How to jump pitfalls in the statistical analysis of field effect studies

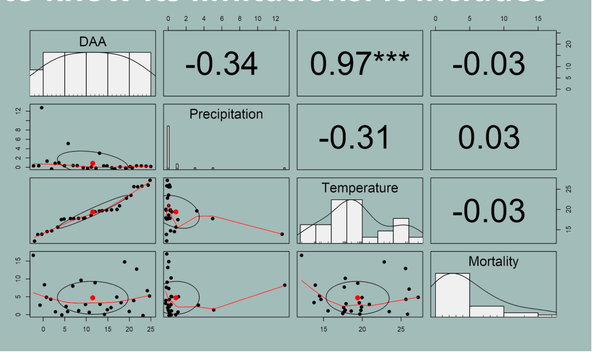
INTRODUCTION

*Anja Ruß**
*Ines Hotopp**
*Olaf Fülling**

Field effect study data holds pitfalls that have to be taken into account in the analysis as they affect the choice of the statistical method. We want to draw attention to some of these pitfalls and show possible approaches on how to jump them. Especially, as the draft guidance on birds and mammals risk assessment using higher tier studies does not include approved methods for the analysis of field effect studies. However, there are several appropriate ways to analyse this kind of data statistically. The concept of Generalized Linear Mixed Models (GLMMs) is one option that is able to cope with many of the peculiarities of field effect study data mentioned here.

Data exploration

Data exploration is a crucial step in analysing any kind of data. It intends to familiarize oneself with the data and get to know its limitations. It includes the investigation of outliers, zero observations, and correlation between covariates.



Data types

Know what kind of data you are dealing with. The numbers itself characterize the underlying data type and distribution. Most data in ecology are Poisson-distributed count data. However, there are other data types such as proportions or ratios. Most statistical methods make assumptions about the data type they are used on and therefore, method and data type should fit together.

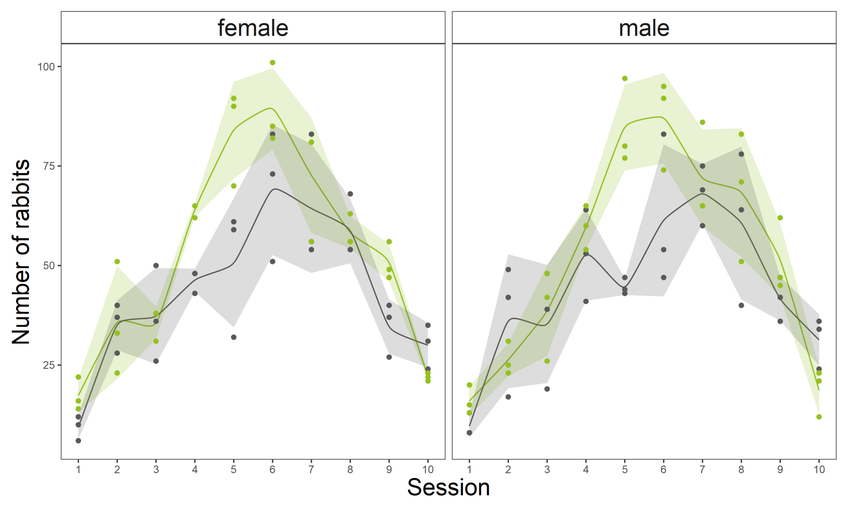
Count data: Positive integer numbers including 0, e.g. number of bees

Ratios: Positive floating point numbers, e.g. ratio of female to male voles

Proportions: Positive floating point numbers between 0 and 1, e.g. proportion of juvenile rabbits

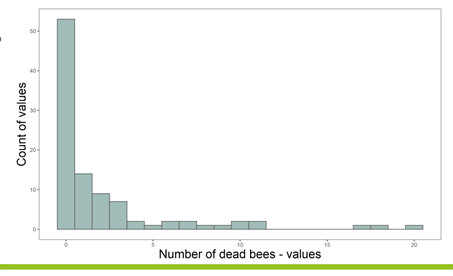
Explanatory variables

The inclusion of more than one explanatory variable can improve the fit of the model to the data, make other effects easier to detect, and allows to look at effects over time by including interactions between variables. However, explanatory variables should not be correlated and overfitting should be avoided.



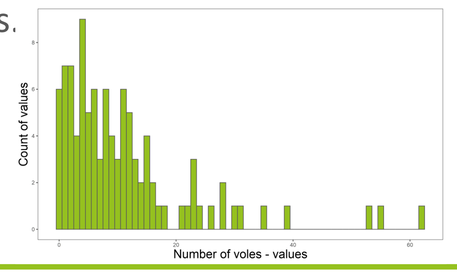
Zero-inflation

- There are more zeros than anything else in the data.
- Specialized GLMMs can deal with this.



Overdispersion

- The data distribution is wider than the data type suggests.
- Specialized GLMMs can deal with this.



Replication and Pseudoreplication

In field effect studies, the data is usually replicated. If a replicate is observed or measured more than once, the data is dependent. Some statistical methods assume independence of the data and are thus not suitable in these cases. GLMMs are made to cope with dependence using the concept of 'random effects'.

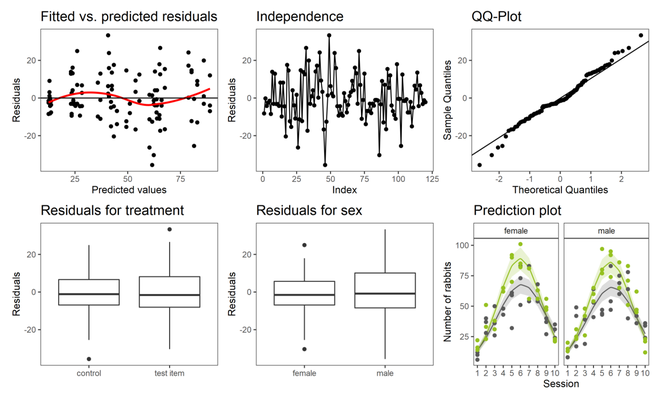
Examples are multiple sessions in a study with small mammals or multiple bee hives located at one field in a honey bee study.

Power and MDDs

- Ideally, a power analysis is conducted a priori
- If not feasible, reporting minimum detectable differences (MDD) and relative MDDs (MDD%) aids the interpretation of study results.
- Evaluation and comparison with other studies should not be based on these MDDs.

Model validation

Validation of fitted models is important to verify that assumptions, such as independence and absence of residual patterns, are not violated, and a prediction plot should accompany the statistical results.



CONCLUSIONS

Ecological data contains many obstacles. Field effect study data shows an increased complexity compared to laboratory data. There are multiple ways to jump the pitfalls adequately, one of them being Generalized Linear Mixed Models. If applied correctly, GLMMs are a powerful tool that are able to handle different data types, dependent data, pseudoreplication, zero-inflation, overdispersion and many more.

* tier3 solutions GmbH,
Kolberger Str. 61-63,
Leverkusen, Germany,
e-mail: anja.russ@tier3.de