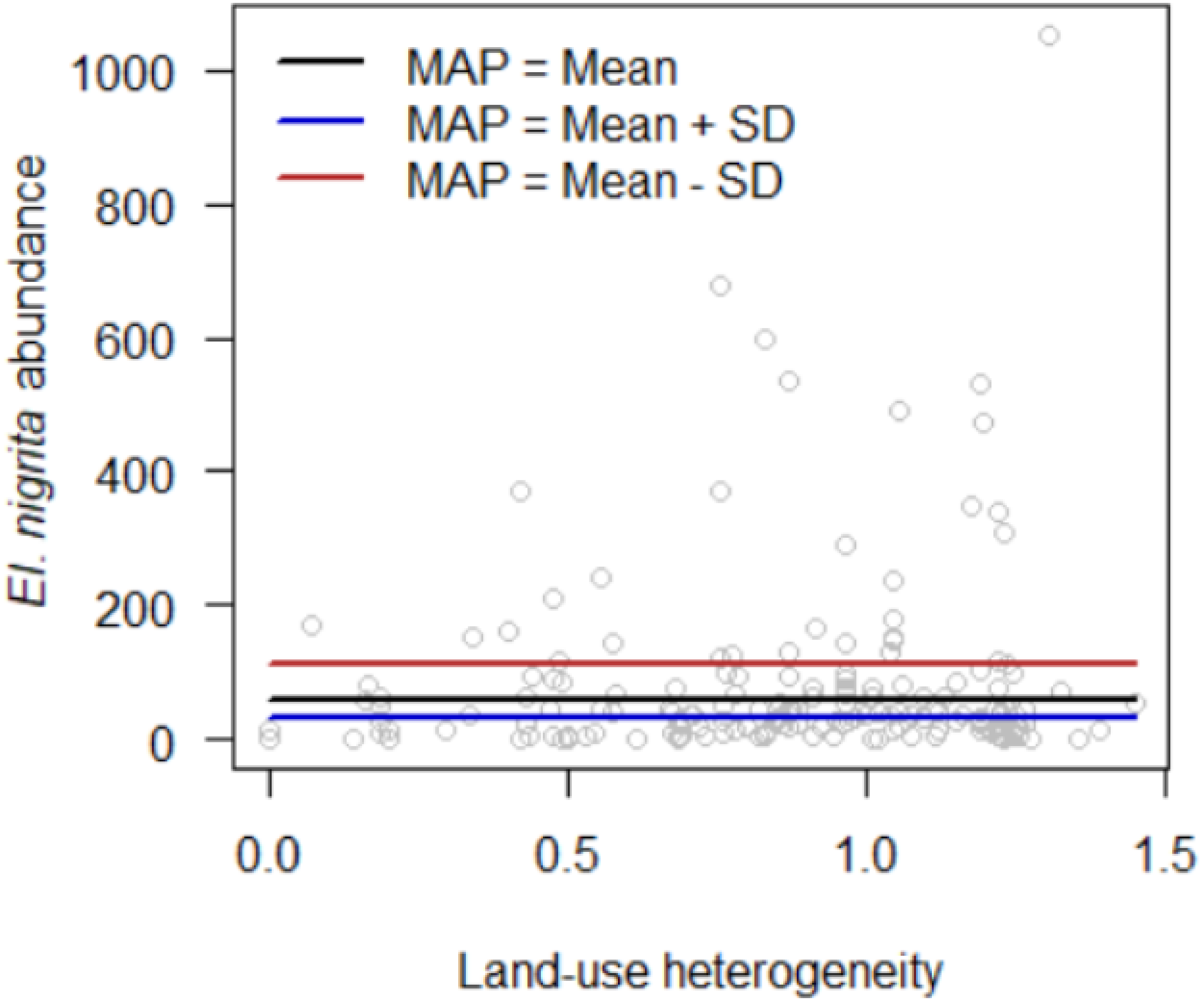
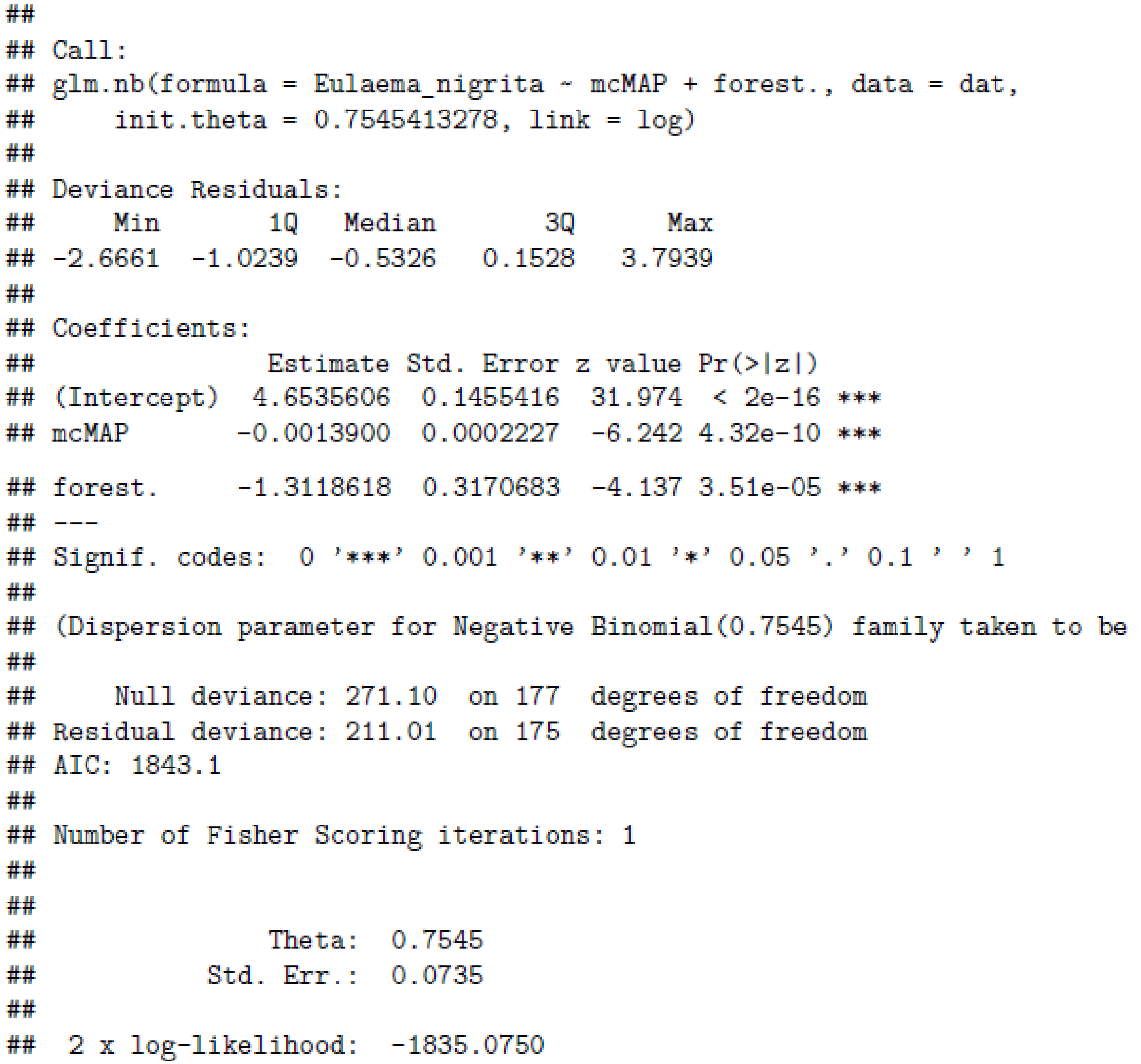


Figure 1. Joint effects of land-use heterogeneity (Shannon diversity of land-use classes) and mean annual precipitation on the number of *Eulaema nigrita* attracted to fragrance baits in the Brazilian Atlantic Forest.

Reporting/Interpreting GLMs

- Table with parameter estimates on the link scale
- For focal parameters, interpretation on the data scale. For example, calculating the predicted value for the mean of the response variable and some reasonable contrast, like +1 SD or +10%



```
##
## Call:
## glm.nb(formula = Eulaema_nigrita ~ mcMAP + forest., data = dat,
##     init.theta = 0.7545413278, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6661  -1.0239  -0.5326   0.1528   3.7939
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.6535606  0.1455416  31.974  < 2e-16 ***
## mcMAP       -0.0013900  0.0002227  -6.242 4.32e-10 ***
## forest.     -1.3118618  0.3170683  -4.137 3.51e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for Negative Binomial(0.7545) family taken to be
##
##     Null deviance: 271.10  on 177  degrees of freedom
## Residual deviance: 211.01  on 175  degrees of freedom
## AIC: 1843.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.7545
##           Std. Err.:  0.0735
##
##  2 x log-likelihood:  -1835.0750
```

Reporting/Interpreting GLMs

- 1. Calculate predicted value for the mean of the predictor (using the model linear predictor and the inverse link function
- 2. Calculate a second predicted value, e.g. for max of predictor
- 3. Interpret biologically

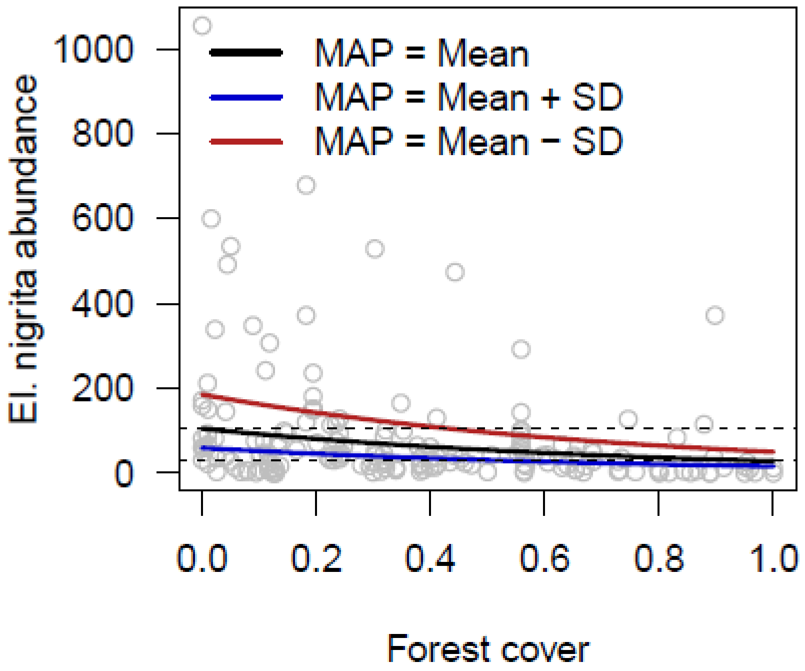


Figure reminders

- (Usually) no headers
- Axis labels sufficiently large
- Explain “everything” in legend (First sentence can be a “title”)

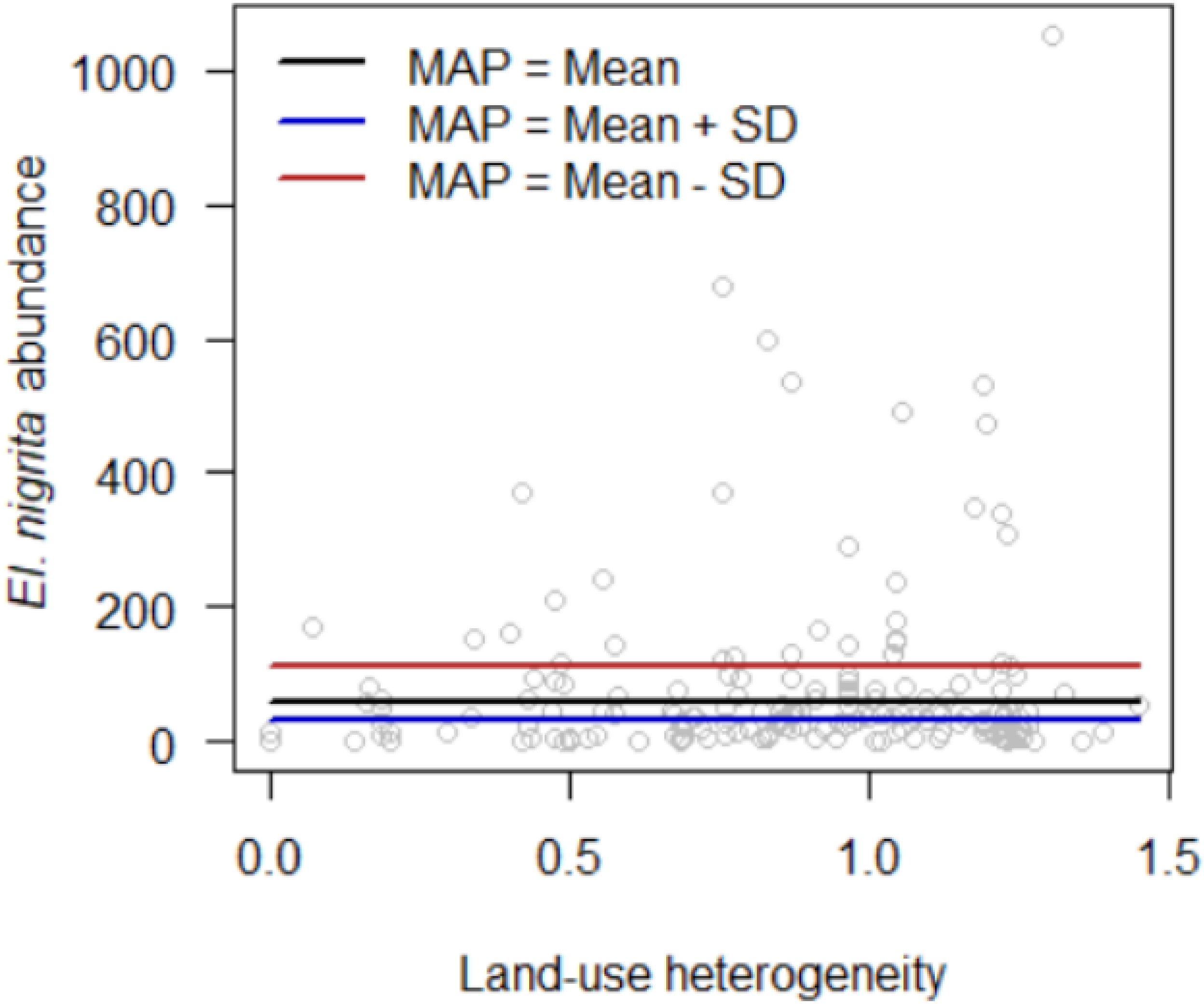


Figure 1. Joint effects of land-use heterogeneity (Shannon diversity of land-use classes) and mean annual precipitation on the number of *Eulaema nigrita* attracted to fragrance baits in the Brazilian Atlantic Forest.