TITLE1

TITLE2

by

EDUARD SZÖCS
from ZĂRNESTI / ROMANIA

Submitted Dissertation thesis for the partial fulfillment of the requirements
for a
Doctor of Natural Sciences
Fachbereich 7: Natur- und Umweltwissenschaften
Universität Koblenz-Landau

August 22, 2016

ACKNOWLEDGMENTS

ABSTRACT

CONTRIBUTIONS

This cumulative dissertation includes four scientific publications:

1. one
2. two
3. three
4. four

This dissertation is based on publications written by multiple authors. Therefore, the first person plural is used throughout the thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

2

ECOTOXICOLOGY IS NOT NORMAL - A COMPARISON OF STATISTICAL APPROACHES FOR ANALYSIS OF COUNT AND PROPORTION DATA IN ECOTOXICOLOGY

Eduard Szöcs[a] & Ralf B. Schäfer[a]

[a]Institute for Environmental Sciences, University Koblenz-Landau, 76829 Landau, Germany

Adapted from the article published 2015 in **Environmental Science and Pollution Research**, 22(18), 13990-13999

2.1 ABSTRACT

Ecotoxicologists often encounter count and proportion data that are rarely normally distributed. To meet the assumptions of the linear model such data are usually transformed or non-parametric methods are used if the transformed data still violate the assumptions. Generalised Linear Models (GLM) allow to directly model such data, without the need for transformation. Here, we compare the performance of two parametric methods, i.e., (1) the linear model (assuming normality of transformed data), (2) GLMs (assuming a Poisson, negative binomial, or binomially distibuted response), and (3) non-parametric methods.

We simulated typical data mimicking low replicated ecotoxicological experiments of two common data types (counts and proportions from counts). We compared the performance of the different methods in terms of statistical power and Type I error for detecting a general treatment effect and determining the lowest observed effect concentration (LOEC). In addition, we outlined differences on a real world mesocosm data set.

For count data, we found that the quasi-Poisson model yielded the highest power. The negative binomial GLM resulted in increased Type I errors, which could be fixed using the parametric bootstrap. For proportions, binomial GLMs performed better than the linear model, except to determine LOEC at extremely low sample sizes. The compared non-parametric methods had generally lower power.

We recommend that counts in one-factorial experiments should be analysed using quasi-Poisson models and proportions from counts by binomial GLMs. These methods should become standard in ecotoxicology.

2.2 INTRODUCTION

Ecotoxicologists perform various kinds of experiments yielding different types of data. Examples are animal counts in mesocosm experiments (non-negative, integer-valued data) or proportions of surviving animals (data bounded between 0 and 1, discrete). These data are typically not normally distributed. Nevertheless, such data are often analysed using methods that assume a normal distribution and variance homogeneity (Wang and Riffel, 2011). To meet these assumptions data are usually transformed. For example, ecotoxicological textbooks (Newman, 2012) and guidelines (EPA, 2002; OECD, 2006) advise that survival data should be transformed using an arcsine square root transformation. For count data from mesocosm experiments a log(Ay + C) transformation is usually applied, where the constants A and C are either chosen arbitrarily or following general recommendations. For example, Brink et al. (2000) suggest to set the term Ay to be 2 for

the lowest abundance value (y) greater than zero and C to 1. Other transformations, like the square root or fourth root transformation, are also commonly applied in community ecology Anderson et al., 2011. Note that there has been little evaluation and advice for practitioners which transformations to use. If the transformed data still do not meet the assumptions of the linear model, non-parametric tests are usually applied Wang and Riffel, 2011.

Generalised linear models (GLM) provide a method to analyse counts or proportions from counts in a statistically sound way Nelder and Wedderburn, 1972. GLMs can handle various types of data distributions, e.g., Poisson or negative binomial (for count data) or binomial (for proportions); the normal distribution being a special case of GLMs. Despite GLMs being available for more than 40 years, ecotoxicologists do not regularly make use of them. Recent studies concluded that the linear model should not be applied on transformed data and GLMs be used as they have better statistical properties (O'Hara and Kotze 2010; Warton 2005 (counts), Warton and Hui 2011 (proportions from counts)).

Ecotoxicological experiments often involve small sample sizes due to practical constraints. For example, extremely low samples sizes ($n < 5$) are common in many mesocosm studies Sanderson, 2002; Szöcs et al., 2015. Small sample sizes lead to low power in statistical hypothesis testing, on which many ecotoxiological approaches (e.g. risk assessment for pesticides) rely. Such an endpoint are L/NOEC values (Lowest / No observed effect concentration). Although their use has been heavily criticized in the past Laskowski, 1995, they are the predominant endpoint in mesocosm experiments Brock et al., 2015; EFSA PPR, 2013.

We explore how GLMs may enhance, when appropriately used, inference in ecotoxicological studies and compared three types of statistical methods (linear model on transformed data, GLM, non-parametric tests). We first illustrate differences between statistical methods using a data set from a mesocosm study. Then we further elaborate differences in detecting a general treatment effect and determining the LOEC using simulations of two common data types in ecotoxicology: counts and proportions from counts.

BIBLIOGRAPHY

Anderson, M. J. et al. (2011). "Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist." In: *Ecology Letters* 14.1, pp. 19–28.

Brink, P. J. van den et al. (2000). "Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions." In: *Aquatic Toxicology* 48.2-3, pp. 251–264.

Brock, T. C. M. et al. (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems." en. In: *Environmental Science and Pollution Research* 22.2, pp. 1160–1174. (Visited on 09/11/2014).

EFSA PPR (2013). "Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters." In: *EFSA Journal* 11.7, p. 3290.

EPA (2002). *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*. U.S. Environmental Protection Agency.

Laskowski, R. (1995). "Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology." In: *Oikos* 73.1, pp. 140–144.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models." In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384. (Visited on 08/27/2014).

Newman, Michael C (2012). *Quantitative ecotoxicology*. English. Boca Raton, FL: Taylor & Francis.

OECD (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application*. Series on Testing and Assessment 54. Paris: OECD.

O'Hara, Robert B. and D. Johan Kotze (2010). "Do not log-transform count data." In: *Methods in Ecology and Evolution* 1.2, pp. 118–122.

Sanderson, Hans (2002). "Pesticide studies." In: *Environmental Science and Pollution Research* 9.6, pp. 429–435. (Visited on 02/01/2014).

Szöcs, Eduard et al. (2015). "Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods." en. In: *Ecotoxicology* 24.4, pp. 760–769. (Visited on 04/17/2015).

Wang, M. and M. Riffel (2011). "Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology." In: *Ecotoxicology and Environmental Safety* 74.4, pp. 684–92.

Warton, David I. (2005). "Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data." In: *Environmetrics* 16.3, pp. 275–289.

Warton, David I. and Francis K. C. Hui (2011). "The arcsine is asinine: the analysis of proportions in ecology." In: *Ecology* 92.1, pp. 3–10.

3 DISCUSSION

AUTHOR'S CONTRIBUTIONS

ARTICLE I

TITLE: Ecotoxicology is not normal - A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology

AUTHORS: Eduard Szöcs and Ralf B. Schäfer

STATUS: Published in 2015 in Environmental Science and Pollution Research, Volume 22, Issue 18, pp 13990-13999

CONTRIBUTION: Szöcs (85%) Designed research, Simulated & Analysed data, Discussed results, Wrote manuscript
Schäfer (15%) Designed research, Discussed results, Edited manuscript

ARTICLE II

TITLE:

AUTHORS:

STATUS:

CONTRIBUTION:

ARTICLE III

TITLE:

AUTHORS: Eduard Szöcs and Ralf B. Schäfer

STATUS:

CONTRIBUTION: Szöcs (90%) Schäfer (10%)

ARTICLE IV

TITLE: taxize: taxonomic search and retrieval in R

AUTHORS: Scott A. Chamberlain and Eduard Szöcs

STATUS: Published in 2013 in F1000Research, Volume 2, Issue 191

CONTRIBUTION: Chamberlain (50%) Designed research, Programmed Software, Wrote manuscript
Szöcs (50%) Designed research, Programmed Software, Wrote manuscript

DECLARATION

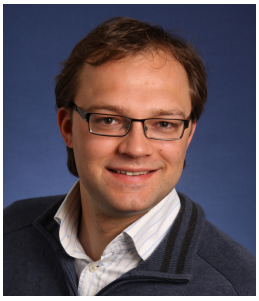
I, the author of this work, certify that this work contains no material which has been accepted or submitted for the award of any other degree at any university or other tertiary institution.

The work has been interdependently prepared. All aids and sources have been clearly specified and the contribution of other authors have been documented and reference lists given.

Neustadt a.d. Weinstraße,
August 22, 2016

Eduard Szöcs

CURRICULUM VITAE



Eduard Szöcs

Personal

Date of birth	16.06.1987
Nationality	german
Marital Status	single
Languages	German (native), English (very good), Romanian (good)

Education

04.2014–present **Ph.D. Environmental Sciences**, *University of Koblenz-Landau*, Landau.
Quantification of large scale effects of pesticides on freshwater ecosystems.

04.2012–03.2014 **M. Sc. Ecotoxicology**, *University of Koblenz-Landau*, Landau.
Thesis: Analysing mesocosm experiments: Principal Response Curves vs. Multivariate Generalized Linear Models.

11.2011 **B. Sc. Umweltwissenschaften**, *University of Koblenz-Landau*, Landau.
Thesis: Effects of salinity and pesticides on community structure of macroinvertebrates in Australian streams.

09.2007–11.2011 **Dipl. Umweltwissenschaften**, *University of Koblenz-Landau*, Landau.

Work Experience and Teaching

02.2016 – present **Research Assistant**, *University of Koblenz-Landau*, Landau.
Field Study in Romania, Data analyses, maintenance of databases and servers, PhD research.

11.2015 – present **Freelance Scientist & Consultant**.
Data sourcing, cleaning and analysis with specialization in Environmental & Ecological data. Courses in ecological statistics with the software *"R"*.

04.2015 – 01.2016 **Research Assistant**, *University of Koblenz-Landau*, Landau.
UBA Project: "PSM in Kleingewässern" (FKZ 3714674040/1). Building, maintaining and analysing a nation-wide german pesticide monitoring database.

05.2014 – 04.2015 **Research Assistant**, *University of Koblenz-Landau*, Landau.
Data analyses, maintenance of databases and servers, PhD research.

Marktplatz 6 – 67433 Neustadt an der Weinstraße
☎ *+49 176 621 927 00* • ✉ *eduardszoecs@posteo.de*
https://edild.github.io

12.2014 – 02.2015 **Teaching Assistant**, *University of Koblenz-Landau*, Landau.
Multivariate Statistics Course.

12.2013 – 02.2014 **Teaching Assistant**, *University of Koblenz-Landau*, Landau.
Multivariate Statistics Course.

12.2013 **Research Assistant**, *University of Koblenz-Landau*, Landau.
Development of a PostgreSQL-database of german physico-chemical data.

12.2012 – 02.2013 **Teaching Assistant**, *University of Koblenz-Landau*, Landau.
Multivariate Statistics Course.

12.2012 **Research Assistant**, *Department System Ecotoxicology, UFZ – Helmholtz Centre for Environmental Research*, Leipzig.
Development of rspear R-package.

05.2012 – 07.2012 **Internship**, *Department System Ecotoxicology, UFZ – Helmholtz Centre for Environmental Research*, Leipzig.
Field Study on the effects of pesticides on macroinvertebrates.

12.2011 – 02.2012 **Teaching Assistant**, *University of Koblenz-Landau*, Landau.
Multivariate Statistics Course.

07.2010 **Teaching Assistant**, *University of Koblenz-Landau*, Landau.
Aquatic Field Course.

06.2006 – 07.2007 **Internship**, *Landschaftspflegeverband Südpfalz e. V.*, Landau.
Freiwilliges Ökologisches Jahr.

Programming Skills

Expert R

Intermediate PostgreSQL, PostGIS, GrassGIS, LaTeX, git, regex, html, xml, xpath, cloud computing

Beginner Python, bash, NetLogo, C++, CDK, openbabel

(Beginner = *"I know the basics and can get the job done"*; Intermediate = *"I can effectively apply these tools"*; Expert = *"I can develop and expand these tools."*)

Software

I have developed or contributed to the following open source software for the R computing environment:

- The **taxize** package (together with Scott Chamberlain) allows taxonomic search, retrieval and handling in R.
- The **webchem** package to retrieve chemical information from the web.
- Contributions to the **vegan** package.
- The **rspear** package calculates $\text{SPEAR}_{\text{pesticides}}$ in R (deprecated).
- A web application to calculate statistical power for population endpoints in mesocosm experiments (currently offline / deprecated).
- Various other R packages and functions related to eco(toxico-)logy.

All software is freely available from my github account (`https://github.com/EDiLD`), homepage or CRAN.

Publications and Conference contributions

Articles

[1] Laurent Lagadic, Ralf B. Schäfer, Marc Roucaute, **Eduard Szöcs**, Sébastien Chouin, Jérôme de Maupeou, Claire Duchet, Evelyne Franquet, Benoit Le Hunsec, Céline Bertrand, and et al. "No association between the use of Bti for mosquito control and the dynamics of non-target aquatic invertebrates in French coastal and continental wetlands". In: *Science of The Total Environment* 553 (2016), 486–494.

[2] **Eduard Szöcs** and Ralf B Schäfer. "Statistical hypothesis testing—To transform or not to transform?" In: *Integrated Environmental Assessment and Management* 12.2 (2016), 398–400.

[3] **Eduard Szöcs** and Ralf B. Schäfer. "webchem: an R Package to retrieve Chemical Information from the Web". In: *Journal of Statistical Software* (under Review, 2016).

[4] John G. Mbaka, **Eduard Szöcs**, and Ralf B. Schäfer. "Meta-analysis on the responses of traits of different taxonomic groups to global and local stressors". In: *Acta Oecologica* 69 (2015), 65–70.

[5] **Eduard Szöcs**, Paul J. van den Brink, Laurent Lagadic, Thierry Caquet, Marc Roucaute, Arnaud Auber, Yannick Bayona, Matthias Liess, Peter Ebke, Alessio Ippolito, Cajo J. F. ter Braak, Theo C. M. Brock, and Ralf B. Schäfer. "Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods". In: *Ecotoxicology* 24.4 (2015), 760–769.

[6] **Eduard Szöcs** and Ralf B. Schäfer. "Ecotoxicology is not normal: A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology". In: *Environmental Science and Pollution Research* 22.18 (2015), 13990–13999.

[7] **Eduard Szöcs**, Eckhard Coring, Jürgen Bäthe, and Ralf B. Schäfer. "Effects of anthropogenic salinization on biological traits and community composition of stream macroinvertebrates". In: *Science of The Total Environment* 468–469 (2014), 943–949.

[8] Scott A. Chamberlain and **Eduard Szöcs**. "taxize: taxonomic search and retrieval in R [v2; ref status: indexed, http://f1000r.es/24v]". In: *F1000Research* 2.191 (2013). DOI: 10.12688/f1000research.2-191.v2.

[9] Ralf B. Schäfer, Mirco Bundschuh, Duncan A. Rouch, **Eduard Szöcs**, Peter C. von der Ohe, Vincent Pettigrove, Ralf Schulz, Dayanthi Nugegoda, and Ben J. Kefford. "Effects of pesticide toxicity, salinity and other environmental variables on selected ecosystem functions in streams and the relevance for ecosystem services". In: *Science of the Total Environment* 415.1 (2012), 69–78.

[10] **Eduard Szöcs**, Ben J. Kefford, and Ralf B. Schäfer. "Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates?" In: *Science of the Total Environment* 437.1 (2012), 121–126.

Poster

[1] **Eduard Szöcs** and Ralf B. Schäfer. "Ecotoxicology is not normal". SETAC Europe; Barcelona, 2015.

[2] **Eduard Szöcs** and Scott A. Chamberlain. "taxize: taxonomic search and retrieval in R". International Statistical Ecology Conference; Montpellier, 2014.

[3] **Eduard Szöcs**, Ben J. Kefford, V. Pettigrove, and Ralf B. Schäfer. "Einfluss von Pestiziden und Salinität auf Makroinvertebratengemeinschaften". SETAC GLB; Landau, 2011.

As a service to the scientific community I performed a total of 4 reviews for the journals *Proceedings of the Royal Society B, PhytoKeys, Zookeys* and *Environmental Toxicology and Chemistry*.

Workshops held

07.2015 **Data analysis in freshwater ecology using R**, *9th Symposium for European Freshwater Sciences*, Geneva.
Workshop held together with Dr. Ralf B. Schäfer and Avit Kumar Bhowmik.
Workshop homepage: `https://github.com/EDiLD/sefs9_Rworkshop`

11.2015 **Data Visualization with ggplot2**, *Workshop held at Young Academics Conference 2015 - Land-Water-Interactions*, Klingenmünster.
Workshop homepage: `https://github.com/EDiLD/r_landau_2015`

Neustadt a.d. Weinstraße, July 20, 2016