STATISTICAL ECO(-TOXICO)LOGY

IMPROVING THE UTILIZATION OF DATA FOR ECOLOGICAL RISK ASSESSMENT

by

EDUARD SZÖCS
from ZĂRNEŞTI / 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

15. November 2016

ACKNOWLEDGMENTS

I thank all the persons that supported me during my studies and this dissertation.
My special thanks go to my supervisor Prof. Dr. Ralf B. Schäfer for his support throughout the last six years. I am thankful for his openness to my ideas and the opportunities given to follow them, for organizing funding throughout this dissertation, for pushing me to sound scientific writing and critical reading, for challenging discussions, not only on statistical eco(-toxico)logy but also outside of the subject area.
Many thanks to Prof. Dr. Ralf Schulz for examining this thesis and his influence on me during my undergrad studies. Moreover, I thank Prof. Dr. Engelbert Niehaus for chairing the dissertation committee.
Without the continuous support of my parents, Anca and Helmut, this thesis would not have been possible - Thank you!
I am grateful to my colleagues, students and other people for asking me tough statistical questions. These questions from a broad range of fields and finding solutions to them widened my expertise in the field.
Special thanks go to Phillip Uhl, Gunnar Oehmichen and Michael Burstert for the discussions during coffee breaks and proofreading this thesis.
I thank all collaborators of projects involved in this thesis, but also past projects. All of you provided help, critical comments and enlightening discussion on my work.
I tried to make this thesis as open and reproducible as possible. I would thank the people developing, maintaining and bug fixing the open software I used throughout this thesis. I thank GitHub for providing me a discount the last three years and a platform for collaboration and version control that was crucial for big parts of this thesis.
Lastly, I cannot thank enough my girlfriend, Anja Loescher, for getting through this stressful time with me. Despite a stressful phase with her profession, she always supported, encouraged and loved me.

SUMMARY

PUBLICATIONS

This cumulative dissertation includes four scientific publications:

1. E. Szöcs and R. B. Schäfer (2015). "Ecotoxicology is not normal". *Environmental Science and Pollution Research* 22 (18), 13990–13999

2. E. Szöcs, M. Brinke, B. Karaoglan, and R. B. Schäfer (2016). "Large scale risks from pesticides in small streams". *Environmental Science & Technology*. submitted.

3. E. Szöcs and R. B. Schäfer (2016). "webchem: An R Package to Retrieve Chemical Information from the Web". *Journal of Statistical Software*. accepted.

4. S. A. Chamberlain and E. Szöcs (2013). "taxize: taxonomic search and retrieval in R". *F1000Research* 2 (191)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION AND OBJECTIVES

THREATS TO FRESHWATER ECOSYSTEMS FROM CHEMICAL POLLUTION

Freshwater ecosystems, like streams, lakes and wetlands, make up only 0.01% of the World's water and cover only 0.8% of Earth's surface (Dudgeon et al., 2006), yet they host an important component of global biodiversity. Freshwaters are a habitat for more than 125,000 species, which represents 10% of global biodiversity and ⅓ of all vertebrate species (Balian et al., 2007; Strayer and Dudgeon, 2010) and provide essential services for human well-being (Aylward et al., 2005). Small water bodies are of particular importance, because of their high abundance (Downing et al., 2012), the high biodiversity they host (Davies et al., 2008) and the ecosystem services they provide (Biggs et al., 2016).

The earth is currently experiencing a functional change driven by human activities which are so far-reaching, that a new geological epoch "Anthropocene" has been proposed (Waters et al., 2016). Consequently, these changes are also associated with biotic changes: 65% of rivers are currently at threat (Vörösmarty et al., 2010) and freshwaters are experiencing the greatest losses of biodiversity (WWF, 2016). A multitude of stressors contribute to this deterioration of freshwater biodiversity including habitat loss and degradation, overexploitation, invasive species and pollution (Dudgeon et al., 2006; Vörösmarty et al., 2010; WWF, 2016). Studies investigating water pollution have mainly focused on nutrient loading, acidification and pollution by organic loading (Schäfer et al., 2016). However, chemicals have become ubiquitous throughout humankind. Currently, more than 100,000 chemicals are registered and in daily use (Schwarzenbach et al., 2010; Schwarzman and Wilson, 2009). These substances will ultimately end somewhere in the environment.

Despite their potential negative effects on biota and humans and their intentional release, pesticides have been neglected in the past by ecological studies investigating threats to freshwaters (Schäfer et al., 2016) and it is unknown how much they contribute to biodiversity loss (Persson et al., 2013; Rockström et al., 2009). However, recent studies indicated that pollution by pesticides may

be a frequent threat to freshwaters that might have been neglected by ecological studies in the past. Malaj et al., (2014) showed that almost half of European water bodies are at risk from pesticides. In the United States, Stone et al., (2014) showed that 61% of assessed agricultural streams exceed aquatic-life benchmarks. On a global scale, Stehle and Schulz, (2015) found that 52.4% of detected insecticide concentrations (n = 11,300) exceeded risk thresholds. The high contact with adjacent land and low water volume of small streams make them particularly vulnerable to pesticide pollution (Biggs et al., 2016), however, there is currently a lack of data on pesticide pollution of small streams (Lorenz et al., 2016).

As a reaction to the degradation of freshwaters, several legal frameworks have been established to safeguard and improve the quality of freshwater ecosystems. In the European Union (EU), the Water Framework Directive (WFD) (European Union, 2000) regulates the protection of aquatic ecosystems and commits the member states to achieve a 'good' status of all water bodies. Knowing of the toxicity of pesticides and their intentional release into the environment, also the introduction and use of new pesticides are highly regulated. Sophisticated environmental risk assessment procedures have been developed and are requested by the EU (European Union, 2009) to ensure that the use of pesticides does not cause unacceptable effects to non-target organism, soil, air and water.

ENVIRONMENTAL RISK ASSESSMENT

Environmental risk assessment (ERA) tries to estimate risks to animals, populations or ecosystems. It investigates if a chemical can be used as intended without causing detrimental impacts to the environment. Moreover, ERA is used as a tool to support decision making under uncertainty (Newman, 2015). Environmental risk is defined as a combination of the severity and the probability of occurrence of a potential adverse effect on the environment (Suter, 2007). Therefore, ERA is based on two components: Effect- and exposure assessment. A combination of both is needed to characterise environmental risks.

Effect assessment characterises the strength of effects using laboratory and semi-field experiments. It establishes relationships between the concentration of a compound and the observed effects. In the European Union a tiered approach with increasing complexity and realism. Lower tier assessment is based on highly standardised single species laboratory experiments, whereas higher tier assessment is refined by testing additional species, extended laboratory ex-

periments or model ecosystem experiments. To address the various uncertainties in effect assessment (e.g. experimental variation, variation between species, variation in environmental conditions etc.) the retrieved toxicity values are multiplied by an assessment factor between 0.01 (lower tier assessment) and 0.5 (higher tier assessment) depending on data quality, which yields to a regulatory acceptable concentration (RAC) (EFSA, 2013).

Exposure Assessment for freshwaters aims to characterise the probability of an adverse effect by deriving a predicted environmental concentration (PEC) in surface waters and sediments (Newman, 2015). It is mainly based on modelling the fate of chemicals in the environment using computer simulations. In the European Union, the FOCUS models are used (EFSA, 2013; FOCUS, 2001). To calculate PECs these models need many compound specific input parameters like the molecular weight, water solubility, partitioning coefficients and dissipation time. Additionally, information on the application regime and crop type is needed. FOCUS models the concentration within edge-of-field streams of 1 meter width (corresponding a catchment size of approx. 7km², see Figure B.11) and 30 cm depth (Erlacher and Wang, 2011). Nevertheless, recent research showed that FOCUS models fail to predict measured field concentrations of pesticides (Knäbel et al., 2014; Knäbel et al., 2012).

The final step in ERA is risk characterisation. It puts together the information gained from effect and exposure assessment. Risk can be expressed in several ways, a quantitative way being the risk quotient approach: A PEC / RAC ratio greater than one indicating potential risks (Amiard-Triquet, 2015; EFSA, 2013; Suter, 2007). Consequently, pesticides can be authorised only if the risk quotient is below one indicating that harmful effects are unlikely.

ENVIRONMENTAL MONITORING

Widespread anthropogenic activities and the induced environmental changes have resulted in concerns about the state of the environment and have led to the development of environmental monitoring programs worldwide (Nichols and Williams, 2006). After authorization, pesticides applied on agricultural fields may enter aquatic ecosystems via diffuse sources like spray-drift, surface run-off or drainage (Liess et al., 1999; Schulz, 2004; Stehle et al., 2013). These entered pesticides may have ecological effects and worsen the chemical status, acting contrary to the goal of the WFD. For monitoring the progress towards the goal of a 'good' status and for assessment of the chemical status of surface

waters the EU WFD established monitoring requirements for all European river basins (European Union, 2000). For chemical monitoring the WFD requires grab sampling and chemical analysis of 21 priority substances (of which 7 are pesticides) every third month and of 24 other pollutants (of which 12 are used as pesticides) every month (European Union, 2013). Additionally, 14 substances (of which 8 are used as pesticides, including all Neonicotinoids) that may pose a significant risk, have an insufficient data basis and are candidates for future priority substances are currently monitored until 2019 (European Union, 2015). Nevertheless, monitoring programs on a national scale might monitor a broader spectrum of chemical substances, e.g. for investigative monitoring. Recent studies indicate that the current sampling and chemical analyses strategy greatly underestimate the pesticide exposure (Moschet et al., 2014; Stehle et al., 2013; Xing et al., 2013).

Environmental monitoring produces humongous amounts of data containing information on pesticide concentrations in the field on a large under many conditions. Therefore, it can be complementary to environmental risk assessment (Suter, 2007). Moreover, data from long-term monitoring programs can be used to study hypotheses about spatial and temporal dynamics and interactions, that are not evident from short term and short scale studies (Gitzen, 2012) and provide insights modelling approaches. If the environmental risk assessment process captured all relevant sources of risk, no concentrations above the derived RAC should be observable in European rivers. Therefore, monitoring data could be used to provide feedback for ERA after approval (Knauer, 2016). However, at present little is known on pesticide concentrations in small streams comparable to those assessed in ERA (Biggs et al., 2016; Lorenz et al., 2016). Monitoring under the WFD is also performed for biological components of freshwaters and a combination with pesticide exposure data might provide valuable insights into large-scale field effects (Schipper et al., 2014).

STATISTICAL ECOTOXICOLOGY

Environmental effect assessment generates data on ecological effects using experiments. The produced datasets range from small univariate datasets (lower tier assessment) to medium sized multivariate datasets (higher tier assessment). In order to extract usable information for assessment, these datasets are analysed using statistical techniques and therefore, statistics are crucial for effect assessment (Newman, 2012). Statistical ecotoxicology combines statistics with

the specific needs and constraints of ecotoxicology. Ecotoxicologists deal generally with low replicated experiments, making statistical inference difficult (Van Der Hoeven, 1998). For example, a recent analysis of eleven mesocosm studies revealed that the sample sizes for these kind of experiments range between two and five. Statistical ecotoxicology aims to provide solutions to statistical challenges in ecotoxicology (Fox and Landis, 2016a), guidance on experimental designs (Johnson et al., 2015) and tools to integrate big data (Van den Brink et al., 2016). The ultimate goal is to improve the accuracy of ERA.

The relationships between the concentration of a compound and the observed effects are usually analysed using dose-response models, which can be used to derive an effective concentration for x% effect (EC_x) (Ritz, 2010). Nevertheless, such relationships cannot always be established from experimental data. For example, mesocosm experiments are conducted to characterise effects on whole biological communities. However, because of multivariate responses and potential indirect effects, there is no clear dose-response relationship and no models for this kind of data available. There are also examples were fitting dose-response models is problematic (Green, 2016). In such cases, there is usually a no-observed-effect concentration (NOEC) computed.

The NOEC is the highest tested concentration that does not lead to significant deviation from the control response and therefore relies on null hypothesis significance testing (NHST). However, the use of NOEC as a toxicity measure in environmental effect assessment has been heavily criticised in the past (Chapman et al., 1996; Fox et al., 2012; Fox and Landis, 2016b; Jager, 2012; Laskowski, 1995; Warne and van Dam, 2008). One such critic is the low statistical power for NHST in common ecotoxicological experiments (Van Der Hoeven, 1998). *A priori* power calculations can provide useful guidance for choosing experimental designs (Johnson et al., 2015), but are rarely used by ecotoxicologists (Newman, 2008).

Instead of conducting experiments, toxicity could be also predicted from molecular structures using quantitative structure-activity relationships (QSAR), which are usually calculated using machine-learning techniques (Cortes-Ciriano, 2016; Murrell et al., 2015). Nevertheless, in order to improve these models to give sufficient prediction accuracy more data from experiments is needed (Kühne et al., 2013). Indeed, a large amount of data is available that could be used for effect and exposure assessment. For example, the US EPA ECOTOX database (U.S. EPA, 2016), the Pesticides Properties Database (Lewis et al., 2016) and ETOX (Umweltbundesamt, 2016) provide toxicity data that could be used

for effect assessment. Databases like Physprop (Howard and Meylan, 2016) and PubChem (Kim et al., 2016) provide chemical properties that are needed as input for exposure models. Monitoring data provides information on realised concentrations, could be used for validation of models and retrospective risk assessment. This "big data" can provide new information and opportunities for ERA (Dafforn et al., 2015). However, it needs to be harmonised, linked and easily accessible in order to be used effectively in ERA.

OBJECTIVES AND OUTLINE OF THE THESIS

The overall goal of this thesis was to contribute to the emerging field of statistical ecotoxicology, environmental risk assessment and environmental monitoring. The main objectives were (i) to scrutinise new methods in statistical ecotoxicology, (ii) explore available monitoring data and (iii) provide tools to deal with big data. Figure 1.1 provides a conceptual overview on ERA and environmental monitoring as outlined in the previous sections, as well as the parts considered in this thesis and the relations between them.

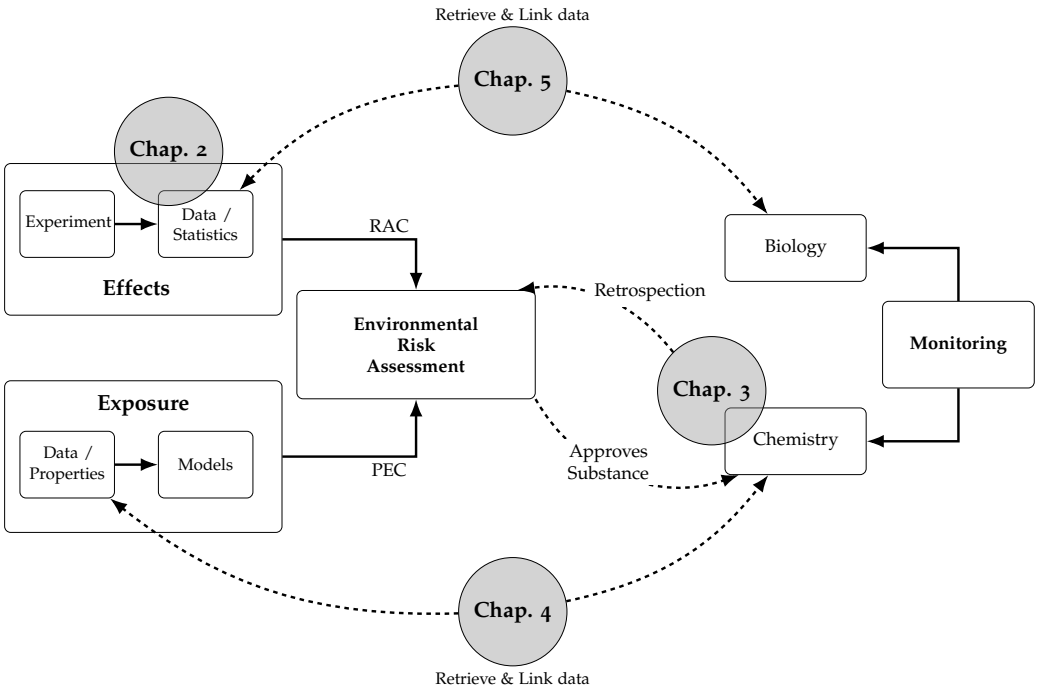


Figure 1.1.: Conceptual overview on environmental risk assessment, environmental monitoring and the parts addressed by this thesis.

The thesis starts with a comparison of statistical methods to analyse ecotoxicological experiments using NHST in effect assessment (Chapter 2). Specific questions addressed were:

- Are newer statistical methods, explicitly considering the type of analysed data, more powerful than currently used methods for NHST?

- How much statistical power do current experimental designs in ecotoxicology exhibit?

Exposure assessment aims at predicting chemical concentrations in small streams. Chapter 3 focuses on measured large-scale environmental concentrations in small streams and the drivers thereof. Specific goals of this study were:

- Compile monitoring data on pesticides in small streams in Germany and check if the available data is suitable to inform ERA.
- Explore the relationship between agricultural land use and stream size and RAC exceedances.
- Scrutinise the annual dynamics of pesticide exposure, as well as the influence of precipitation on measured pesticide concentrations.
- We use RACs derived from ERA to assess the current pollution in German streams and identify pesticides exhibiting currently a risk to freshwaters.

The compilation of monitoring data from different data sources in Chapter 3, resulted in a big inhomogeneous amount of data. Moreover, Biologists, Chemists and ecotoxicologists face similar problems with the need to identify and harmonise their biological and chemical data. Chapters 4 (chemical data) and 5 (biological data) describe software solutions to simplify and accelerate the workflow of:

- validating and harmonising chemical and taxonomic data
- linking datasets from different databases
- retrieving properties and identifiers

REFERENCES

Amiard-Triquet, C. (2015). *Aquatic ecotoxicology: advancing tools for dealing with emerging risks*. Boston, MA: Elsevier.

Aylward, B., J. Bandyopadhyay, J.-C. Belausteguigotia, P. Borkey, A. Z. Cassar, L. Meadors, L. Saade, M. Siebentritt, R. Stein, S. Tognetti, et al. (2005). "Freshwater ecosystem services". *Ecosystems and human well-being: policy responses* 3, 213–256.

Balian, E. V., H. Segers, C. Lévèque, and K. Martens (2007). "The Freshwater Animal Diversity Assessment: an overview of the results". *Hydrobiologia* 595 (1), 627–637.

Biggs, J., S. von Fumetti, and M. Kelly-Quinn (2016). "The importance of small waterbodies for biodiversity and ecosystem services: implications for policy makers". *Hydrobiologia*.

Chapman, P., P. Chapman, and R. Caldwell (1996). "A warning: NOECs are inappropriate for regulatory use". *Environmental Toxicology and Chemistry* 15 (2), 77–79.

Cortes-Ciriano, I. (2016). "Bioalerts: a python library for the derivation of structural alerts from bioactivity and toxicity data sets". *Journal of Cheminformatics* 8 (1).

Dafforn, K. A., E. L. Johnston, A. Ferguson, C. Humphrey, W. Monk, S. J. Nichols, S. L. Simpson, M. G. Tulbure, and D. J. Baird (2015). "Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems." *Marine and Freshwater Research*.

Davies, B., J. Biggs, P. Williams, M. Whitfield, P. Nicolet, D. Sear, S. Bray, and S. Maund (2008). "Comparative biodiversity of aquatic habitats in the European agricultural landscape". *Agriculture, Ecosystems & Environment* 125 (1-4), 1–8.

Downing, J. A., J. J. Cole, C. A. Duarte, J. J. Middelburg, J. M. Melack, Y. T. Prairie, P. Kortelainen, R. G. Striegl, W. H. McDowell, and L. J. Tranvik (2012). "Global abundance and size distribution of streams and rivers". *Inland waters* 2 (4), 229–236.

Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. I. Kawabata, D. J. Knowler, C. Leveque, R. J. Naiman, A. H. Prieur-Richard, D. Soto, M. L. J. Stiassny, and C. A. Sullivan (2006). "Freshwater biodiversity: importance, threats, status and conservation challenges". *Biological Reviews* 81 (2), 163–182.

EFSA (2013). "Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters". *EFSA Journal* 11 (7), 3290.

Erlacher, E. and M. Wang (2011). "Regulation (EC) No. 1107/2009 and upcoming challenges for exposure assessment of plant protection products – Harmonisation or national modelling approaches?" *Environmental Pollution* 159 (12), 3357–3363.

European Union (2000). "Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy". *Official Journal of the European Union* L 327, 1–73.

European Union (2009). "Regulation (EC) No 1107/2009 of the European Parliament and of the Council of 21 October 2009 concerning the placing of plant protection products on the market and repealing Council Directives 79/117/EEC and 91/414/EEC". *Official Journal of the European Union* L 309, 1–50.

European Union (2013). "Directive 2013/39/EU of the European Parliament and of the Council of 12 August 2013 amending Directives 2000/60/EC and 2008/105/EC as regards priority substances in the field of water policy". *Official Journal of the European Union* L226, 1–17.

European Union (2015). "Commission Implementing Decision (EU) 2015/495 of 20 March 2015 establishing a watch list of substances for Union-wide monitoring in the field of water policy pursuant to Directive 2008/105/EC of the European Parliament and of the Council (notified under document C(2015) 1756)". *Official Journal of the European Union* L28, 40–42.

FOCUS (2001). *FOCUS Surface Water Scenarios in the EU Evaluation Process under 91/414/EEC.* EC Document Reference SANCO/4802/2001-rev.2.

Fox, D. R., E. Billoir, S. Charles, M. L. Delignette-Muller, and C. Lopes (2012). "What to do with NOECS/NOELS—prohibition or innovation?" *Integrated Environmental Assessment and Management* 8 (4), 764–766.

Fox, D. R. and W. G. Landis (2016a). "Comment on ET&C perspectives, November 2015-A holistic view". *Environmental Toxicology and Chemistry* 35 (6), 1337–1339.

Fox, D. R. and W. G. Landis (2016b). "Don't be fooled-A no-observed-effect concentration is no substitute for a poor concentration-response experiment: NOEC and a poor concentration-response experiment". *Environmental Toxicology and Chemistry* 35 (9), 2141–2148.

Gitzen, R. A., ed. (2012). *Design and analysis of long-term ecological monitoring studies*. Cambridge ; New York: Cambridge University Press.

Green, J. W. (2016). "Issues with using only regression models for ecotoxicity studies". *Integrated Environmental Assessment and Management* 12 (1), 198–199.

Howard, P. H. and W. Meylan (2016). *Physical and Chemical Property Database*. URL: http://www.srcinc.com/what-we-do/environmental/scientific-databases.html.

Jager, T. (2012). "Bad habits die hard: The NOEC's persistence reflects poorly on ecotoxicology". *Environmental Toxicology and Chemistry* 31 (2), 228–229.

Johnson, P. C. D., S. J. E. Barry, H. M. Ferguson, and P. Müller (2015). "Power analysis for generalized linear mixed models in ecology and evolution". *Methods in Ecology and Evolution* 6 (2), 133–142.

Kim, S., P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant (2016). "PubChem Substance and Compound databases". *Nucleic Acids Research* 44 (D1), D1202–D1213.

Knäbel, A., K. Meyer, J. Rapp, and R. Schulz (2014). "Fungicide Field Concentrations Exceed FOCUS Surface Water Predictions: Urgent Need of Model Improvement". *Environmental Science & Technology* 48 (1), 455–463.

Knäbel, A., S. Stehle, R. B. Schäfer, and R. Schulz (2012). "Regulatory FOCUS Surface Water Models Fail to Predict Insecticide Concentrations in the Field". *Environmental Science & Technology* 46 (15), 8397–8404.

Knauer, K. (2016). "Pesticides in surface waters: a comparison with regulatory acceptable concentrations (RACs) determined in the authorization process and consideration for regulation". *Environmental Sciences Europe* 28 (13).

Kühne, R., R.-U. Ebert, P. C. von der Ohe, N. Ulrich, W. Brack, and G. Schüürmann (2013). "Read-Across Prediction of the Acute Toxicity of Organic Compounds toward the Water Flea Daphnia magna". *Molecular Informatics* 32 (1), 108–120.

Laskowski, R. (1995). "Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology". *Oikos* 73 (1), 140–144.

Lewis, K. A., J. Tzilivakis, D. J. Warner, and A. Green (2016). "An international database for pesticide risk assessments and management". *Human and Ecological Risk Assessment: An International Journal* 22 (4), 1050–1064.

Liess, M., R. Schulz, M.-D. Liess, B. Rother, and R. Kreuzig (1999). "Determination of insecticide contamination in agricultural headwater streams". *Water Research* 33 (1), 239–247.

Lorenz, S., J. J. Rasmussen, A. Süß, T. Kalettka, B. Golla, P. Horney, M. Stähler, B. Hommel, and R. B. Schäfer (2016). "Specifics and challenges of assessing exposure and effects of pesticides in small water bodies". *Hydrobiologia*, 1–12.

Malaj, E., P. C. v. d. Ohe, M. Grote, R. Kühne, C. P. Mondy, P. Usseglio-Polatera, W. Brack, and R. B. Schäfer (2014). "Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale". *Proceedings of the National Academy of Sciences* 111 (26), 9549–9554.

Moschet, C., I. Wittmer, J. Simovic, M. Junghans, A. Piazzoli, H. Singer, C. Stamm, C. Leu, and J. Hollender (2014). "How a Complete Pesticide Screening Changes the Assessment of Surface Water Quality". *Environmental Science & Technology* 48 (10), 5423–5432.

Murrell, D. S., I. Cortes-Ciriano, G. J. P. van Westen, I. P. Stott, A. Bender, T. E. Malliavin, and R. C. Glen (2015). "Chemically Aware Model Builder (camb):

an R package for property and bioactivity modelling of small molecules". *Journal of Cheminformatics* 7 (1).

Newman, M. C. (2008). ""What exactly are you inferring?" - A closer look at hypothesis testing". *Environmental Toxicology and Chemistry* 27 (7). Newman, M. C., 1633–1633.

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

Newman, M. C. (2015). *Fundamentals of ecotoxicology: the science of pollution*. Boca Raton: CRC Press, Taylor & Francis Group.

Nichols, J. and B. Williams (2006). "Monitoring for conservation". *Trends in Ecology & Evolution* 21 (12), 668–673.

Persson, L. M., M. Breitholtz, I. T. Cousins, C. A. de Wit, M. MacLeod, and M. S. McLachlan (2013). "Confronting unknown planetary boundary threats from chemical pollution". *Environmental science & technology* 47 (22), 12619–12622.

Ritz, C. (2010). "Toward a unified approach to dose-response modeling in ecotoxicology". *Environmental Toxicology and Chemistry* 29 (1), 220–229.

Rockström, J., W. Steffen, K. Noone, A. Persson, 3. Chapin F. S., E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sorlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. A. Foley (2009). "A safe operating space for humanity". *Nature* 461 (7263), 472–5.

Schäfer, R. B., B. Kühn, E. Malaj, A. König, and R. Gergs (2016). "Contribution of organic toxicants to multiple stress in river ecosystems". *Freshwater Biology*. DOI: 10.1111/fwb.12811.

Schipper, A. M., L. Posthuma, D. de Zwart, and M. A. J. Huijbregts (2014). "Deriving Field-Based Species Sensitivity Distributions (f-SSDs) from Stacked Species Distribution Models (S-SDMs)". *Environmental Science & Technology* 48 (24), 14464–14471.

Schulz, R. (2004). "Field Studies on Exposure, Effects, and Risk Mitigation of Aquatic Nonpoint-Source Insecticide Pollution: A Review". *Journal of Environmental Quality* 33 (2), 419–448.

Schwarzenbach, R. P., T. Egli, T. B. Hofstetter, U. v. Gunten, and B. Wehrli (2010). "Global Water Pollution and Human Health". *Annual Review of Environment and Resources* 35 (1), 109–136.

Schwarzman, M. R. and M. P. Wilson (2009). "New Science for Chemicals Policy". *Science* 326 (5956), 1065–1066.

Stehle, S., A. Knäbel, and R. Schulz (2013). "Probabilistic risk assessment of insecticide concentrations in agricultural surface waters: a critical appraisal". *Environmental Monitoring and Assessment* 185 (8), 6295–6310.

Stehle, S. and R. Schulz (2015). "Pesticide authorization in the EU—environment unprotected?" *Environmental Science and Pollution Research* 22 (24), 19632–19647.

Stone, W. W., R. J. Gilliom, and K. R. Ryberg (2014). "Pesticides in U.S. Streams and Rivers: Occurrence and Trends during 1992–2011". *Environmental Science & Technology* 48 (19), 11025–11030.

Strayer, D. L. and D. Dudgeon (2010). "Freshwater biodiversity conservation: recent progress and future challenges". *Journal of the North American Benthological Society* 29 (1), 344–358.

Suter, G. W., ed. (2007). *Ecological risk assessment*. Boca Raton: CRC Press/Taylor & Francis.

Umweltbundesamt (2016). *ETOX: Information System Ecotoxicology and Environmental Quality Targets*. URL: http://webetox.uba.de/webETOX/index.do.

U.S. EPA (2016). *The ECOTOXicology knowledgebase (ECOTOX)*. URL: http://cfpub.epa.gov/ecotox/.

Van den Brink, P. J., C. B. Choung, W. Landis, M. Mayer-Pinto, V. Pettigrove, P. Scanes, R. Smith, and J. Stauber (2016). "New approaches to the ecological risk assessment of multiple stressors". *Marine and Freshwater Research* 67 (4), 429.

Van Der Hoeven, N. (1998). "Power analysis for the NOEC: What is the probability of detecting small toxic effects on three different species using the appropriate standardized test protocols?" *Ecotoxicology* 7 (6), 355–361.

Vörösmarty, C. J., P. B. McIntyre, M. O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S. E. Bunn, C. A. Sullivan, C. R. Liermann, and P. M. Davies (2010). "Global threats to human water security and river biodiversity". *Nature* 467 (7315), 555–561.

Warne, M. S. J. and R. van Dam (2008). "NOEC and LOEC data should no longer be generated or used". *Australasian Journal of Ecotoxicology* 14, 1–5.

Waters, C. N., J. Zalasiewicz, C. Summerhayes, A. D. Barnosky, C. Poirier, A. Galuszka, A. Cearreta, M. Edgeworth, E. C. Ellis, M. Ellis, et al. (2016). "The Anthropocene is functionally and stratigraphically distinct from the Holocene". *Science* 351 (6269), aad2622.

WWF (2016). *Living Planet Report 2016 - Risk and resilience in a new era*. URL: http://wwf.panda.org/about_our_earth/all_publications/lpr_2016/.

Xing, Z., L. Chow, H. Rees, F. Meng, S. Li, B. Ernst, G. Benoy, T. Zha, and L. M. Hewitt (2013). "Influences of Sampling Methodologies on Pesticide-Residue Detection in Stream Water". *Archives of Environmental Contamination and Toxicology* 64 (2), 208–218.

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, Landau, Germany

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

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.

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 transforma-

tion. 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, van den 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.

METHODS

Models for count data

Linear model for transformed data

To meet the assumptions of the standard linear model, count data usually needs to be transformed. We followed the recommendations of van den Brink et al., (2000) and used a log(Ay + 1) transformation (eqn. 2.1):

$$Y_{new\ i} = \log(AY_i + 1) \tag{2.1}$$

, where Y_i is the measured and $Y_{new\ i}$ the transformed abundance of the ith observation. The factor A was chosen in such way that AY equals 2 for the lowest non-zero abundance value (Y).

Then we fitted the linear model to the transformed abundances (hereafter LM):

$$\begin{gathered} Y_{new\ i} \sim N(\mu_i, \sigma^2) \\ E(Y_{new\ i}) = \mu_i \text{ and } var(Y_{new\ i}) = \sigma^2 \\ \mu_i = \beta \times X_i \end{gathered} \tag{2.2}$$

This model assumes a normal distribution of the transformed abundances. The expected value for each observation i is given by its mean (μ_i) and the variance (σ^2) is constant between treatments. We allow this mean to vary between treatments (X_i codes the treatments) and β are the estimated coefficients related to these changes in transformed abundances between treatments (eqn. 2.2).

Generalised Linear Models

GLMs extend the linear model to variables that are not normally distributed. Instead of transforming the response variable, the counts could be directly modeled by a Poisson GLM (GLM_p):

$$\begin{aligned} Y_i &\sim P(\mu_i) \\ E(Y_i) = var(Y_i) &= \mu_i \\ log(\mu_i) &= \beta \times X_i \end{aligned} \tag{2.3}$$

This model assumes Poisson distributed abundances with mean $\mu_i \geqslant 0$. The expected value for each observation i is given by its mean. Moreover, this model assumes that mean and variance are equal. We are modeling the mean as a function of treatment membership (X_i). However, to avoid negative values of the mean this is done on a log scale. Therefore, β also describes the differences between treatments on a log scale (eqn. 2.3).

The assumption of equal mean and variance is rarely met with ecological data, which is typically characterized by greater variance than the mean (overdispersion). To overcome this problem a quasi-Poisson model (GLM_{qp}) could be used, which models the variance as a linear function of the mean (eqn. 2.4):

$$var(Y_i) = \phi\mu_i \tag{2.4}$$

Here, ϕ is used to account for additional variation and is known as overdispersion parameter. The quasi-Poisson model is a post hoc method, meaning that first a Poisson model is estimated (eqn. 2.3) and than the standard errors are scaled by the degree of overdispersion (Hilbe, 2014).

Another possibility to deal with overdispersion is to model abundances by a negative binomial distribution (GLM_{nb}, eqn. 2.5):

$$\begin{aligned} Y_i &\sim NB(\mu_i, \kappa) \\ E(Y_i) = \mu_i \text{ and } var(Y_i) &= \mu_i + \mu_i^2/\kappa \\ log(\mu_i) &= \beta \times X_i \end{aligned} \tag{2.5}$$

This models assumes that abundances are negative binomially distributed, with a mean of $\mu_i \geqslant 0$ and a variance $\mu_i + \mu_i^2/\kappa$. Similar to the Poisson model we use a log link between mean and treatments. Note, that the quasi-Poisson model

assumes a linear mean-variance relationship (eqn. 2.4), whereas the negative binomial model assumes a quadratic relationship (eqn. 2.5).

The above described models are most commonly used in ecology (Ver Hoef and Boveng, 2007), although other distributions for count data are possible, like the negative binomial model with a linear mean-variance relationship (also known as NB1) or the poisson inverse gaussian model (Hilbe, 2014).

Models for binomial data

A binomial variable counts how often an event x occurs in a fixed number of independent trials N (e.g. *"5 out of 10 fish survived"*), with an equal probability of occurrence π between trials. The number of times an event occurs can also be calculated as proportion x/N.

Linear model for transformed data

To accommodate the assumptions for the standard linear model with such proportions, a special arcsine square root transformation (eqn. 2.6) is suggested (EPA, 2002; Newman, 2012):

$$Y_{new\ i} = \begin{cases} \arcsin(1) - \arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } Y_i = 1 \\ \arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } Y_i = 0 \\ \arcsin(\sqrt{Y_i}) & \text{, otherwise} \end{cases} \tag{2.6}$$

, where Y_i are the untransformed proportions, $Y_{new\ i}$ are the transformed proportions, and n is the total number of exposed animals per treatment. The transformed proportions are then analysed using the standard linear model (LM, eqn. 2.2). Note, that the coefficients of the linear model are not directly interpretable due to transformation.

Generalised Linear Models

A more natural way to model such data is the binomial distribution with parameters N and π (GLM_{bin}):

$$\begin{aligned} Y_i &\sim Bin(N, \pi_i) \\ E(Y_i) = \pi_i \times N \text{ and } var(Y_i) &= \pi_i(1-\pi_i)/N \\ logit\ (\pi_i) &= \beta \times X_i \end{aligned} \tag{2.7}$$

This model assumes that the number of occurrences (Y_i) are binomially distributed, where N = number of trials (e.g. exposed animals) and π_i is the probability of occurrences (fish survived), which together give the expected number of occurrences. The variance of the binomial distribution is a quadratic function of the mean. We are modeling the probability of occurrence as function of treatment membership (X_i) and to ensure that $0 < \pi_i < 1$ we do this on a logit scale (eqn. 2.7). The estimated coefficients (β) of this model are directly interpretable as changes in log odds between treatments.

Non-independent trials (e.g. fish are grouped in aquaria) may lead to overdispersion (Williams, 1982). Methods to deal with overdispersed binomial data are for example quasi methods (see above) or Generalized Linear Mixed models (GLMM). However, these are not further investigated in this paper (see Warton and Hui, (2011) for a comparison).

Statistical Inference

After model fitting the next step is statistical inference. Ecotoxicologists are generally interested in two hypotheses: (i) is there any treatment related effect? and (ii) which treatments show a treatment effect (to determine the LOEC)?

Following general recommendations (Bolker et al., 2009; Faraway, 2006), we used F-tests (LM and GLM_{qp}) and Likelihood-Ratio (LR) tests (GLM_{p}, GLM_{nb} and GLM_{bin}) to test the first hypothesis. However, it is well known that the LR test is unreliable with small sample sizes (Wilks, 1938). Therefore, we additionally explored the parametric bootstrap (Faraway, 2006) to assess the significance of the LR. Bootstrapping is computationally very intensive and for this reason we applied it only for the LR test of the negative binomial models (using 500 bootstrap samples, denoted as GLM_{npb}).

To assess the LOEC we used Dunnett contrasts (Dunnett, 1955) with one-sided Wald t tests (normal and quasi-Poisson models) and one-sided Wald Z tests (Poisson, negative binomial and binomial models). Beside these parametric methods we also applied two, in ecotoxicology commonly used, non-parametric methods: The Kruskal-Wallis test (*KW*) to test for a general treatment effect and a pairwise Wilcoxon test (*WT*) to determine the LOEC. We adjusted for multiple testing using the method of Holm, (1979).

Case study

Brock et al., (2015) presents a typical example of data from mesocosm studies, which we use to demonstrate differences between methods. The data are mayfly larvae counts on artificial substrate samplers at one sampling date. A total of 18 mesocosms have been sampled from 6 treatments (Control (n = 4), 0.1, 0.3, 1, 3 mg/L (n = 3) and 10 mg/L (n = 2)) (Figure 2.1).

Simulations

Count data

To further scrutinise the differences between methods we simulated data sets with known properties. We simulated count data that mimics the data of the case study with five treatments (T1 - T5) and one control group (C). Counts were drawn from a negative binomial distribution with overdispersion at all treatments ($\kappa = 4$, eqn. 2.5). We simulated data sets with different number of replicates (N = {3, 6, 9}) and different abundances in control treatments (μ_C = {2, 4, 8, 16, 32, 64, 128}). For Type I error estimation mean abundance was equal between treatments. For power estimation, mean abundance in treatments T2 - T5 was reduced to half of control and T1 ($\mu_{T_2} = ... = \mu_{T_5} = 0.5\ \mu_C = 0.5\ \mu_{T_1}$), resulting in a theoretical LOEC at T2. We generated 1000 data sets for each combination of N and μ_C and analysed these using the models outlined in section 2.3.1.

Binomial data

We simulated data from a commonly used design as described in Weber et al., (1989), with 5 treated (T1 - T5) and one control group (C). Proportions were

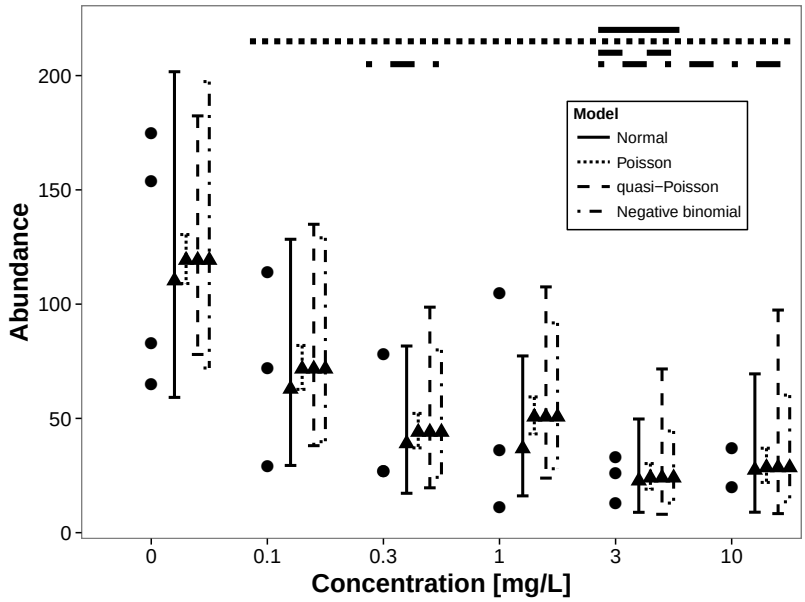


Figure 2.1.: Data from Brock et al., (2015) (dots). Predicted values (triangles) and 95% Wald Z or t confidence intervals from the fitted models (vertical lines) are given beside. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts). The data showed considerable overdispersion ($\kappa = 3.91, \phi = 22.41$) and therefore, the Poisson model underestimates the width of confidence intervals.

drawn from a Bin(10, π) distribution, with varying probability of survival (π = {0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95}) and varying number of replicates (N = {3, 6, 9}). For Type I error estimation, π was equal between treatments. For power estimation π was fixed at 0.95 in C and T1 and varied only in treatments T2 - T5. For each combination we simulated 1000 data sets and analysed these using the models outlined in section 2.3.2.

Data Analysis

We analysed the case study and the simulated data using the outlined methods. We compared the methods and models in terms of Type I error (detection of an effect when there is none) and power (ability to detect an effect when it is present) at a significance level of $\alpha = 0.05$.

All simulations were done in R (Version 3.1.2) (R Core Team, 2014) on an Amazon EC2 virtual Linux server (64bit, 15GB RAM, 8 cores, 2.8 GHz). Source code to reproduce the simulations and paper is available online at `https://github.com/EDiLD/usetheglm`. Moreover, Supplement A.2 provides worked examples of the data of Brock et al., (2015) and Weber et al., (1989).

RESULTS

Case study

The data set showed considerably higher variance then expected by the Poisson model (ϕ = 22.41 (eqn. 2.4), κ = 3.91 (eqn. 2.5)). Therefore, the Poisson model did not fit to this data and led to underestimated standard errors and confidence intervals, as well as overestimated statistical significance (Figure 2.1). In this case, inferences on the Poisson model are not valid and we do not further discuss its results. The normal (F = 2.57, p = 0.084) and quasi-Poisson model (F = 2.90, p = 0.061), as well as the Kruskal test (p = 0.145) did not show a statistically significant treatment effects. By contrast, the LR test and parametric bootstrap of the negative binomial model indicated a treatment-related effect (LR = 13.99, p = 0.016, bootstrap: p = 0.042).

All methods predicted similar values, except the normal model predicting always lower abundances (Figure 2.1). 95% confidence intervals (CI) were most narrow for the negative binomial model and widest for the quasi-Poisson model

- especially at lower estimated abundances. Consequently, the LOECs differed (Normal and quasi-Poisson: 3 mg/L, negative binomial: 0.3 mg/L). The pairwise Wilcoxon test did not detect any treatment different from control.

Simulations

Count data

For detecting a general treatment effect, GLM_{nb} and GLM_{p} showed inflated Type I error rates, whereas KW was conservative at low sample sizes. However, using the parametric bootstrap for the negative binomial model (GLM_{npb}), as well as LM and GLM_{qp} resulted in appropriate Type I error rates. For detecting a treatment effect,GLM_{qp} had the highest power, followed by GLM_{npb}, LM and KW, the latter having least power (Figure 2.2). For our simulation design (reduction in abundance by 50%) a sample size per treatment of n = 9 was needed to achieve a power greater than 80%. At small sample sizes (n = 3, 6) and low abundances (μ_C = 2, 4) many of the negative binomial models (GLM_{nb} and GLM_{npb}) did not converge to a solution (convergence rate <85% of the simulations, Supplement A.1).

For LOEC determination GLM_{nb} and GLM_{p} showed an increased Type I error and all other methods were slightly conservative. The inferences on LOEC generally showed less power. LM showed a mean reduction of 20.7% and GLM_{qp} of 24.3 %. Power to detect the LOEC was highest for GLM_{qp}. LM and WT showed less power, with WT having no power to detect the LOEC at low sample sizes (Figure 2.3).

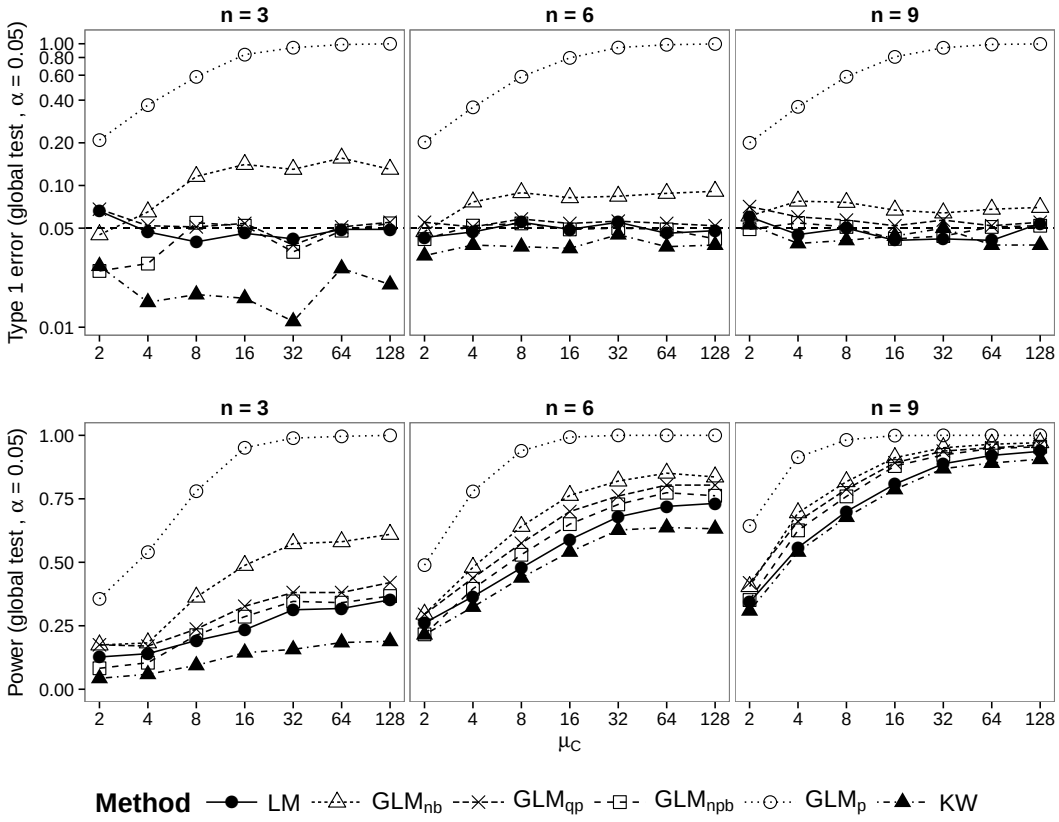


Figure 2.2.: Count data simulations: Type I error (top) and Power (bottom) for the test of a treatment effect. Type I errors are displayed on a logarithmic scale. Power levels for models with inflated Type I errors (GLM_P and GLM_{qp}) are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} and GLM_{npb} models did converge. Dashed horizontal line denotes the nominal I error rate at $\alpha = 0.05$.

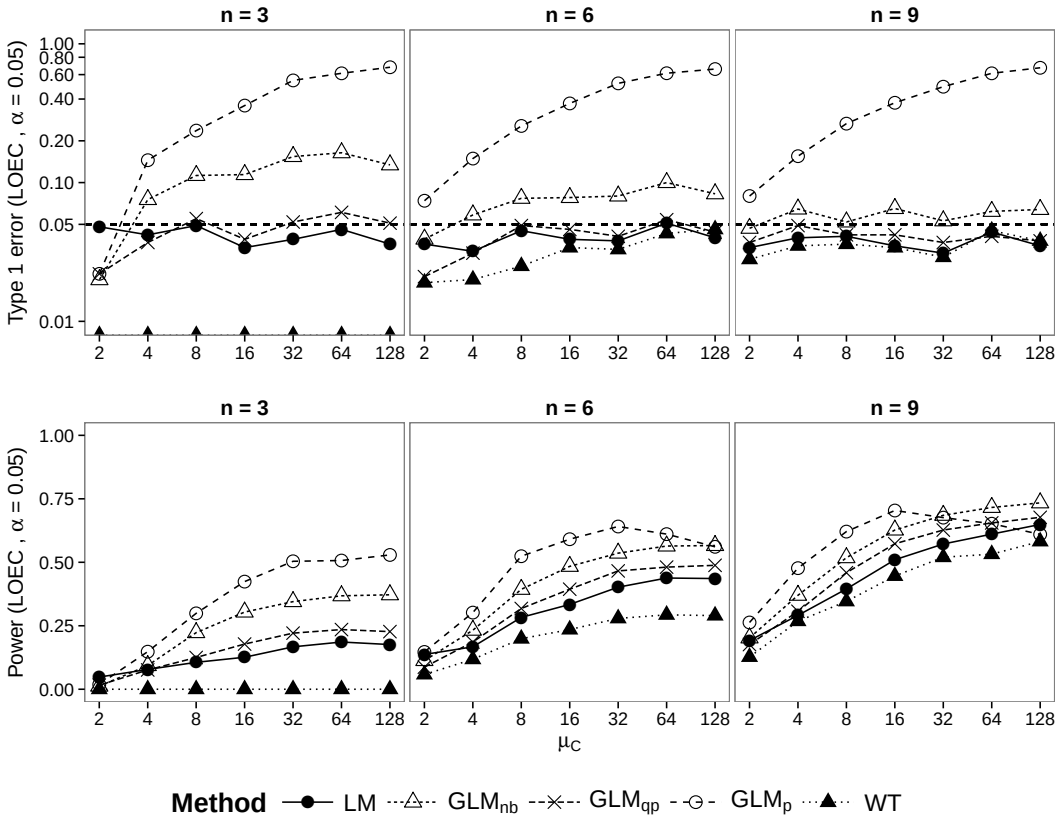


Figure 2.3.: Count data simulations: Type I error (top) and Power (bottom) for determination of LOEC. Type I errors are displayed one a logarithmic scale. Power levels for models with inflated Type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} models did converge. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

Binomial data

GLM_{bin} showed slightly increased Type I error rates at low sample sizes and small effect sizes. *KW* was more conservative than LM and GLM_{bin}. In addition, GLM_{bin} exhibited the greatest power for testing the treatment effect. This was especially apparent at low sample sizes (n = 3), with up to 27% higher power compared to LM. However, the differences between methods quickly vanished with increasing samples sizes (Figure 2.4).

For inference on LOEC we found that all methods were slightly conservative. *WT* was generally more conservative and GLM_{bin} especially at low effect sizes ($p_E > 0.7$). Inference on LOEC was not as powerful as inference on the general treatment effect. Contrary to the general treatment effect, LM showed the higher power than GLM_{bin} at small sample sizes (n = 3, 6). *WT* had no power for n = 3 and showed less power in the other simulation runs (Figure 2.5).

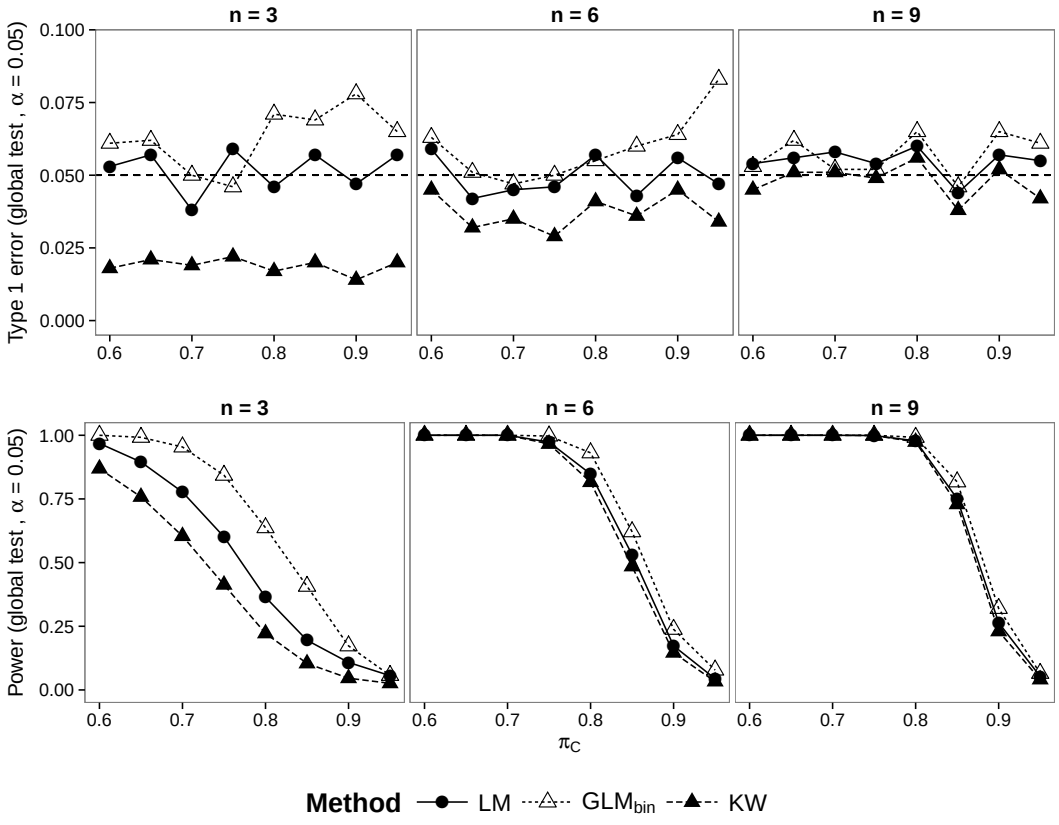


Figure 2.4.: Binomial data simulations: Type I error (top) and power (bottom) for the test of a treatment effect. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

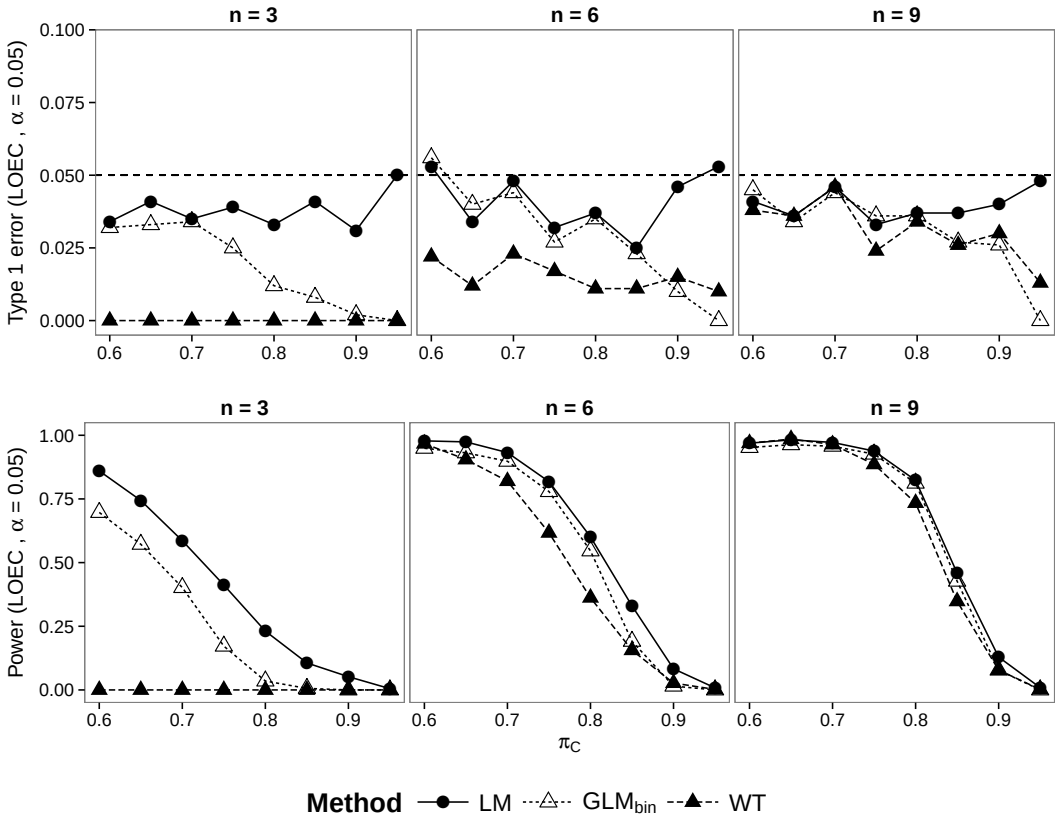


Figure 2.5.: Binomial data simulations: Type I error (top) and power (bottom) for the test for determination of LOEC. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

DISCUSSION

Case study

The outlined case study demonstrates that the choice of the statistical model and procedure can have substantial impact on ecotoxicological inferences and endpoints like the LOEC. Therefore, ecotoxicologists should not base their inferences solely on statistical significance tests, but also on model estimates, their uncertainty and importance (Gelman and Stern, 2006). O'Hara and Kotze, (2010) showed that the linear model on log transformed data gave unreliable and biased estimates, whereas GLMs performed well with little bias. Bias occurs also when back-transforming fitted means to the original scale, which explains the lower predicted means by LM in Figure 2.1 (Rothery, 1988) and should be corrected for (Newman, 1993). When applied to non-transformed data, the linear model would predict identical treatment means as GLMs, because for a categorical predictor the predicted means of the LM and GLM are identical. When applied to non-transformed data, the linear model would result in identical predicted treatment means as GLMs. However, predictions would differ with continuous predictors and GLMs are particularly advantageous in this case.

This is further highlighted by the fact that for the same model (linear model applied to transformed data), Brock et al., (2015) reported a 10-fold lower LOEC (0.3 mg/L) then found in our study (3 mg/L, Figure 2.1). The reasons are manifold: (i) Brock et al., (2015) used a $\log(2\ y + 1)$ transformation, whereas we used a $\log(A\ y + 1)$ transformation, where A = 2 / 11 = 0.182 (van den Brink et al., 2000). (ii) We adjusted for multiple testing using Holm's (1979) method. (iii) Brock et al., (2015) used a one-sided Williams test (Williams, 1972), whereas we used one-sided comparisons to the control (Dunnett contrasts). The choice of transformation contributed only little to the differences. If the assumptions of Williams test are met it has strictly greater power than Dunnett contrasts (Jaki and L. A. Hothorn, 2013), which explains the differences in the case study. A generalisation of the Williams test as multiple contrast test (MCT) can be used in a GLM framework (T. Hothorn et al., 2008). Nevertheless, such a Williams-type MCT is not a panacea (L. A. Hothorn, 2014) and our simulated semi-concave dose-response relationship is a situation where it fails and likely underestimates the LOEC (Kuiper et al., 2014).

Overdispersion is common for ecological datasets (Warton, 2005) and the case study illustrates the potential effects of overdispersion that is not accounted

for: standard errors will be underestimated and significance overestimated (Figures 2.1). This is also shown by our simulations (Figures 2.2, 2.3) where GLM_p showed increased Type I error rates because of overdispersed simulated data. However, in factorial designs the mean-variance relationship can be easily checked by plotting mean versus variance of the treatment groups or by inspecting residual versus fitted values plots (see Supplement A.2). Our simulations revealed that the LR test for GLM_{nb} is invalid because of increased Type I errors. This explains why it had the lowest p-value in the case study.

In the introduction we pointed out that there is little advice how to choose between the plenty of possible transformations - how do GLMs simplify this problem? The distribution modeled can be chosen using knowledge about the data (e.g. bounds, integer or continuous data etc). Knowing what type of data is modeled (see Methods section), the model selection process can be completely guided by the data and diagnostic tools. Therefore, choosing an appropriate model is easier than choosing between possible transformations.

Simulations

Our simulations showed that GLMs have generally greater power than the linear model applied to transformed data. However, the simulations also suggest that the power at the population level in common mesocosm experiments is low. For common samples sizes ($n \leqslant 4$) and a reduction in abundance of 50% we found a low power to detect any treatment-related effect (<50% for methods with appropriate Type I error, Figure 2.2). Statistical power to detect the correct LOEC was even lower (less than 25%), which can be attributed to multiple testing. The low power of all methods to detect significant treatment levels such as the LOEC or NOEC suggests that these endpoints from ecotoxicological studies should be interpreted with caution and underpins their criticism (Landis and Chapman, 2011; Laskowski, 1995).

Mesocosm studies allow also for inferences on the community level. For community analyses *GLM for multivariate data* (Warton et al., 2012) have been proposed as alternative to Principal Response Curves (PRC) and yielded similar inferences, but better indication of responsive taxa (Szöcs et al., 2015). However, ter Braak and Šmilauer, (2015) argue to use data transformations with community data because of their simplicity and robustness. Although our simulations covered only simple experimental designs at the population level, findings may also extend to more complex situations. Nested or repeated designs with

non-normal data could be analysed using Generalised Linear Mixed Models (GLMM) and may have advantages with respect to power (Stroup, 2015).

To counteract the problems with low power at the population level Brock et al., (2015) proposed to take the Minimum Detectable Difference (MDD), a method to assess statistical power *a posteriori*, for inference into account. However, *a priori* power analyses can be performed easily using simulations, even for complex experimental designs (Johnson et al., 2015), and might help to design, interpret and evaluate ecotoxicological studies. Moreover, Brock et al., (2015) proposed that statistical power of mesocosm experiments can be increased by reducing sampling variability through improved sampling techniques and quantification methods, though they also caution against depleting populations through more exhaustive sampling. As we showed, using GLMs can enhance the power at no extra costs.

Wang and Riffel, (2011) advocated that in the typical case of small sample sizes (n <20) and non-normal data, non-parametric tests perform better than parametric tests assuming normality. In contrast, our results showed that the often applied KW and WT have less power compared to LM. Moreover, GLMs always performed better than non-parametric tests. Though more powerful non-parametric tests may be available (Konietschke et al., 2012), these are focused on hypothesis testing and do not provide estimation of effect sizes. Additionally to testing, GLMs allow the estimation and interpretation of effects that might not be statistically significant, but ecologically relevant. Therefore, we advise using GLMs instead of non-parametric tests for non-normal data.

We found an increased Type-I error for $\mathrm{GLM}_{\mathrm{nb}}$ at low sample sizes. However, it is well known that the LR statistic is not reliable at small sample sizes (Bolker et al., 2009; Wilks, 1938). Parametric bootstrap ($\mathrm{GLM}_{\mathrm{npb}}$) is a valuable alternative in such situations and maintains appropriate levels (Figure 2.2). Moreover, at small sample sizes and low abundances a significant amount of negative binomial models did not converge. We used an iterative algorithm to fit these models (Venables and Ripley, 2002) and other methods assessing the likelihood directly may perform better.

$\mathrm{GLM}_{\mathrm{qp}}$ showed higher statistical power than $\mathrm{GLM}_{\mathrm{npb}}$ (Figure 2.2, bottom). This could be explained by the simpler mean-variance relationship of $\mathrm{GLM}_{\mathrm{qp}}$ (eqn. 2.4 and 2.5), because at small samples sizes, low abundances or few treatment groups it is difficult to determine the mean-variance relationship. Our results are similar to Ives, (2015), who compared GLMs to LM applied to transformed data for testing regression coefficients. Because of inflated Type I errors

for GLM_{nb} and, in the case of multiple explanatory variables in the model, inflated Type I errors of GLM_{qp} he considered the LM on transformed data as most robust and recommended its preferred use. However, we showed that the parametric bootstrap LR test of GLM_{nb} provides appropriate Type I errors and bootstrapping might be an alternative for testing coefficients. Nevertheless, bootstrapping is computationally very intensive and we found no gains in power compared to GLM_{qp} (Figure 2.2). Given the higher power, appropriate Type I errors, stable convergence and reduced bias (O'Hara and Kotze, 2010) we suggest that count data in one factorial experiments should be analysed using the quasi-Poisson model.

Binomial data are often collected in lab trials, where increasing the sample size may be relatively easy to accomplish. We found notable differences in power to detect a treatment effect for all simulated sample sizes. Similarly, Warton and Hui, (2011) also found that GLMs have higher power than arcsine transformed linear models. Though we did not simulate overdispersed binomial data, this should be checked and accounted for. In such situations a GLMM may offer an appealing alternative (Warton and Hui, 2011). At low effect sizes GLM_{bin} became conservative with increasing π_C, although this effect lessened as sample size increased (Figure 2.5). This is because π approaches its boundary and is also known as the *Hauck-Donner effect* (Hauck and Donner, 1977). A LR-Test or parametric bootstrap may provide an alternative in such situations (Bolker et al., 2009). This can also explain why LM performed better for deriving LOECs at low sample sizes.

GLMs can be fitted with several statistical software packages and many textbooks are available to introduce ecotoxicologists to these models (e.g. Zuur 2013 or Quinn and Keough 2009). We recommend that ecotoxicologists should change their models instead of their data. GLMs should become a standard method in ecotoxicology and incorporated into respective guidelines.

REFERENCES

Anderson, M. J., T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen, and N. G. Swenson (2011). "Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist". *Ecology Letters* 14 (1), 19–28.

Bolker, B., M. Brooks, C. Clark, S. Geange, J. Poulsen, M. Stevens, and J. White (2009). "Generalized linear mixed models: a practical guide for ecology and evolution". *Trends in Ecology & Evolution* 24 (3), 127–135.

Brock, T. C. M., M. Hammers-Wirtz, U. Hommen, T. G. Preuss, H.-T. Ratte, I. Roessink, T. Strauss, and P. J. Van den Brink (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems". *Environmental Science and Pollution Research* 22 (2), 1160–1174.

Dunnett, C. W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control". *Journal of the American Statistical Association* 50 (272), 1096–1121.

EFSA PPR (2013). "Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters". *EFSA Journal* 11 (7), 3290.

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

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton: Chapman & Hall.

Gelman, A. and H. Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant". *The American Statistician* 60 (4), 328–331.

Hauck, W. W. and A. Donner (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis". *Journal of the American Statistical Association* 72 (360), 851.

Hilbe, J. M. (2014). *Modeling Count Data*. New York, NY: Cambridge University Press.

Holm, S. (1979). "A simple sequentially rejective multiple test procedure". *Scandinavian journal of statistics* 6 (2), 65–70.

Hothorn, L. A. (2014). "Statistical evaluation of toxicological bioassays – a review". *Toxicol. Res.* 3 (6), 418–432.

Hothorn, T., F. Bretz, and P. Westfall (2008). "Simultaneous inference in general parametric models". *Biometrical Journal* 50 (3), 346–363.

Ives, A. R. (2015). "For testing the significance of regression coefficients, go ahead and log-transform count data". *Methods in Ecology and Evolution* 6 (7), 828–835.

Jaki, T. and L. A. Hothorn (2013). "Statistical evaluation of toxicological assays: Dunnett or Williams test—take both". *Archives of Toxicology* 87 (11), 1901–1910.

Johnson, P. C. D., S. J. E. Barry, H. M. Ferguson, and P. Müller (2015). "Power analysis for generalized linear mixed models in ecology and evolution". *Methods in Ecology and Evolution* 6 (2), 133–142.

Konietschke, F., L. A. Hothorn, and E. Brunner (2012). "Rank-based multiple test procedures and simultaneous confidence intervals". *Electronic Journal of Statistics* 6, 738–759.

Kuiper, R. M., D. Gerhard, and L. A. Hothorn (2014). "Identification of the Minimum Effective Dose for Normally Distributed Endpoints Using a Model Selection Approach". *Statistics in Biopharmaceutical Research* 6 (1), 55–66.

Landis, W. G. and P. M. Chapman (2011). "Well past time to stop using NOELs and LOELs". *Integrated Environmental Assessment and Management* 7 (4), vi–viii.

Laskowski, R. (1995). "Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology". *Oikos* 73 (1), 140–144.

Nelder, J. A. and R. W. M. Wedderburn (1972). "Generalized Linear Models". *Journal of the Royal Statistical Society. Series A (General)* 135 (3), 370–384.

Newman, M. C. (1993). "Regression analysis of log-transformed data: Statistical bias and its correction". *Environmental Toxicology and Chemistry* 12 (6), 1129–1133.

Newman, M. C. (2012). *Quantitative ecotoxicology*. 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, R. B. and D. J. Kotze (2010). "Do not log-transform count data". *Methods in Ecology and Evolution* 1 (2), 118–122.

Quinn, G. P. and M. J. Keough (2009). *Experimental design and data analysis for biologists*. Cambridge: Cambridge Univ. Press.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: http://www.R-project.org/.

Rothery, P. (1988). "A cautionary note on data transformation: bias in back–transformed means". *Bird Study* 35 (3), 219–221.

Sanderson, H. (2002). "Pesticide studies". *Environmental Science and Pollution Research* 9 (6), 429–435.

Stroup, W. W. (2015). "Rethinking the analysis of non-normal data in plant and soil science". *Agronomy Journal* 107 (2), 811–827.

Szöcs, E., P. J. v. d. Brink, L. Lagadic, T. Caquet, M. Roucaute, A. Auber, Y. Bayona, M. Liess, P. Ebke, A. Ippolito, C. J. F. t. Braak, T. C. M. Brock, and R. B. Schäfer (2015). "Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods". *Ecotoxicology* 24 (4), 760–769.

Ter Braak, C. J. and P. Šmilauer (2015). "Topics in constrained and unconstrained ordination". *Plant Ecology* 216 (5), 683–696.

Van den Brink, P. J., J. Hattink, T. C. M. Brock, F. Bransen, and E. van Donk (2000). "Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions". *Aquatic Toxicology* 48 (2-3), 251–264.

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

Ver Hoef, J. M. and P. L. Boveng (2007). "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" *Ecology* 88 (11), 2766–2772.

Wang, M. and M. Riffel (2011). "Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology". *Ecotoxicology and Environmental Safety* 74 (4), 684–92.

Warton, D. I. (2005). "Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data". *Environmetrics* 16 (3), 275–289.

Warton, D. I. and F. K. C. Hui (2011). "The arcsine is asinine: the analysis of proportions in ecology". *Ecology* 92 (1), 3–10.

Warton, D. I., S. T. Wright, and Y. Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects". *Methods in Ecology and Evolution* 3 (1), 89–101.

Weber, C. I., W. H. Peltier, T. J. Norbert-King, W. B. Horning, F. Kessler, J. R. Menkedick, T. W. Neiheisel, P. A. Lewis, D. J. Klemm, Q. Pickering, E. L. Robinson, J. M. Lazorchak, L. Wymer, and R. W. Freyberg (1989). "Short-term methods for estimating the chronic toxicity of effluents and receiving waters to fresh- water organisms". (EPA/600/4–89/001).

Wilks, S. S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". *The Annals of Mathematical Statistics* 9 (1), 60–62.

Williams, D. A. (1972). "The comparison of several dose levels with a zero dose control". *Biometrics*, 519–531.

Williams, D. A. (1982). "Extra-Binomial Variation in Logistic Linear Models". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31 (2), 144–148.

Zuur, A. F. (2013). *A beginner's guide to GLM and GLMM with R: a frequentist and Bayesian perspctive for ecologists*. Newburgh: Highland Statistics.

3 LARGE SCALE RISKS FROM PESTICIDES IN SMALL STREAMS

Eduard Szöcs[a], Marvin Brinke[b], Bilgin Karaoglan[c] & Ralf B. Schäfer[a]

[a]Institute for Environmental Sciences, University Koblenz-Landau, Landau, Germany
[b]German Federal Institute of Hydrology (BfG), Koblenz, Germany
[c]German Environment Agency (UBA), Dessau-Roßlau, Germany

Submitted to *Environmental Science & Technology* in 2016

ABSTRACT

Small streams are important refugia for biodiversity. In agricultural areas they may be at high risk from pesticide pollution. However, most related studies have been limited to a few streams on the regional level, hampering extrapolation to larger scales. We used data from German governmental water quality monitoring to quantify the drivers of pesticide risk and to assess pesticide risk in small streams on a large scale. The data set comprised of 1,766,104 measurements of 478 pesticides (including metabolites) related to 24,743 samples from 2,301 sampling sites. We investigated the influence of agricultural land use, catchment size, as well as precipitation and seasonal dynamics on pesticide risk using new statistical modelling techniques that explicitly consider the limit of quantification. Agricultural land use lead to a 3.7-fold increase in exceedance of risk thresholds when the proportion of agriculture in a catchment exceeded 28 percent. Precipitation increased pesticide risk by 36% and risk was the highest during summer months. Risk thresholds were exceeded in 26% of streams, with the highest risk related to neonicotinoid insecticides. We conclude that pesticides from agricultural land use are a major threat to small streams and their biodiversity and that a realistic pesticide sampling would be driven by precipitation events.

INTRODUCTION

More than 50% of the total land area in Germany is used by agriculture (Statistisches Bundesamt, 2014). In the year 2014 more than 45,000 tonnes of 776 authorised plant protection products were sold for application on this area (Bundesamt für Verbraucherschutz und Lebensmittelsicherheit, 2015). The applied pesticides may enter surface waters via spray-drift, edge-of-field run-off or drainage (Liess et al., 1999; Schulz, 2001; Stehle et al., 2013). Once entered the surface waters they may have adverse effects on biota and ecosystem functioning (Schäfer et al., 2012). Although it is known that pesticide pollution and its ecological effects increase with the fraction of agricultural land use in the catchment (Schulz, 2004), the shape of the relationship is unknown and studies on potential thresholds are lacking.

Two recent studies indicate that pesticides might threaten freshwater biodiversity in the European union. Malaj et al., (2014) analysed data supplied to the European Union (EU) in the context of the Water Framework Directive (WFD)

and showed that almost half of European water bodies are at risk from pesticides. Stehle and Schulz, (2015b) compiled 1,566 measured concentrations of 23 insecticides in the EU from scientific publications. They found that many of these measurements exceed regulatory acceptable concentrations (RAC). However, these studies reflect only a small amount of potentially available data (173 sites in predominantly mid-sized and large rivers in Malaj et al., (2014) and 138 measurements in Stehle and Schulz, (2015b)), and it is unclear how representative they are for Germany. Much more comprehensive data on thousands of sites are available from national monitoring programs that are setup for the surveillance of water quality, which is done independently by the federal states in Germany in compliance with the WFD (Quevauviller et al., 2008) and additional state-specific needs. Despite that these data are providing the opportunity to study pesticide risks and other research questions on a large scale with high spatial density, to date these data have not been compiled and related analyses are lacking.

Small streams comprise a major fraction of streams (Nadeau and Rains, 2007), accommodate a higher proportion of biodiversity compared to larger freshwater systems (Biggs et al., 2014; Davies et al., 2008) and play an important role in the recolonization of disturbed downstream reaches (Liess and von der Ohe, 2005; Orlinskiy et al., 2015). Nevertheless, a clear definition of small streams in terms of catchment or stream size is currently lacking (Lorenz et al., 2016). For example, the WFD defines small streams with a catchment size between 10 and 100 km², without further categorisation of streams <10km² and Lorenz et al., (2016) defines small streams with catchment size <10km². Moreover, small streams might particularly be at high risk of pesticide contamination in case of adjacent agricultural areas given their low dilution potential (Liess et al., 1999; Schulz, 2004). Indeed, meta-analyses using data from studies with a few sites reported higher pesticide pollution in smaller streams compared to bigger streams (Schulz, 2004; Stehle and Schulz, 2015b). Despite their ecological relevance and potentially higher pesticide exposure, a recent analysis of pesticide studies showed that a disproportionally small fraction of studies was conducted in small water bodies, and these were largely limited to a few sites (Lorenz et al., 2016). Consequently, knowledge on the pesticide pollution of small streams on larger scales is scant. In European law, the Directive 2009/128/EC (European Union, 2009) places an obligation on the EU Member States to adopt National Action Plans (NAP) for the Sustainable Use of Plant Protection Products and the

German NAP also addresses the knowledge gap concerning pesticide impact on small streams, specifically including those with catchment size <10km².

In this study, we compiled and analysed large-scale chemical monitoring data from small streams in Germany. First, we analysed the shape of the relationship between pesticide risk, agricultural land use, and catchment size and examined whether related thresholds for pesticide risks can be derived. Second, we investigated the influence of precipitation and seasonal dynamics on pesticide detections, given that precipitation proved an important driver of pesticide exposure in several small-scale studies (Schulz, 2004; Wittmer et al., 2010), but it is unknown whether a precipitation signal prevails on large scales. Finally, we quantified the current risks from pesticides in small streams in Germany and the compounds accountable for the risk.

METHODS

Data compilation

We queried pesticide monitoring data from sampling sites that can be classified as small streams (catchment sizes < 100 km² according to the WFD) from all 13 non-city federal states of Germany (see Supplemental Table S1 for the abbreviations of federal state names) for 2005 to 2015. We homogenised and unified all data provided by the federal states into a database and implemented a robust data-cleaning workflow (see Supplemental Figure S1 for details) (Poisot, 2015).

We identified precipitation at sampling sites by a spatio-temporal intersection of sampling events with gridded daily precipitation data (60×30 arcsec resolution) available from the German Meteorological Service (DWD). This data spatially interpolates daily precipitation values from local weather stations (Rauthe et al., 2013). We performed the intersection for the actual sampling date and the day before and extracted precipitation during and up to 48 hours before sampling.

Characterization of catchments

We compiled a total of 2,369 sampling sites in small streams with pesticide measurements. Alongside, we also queried catchment sizes and agricultural land use within the catchment for the sampling sites from the federal states. Catch-

ment size was provided for 59% of sites. Additionally, we delineated upstream catchments for each of the sampling sites using (i) a digital elevation model (DEM) (EEA, 2013) and the multiple flow direction algorithm (Holmgren, 1994) as implemented in GRASS GIS 7 (Neteler et al., 2012) and (ii) from drainage basins provided by the Federal Institute of Hydrology (BfG). Delineated catchments were visually checked for accuracy by comparison with state stream networks and derived information amalgamated with existing data. Thus, catchment size information was available for 99% of all sites (59% from authorities, 24% from DEM and 16% from drainage basins).

For each derived catchment (either from DEM or drainage basins) we calculated the % agricultural land-use within the catchment based on the Authoritative Topographic-Cartographic Information System (ATKIS) of the land survey authorities (AdV, 2016). Thus, agricultural land use information was available for 98% of all sites (24% from authorities, 52% from DEM and 22% from drainage basins). 68 sites (3%) that lacked catchment size or land use information were omitted from the analysis, resulting in 2301 sites used in the analyses outlined below.

Characterization of pesticide pollution

We characterised pesticide pollution using regulatory acceptable concentrations (RAC) (Brock et al., 2010). RACs are derived during pesticide authorisation as part of the ecological risk assessment. No unacceptable ecological effects are expected if the environmental concentration remains below this concentration. Stehle and Schulz, (2015b) showed that RAC exceedances reflect a decrease in biodiversity and from this perspective are ecologically relevant indicators. The German Environment Agency (UBA) provided RACs for 107 compounds, including those with the highest detection rates (Supplemental Table S2). Based on theses RACs, we calculated Risk Quotients (RQ):

$$RQ_i = \frac{C_i}{RAC_i} \tag{3.1}$$

where C_i is the concentration of a compound i in a sample and RAC_i the respective RAC.

Statistical analyses

All data-processing and analyses were performed using R (R Core Team, 2016). To display differences in the spectra of analysed compounds between federal states we used Multidimensional Scaling (MDS) based on Jaccard dissimilarity in conjunction with complete linkage hierarchical clustering using the vegan package (Oksanen et al., 2016). We determined the optimum number of clusters using the average silhouette width (Rousseeuw, 1987).

We expected non-linear responses to agriculture and catchment size and therefore, used generalised additive models (GAM) to establish relationships (Fewster et al., 2000). We modelled the number of RAC exceedances (RQ >1) at a site as:

$$\begin{aligned} No(RQ > 1)_i &\sim NB(\mu_i, \kappa) \\ log(\mu_i) &= \beta_0 + f_1(agri_i) + f_2(size_i) + log(n_i) \end{aligned} \tag{3.2}$$

where $No(RQ > 1)_i$ is the observed number of RAC exceedances at site i. We modelled $No(RQ > 1)_i$ as resulting from a negative binomial distribution (NB) with mean μ_i and a quadratic mean-variance-relationship ($Var(No(RQ > 1)_i) = \mu_i + \frac{\mu_i^2}{\kappa}$). The proportion of agriculture within the catchment ($agri_i$) and the catchment size of the site ($size_i$) were used as predictors of the number of RAC exceedances. β_0 is the intercept and f_1 and f_2 are smoothing functions using penalized cubic regression splines (Wood, 2006). The degree of smoothness was estimated using restricted maximum likelihood (REML) during the model fitting process (Wood, 2011). The number of measurements per site (n_i) was used as an offset to account for differences in sampling efforts (sampling interval and analysed compound spectrum) at a site and is equivalent to modelling the rate of exceedances. We used point-wise 95% Confidence Intervals (CI) of the first derivative of the fitted smooth to identify regions of statistically significant changes. GAMs were fitted using the mgcv package (Wood, 2011).

To assess the influence of precipitation and seasonality, we modelled the RQ of individual compounds as the response variable. RQ and concentrations show a skewed distribution with an excess of zeros (no pesticides detected and quantified). Therefore, we modelled these as two processes (one generating values below the limit of quantification (LOQ) and one generating values above LOQ) using a Zero-Adjusted Gamma (ZAGA) distribution (Rigby and D. M. Stasinopoulos, 2005; M. Stasinopoulos et al., 2016) (Equation 3.3). These two processes can

be interpreted as changes in the mean value of RQ (change in μ) and changes in the probability of exceeding LOQ and showing any risk (change in ν).

$$RQ_i \sim ZAGA(\mu_i, \sigma, \nu_i) = \begin{cases} (1-\nu_i) & \text{if } y < LOQ \\ \nu_i \times f_{Gamma}(\mu_i, \sigma) & \text{if } y \geqslant LOQ \end{cases} \tag{3.3}$$

ν_i denotes the probability of a measurement i being above LOQ and f_{Gamma} denotes the gamma function and is used for values equal to or greater LOQ, with μ being the mean and σ the standard deviation of RQ. We used the $\log(x + 0.05)$ transformed precipitation at sampling date ($\log prec_0$) and the day before ($\log prec_{-1}$), as well as quarters of the year ($Q1 - Q4$) as linear predictors for μ and ν. We used appropriate link functions for μ and ν and assumed σ to be constant. Equation 3.4 summarises the deterministic part of the model for a measurement i.

$$\begin{aligned} \log(\mu_i) &= \log prec_{0i} + \log prec_{-1i} + Q1_i + Q2_i + Q3_i + Q4_i \\ logit(\nu_i) &= \log prec_{0i} + \log prec_{-1i} + Q1_i + Q2_i + Q3_i + Q4_i \end{aligned} \tag{3.4}$$

To account for temporal autocorrelation and differences between federal states we used $site$ nested within $state$ as random intercepts. We implemented this model using the gamlss package (D. M. Stasinopoulos and Rigby, 2007).

We fitted this model separately to each compound with a RAC, measured in at least 1000 samples and with more than 5% of values above LOQ (n = 22 compounds, see Supplemental Table S3 for a list of compounds). To summarise the coefficients across the 22 modelled compounds we used a random effect meta-analysis for each model coefficient separately (Harrison, 2011), resulting in an averaged effect of the 22 compounds. The results of individual compounds are provided in the Supplemental Table S4 and Figure S7. The meta-analysis was performed using the metafor package (Viechtbauer, 2010).

RESULTS

Overview of the compiled data

The compiled dataset used for analysis comprised 1,766,104 pesticide measurements in 24,743 samples from 2,301 sampling sites in small streams. These samples were all taken via grab sampling. We found large differences between federal states in the number of sampling sites and their spatial distribution (Figure 3.1 and Supplemental Table S1). The number of small stream sampling sites per state ranged from 1 (Lower Saxonia, NI) to 1139 (North Rhine-Westphalia, NW). No data were available from Brandenburg.

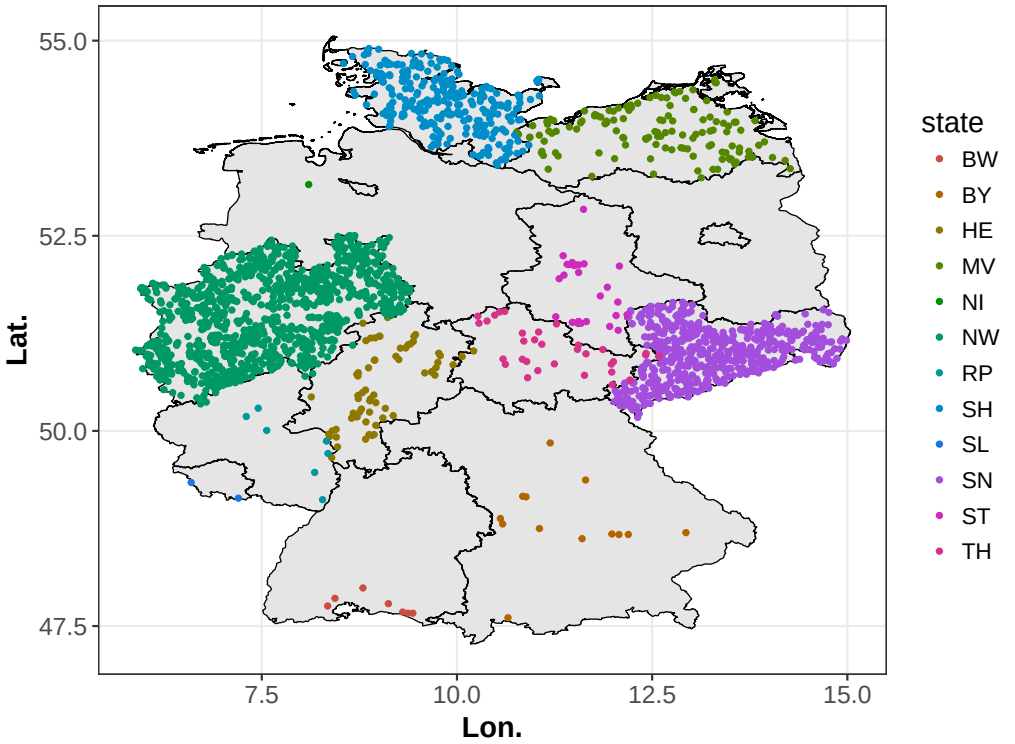


Figure 3.1.: Spatial distribution of the 2,301 small stream sampling sites. Colour codes different federal states (see Supplemental Table S1 for abbreviations).

In total 478 different compounds used as pesticides and their metabolites were measured at least once (Supplemental Table S2). Most of the compounds were herbicides (179), followed by insecticides (117) and fungicides (109). Most samples were taken in the months April till October, while fewer samples were taken during winter (see Supplemental Figure S2). We found substantial differences in the spectra of analysed pesticides between federal states (Figure 3.2). The number of different pesticides per state ranged from 57 (SL) to 236 (RP) (Supplemental Table S1). Hierarchical clustering revealed that RP and NI analysed distinct compound spectra compared to the cluster of other states. However, it has to be noted that both states surveyed these distinct spectra in special monitoring programs from only a few sites. Although there was high variability within the remaining cluster, this could not be further split (Figure 3.2, also Supplemental Figures S3 and S4). 4% (=71,113) of all measurements were concentrations above LOQ.

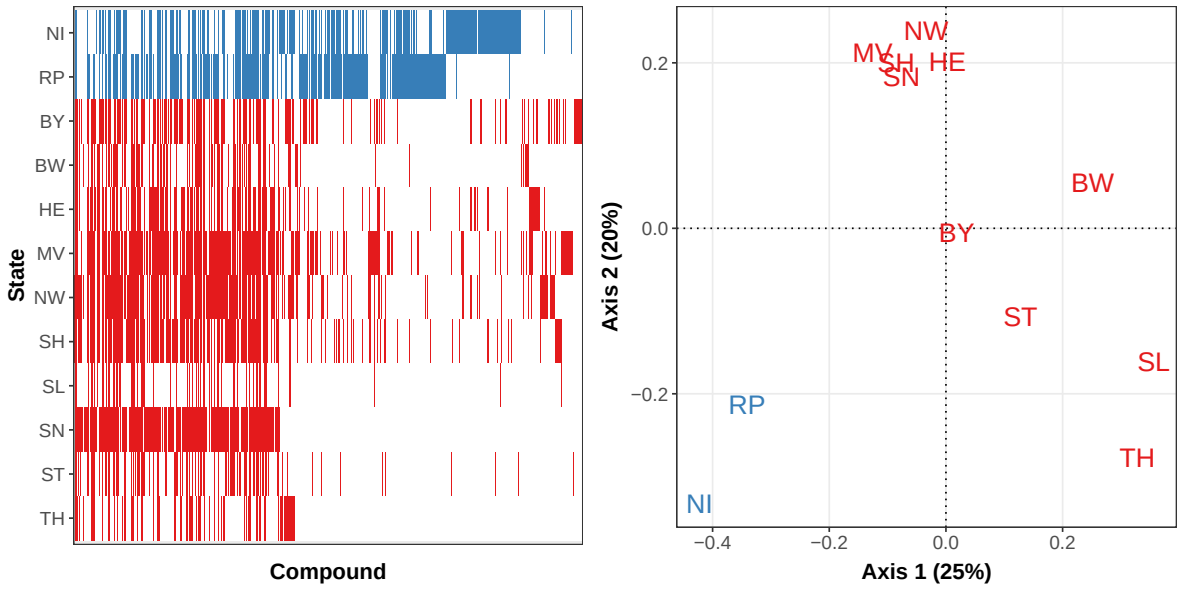


Figure 3.2.: Compound spectra of the different federal states. Left: Barcode plot - each vertical line is an analysed compound. Right: MDS ordination. Colors according to two clusters determined by hierarchical clustering (see Supplemental Figure S3 and S4).

The distribution of sampling sites across catchment sizes indicated a disproportionally low number of sites with catchments below 10 km^2, with most sampling sites having catchment sizes between 10 and 25 km^2 (Figure 3.3).

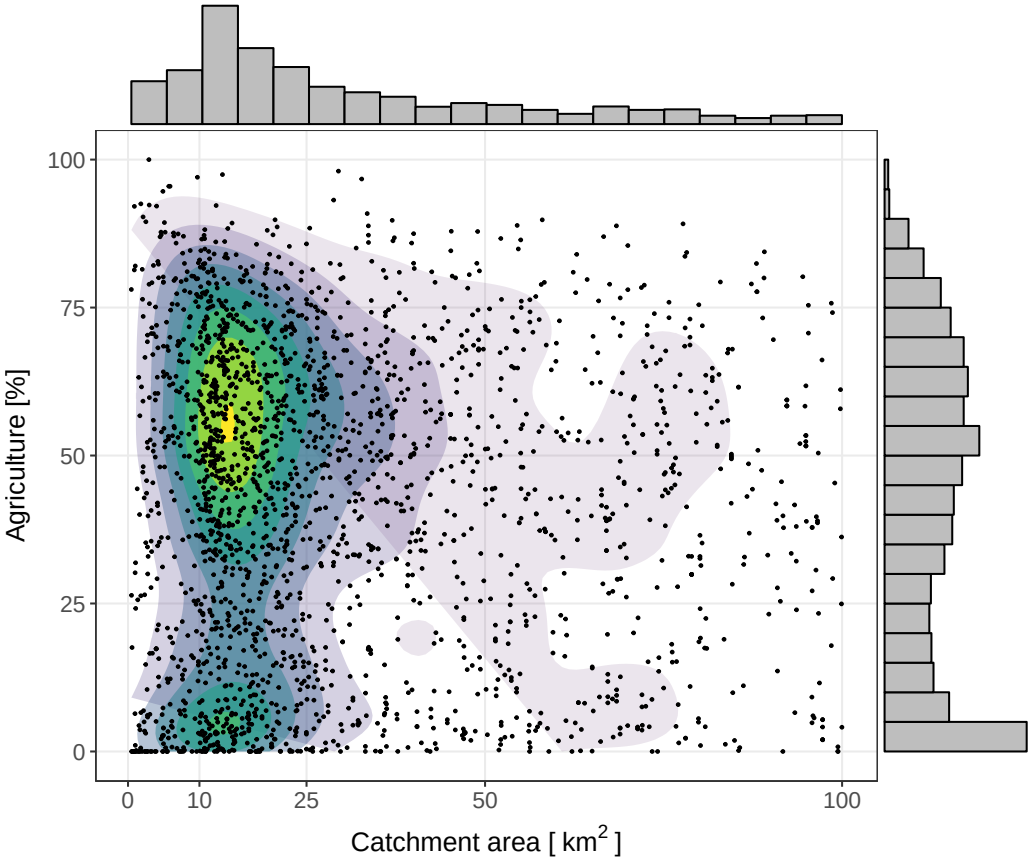


Figure 3.3.: Distribution of catchment area and agriculture within the catchment area across the sampling sites. Colour codes the 2-dimensional density of points.

Influence of agricultural land use and catchment size

The number of RAC exceedances increased strongly and statistically significant up to 28% agriculture within the catchment. The mean number of RAC exceedances per site increased 3.7-fold from 0.10 (no agriculture) to 0.39 (28% agriculture within the catchment). Above this threshold the exceedances levelled. Above 75% agriculture within the catchment the number of exceedances further increased, but the increase was not statistically significant (Figure 3.4,

left). Catchment size had no statistically significant effect on the number of RAC exceedances (Figure 3.4, right). We also could not detect a statistically significant interaction between catchment size and agriculture.

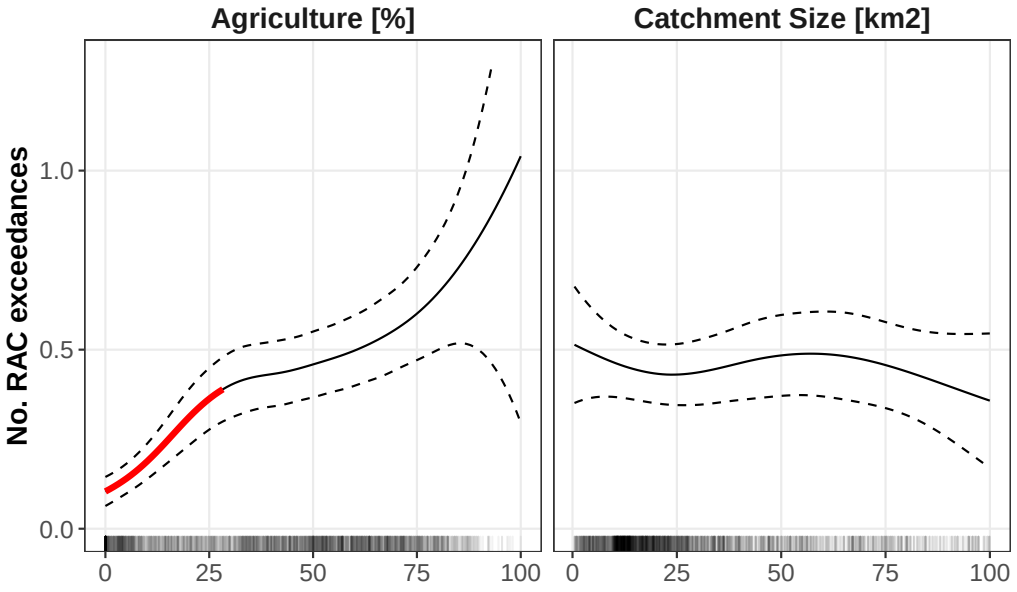


Figure 3.4.: Effect of percent agriculture within the catchment (left) and catchment size (right) on the number of RAC exceedances. Red line marks statistically significant changes. Dashed lines denote 95% point-wise Confidence Intervals.

Effect of precipitation on pesticide risk

The spatio-temporal intersection revealed that most samples were taken during periods of low precipitation. For example, only 5% of the samples were taken at or after days with rainfall events greater than 10mm / day that may lead to run-off (Supplemental Figure S6).

$prec_0$ and $prec_{-1}$ increased the probability of exceeding LOQ and RQ. In Q2 an increase from 0.1 mm to 10 mm of precipitation before sampling ($prec_{-1}$) lead on average to a 36% higher mean RQ of 0.05. The probability to exceed LOQ increases 1.6-fold from 8.7% to 13.5% (Figure 3.5, top). Effects differed between individual compounds and are provided in the Supplemental Table S4. Precipitation before sampling ($prec_{-1}$) had a stronger effect than precipitation

during sampling ($prec_0$) on the probability of exceeding LOQ. This difference was less pronounced for the mean value of RQ (Figure 3.5, top).

The first quarter showed the lowest RQ and probability of exceeding LOQ. Both increased during summer months and decreased towards the end of the year. There was a 2.5-fold higher probability of exceeding LOQ in Q2 (10.6%) than in Q1 (4.6%). The differences were less pronounced for the mean value of RQ and with less precision (Figure 3.5, bottom). Individual compounds showed different temporal patterns (see Supplemental Table S4).

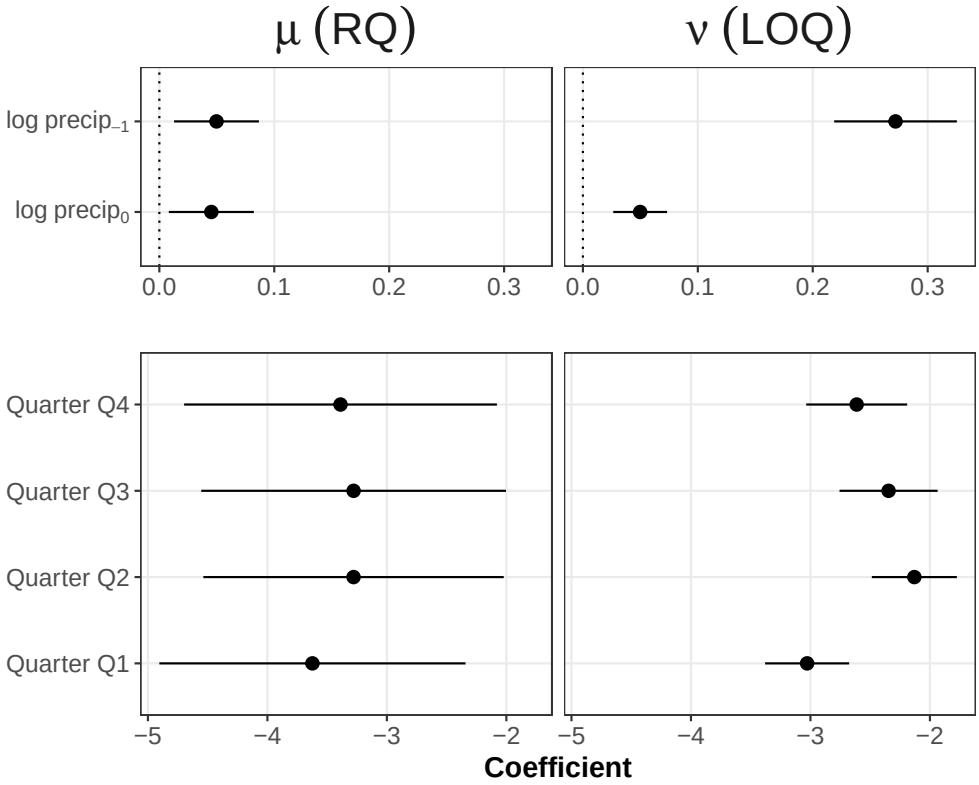


Figure 3.5.: Summarised coefficients (and their 95% CI) for precipitation (top row) and quarter (bottom row) from a meta-analysis of the 22 modelled compounds. Left: coefficients for mean RQ (μ), right: coefficients for probability of exceeding LOQ (ν). Coefficients are shown on the link scale (see Eq. 3.4). Single compound coefficients are provided in Supplemental Table S4 and Figure S7).

Pesticide risk in small streams

We found RAC exceedances in 25.5% of sampling sites and RQ > 0.1 in 54% of sites. In 23% of sites none of the chemicals, for which RACs were available, were detected (see also Supplemental Figure S8). Neonicotinoid insecticides and Chlorpyrifos showed the highest RQ (Figure 3.6). For Thiacloprid and Chlorpyrifos the RAC was equal or less than LOQ, therefore, all detections have a RQ $\geqslant$ 1. The herbicides Nicosulfuron and Diflufenican, as well as the fungicide Dimoxystrobin also showed high exceedances of RQ (26.7, 14.1 and 21.1 % of measurements > LOQ), see also Supplemental Table S5). RAC exceedances were found in 14% of samples with concentrations >LOQ (and 7.3% of all samples).

The highest RQs were observed for Chlorpyrifos (max(RQ) = 220), Clothianidin (max(RQ) = 157), Dimoxystrobin(max(RQ) = 117) and Isoproturon (max(RQ) = 80). Where analysed, metabolites exhibited the highest detection rates (for example, Metazachlor sulfonic acid was detected in 84% of all samples where it was analysed (n = 3038, see also Supplemental Figure S9). Glyphosate was the compound with the highest detection rates (41%, n = 3557 samples), followed by Boscalid (23%, n = 9886) and Isoproturon (22%, n = 19112). However, only the latter showed RAC exceedances (Figure 3.6). In 45.9% of samples more than one compound was quantified, with a maximum of 54 different compounds in one sample (Supplemental Figure S10).

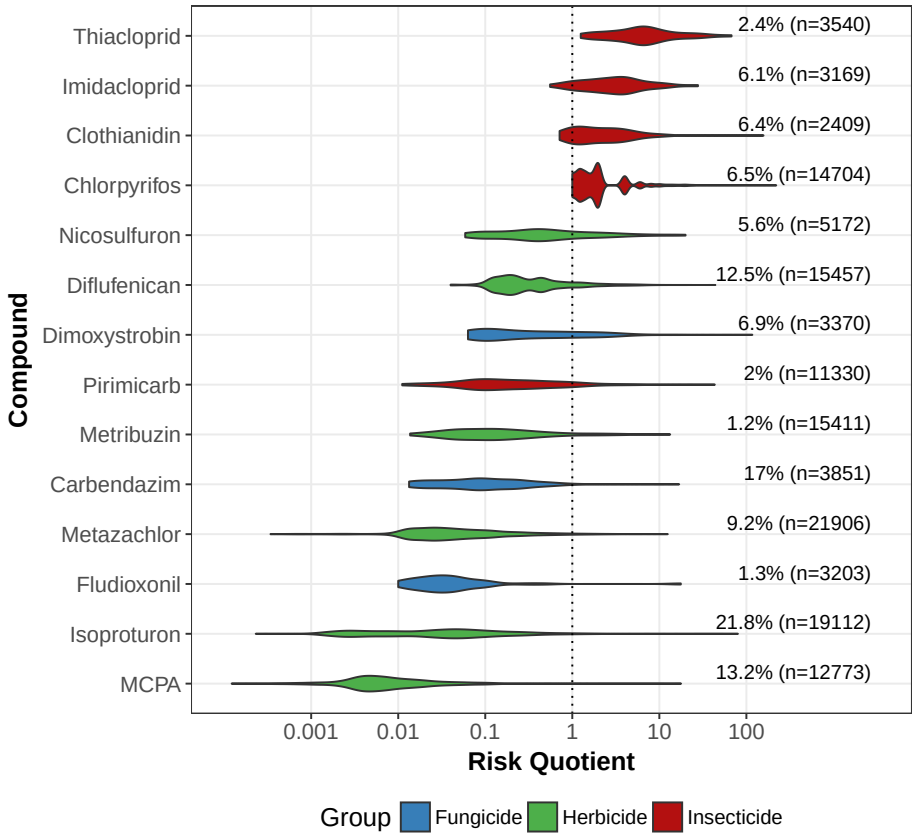


Figure 3.6.: 15 compounds with the highest risk quotients in small streams. Non-detects are not shown due to the logarithmic axis. Numbers on the right give the percentage of values >LOQ and the total number of samples were the compound was analysed.

DISCUSSION

Overview on the compiled dataset

The compiled dataset of governmental monitoring data, with a particular focus on small streams, represents currently the most comprehensive available for Germany. Similar nationwide datasets have been compiled for the Netherlands (Vijver et al., 2008), Switzerland (Munz and Leu, 2011) and the United States (Stone et al., 2014). While the compilations from Europe are of similar quantity and quality to the data compiled and analysed here, the compilation used in

Stone et al., (2014) is much smaller, though these data may be complemented by more data in future analyses.

A nationwide assessment of pesticide pollution is hampered by inhomogeneous data across federal states: Beside large differences in the spatial distribution and quantity of sampling sites (Figure 3.1), the spectrum of analysed compounds (Figure 3.2) and the quality of chemical analyses differed between states. Despite the outlined differences between states, all ecoregions occurring in Germany (Abell et al., 2008; Illies, 1978) were covered by the presented dataset and thus it might nonetheless represent a sample covering all types of small streams in Germany. For Thiacloprid and Chlorpyrifos the LOQs were above the RAC, which means that exceedances are likely underestimated. For these compounds a lowering of LOQ through an improvement of chemical analysis is essential for reliable assessment. Moreover, a nationwide assessment would benefit from a harmonised spectrum of analysed compounds between federal states.

Given their high abundance in the landscape (Nadeau and Rains, 2007) small streams below 10 km² are disproportionally less sampled in current monitoring (Figure 3.3), which may be attributed to the missing categorisation in the WFD. Clearly, there is currently a lack of knowledge on stressor effects on small streams. We analysed only data from small streams, however, for lentic small water bodies this lack might be even greater (Lorenz et al., 2016).

Influence of agricultural land use and catchment size

We found a strong influence of agriculture on the pollution of streams. Above 25% agriculture within a catchment, it is likely that a RAC will be exceeded, with a further increase in entirely agricultural catchments (above 75 % agriculture). To our knowledge, this is the first study investigating such thresholds of pesticide risk. Previous studies examined thresholds for the percentage of agricultural land use with respect to the response of biological communities, integrating different agricultural stressors. Feld, (2013) found change points of biological community metrics at 40% agricultural land use in lowland streams in Europe. Similarly, Waite, (2014) found a threshold for aquatic diatoms at 40% agricultural land use in wadeable streams in the United States. Our results coincide with these thresholds and suggest that pesticides might contribute to the observed biological changes.

We did not find a statistically significant relationship between pesticide pollution and catchment size. However, previous studies showed that small streams

are more polluted than bigger streams (Knauer, 2016; Schulz, 2004; Stehle and Schulz, 2015b). This can be explained by the relatively short gradient of catchment sizes in our dataset, with most of the streams with catchments above 10 km^2 and below 100 km^2 (Figure 3.3, top). For example, the gradient of Schulz, (2004) covered 6 orders of magnitude.

Effect of precipitation on pesticide risk

Our results revealed that pesticide sampling for chemical monitoring in Germany is mainly performed when no precipitation occurs. Nevertheless, we found a 36% higher RQ if samples were taken after rainfall events. Samples taken on the day of a rainfall event showed less risk than samples taken one day after a rainfall event. This could be explained by the sampling preceding the rainfall event and the delay between the start of a rain event and the peak in discharge or runoff. The effects of precipitation were more pronounced for the probability to exceed LOQ, with smaller effect sizes for the absolute value of RQ. This may be explained by a higher variability of absolute concentrations. Overall, our results indicate that current pesticide monitoring relying on grab sampling, largely disconnected from precipitation events, underestimates pesticide risks. Automatic event-driven samplers (Stehle et al., 2013) and passive samplers (Fernández et al., 2014; Moschet et al., 2015) may help overcome these shortcomings and provide a better representation of risks, especially for small water bodies (Lorenz et al., 2016).

We found the highest the probability of exceeding LOQ during summer (10% for Q2) and lowest in the first quarter of the year (4%, Figure 3.5, bottom right). This annual pattern coincides with the main application season for pesticides in Central Europe. Nevertheless, there are compound-specific differences in the annual pattern, which explains the wide CI for the absolute RQ (Figure 3.5, bottom left). For example, the herbicide Diflufenican showed the highest RQ and the highest probability of exceeding LOQ during the winter quarters Q1 and Q4 (Supplemental Table S4), which coincides with the application period it is registered for in Germany (BVL, 2016). Our study suggests that pesticide risks display compound specific spatio-temporal dynamics. Currently, little is known about these and further research on those might provide useful information for future ecological risk assessment. For example, the sensitivity of organisms is often life stage dependent (Hutchinson et al., 1998) and knowledge on temporal dynamics could inform on concurrent exposure to multiple pesticides, as well

as assist to parameterise toxicokinetic and toxicodynamic models (Ashauer et al., 2016). Moreover, our results show that analysing absolute concentrations and probabilities of LOQ together might deliver valuable insights into risk dynamics.

Pesticides in small streams

Our results suggest that small streams are frequently exposed to ecologically relevant pesticide concentrations. In one-quarter of small streams RACs were exceeded at least once. Stehle and Schulz, (2015b) found the highest percentage of RAC exceedances for organophosphate insecticides. By contrast, we found that neonicotinoid insecticides have highest exceedances of RACs, followed by the organophosphate chlorpyrifos. This difference can be attributed to the low sample size for neonicotinoid insecticides in their study (n = 33) compared to the dataset presented here (for example 3,540 samples of Thiacloprid, Figure 3.6). Overall, our results suggest that neonicotinoids may currently pose a high risk to freshwater ecosystems. Moreover, our results add further evidence to the growing literature on the risks arising from neonicotinoids for aquatic (Morrissey et al., 2015) and terrestrial (Pisa et al., 2015) ecosystems.

Compared to Stehle and Schulz, (2015b) we found higher rates of RAC exceedances for insecticides. They found exceedances in 37.1% of insecticide measurements >LOQ (n = 1352, 23 insecticides), whereas, we found exceedances in 67% of insecticide measurements with RACs >LOQ (n = 1855, 22 insecticides). This could be attributed to different insecticides considered and different underlying RACs. Our study has only 7 insecticides with RACs in common with the insecticides investigated by Stehle and Schulz, (2015b). Moreover, all RACs were lower in our study (average difference = -0.71 µg/L, range = [-2.757; -0.005]). Nevertheless, it must be noted that the dataset compiled here comprised only samples from grab sampling, which may considerably underestimate pesticide exposure (Stehle et al., 2013; Xing et al., 2013).

By contrast, Knauer, (2016) found exceedances from monitoring data mainly for herbicides and fungicides and only one insecticide Chlorpyrifos-methyl. Moreover, RAC exceedances in Switzerland were generally lower and less abundant (for example 6 exceedances (=0.2%) for Isoproturon with a maximum RQ of 2) compared to our results for Germany. This might reflect differences in pesticide use between countries, ecoregions and RACs used. From the definition of RAC it follows that if the concentration of a compound exceeds its RAC eco-

logical effects are expected. Indeed, Stehle and Schulz, (2015a) found that the biological diversity of stream invertebrates was significantly reduced by 30% at RQ = 1.12 and by 10% at 1/10 of RAC. We found RQ values greater than 1.12 in 25% of small streams and RQ at 1/10 of RAC in 54% of small streams. Consequently, we conclude that agricultural pesticides are on a large scale a major threat to small streams, the biodiversity they host and the services they provide. This threat may exacerbate because pesticides often occur in mixtures (Schreiner et al., 2016) and may co-occur with other stressors (Schäfer et al., 2016).

Monitoring data, despite the outlined limitations, provides an opportunity to study large-scale environmental occurrence patterns of pesticides. Furthermore, such nationwide compilations, may not only be used for governmental surveillance, but also to answer other questions, like validation of exposure modelling (Knäbel et al., 2014), retrospective evaluation of regulatory risk assessment (Knauer, 2016; Stehle and Schulz, 2015b)or occurrences of pesticide mixtures (Schreiner et al., 2016). However, the sampling design needs to account for precipitation events to provide robust data. Our results suggest that exceedances of RACs are landscape dependent and therefore, pesticide regulation should account for landscape features. Moreover, the high exceedances of RACs indicate that greater efforts are needed to describe causal links, which may lead to further developments of the current authorisation procedure.

REFERENCES

Abell, R., M. L. Thieme, C. Revenga, M. Bryer, M. Kottelat, N. Bogutskaya, B. Coad, N. Mandrak, S. C. Balderas, W. Bussing, M. L. J. Stiassny, P. Skelton, G. R. Allen, P. Unmack, A. Naseka, R. Ng, N. Sindorf, J. Robertson, E. Armijo, J. V. Higgins, T. J. Heibel, E. Wikramanayake, D. Olson, H. L. López, R. E. Reis, J. G. Lundberg, M. H. Sabaj Pérez, and P. Petry (2008). "Freshwater Ecoregions of the World: A New Map of Biogeographic Units for Freshwater Biodiversity Conservation". *BioScience* 58 (5), 403–414.

AdV (2016). *ATKIS - Amtliche Geobasisdaten*. Arbeitsgemeinschaft der Vermessungsverwaltungen der Länder der Bundesrepublik Deutschland. URL: http://www.adv-online.de/AAA-Modell/ATKIS/.

Ashauer, R., C. Albert, S. Augustine, N. Cedergreen, S. Charles, V. Ducrot, A. Focks, F. Gabsi, A. Gergs, B. Goussen, T. Jager, A.-M. Nyman, V. Poulsen,

S. Reichenberger, R. B. Schäfer, P. J. Van den Brink, K. Veltman, S. Vogel, E. I. Zimmer, and T. G. Preuss (2016). "Modelling survival: exposure pattern, species sensitivity and uncertainty". *Scientific Reports* 6, 29178.

Biggs, J., P. Nicolet, M. Mlinaric, and T. Lalanne (2014). *Report of the Workshop on the Protection and Management of Small Water Bodies*. Brussels. The European Environmental Bureau (EEB) and the Freshwater Habitats Trust. Brussels. The European Environmental Bureau (EEB) and the Freshwater Habitats Trust.

Brock, T. C. M., A. Alix, C. D. Brown, E. Capri, B. E. Gottesbüren, F. Heimbach, C. M. Lythgo, R. Schulz, and M. Streloke, eds. (2010). *Linking aquatic exposure and effects: risk assessment of pesticides: EU and SETAC Europe workshop ELINK, Bari, Italy, and Wageningen, Netherlands*. Boca Raton: CRC Press.

Bundesamt für Verbraucherschutz und Lebensmittelsicherheit (2015). *Absatz an Pflanzenschutzmitteln in der Bundesrepublik Deutschland - Ergebnisse der Meldungen gemäß § 64 Pflanzenschutzgesetz für das Jahr 2014*.

BVL (2016). *Online Datenbank für zugelassene Pflanzenschutzmittel*. URL: www.bvl.bund.de/DE/04_Pflanzenschutzmittel/01_Aufgaben/02_ZulassungPSM/01_ZugelPSM/01_OnlineDatenbank/psm_onlineDB_node.html.

Davies, B. R., J. Biggs, P. J. Williams, J. T. Lee, and S. Thompson (2008). "A comparison of the catchment sizes of rivers, streams, ponds, ditches and lakes: implications for protecting aquatic biodiversity in an agricultural landscape". *Hydrobiologia* 597 (1), 7–17.

EEA (2013). *Digital Elevation Model over Europe (EU-DEM)*. URL: http://www.eea.europa.eu/data-and-maps/data/eu-dem#tab-metadata.

European Union (2009). "Directive 2009/128/EC of the European Parliament and of the Council of 21 October 2009 establishing a framework for Community action to achieve the sustainable use of pesticides". *OJ L* 309, 71–86.

Feld, C. K. (2013). "Response of three lotic assemblages to riparian and catchment-scale land use: implications for designing catchment monitoring programmes: *Response of three lotic assemblages to land use*". *Freshwater Biology* 58 (4), 715–729.

Fernández, D., E. L. Vermeirssen, N. Bandow, K. Muñoz, and R. B. Schäfer (2014). "Calibration and field application of passive sampling for episodic

exposure to polar organic pesticides in streams". *Environmental Pollution* 194, 196–202.

Fewster, R. M., S. T. Buckland, G. M. Siriwardena, S. R. Baillie, and J. D. Wilson (2000). "Analysis of population trends for farmland birds using generalized additive models". *Ecology* 81 (7), 1970–1984.

Harrison, F. (2011). "Getting started with meta-analysis". *Methods in Ecology and Evolution* 2 (1), 1–10.

Holmgren, P. (1994). "Multiple flow direction algorithms for runoff modelling in grid based elevation models: An empirical evaluation". *Hydrological Processes* 8 (4), 327–334.

Hutchinson, T. H., J. Solbe, and P. J. Kloepper-Sams (1998). "Analysis of the ecetoc aquatic toxicity (EAT) database III—comparative toxicity of chemical substances to different life stages of aquatic organisms". *Chemosphere* 36 (1), 129–142.

Illies, J. (1978). "Limnofauna Europaea".

Knäbel, A., K. Meyer, J. Rapp, and R. Schulz (2014). "Fungicide Field Concentrations Exceed FOCUS Surface Water Predictions: Urgent Need of Model Improvement". *Environmental Science & Technology* 48 (1), 455–463.

Knauer, K. (2016). "Pesticides in surface waters: a comparison with regulatory acceptable concentrations (RACs) determined in the authorization process and consideration for regulation". *Environmental Sciences Europe* 28 (13).

Liess, M., R. Schulz, M.-D. Liess, B. Rother, and R. Kreuzig (1999). "Determination of insecticide contamination in agricultural headwater streams". *Water Research* 33 (1), 239–247.

Liess, M. and P. C. von der Ohe (2005). "Analyzing effects of pesticides on invertebrate communities in streams". *Environmental Toxicology and Chemistry* 24 (4), 954–965.

Lorenz, S., J. J. Rasmussen, A. Süß, T. Kalettka, B. Golla, P. Horney, M. Stähler, B. Hommel, and R. B. Schäfer (2016). "Specifics and challenges of assessing exposure and effects of pesticides in small water bodies". *Hydrobiologia*, 1–12.

Malaj, E., P. C. v. d. Ohe, M. Grote, R. Kühne, C. P. Mondy, P. Usseglio-Polatera, W. Brack, and R. B. Schäfer (2014). "Organic chemicals jeopardize the health of freshwater ecosystems on the continental scale". *Proceedings of the National Academy of Sciences* 111 (26), 9549–9554.

Morrissey, C. A., P. Mineau, J. H. Devries, F. Sanchez-Bayo, M. Liess, M. C. Cavallaro, and K. Liber (2015). "Neonicotinoid contamination of global surface waters and associated risk to aquatic invertebrates: A review". *Environment International* 74, 291–303.

Moschet, C., E. L. Vermeirssen, H. Singer, C. Stamm, and J. Hollender (2015). "Evaluation of in-situ calibration of Chemcatcher passive samplers for 322 micropollutants in agricultural and urban affected rivers". *Water Research* 71, 306–317.

Munz, N. and C. Leu (2011). "Pestizidmessungen in Fliessgewässern - schweizweite Auswertung". *Aqua & Gas* 11, 32–41.

Nadeau, T.-L. and M. C. Rains (2007). "Hydrological Connectivity Between Headwater Streams and Downstream Waters: How Science Can Inform Policy1: Hydrological Connectivity Between Headwater Streams and Downstream Waters: How Science Can Inform Policy". *JAWRA Journal of the American Water Resources Association* 43 (1), 118–133.

Neteler, M., M. H. Bowman, M. Landa, and M. Metz (2012). "GRASS GIS: A multi-purpose open source GIS". *Environmental Modelling & Software* 31, 124–130.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner (2016). *vegan: Community Ecology Package*. R package version 2.4-1. URL: https://CRAN.R-project.org/package=vegan.

Orlinskiy, P., R. Münze, M. Beketov, R. Gunold, A. Paschke, S. Knillmann, and M. Liess (2015). "Forested headwaters mitigate pesticide effects on macroinvertebrate communities in streams: Mechanisms and quantification". *Science of The Total Environment* 524, 115–123.

Pisa, L. W., V. Amaral-Rogers, L. P. Belzunces, J. M. Bonmatin, C. A. Downs, D. Goulson, D. P. Kreutzweiser, C. Krupke, M. Liess, M. McField, C. A. Morrissey, D. A. Noome, J. Settele, N. Simon-Delso, J. D. Stark, J. P. Van der Sluijs, H. Van Dyck, and M. Wiemers (2015). "Effects of neonicotinoids and fipronil on non-target invertebrates". *Environmental Science and Pollution Research* 22 (1), 68–102.

Poisot, T. (2015). "Best publishing practices to improve user confidence in scientific software". *Ideas in Ecology and Evolution* 8.

Quevauviller, P., U. Borchers, C. Thompson, and T. Simonart (2008). *The Water Framework Directive: Ecological and Chemical Status Monitoring*. John Wiley & Sons.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. URL: https://www.R-project.org/.

Rauthe, M., H. Steiner, U. Riediger, A. Mazurkiewicz, and A. Gratzki (2013). "A Central European precipitation climatology – Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS)". *Meteorologische Zeitschrift* 22 (3), 235–256.

Rigby, R. A. and D. M. Stasinopoulos (2005). "Generalized additive models for location, scale and shape". *Applied Statistics* 54, 507–554.

Rousseeuw, P. J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of computational and applied mathematics* 20, 53–65.

Schäfer, R. B., B. Kühn, E. Malaj, A. König, and R. Gergs (2016). "Contribution of organic toxicants to multiple stress in river ecosystems". *Freshwater Biology*. DOI: 10.1111/fwb.12811.

Schäfer, R. B., P. v. d. Ohe, J. Rasmussen, J. B. Kefford, M. Beketov, R. Schulz, and M. Liess (2012). "Thresholds for the effects of pesticides on invertebrate communities and leaf breakdown in stream ecosystems". *Environmental Science & Technology* 46 (9), 5134–5142.

Schreiner, V. C., E. Szöcs, A. K. Bhowmik, M. G. Vijver, and R. B. Schäfer (2016). "Pesticide mixtures in streams of several European countries and the USA". *Science of The Total Environment* 573, 680–689.

Schulz, R. (2001). "Comparison of spray drift-and runoff-related input of azinphos-methyl and endosulfan from fruit orchards into the Lourens River, South Africa". *Chemosphere* 45 (4), 543–551.

Schulz, R. (2004). "Field Studies on Exposure, Effects, and Risk Mitigation of Aquatic Nonpoint-Source Insecticide Pollution: A Review". *Journal of Environmental Quality* 33 (2), 419–448.

Stasinopoulos, D. M. and R. A. Rigby (2007). "Generalized additive models for location scale and shape (GAMLSS) in R". *Journal of Statistical Software* 23 (7), 1–46.

Stasinopoulos, M., B. R. w. c. f. C. Akantziliotou, G. Heller, R. Ospina, N. Motpan, F. McElduff, V. Voudouris, M. Djennad, M. Enea, and A. Ghalanos (2016). *gamlss.dist: Distributions to be Used for GAMLSS Modelling*. R package version 4.3-6. URL: https://CRAN.R-project.org/package=gamlss.dist.

Statistisches Bundesamt (2014). *Bodenfläche nach Art der tatsächlichen Nutzung*. Fachserie 3 Reihe 5.1.

Stehle, S., A. Knäbel, and R. Schulz (2013). "Probabilistic risk assessment of insecticide concentrations in agricultural surface waters: a critical appraisal". *Environmental Monitoring and Assessment* 185 (8), 6295–6310.

Stehle, S. and R. Schulz (2015a). "Agricultural insecticides threaten surface waters at the global scale". *Proceedings of the National Academy of Sciences* 112 (18), 5750–5755.

Stehle, S. and R. Schulz (2015b). "Pesticide authorization in the EU—environment unprotected?" *Environmental Science and Pollution Research* 22 (24), 19632–19647.

Stone, W. W., R. J. Gilliom, and K. R. Ryberg (2014). "Pesticides in US streams and rivers: occurrence and trends during 1992–2011". *Environmental science & technology* 48 (19), 11025–11030.

Viechtbauer, W. (2010). "Conducting meta-analyses in R with the metafor package". *Journal of Statistical Software* 36 (3), 1–48.

Vijver, M. G., M. Van 'T Zelfde, W. L. Tamis, K. J. Musters, and G. R. De Snoo (2008). "Spatial and temporal analysis of pesticides concentrations in surface water: Pesticides atlas". *Journal of Environmental Science and Health, Part B* 43 (8), 665–674.

Waite, I. R. (2014). "Agricultural disturbance response models for invertebrate and algal metrics from streams at two spatial scales within the US". *Hydrobiologia* 726 (1), 285–303.

Wittmer, I. K., H. P. Bader, R. Scheidegger, H. Singer, A. Luck, I. Hanke, C. Carlsson, and C. Stamm (2010). "Significance of urban and agricultural land use for biocide and pesticide dynamics in surface waters". *Water Research* 44 (9), 2850–2862.

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Texts in statistical science. Boca Raton and Fla: Chapman & Hall/CRC.

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73 (1), 3–36.

Xing, Z., L. Chow, H. Rees, F. Meng, S. Li, B. Ernst, G. Benoy, T. Zha, and L. M. Hewitt (2013). "Influences of Sampling Methodologies on Pesticide-Residue Detection in Stream Water". *Archives of Environmental Contamination and Toxicology* 64 (2), 208–218.

4 WEBCHEM: AN R PACKAGE TO RETRIEVE CHEMICAL INFORMATION FROM THE WEB

Eduard Szöcs[a] & Ralf B. Schäfer[a]

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

Accepted in *Journal of Statistical Software*, 2016.

ABSTRACT

A wide range of chemical information is freely available online, including identifiers, experimental and predicted chemical properties. However, these data are scattered over various data sources and not easily accessible to researchers. Manual searching and downloading of such data is time-consuming and error-prone. We developed the open-source R package webchem that allows users to automatically query chemical data from currently 11 web sources. These cover a broad spectrum of information. The data are automatically imported into an R object and can directly be used in subsequent analyses. webchem enables easy, structured and reproducible data retrieval and usage from publicly available web sources. In addition, it facilitates data cleaning, identification and reporting of substances. Consequently, it reduces the time researchers need to spend on chemical data compilation.

INTRODUCTION

Before each statistical analysis, data cleaning is often required to ensure good data quality. Data cleaning is the process of detecting errors and inconsistencies in data sets (Chapman, 2005). In practice, the data cleaning step is often more time consuming than the subsequent statistical analysis, particularly, when the analysis relies on the joining of multiple data sources.

When dealing with chemical data sets (e.g. environmental monitoring data, toxicological data), a first step is often to validate the names of chemicals or to link them to unique codes that simplify subsequent querying and appending of compound-related physico-chemical or toxicological information. Several web sources provide chemical names or link them to unique codes (see also section *Data sources* below). However, manual searching for each compound, often through a graphical web interface, is tedious, error-prone and not reproducible (Peng, 2009).

To simplify, robustify and automate this task, i.e. to search and retrieve chemical information from the web, we created the webchem package for the free and open source R language (R Core Team, 2016; Wehrens, 2011). R is one of the most widely used software environments for data cleaning, analysing and visualising data, and supports full reproducibility of each step (Marwick, 2016).

In the following, we describe the basic functionality of the package and demonstrate with a few use cases how to clean and retrieve new data with webchem.

IMPLEMENTATION AND DESIGN DETAILS

The webchem package is written entirely in R and available under a MIT license. The development repository is hosted on GitHub, (2016) and a stable version is released on the official R repository (CRAN, 2016). webchem is part of the rOpenSci project (Boettiger et al., 2015), which aims at fully reproducible data analysis.

webchem follows best practices for scientific software (Poisot, 2015; Wilson et al., 2014), namely: (i) a public available repository with easy collaboration and an issue tracker (via GitHub), (ii) a non-restrictive license, version control (git), (iii) an elaborate test-suite covering more than 90% of the relevant lines of code (currently approximately 1500 lines, using testthat (Wickham, 2011)), (iv) continuous integration (via Travis-CI, (2016) and AppVeyor, (2016); testing on Linux & Windows with current and development R versions), (v) in-source documentation (using roxygen2 (Wickham et al., 2015)) and (vi) compliance with a style guide (Wickham, 2015a).

webchem builds on top of the following R packages: RCurl (Lang and Team, 2016) and httr (Wickham, 2016) for data transfer, stringr (Wickham, 2015c) for string handling, xml2 (Wickham, 2015d) and rvest (Wickham, 2015b) for parsing HTML and XML, jsonlite (Ooms, 2014) for parsing JSON, rcdk (Guha, 2007) for parsing SMILES. For parsing molfiles we use a lightweight implementation of (Grabner et al., 2012).

Some data sources provide application programming interfaces (API). Web APIs define functions that allow accessing services and data via http and return data in a specific way. webchem uses the API of a data source provider, where available. For sources where an API is lacking, data is directly searched and extracted from the web pages, analogous to manual interaction with a website.

Only few design decisions have been made: Each function name has a prefix and suffix separated by an underscore (Chamberlain and Szöcs, 2013). They follow the format of source_function, e.g. cs_compinfo uses ChemSpider as source (see next section) to retrieve compound information. Some functions require querying first a unique identifier from the data source and then use this identifier to query further information. The prefix get is used to denote these functions, e.g. get_csid to retrieve the identifier used in ChemSpider.

webchem is friendly to the resources of data providers. Between each request there is a time-out of 0.3 to 2 seconds depending on the data source. Therefore, processing of larger data sets can take some time, but still represents a major

improvement compared to manual lookup. We provide a link to the *Terms of Use* of data providers in the documentation of each function and we encourage the users to read these before using webchem. Moreover, all functions return an URL of the source, which can be used for (micro-)attribution.

DATA SOURCES

The backbone of webchem are data sources providing their data and functionality to the public. Currently, data can be retrieved from 11 sources. These cover a broad spectrum of available data, like identifiers, experimental and predicted properties and regulatory information (Figure 4.1, a detailed overview of all sources is included as supplement):

NIH CHEMICAL IDENTIFIER RESOLVER (CIR) A web service that converts from and to various chemical identifiers (NIH, 2016).

CHEMICAL TRANSLATION SERVICE (CTS) A web service that converts from and to various chemical identifiers (Wohlgemuth et al., 2010).

ETOX Information System Ecotoxicology and Environmental Quality Targets by the German Federal Environmental Agency. Provides basic identifiers, synonyms, ecotoxicological data and quality targets for different countries (UBA, 2016).

PAN PESTICIDE DATABASE Information on pesticides - provides basic identifiers, ecotoxicological data and chemical properties (PAN, 2016).

SRC PHYSPROP Contains physical properties for over 41,000 chemicals. Physical properties collected from a wide variety of sources including experimental and modeled values(Howard and Meylan, 2016).

PUBCHEM PubChem is a public repository for information on chemical substances, providing identifiers, properties and synonyms (Kim et al., 2016). We use an interface to the PUG-REST web service (Kim et al., 2015).

WIKIDATA Wikipedia contains information for over 15,000 chemicals (Ertl et al., 2015; Wikipedia, 2016). Currently webchem can only query chemical identifiers.

COMPENDIUM OF PESTICIDE COMMON NAMES The compendium provides information on pesticide common names, identifiers and classification (Wood, 2016).

CHEMID*plus* is a large web-based database provided by the National Library of Medicine (NLM). It provides identifiers, synonyms, toxicological data and chemical properties (Tomasulo, 2002).

CHEMSPIDER is a free chemical structure database providing access to over 40 million structures. It provides identifiers, properties and can also be used to convert identifiers (Pence and Williams, 2010).

OPSIN The Open Parser for Systematic IUPAC nomenclature is a chemical name interpreter and provides InChI and SMILES identifiers (Lowe et al., 2011).

Though the data sources exhibit some overlap in the provided information, each has been selected because it also provides unique information and we encourage the interested reader to consult the related source for details. However, we provide a brief overview in the Supporting Information.

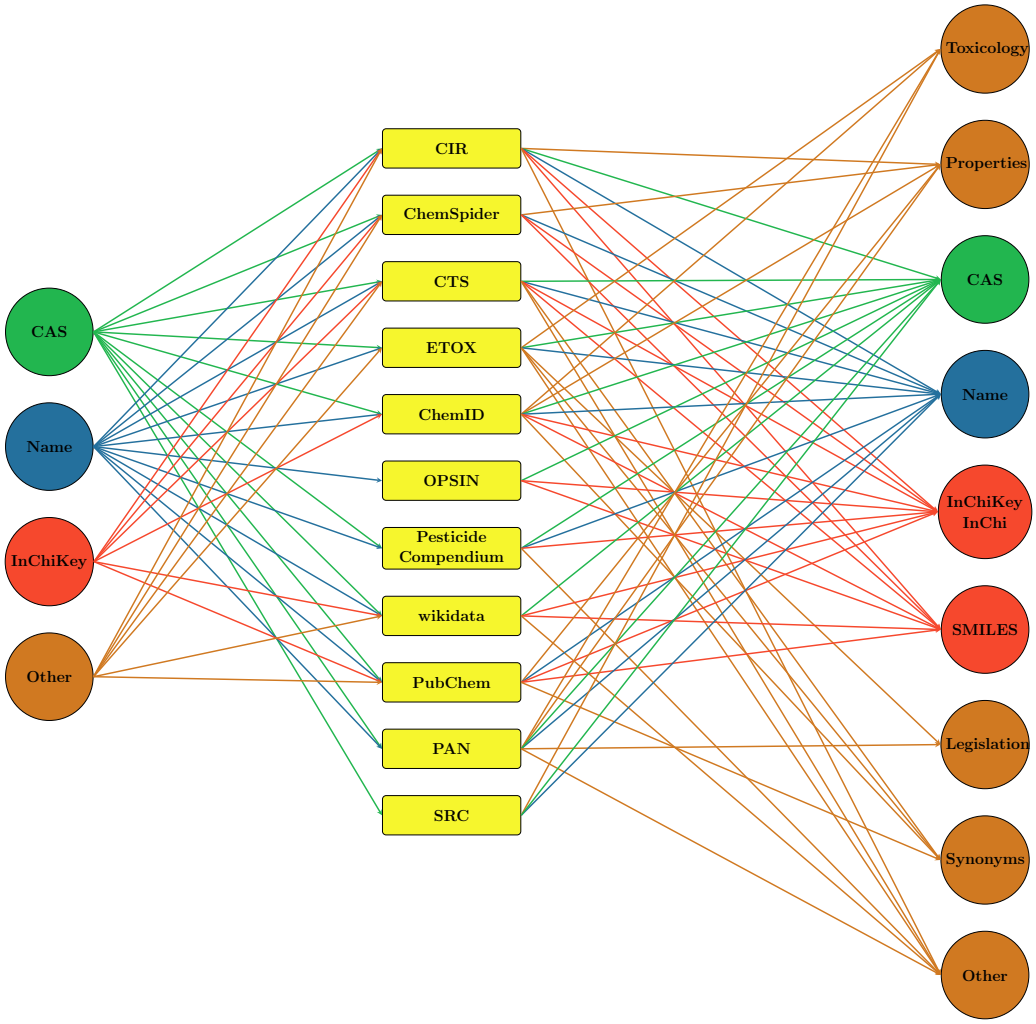


Figure 4.1.: Overview of current data sources. Input and output possibilities currently implemented in the package.

USE CASES

Installation

webchem can be easily installed and loaded from CRAN:

```
R> install.packages("webchem")
R> library("webchem")
```

The package is under active development. The latest development version is available from GitHub and also permanently available at Zenodo, (2016). This document has been created using webchem version 0.1.

Sample data sets

To demonstrate the capabilities of webchem we use two small publicly available real world data sets. The data sets are only used for purpose of demonstration, have been slightly preprocessed (not shown) and are available through the package.

(i) jagst: This data set comprises environmental monitoring data of organic substances in the river Jagst, Germany, sampled in 2013. The data is publicly available and can be retrieved from LUBW, (2016). It comprises concentrations (in µg / L) of 34 substances on 13 sampling occasions. First we load the data set and inspect the first six rows:

```
R> data("jagst")
R> head(jagst)
```

```
##         date          substance value qual
## 1 2013-01-04 2,4-Dimethylphenol 0.006    <
## 2 2013-01-29 2,4-Dimethylphenol 0.006    <
## 3 2013-02-26 2,4-Dimethylphenol 0.006    <
## 4 2013-03-26 2,4-Dimethylphenol 0.006    <
## 5 2013-04-23 2,4-Dimethylphenol 0.006    <
## 6 2013-05-22 2,4-Dimethylphenol 0.006    <
```

This data set identifies substances only by substance names. Values below the limit of quantification (LOQ) are indicated by a qualifier column.

(ii) lc50: This data consists of median acute lethal concentration for the water flea *Daphnia magna* in 48 h tests ($LC_{50,D.magna,48h}$) of 124 insecticides. The data has been retrieved from the EPA ECOTOX database (U.S. EPA, 2016).

```
R> data("lc50")
R> head(lc50)
```

```
##        cas        value
## 4  50-29-3    12.415277
## 12 52-68-6     1.282980
## 15 55-38-9    12.168138
## 18 56-23-5 35000.000000
## 21 56-38-2     1.539119
## 36 57-74-9    98.400000
```

This data set identifies the substances only by CAS numbers.

Query identifiers

The jagst data set covers 34 substances that are identified by (German) names. Merging and linking these to other tables is hampered by differences and ambiguity in compound names.

One possibility to resolve this, is to use different chemical identifiers allowing easy identification. There are several identifiers available, e.g. registry numbers like CAS or EC, database identifiers like PubChemCID (Kim et al., 2016) or ChemSpiderID (Pence and Williams, 2010), line notations like SMILES (Weininger, 1990), InChI and InChiKey (Heller et al., 2015). In this first example we query several identifiers to create a table that can be used as (i) supplemental information to a research article or (ii) to facilitate subsequent matching with other data.

As we are are dealing with German substance names we start to query ETOX for CAS registry numbers. A common work flow when dealing with web resources is to 1) query a unique identifier of the source, 2) use this identifier to retrieve additional information and 3) extract the parts that are needed from the R object (Chamberlain and Szöcs, 2013).

First we search for ETOX internal ID numbers using the substance names:

```
R> subs <- unique(jagst$substance)
R> ids <- get_etoxid(subs, match = 'best')
R> head(ids)
```

```
##   etoxid                            match distance                 query
## 1   8668     2,4-Dimethylphenol ( 8668 )        0     2,4-Dimethylphenol
## 2   8494 4-Chlor-2-methylphenol ( 8494 )        0 4-Chlor-2-methylphenol
## 3   <NA>                              <NA>     <NA>     4-para-nonylphenol
## 4   8397                Atrazin ( 8397 )        0                Atrazin
## 5   7240                 Benzol ( 7240 )        0                 Benzol
## 6   7331        Desethylatrazin ( 7331 )        0        Desethylatrazin
```

Only three substances could not be found in ETOX. Here we specify that only the *'best'* match (in terms of the Levenshtein distance between query and results) is returned. A manual check confirms appropriate matches. Other options include: *'all'* - returns all matches; *'first'* - returns only the first match (not necessarily the best match); *'ask'* - this enters an interactive mode, where the user is asked for a choice if multiple matches are found and *'na'* which returns NA in case of multiple matches.

We use these data to retrieve basic information on the substances.

```
R> etox_data <- etox_basic(ids$etoxid)
```

webchem always returns a named list (one entry for each substance) and the available information content can be very voluminous. Therefore, we provide extractor functions for the common identifiers: CAS, SMILES and InChIKeys.

```
R> etox_cas <- cas(etox_data)
R> head(etox_cas)
```

```
##        8668        8494        <NA>        8397        7240        7331
##  "105-67-9" "1570-64-5"          NA "1912-24-9"   "71-43-2" "6190-65-4"
```

A variety of data are available and we cannot provide extractor functions for each of those. Therefore, if users need to extract other data, they have to write simple extractor functions (see following examples).

In the same manner, we can now query other identifiers from another source using these CAS numbers (Figure 4.1), like PubChem

```
R> cids <- get_cid(etox_cas)
R> pc_data <- pc_prop(cids, properties = c('CanonicalSMILES'))
R> pc_smiles <- smiles(pc_data)
```

or ChemSpider

```
R> csids <- get_csid(etox_cas, token = token)
R> cs_data <- cs_compinfo(csids, token = token)
R> cs_inchikey <- inchikey(cs_data)
```

Finally, we combine the queried data into one data.frame

```
R> res <- data.frame(name = subs, cas = etox_cas, smiles = pc_smiles,
  cid = pc_data$CID, inchikey = cs_inchikey, csid = cs_data$csid,
  stringsAsFactors = FALSE)
```

Note that in order to use the ChemSpider functions, a personal authentication key (token) is needed, which can be retrieved from the ChemSpider web page. Finally, we obtain a compound table containing many different identifiers (Table 4.1), allowing easy identification and merging with other data sets, e.g. the lc50 data set based on CAS.

Name	CAS	SMILES	CID	InChIKey	CSID
2,4-Dimethylphenol	105-67-9	CC1=CC(...	7771	KUFFULV...	13839123
4-Chlor-2-methylphenol	1570-64-5	CC1=C(C...	14855	RHPUJHQ...	14165
4-para-nonylphenol	-	-	-	-	-
Atrazin	1912-24-9	CCNC1=N...	2256	MXWJVTO...	2169
Benzol	71-43-2	C1=CC=C...	241	UHOVQNZ...	236
Desethylatrazin	6190-65-4	CC(C)NC...	22563	DFWFIQK...	21157

Table 4.1.: Identifiers for the jagst data sets as queried with webchem. Only the first 6 entries are shown. For SMILES and InChIKey only the first 7 characters are shown. - = not found.

Toxicity of different pesticide groups

Another question we might ask is *How does toxicity vary between insecticide groups?* Answering this question would require tedious lookup of insecticide groups

for each of the 124 CAS numbers in the lc50 data set. The Compendium of Pesticide Common Names (Wood, 2016) contains such information and can be easily queried using CAS numbers with webchem:

```
R> aw_data <- aw_query(lc50$cas, type = 'cas')
```

To extract the chemical group from the retrieved data set, we write a simple extractor function and apply this to the retrieved data:

```
R> igroup <- sapply(aw_data, function(y) y$subactivity[1])
R> igroup[1:3]
```

```
##                                  50-29-3
##             "organochlorine insecticides"
##                                  52-68-6
##                "phosphonate insecticides"
##                                  55-38-9
## "phenyl organothiophosphate insecticides"
```

Figure 4.2 displays the result after additional data cleaning (see supplement for full code). Overall, it took only 5 R statements to retrieve, clean and plot the data using ggplot2 (Wickham, 2009).

Querying partitioning coefficients

Some data sources also provide data on chemical properties that can be queried. Here we query for the lc50 data the log $P_{oct/wat}$ from the SRC PHYSPROP database to build a simple quantitative structure–activity relationship (QSAR) to predict toxicity.

```
R> pp_data <- pp_query(lc50$cas)
```

The database contains predicted and experimental values. Extracting log $P_{oct/wat}$ from the data object is slightly more complicated, because i) for some compounds no data could be found and ii) the data-object has a more complex structure (a data frame within a list).

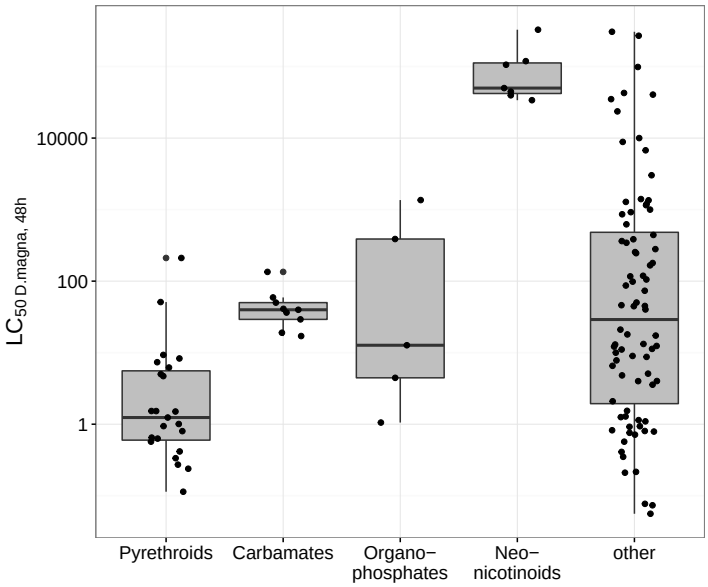


Figure 4.2.: Toxicity of different pesticide groups. LC_{50} values have been retrieved from EPA ECOTOX database, chemical groups from the Compendium of Pesticide Common Names.

```
R> lc50$logp <- sapply(pp_data, function(y) {
  if (length(y) == 1 && is.na(y))
    return(NA)
  y$prop$value[y$prop$variable == 'Log P (octanol-water)']
})
```

We opted for this more complex approach, because the information available is very diverse and we cannot provide an extractor function for each purpose. Moreover, it provides users with high flexibility regarding organisation of their data. Nevertheless, in the documentation of each function we provide examples on how to extract more complicated parts of the data. The resulting data and model is displayed in Figure 4.3.

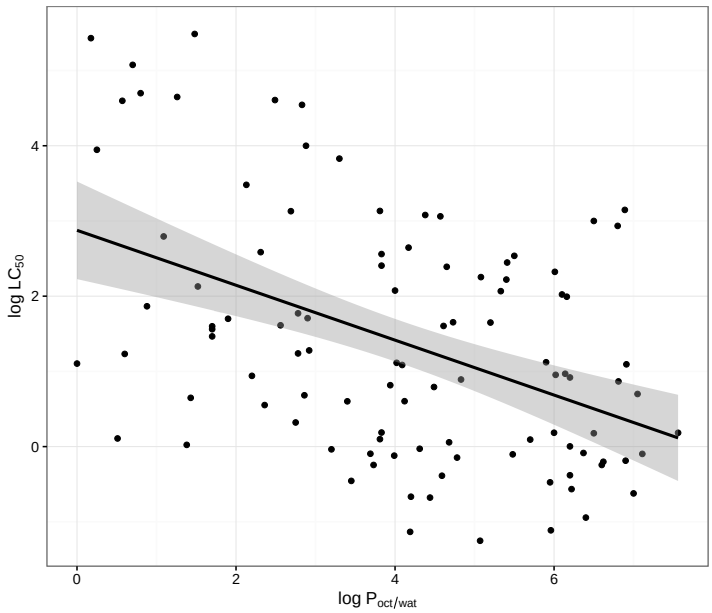


Figure 4.3.: Simple QSAR for predicting log LC_{50} of pesticides by log P. Log P values have been retrieved from SRC Physprop database (97 experimental data, 9 estimated data and 18 substances without data). The line indicates the regression model (log $LC_{50} = 2.88 - 0.37$ logP, RMSE = 1.45).

Regulatory information

Regulatory information is of particular interest if concentrations exceed national thresholds. In the European Union (EU) the Water Framework Directive (WFD, EU-WFD, (2000)) defines Environmental Quality Standards (EQS). Similarly, the U.S. and Canadian EPA and the WHO define Quality Standards. Information on these standards can be queried with webchem from the PAN Pesticide Database (using pan_query()) and from ETOX (using etox_targets()).

In this example we search for the minimum EQS for the EU for the compounds in the jagst data set, join these with measured concentrations and evaluate wether exceedances occurred..

We re-use the above queried ETOX-IDs to obtain further information from ETOX, namely the MAC-EQS:

```
R> eqs <- etox_targets(ids$etoxid)
R> ids$mac <- sapply(eqs, function(y){
  if (length(y) == 1 && is.na(y)) {
    return(NA)
  } else {
    res <- y$res
    min(res[res$Country_or_Region == 'EEC / EU' &
              res$Designation == 'MAC-EQS', 'Value_Target_LR'])
  }
})
```

Again, the returned information is humongous and we encourage users to study the returned objects and description of the data source. Here, the column Designation defines the type of EQS and Value_Target_LR contains the value. Unfortunately, we only found MAC-EQS values for 5 substances:

```
R> (mac <- with(ids, ids[!is.na(mac) & is.finite(mac),
                     c('etoxid', 'query', 'mac')]))
```

```
##    etoxid       query    mac
## 4    8397     Atrazin  2.000
## 5    7240      Benzol 50.000
## 11   8836     Irgarol  0.016
## 12   7442 Isoproturon  1.000
## 29   8756   Terbutryn  0.034
```

The get_etoxid() function used to search ETOX-IDs returns also the original substance name (query), so that we can easily join the table with MAC values with the measurements table :

```
R> jagst_eqs <- merge(jagst, mac, by.x = 'substance', by.y = 'query')
R> head(jagst_eqs)
```

```
##   substance       date  value qual etoxid mac
## 1   Atrazin 2013-09-10 0.0068    =   8397   2
## 2   Atrazin 2013-10-08 0.0072    =   8397   2
## 3   Atrazin 2013-03-26 0.0040    =   8397   2
## 4   Atrazin 2013-04-23 0.0048    =   8397   2
```

```
## 5   Atrazin 2013-11-05 0.0036    =   8397   2
## 6   Atrazin 2013-07-16 0.0052    =   8397   2
```

Finally, we can compare the measured value to the MAC, which reveals that there have been no exceedances of these 5 compounds.

Utility functions

Furthermore, webchem provides also basic functions to check identifiers that can be used for data quality assessment. The functions either use simple formatting rules,

```
R> is.inchikey('BQJCRHHNABKAKU-KBQPJGBKS-AN')

## Hyphens not at position 15 and 26.

## [1] FALSE

R> is.cas('64-17-6')

## Checksum is not correct!  5 vs.  6

## [1] FALSE
```

or web resources like ChemSpider

```
R> is.inchikey('BQJCRHHNABKAKU-KBQPJGBKSA-5',
  type = 'chemspider')

## [1] FALSE
```

DISCUSSION

Related software

Within the R ecosystem, there are only a few similar projects: rpubchem (Guha, 2014) provides an interface to PubChem. Similarly, ChemmineR (Cao et al., 2008), a mature chemo-informatics package, provides an interface to Pubchem. webchem does not provide any chemo-informatic functionality, but integrates access to many data sources. WikidataR (Keyes and Graul, 2016) provides an interface to wikidata that could be used to retrieve chemical data from Wikipedia. However, it does not provide predefined methods for chemical data like webchem. Within the Python ecosystem the libraries PubChempy (Swain, 2015b), ChemSpiPy (Swain, 2015a) and CIRpy (Swain, 2016) are available for similar tasks as those outlined here. webchem is not specialized and tries to integrate many data sources and for some of these it provides a unique programmatic interface. The Chemical Translation Service (Wohlgemuth et al., 2010), which is also one of the sources that can be queried, allows batch conversion of chemical identifiers. However, it does not provide access to other data (experimental, modeled or regulatory data).

Open Science

An increasing number of scientific data is becoming publicly available (Gewin, 2016; O'Boyle et al., 2011; Reichman et al., 2011), either in public data repositories or as supplement to publications. To be usable for other researchers chemical compounds should be properly identified, not only by chemical names but also with accompanying identifiers like InChIKey, SMILES and authority-assigned identifiers. webchem provides an easy way to create such meta tables as shown in Table 4.1 and facilitates chemical data availability to researchers. However, good quality of data is crucial for every analysis (Stieger et al., 2014) and additional effort and methods are needed to validate data quality.

Further development

We have outlined only a few use cases that will likely be useful for many researchers. Given the huge amount of publicly available information, many other

possibilities can be envisioned. webchem is currently under active development and several other data sources have not been implemented yet but may be in the future. GitHub makes contributing easy and we strongly encourage contribution to the package. Moreover, comments, feedback and feature requests are highly welcome.

CONCLUSIONS

Researchers need to have easy access to global knowledge on chemicals. webchem can save *"hundreds of working hours"* gathering this knowledge (Münch and Galizia, 2016), so that researchers can focus on other tasks.

REFERENCES

AppVeyor (2016). URL: https://www.appveyor.com/.

Boettiger, C., S. Chamberlain, E. Hart, and K. Ram (2015). "Building Software, Building Community: Lessons from the ROpenSci Project". *Journal of Open Research Software* 3 (1).

Cao, Y, Charisi, A, Cheng, L. C, Jiang, T, Girke, and T (2008). "ChemmineR: A Compound Mining Framework for R". *Bioinformatics* 24 (15), 1733–1734.

Chamberlain, S. A. and E. Szöcs (2013). "taxize: Taxonomic Search and Retrieval in R". *F1000Research* 2 (191).

Chapman, A. (2005). *Principles and Methods of Data Cleaning*. Report for the Global Biodiversity Information Facility, Copenhagen. GBIF. URL: http://www.gbif.org/orc/?doc_id=1262.

CRAN (2016). *webchem: Retrieve Chemical Information from the Web*. URL: https://CRAN.R-project.org/package=webchem.

Ertl, P., L. Patiny, T. Sander, C. Rufener, and M. Zasso (2015). "Wikipedia Chemical Structure Explorer: Substructure and Similarity Searching of Molecules from Wikipedia". *Journal of Cheminformatics* 7 (1).

Gewin, V. (2016). "Data sharing: An Open Mind on Open Data". *Nature* 529 (7584), 117–119.

GitHub (2016). *webchem: Retrieve Chemical Information from the Web*. URL: https://github.com/ropensci/webchem.

Grabner, M., K. Varmuza, and M. Dehmer (2012). "RMol: A Toolset for Transforming SD/Molfile Structure Information into R Objects". *Source Code for Biology and Medicine* 7, 12.

Guha, R. (2007). "Chemical Informatics Functionality in R". *Journal of Statistical Software* 18 (5), 1–16.

Guha, R. (2014). *rpubchem: Interface to the PubChem Collection*. R package version 1.5.0.2. URL: https://CRAN.R-project.org/package=rpubchem.

Heller, S. R., A. McNaught, I. Pletnev, S. Stein, and D. Tchekhovskoi (2015). "InChI, the IUPAC International Chemical Identifier". *Journal of Cheminformatics* 7 (1).

Howard, P. H. and W. Meylan (2016). *Physical / Chemical Property Database (PHYSPROP)*. URL: http://www.srcinc.com/what-we-do/environmental/scientific-databases.html.

Keyes, O. and C. Graul (2016). *WikidataR: API Client Library for Wikidata*. R package version 1.0.1. URL: https://CRAN.R-project.org/package=WikidataR.

Kim, S., P. A. Thiessen, E. E. Bolton, and S. H. Bryant (2015). "PUG-SOAP and PUG-REST: Web Services for Programmatic Access to Chemical Information in PubChem". *Nucleic Acids Research* 43 (W1), W605–W611.

Kim, S., P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, and et al. (2016). "PubChem Substance and Compound Databases". *Nucleic Acids Research* 44 (D1), D1202–D1213.

Lang, D. T. and t. C. Team (2016). *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. R package version 1.95-4.8. URL: http://CRAN.R-project.org/package=RCurl.

Lowe, D. M., P. T. Corbett, P. Murray-Rust, and R. C. Glen (2011). "Chemical Name to Structure: OPSIN, an Open Source Solution". *Journal of Chemical Information and Modeling* 51 (3), 739–753.

LUBW - Landesanstalt für Umwelt, M. u. N. B.-W. (2016). *Jahresdatenkatalog Fließgewässer 2013*. URL: http://jdkfg.lubw.baden-wuerttemberg.de/servlet/is/300/.

Marwick, B. (2016). "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation". *Journal of Archaeological Method and Theory*, 1–27.

Münch, D. and C. G. Galizia (2016). "DoOR 2.0 - Comprehensive Mapping of Drosophila Melanogaster Odorant Responses". *Scientific Reports* 6, 21841.

NIH (2016). *NIH Chemical Identifier Resolver*. URL: http://cactus.nci.nih.gov/chemical/structure.

O'Boyle, N. M., R. Guha, E. L. Willighagen, S. E. Adams, J. Alvarsson, J.-C. Bradley, I. V. Filippov, R. M. Hanson, M. D. Hanwell, G. R. Hutchison, and et al. (2011). "Open Data, Open Source and Open Standards in Chemistry: The Blue Obelisk Five Years On." *Journal of Cheminformatics* 3, 37.

Ooms, J. (2014). "The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects". *arXiv preprint*. URL: http://arxiv.org/abs/1403.2805.

PAN (2016). *Pesticide Action Network(PAN) Pesticide Database*. URL: http://www.pesticideinfo.org/.

Pence, H. E. and A. Williams (2010). "ChemSpider: An Online Chemical Information Resource". *Journal of Chemical Education* 87 (11), 1123–1124.

Peng, R. D. (2009). "Reproducible Research and Biostatistics". *Biostatistics* 10 (3), 405–408.

Poisot, T. (2015). "Best publishing practices to improve user confidence in scientific software". *Ideas in Ecology and Evolution* 8.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Reichman, O. J., M. B. Jones, and M. P. Schildhauer (2011). "Challenges and Opportunities of Open Data in Ecology". *Science* 331 (6018), 703–5.

Stieger, G., M. Scheringer, C. A. Ng, and K. Hungerbühler (2014). "Assessing the Persistence, Bioaccumulation Potential and Toxicity of Brominated Flame Retardants: Data Availability and Quality for 36 Alternative Brominated Flame Retardants". *Chemosphere* 116, 118–123.

Swain, M. (2015a). *ChemSpiPy*. URL: https://github.com/mcs07/ChemSpiPy.

Swain, M. (2015b). *PubChemPy*. URL: https://github.com/mcs07/PubChemPy.

Swain, M. (2016). *CIRpy*. URL: https://github.com/mcs07/CIRpy.

Tomasulo, P. (2002). "ChemIDplus - Super Source for Chemical and Drug Information". *Medical Reference Services Quarterly* 21 (1), 53–59.

Travis-CI (2016). URL: https://travis-ci.org/.

UBA (2016). *ETOX: Information System Ecotoxicology and Environmental Quality Targets*. URL: https://webetox.uba.de/webETOX/index.do.

U.S. EPA (2016). *ECOTOX database*. URL: http://cfpub.epa.gov/ecotox/.

Wehrens, R. (2011). *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer.

Weininger, D. (1990). "SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures". *Journal of Chemical Information and Computer Sciences* 30(3), 237–243.

EU-WFD (2000). "Directive 2000/60/EC of the European Parliament and of the Council Establishing a Framework for the Community Action in the Field of Water Policy". *The European Parliament and Council* (L327/1).

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.

Wickham, H. (2011). "testthat: Get Started with Testing". *The R Journal* 3, 5–10.

Wickham, H. (2015a). *Advanced R*. The R Series. CRC Press.

Wickham, H. (2015b). *rvest: Easily Harvest (Scrape) Web Pages*. R package version 0.3.1. URL: https://CRAN.R-project.org/package=rvest.

Wickham, H. (2015c). *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.0.0. URL: http://CRAN.R-project.org/package=stringr.

Wickham, H. (2015d). *xml2: Parse XML*. R package version 0.1.2. URL: https://CRAN.R-project.org/package=xml2.

Wickham, H. (2016). *httr: Tools for Working with URLs and HTTP*. R package version 1.1.0. URL: https://CRAN.R-project.org/package=httr.

Wickham, H., P. Danenberg, and M. Eugster (2015). *roxygen2: In-Source Documentation for R*. R package version 5.0.1. URL: http://CRAN.R-project.org/package=roxygen2.

Wikipedia (2016). *WikiProject Chemistry*. URL: https://www.wikidata.org/wiki/Wikidata:WikiProject_Chemistry.

Wilson, G., D. A. Aruliah, C. T. Brown, N. P. Chue Hong, M. Davis, R. T. Guy, S. H. D. Haddock, K. D. Huff, I. M. Mitchell, M. D. Plumbley, B. Waugh, E. P. White, and P. Wilson (2014). "Best Practices for Scientific Computing". *PLoS Biology* 12 (1). Ed. by J. A. Eisen, e1001745.

Wohlgemuth, G., P. K. Haldiya, E. Willighagen, T. Kind, and O. Fiehn (2010). "The Chemical Translation Service – a Web-Based Tool to Improve Standardization of Metabolomic Reports". *Bioinformatics* 26 (20), 2647–2648.

Wood, A. (2016). *Compendium of Pesticide Common Names*. URL: http://www.alanwood.net/pesticides/index.

Zenodo (2016). *webchem: Retrieve Chemical Information from the Web*. URL: http://dx.doi.org/10.5281/zenodo.33823.

5 TAXIZE: TAXONOMIC SEARCH AND RETRIEVAL IN R

Scott A. Chamberlain[a] & Eduard Szöcs[b]

[a]Biology Department, Simon Fraser University, Burnaby, BC, Canada,
[b]Institute for Environmental Sciences, University Koblenz-Landau, Landau, Germany

Adapted from the article published in 2013 in *F1000Research*, 2.191.

This chapter reflects the software state in 2013. In the meantime there have been many changes to taxize, so that not all parts presented here and thre respective supplementary materials work anymore. For a more recent description please visit the project homepage https://github.com/ropensci/taxize.

ABSTRACT

All species are hierarchically related to one another, and we use taxonomic names to label the nodes in this hierarchy. Taxonomic data is becoming increasingly available on the web, but scientists need a way to access it in a programmatic fashion that's simple and reproducible. We have developed taxize, an open-source software package for the R language (freely available from http://cran.r-project.org/web/packages/taxize). taxize provides simple, programmatic access to taxonomic data for 13 data sources around the web. We discuss the need for a taxonomic toolbelt in R, and outline a suite of use cases for which taxize is ideally suited (including a full workflow as an appendix). The taxize package facilitates open and reproducible science by allowing taxonomic data collection to be done in the open-source R platform.

INTRODUCTION

Evolution by natural selection has led to a hierarchical relationship among all living organisms. Thus, species are categorized using a taxonomic hierarchy, starting with the binomial species name (e.g, *Homo sapiens*), moving up to genus (*Homo*), then family (*Hominidae*), and on up to Domain (*Eukarya*). Although taxonomic classifications are human constructs created to understand the real phylogeny of life (Benton, 2000), they are nonetheless essential to organize the vast diversity of organisms. Biologists, whether studying organisms at the cell, organismal, or community level, can put their study objects into taxonomic context, allowing them to infer close and distant relatives, find relevant literature, and more.

The use of taxonomic names is, unfortunately, not straightforward. Taxonomic names often vary due to name revisions at the generic or specific levels, lumping or splitting lower taxa (genera, species) among higher taxa (families), and name spelling changes. For example, a study found that a compilation of 308,000 plant observations from 51 digitized herbarium records had 22,100 unique taxon names, of which only 13,000 were accepted names (Boyle et al., 2013; Weiser et al., 2007). In addition, there is no one authoritative source of taxonomic names for all taxa - although, there are taxon specific sources that are used by many scientists. Different sources (e.g., uBio [Universal Biological Indexer and Organizer], Tropicos, ITIS [Integrated Taxonomic Information Service]) may use different accepted names for the same taxon. For exam-

ple, while ITIS has *Helianthus x glaucus* as an accepted name, The Plant List (http://www.theplantlist.org) gives that name as unresolved. But *Helianthus glaucus* is an accepted name in The Plant List, while ITIS does not list this name.

One attempt to help inconsistencies in taxonomy is the use of numeric codes. For example, ITIS assigns a Taxonomic Serial Number (TSN) to each taxon, while uBio assigns each taxon a NameBank identifier (namebankID), and Tropicos assigns their own identifier to each taxon. Codes are helpful within a database as they can easily refer to, for example, *Helianthus annuus* with a code like 123456 instead of its whole name. However, each database uses their own code; in this case for *Helianthus annuus*, ITIS uses 36616, uBio uses 2658020, and Tropicos uses 40022652. As there are no universal codes for taxa across databases, this can lead to additional confusion. Last, name comparisons across databases have to be done with the actual names, not the codes.

Taxonomic data is getting easier to obtain through the web (e.g., http://eol.org/). However, there are a number of good reasons to obtain taxonomic information programatically rather than through a web interface. First, if you have more than a few names to look up on a website, it can take quite a long time to enter each name, get data, and repeat for each species. Programatically getting taxonomic names solves the problem by looping over a list of names. In addition, doing taxonomic searching, etc. becomes reproducible. With increasing reports of irreproducibility in science (Stodden, 2010; Zimmer, 2012), it is extremely important to make science workflows repeatable.

The R language is widely used by biologists, and now has over 5,000 packages on the Comprehensive R Archive Network (CRAN) to extend R. R is great for manipulating, visualizing and fitting statistical models to data. Gentleman et al. Gentleman et al., (2004) give a detailed discussion of advantages of R in computational biology. Getting data from the web will be increasingly common as more and more data gets moved to the cloud. Therefore, there is a need to get data from the web directly into R. Increasingly, data is available from the web via application programming interfaces (API). These allow computers to talk to one another using code that is not human readable, but is machine readable. Web APIs often define a number of methods that allow users to search for a species name, or retrieve the synonyms for a species name, for example. A further advantage of APIs is that they are language agnostic, meaning that data can be consumed in almost any computing context, allowing users to interact with the web API without having to know the details of the code. Moreover

data can be accessed from every computer, whereas for example an Excel file can only be opened in a few programs.

The goal of taxize is to make many use cases that involve retrieving and resolving taxonomic names easy and reproducible. In taxize, we have written a suite of R functions that interact with many taxonomic data sources via their web APIs (Table 5.1). The interface to each function is usually a simple list of species names, just as a user would enter when interacting with a website. Therefore, we hope that moving from a web to an R interface for taxonomic names will be relatively seamless (if one is already nominally familiar with R).

Here, we justify the need for programmatic taxonomic resolution tools like taxize, discuss our data sources, and run through a suite of use cases to demonstrate the variety of ways that users can use taxize.

Table 5.1.: Some key functions in taxize, what they do, and their data sources

Function name	What it does	Source
apg_lookup()	Changes names to match the APGIII list	Angiosperm Phylogeny Group http://www.mobot.org/MOBOT/research/APweb/
classification()	Upstream classification	Various
col_children()	Direct children	Catalogue of Life http://www.catalogueoflife.org/
col_downstream()	Downstream taxa to specified rank	Catalogue of Life http://www.catalogueoflife.org/
eol_hierarchy()	Upstream classification	Encyclopedia of Life http://eol.org/
eol_search()	Search EOL taxon information	Encyclopedia of Life http://eol.org/
get_seqs()	Get NCBI sequences	National Center for Biotechnology Information (Federhen, 2012)
get_tsn()	Get ITIS TSN	Integrated Taxonomic Information System http://www.itis.gov/
get_uid()	Get NCBI UID	National Center for Biotechnology Information (Federhen, 2012)

Table 5.1 – *Cont.*

Function name	What it does	Source
gisd_isinvasive()	Invasiveness status	Global Invasive Species Database http://www.issg.org/database/welcome/
gni_parse()	Parse scientific names into components	Global Names Index http://gni.globalnames.org/
gni_search()	Search EOL's global names index	Global Names Index http://gni.globalnames.org/
gnr_resolve()	Resolve names using EOL's global names index	Global Names Resolver http://resolver.globalnames.org/
itis_downstream()	Downstream taxa to specified rank	Integrated Taxonomic Information System http://www.itis.gov/
iucn_status()	IUCN status	IUCN Red List http://www.iucnredlist.org
phylomatic_tree()	Get a plant Phylogeny	Phylomatic (Webb and Donoghue, 2005)
plantminer()	Search Plantminer	Plantminer (Carvalho et al., 2010)
searchby-commonname()	Search ITIS by common name	Integrated Taxonomic Information System http://www.itis.gov/
searchby-scientificname()	Search ITIS by scientific name	Integrated Taxonomic Information System http://www.itis.gov/
tax_name()	Get taxonomic name for specific rank	Various
tax_rank()	Get rank of a taxonomic name	Various
tnrs()	Resolve names using iPlant	iPlant Taxonomic Name Resolution Service http://tnrs.iplantcollaborative.org/
tp_accepted-names()	Check for accepted names using Tropicos	Tropicos http://www.tropicos.org/
tpl_search()	Search the Plant List	The Plant List http://www.theplantlist.org

Table 5.1 – *Cont.*

Function name	What it does	Source
ubio_namebank()	Search uBio	uBio http://www.ubio.org/index.php?pagename=sample_tools

WHY DO WE NEED TAXIZE?

There is a large suite of applications developed around the problem of searching for, resolving, and getting higher taxonomy for species names. For example, Linnaeus (http://linnaeus.sourceforge.net/) provides the ability to search for taxonomic names in documents and normalize those names found. In addition, there are many web interfaces to search for and normalize names such as Encyclopedia of Life's Global Names Resolver (http://resolver.globalnames.org/), uBio tools (www.ubio.org/index.php?pagename=sample_tools), and iPlant's Taxonomic Name Resolution Service (http://tnrs.iplantcollaborative.org/).

All of these data repositories provide ways to search for taxonomic names and resolve them in some cases. However, scientists ideally need a tool that is free and can be used programmatically, thereby facilitating reproducible research. The goal of taxize is to facilitate the creation of reproducible and easy to use workflows for searching for taxonomic names, resolving them, getting higher taxonomic names, and other tasks related to research dealing with species.

DATA SOURCES AND PACKAGE DETAILS

taxize uses many data sources (Table 5.1), and more can be easily added. There are two common tasks provided by the data sources: name search and name resolution. Other functionality in taxize includes retrieving a classification tree for a species, or retrieving child taxa of a focal taxon. O ne of the data sources (Phylomatic) returns phylogenies, while another (NCBI) returns genetic sequence data. However, there are other R packages that are focused solely on sequence data, such as rsnps (Chamberlain and Ushey, 2013), rentrez (Winter, 2013), BoSSA (Lefeuvre, 2010), and ape (Paradis et al., 2004), so taxize does not venture deeply into these other domains.

Some of the data sources taxize interacts with require authentication. That is, in addition to the search terms the user provides (e.g., *Homo sapiens*), the data provider requires an alphanumeric identification key. This is necessary in some cases so that API providers can 1) better prevent databases crashing from too many requests, 2) collect analytics on requests to their API to provide better performance, etc., and 3) provide user level modification of rules for interacting with the API. The services that require an API key in taxize are: Encyclopedia of Life (EOL) (`http://eol.org/`), the Universal Biological Indexer and Organizer (uBio) (`http://www.ubio.org/index.php?pagename=sample_tools`), Tropicos (`http://www.tropicos.org/`), and Plantminer (Carvalho et al., 2010). One can easily obtain API keys by visiting the website of each service (see Table 5.1 for links to each site). There are two typical ways of using API keys. First, you can pass in your API key in a function call (e.g., *ubio_namebank(srchName='Ursus americanus', key='your_alphanumeric_key')*). Second, you can store your key in the .Rprofile file, which is a common place to store settings. We recommend the second option as it simplifies function calls as taxize detects the stored keys.

taxize would not have been possible without the work of others. taxize uses httr (Wickham, 2012a) and RCurl (Lang, 2013a) for performing calls to web APIs, XML (Lang, 2013c) for parsing XML, RJSONIO (Lang, 2013b) for parsing JSON, and stringr (Wickham, 2012b) and plyr (Wickham, 2011) for manipulating data.

New data sources can be added: for example, we plan to add the following sources: Wikispecies (`https://species.wikimedia.org`) and The Tree of Life (`http://tolweb.org/tree/`). A connection to `www.freshwaterecology.info` (a database with autecological characteristics, ecological preferences and biological traits as well as distribution patterns of more than 12,000 European freshwater organisms belonging to fish, macro-invertebrates, macrophytes, diatoms and phytoplankton) will be finished when their new API is released. In addition, the authors welcome further suggestions of data sources to be added.

USE CASES

First, install taxize

First, one must install and load taxize into the R session.

```
R> install.packages("taxize")
R> library(taxize)
```

Advanced users can also download and install the latest development copy from GitHub https://github.com/ropensci/taxize, also permanently available at http://dx.doi.org/10.5281/zenodo.7097.

Resolve taxonomic names

This is a common task in biology. We often have a list of species names and we want to know a) if we have the most up to date names, b) if our names are spelled correctly, and c) the scientific name for a common name. One way to resolve names is via the Global Names Resolver (GNR) service provided by the Encyclopedia of Life (http://resolver.globalnames.org/). Here, on can search for two misspelled names:

```
R> temp <- gnr_resolve(names = c("Helianthos annus",
                                 "Homo saapiens"))
R> temp[ , -c(1,4)]

#                  matched_name       data_source_title
# 1        Helianthus annuus L.       Catalogue of Life
# 2            Helianthus annus GBIF Taxonomic Backbone
# 3            Helianthus annus                     EOL
# 4         Helianthus annus L.                     EOL
# 5            Helianthus annus           uBio NameBank
# 6 Homo sapiens Linnaeus, 1758       Catalogue of Life
```

The correct spellings are *Helianthus annuus* and *Homo sapiens*. Another approach uses the Taxonomic Name Resolution Service via the Taxosaurus API (http://taxosaurus.org/) developed by iPLant and the Phylotastic organization. In this example is a list of species names, some of which are misspelled, and then call the API with the *tnrs* function.

```
R> mynames <- c("Helianthus annuus", "Pinus contort",
              "Poa anua", "Abis magnifica", "Rosa california",
              "Festuca arundinace", "Sorbus occidentalos",
              "Madia sateva")
tnrs(query = mynames)[ , -c(5:7)]

#          submittedName        acceptedName    sourceId score
# 9    Helianthus annuus   Helianthus annuus iPlant_TNRS     1
# 10   Helianthus annuus   Helianthus annuus        NCBI     1
```

```
# 4        Pinus contort      Pinus contorta iPlant_TNRS  0.98
# 5             Poa anua           Poa annua iPlant_TNRS  0.96
# 3       Abis magnifica     Abies magnifica iPlant_TNRS  0.96
# 7      Rosa california    Rosa californica iPlant_TNRS  0.99
# 8      Rosa california          California        NCBI     1
# 2   Festuca arundinace Festuca arundinacea iPlant_TNRS  0.99
# 1  Sorbus occidentalos Sorbus occidentalis iPlant_TNRS  0.99
# 6         Madia sateva        Madia sativa iPlant_TNRS  0.97
```

It turns out there are a few corrections: e.g., *Madia sateva* should be *Madia sativa*, and *Rosa californica* should be *Rosa californica*. Note that this search worked because fuzzy matching was employed to retrieve names that were close, but not exact matches. Fuzzy matching is only available for plants in the TNRS service, so we advise using EOL's Global Names Resolver if you need to resolve animal names.

taxize takes the approach that the user should be able to make decisions about what resource to trust, rather than making the decision on behalf of the user. Both the EOL GNR and the TNRS services provide data from a variety of data sources. The user may trust a specific data source, and thus may want to use the names from that data source. In the future, we may provide the ability for taxize to suggest the best match from a variety of sources.

Another common use case is when there are many synonyms for a species. In this example, there are six synonyms of the currently accepted name for a species.

```
R> library(plyr)
R> mynames <- c("Helianthus annuus ssp. jaegeri",
              "Helianthus annuus ssp. lenticularis",
              "Helianthus annuus ssp. texanus",
              "Helianthus annuus var. lenticularis",
              "Helianthus annuus var. macrocarpus",
              "Helianthus annuus var. texanus")
R> tsn <- get_tsn(mynames)
R> ldply(tsn, itis_acceptname)

#   submittedTsn      acceptedName acceptedTsn
# 1       525928 Helianthus annuus       36616
# 2       525929 Helianthus annuus       36616
# 3       525930 Helianthus annuus       36616
# 4       536095 Helianthus annuus       36616
```

```
# 5       536096 Helianthus annuus       36616
# 6       536097 Helianthus annuus       36616
```

Retrieve higher taxonomic names

Another task biologists often face is getting higher taxonomic names for a taxa list. Having the higher taxonomy allows you to put into context the relationships of your species list. For example, you may find out that species A and species B are in Family C, which may lead to some interesting insight, as opposed to not knowing that Species A and B are closely related. This also makes it easy to aggregate/standardize data to a specific taxonomic level (e.g., family level) or to match data to other databases with different taxonomic resolution (e.g., trait databases).

A number of data sources in taxize provide the capability to retrieve higher taxonomic names, but we will highlight two of the more useful ones: Integrated Taxonomic Information System (ITIS) (http://www.itis.gov/) and National Center for Biotechnology Information (NCBI) (Federhen, 2012). First, search for two species, *Abies procera* and *Pinus contorta* within ITIS.

```
R> specieslist <- c("Abies procera", "Pinus contorta")
R> classification(specieslist, db = "itis")
```

```
# $'Abies procera'
#         rankName       taxonName    tsn
# 1        Kingdom         Plantae 202422
# 2     Subkingdom  Viridaeplantae 846492
# 3   Infrakingdom    Streptophyta 846494
# 4       Division    Tracheophyta 846496
# 5    Subdivision Spermatophytina 846504
# 6  Infradivision    Gymnospermae 846506
# 7          Class       Pinopsida 500009
# 8          Order         Pinales 500028
# 9         Family        Pinaceae  18030
# 10         Genus           Abies  18031
# 11       Species   Abies procera 181835
#
# $'Pinus contorta'
#         rankName       taxonName    tsn
# 1        Kingdom         Plantae 202422
```

```
# 2     Subkingdom  Viridaeplantae 846492
# 3   Infrakingdom    Streptophyta 846494
# 4       Division    Tracheophyta 846496
# 5    Subdivision Spermatophytina 846504
# 6  Infradivision    Gymnospermae 846506
# 7          Class       Pinopsida 500009
# 8          Order         Pinales 500028
# 9         Family        Pinaceae  18030
# 10         Genus           Pinus  18035
# 11       Species  Pinus contorta 183327
```

It turns out both species are in the family Pinaceae. You can also get this type of information from the NCBI by excuting the following code in R: *classification(specieslist, db = 'ncbi')*.

Instead of a full classification, you may only want a single name, say a family name for your species of interest. The function *tax_name* is built just for this purpose. As with the *classification*-function you can specify the data source with the *db* argument, either ITIS or NCBI.

```
R> tax_name(query = "Helianthus annuus", get = "family",
         db = "itis")

#       family
# 1 Asteraceae

R> tax_name(query = "Helianthus annuus", get = "family",
         db = "ncbi")

#       family
# 1 Asteraceae
```

If a data source does not provide information on the queried species, the result could be taken from another source and the results from the different sources could be pooled.

Interactive name selection

As mentioned previously most databases use a numeric code to reference a species. A general workflow in taxize is: Retrieve Code for the queried species and then use this code to query more data/information. Below are a few examples. When you run these examples in R, you are presented with a command

prompt asking for the row that contains the name you would like back; that output is not printed below for brevity. In this example, the search term has many matches. The function returns a data.frame of the matches, and asks for the user to input which row number to accept.

```
R> get_tsn(searchterm = "Heliastes", searchtype = "sciname")

#            combinedname    tsn
# 1     Heliastes bicolor 615238
# 2   Heliastes chrysurus 615250
# 3     Heliastes cinctus 615573
# 4  Heliastes dimidiatus 615257
# 5  Heliastes hypsilepis 615273
# 6 Heliastes immaculatus 615639
# 7 Heliastes opercularis 615300
# 8      Heliastes ovalis 615301
#  1
# NA
# attr(,"class")
# [1] "tsn"
```

In another example, you can pass in a long character vector of taxonomic names:

```
R> splist <- c("annona cherimola", 'annona muricata',
            "quercus robur", "shorea robusta",
            "pandanus patina", "oryza sativa",
            "durio zibethinus")
R> get_tsn(searchterm = splist, R> searchtype = "sciname")

# [1] "506198" "18098"  "19405"  "506787" "507376" "41976"
# [7] "506099"
# attr(,"class")
# [1] "tsn"
```

In another example, note that no match at all returns an NA:

```
R> get_uid(sciname = c("Chironomus riparius", "aaa vva"))

# [1] "315576" NA
# attr(,"class")
# [1] "uid"
```

Retrieve a phylogeny

Ecologists are increasingly taking a phylogenetic approach to ecology, applying phylogenies to topics such as the study of community structure (Webb et al., 2002), ecological networks (Rafferty and Ives, 2013), functional trait ecology (Poff et al., 2006). Yet, Many biologists are not adequately trained in reconstructing phylogenies. Fortunately, there are some sources for getting a phylogeny without having to know how to build one; one of these is for angiosperms, called Phylomatic (Webb and Donoghue, 2005). We have created a workflow in taxize that accepts a species list, and taxize works behind the scenes to get higher taxonomic names, which are required by Phylomatic to get a phylogeny. Here is a short example, producing the tree in figure 5.1.

```
R> taxa <- c("Poa annua", "Abies procera", "Helianthus annuus")
R> tree <- phylomatic_tree(taxa = taxa)
R> tree$tip.label <- capwords(tree$tip.label)
R> plot(tree, cex = 1)
```

Behind the scenes the function *phylomatic_tree* retrieves a Taxonomic Serial Number (TSN) from ITIS for each species name, then a string is created for each species like this *poaceae/oryza/oryzasativa* (with format "family/genus/genus_epithet"). These strings are submitted to the Phylomatic API, and if no errors occur, a phylogeny in newick format is returned. The *phylomatic_tree()* function also cleans up the newick string and converts it to an ape *phylo* object, which can be used for plotting and phylogenetic analyses. Be aware that Phylomatic has certain limitations - refer to the paper describing Phylomatic (Webb and Donoghue, 2005) and the website http://phylodiversity.net/phylomatic/.

What taxa are children of the taxon of interest?

If someone is not a taxonomic specialist on a particular taxon they probably do not know what children taxa are within a family, or within a genus. This task becomes especially unwieldy when there are a large number of taxa downstream. You can of course go to a website like Wikispecies (http://species.wikimedia.org) or Encyclopedia of Life (http://eol.org/) to get downstream names. However, taxize provides an easy way to programatically search for downstream taxa, both for the Catalogue of Life (CoL) (http://www.catalogueoflife.org/) and the Integrated Taxonomic Information System (http://www.itis.gov/).

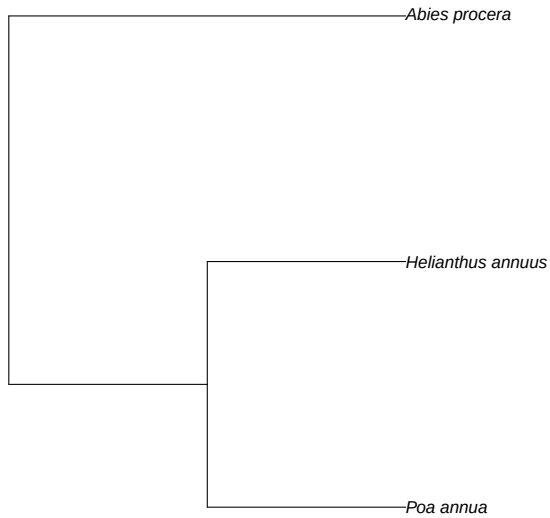


Figure 5.1.: A phylogeny for three species. This phylogeny was produced using the *phylomatic_tree* function, which queries the Phylomatic database, and prunes a previously created phylogeny of plants.

Here is a short example using the CoL in which we want to find all the species within the genus *Apis* (honey bees).

```
R> col_downstream(name = "Apis", downto = "Species")[[1]]

#   childtaxa_id     childtaxa_name childtaxa_rank
# 1      6971712 Apis andreniformis        Species
# 2      6971713        Apis cerana        Species
# 3      6971714       Apis dorsata        Species
# 4      6971715        Apis florea        Species
# 5      6971716 Apis koschevnikovi        Species
# 6      6845885     Apis mellifera        Species
# 7      6971717   Apis nigrocincta        Species
```

The result from the above call to *col_downstream()* is a data.frame that gives a number of columns of different information.

IUCN Status

There are a number of things a user can do once they have the correct taxonomic names. One thing a user can do is ask about the conservation status of a species (IUCN Red List of Threatened Species (http://www.iucnredlist.org)). We have provided a set of functions, *iucn_summary* and *iucn_status*, to search for species names, and extract the status information, respectively. Here, you can search for the panther and lynx.

```
R> ia <- iucn_summary(c("Panthera uncia", "Lynx lynx"))
R> iucn_status(ia)

# Panthera uncia      Lynx lynx
#           "EN"           "LC"
```

It turns out that the panther has a status of endangered (EN) and the lynx has a status of least concern (LC).

Search for available genes in GenBank

Another use case available in taxize deals with genetic sequences. taxize has three functions to interact with GenBank to search for available genes (*get_genes_avail*), download genes by GenBank ID (*get_genes*), and download genes via taxonomic name search, including retrieving a congeneric if the searched taxon does not exist in the database (*get_seqs*). In this example, one can search for gene sequences for *Umbra limi*.

```
R> out <- get_genes_avail(taxon_name = "Umbra limi",
                          seqrange = "1:2000",
                          getrelated = FALSE)
```

Then one can ask if 'RAG1' exists in any of the gene names.

```
R> out[grep("RAG1", out$genesavail, ignore.case = TRUE), -3]

#         spused length access_num       ids
# 413 Umbra limi    732   JX190826 394772608
# 427 Umbra limi    959   AY459526  45479841
# 434 Umbra limi   1631   AY380548  38858304
```

It turns out that there are 430 different unique records found. However, this doesn't mean that there are 430 different genes found as the API does not provide metadata to classify genes. You can use regular expressions (e.g., *grep*) to search for the gene of interest.

Matching species tables with different taxonomic resolution

Biologists often need to match different sets of data tied to species. For example, trait-based approaches are a promising tool in ecology (Statzner and Bêche, 2010). One problem is that abundance data must be matched with trait databases such as the NCBI Taxonomy database (Usseglio-Polatera et al., 2000). These two data tables may contain species information on different taxonomic levels and data might have to be aggregated to a joint taxonomic level, so that the data can be merged. taxize can help in this data-cleaning step, providing a reproducible workflow.

A user can use the mentioned *classification*-function to retrieve the taxonomic hierarchy and then search the hierarchies up- and downwards for matches. Here is an example to match a species (A) with names of on different taxonomic levels (B1 & B2).

```
R> A <- "gammarus roeseli"
R> B1 <- "gammarus"
R> B2 <- "gammaridae"
R> A_clas <- classification(A, db = 'ncbi')
R> B1_clas <- classification(B1, db = 'ncbi')
R> B2_clas <- classification(B2, db = 'ncbi')
R> A_clas[[1]]$Rank[tolower(A_clas[[1]]$ScientificName) %in% B1]

# [1] "genus"

R> A_clas[[1]]$Rank[tolower(A_clas[[1]]$ScientificName) %in% B2]

# [1] "family"
```

If one finds a direct match (here *Gammarus roeseli*), they will be lucky. However, Gammaridae can also be matched with *Gammarus roeseli*, but on a lower taxonomic level. A more comprehensive and realistic example (matching a trait table with an abundance table) is given in the supplemental materials.

Aggregating data to a specific taxonomic rank

In biology, one can ask questions at varying taxonomic levels. This use case is easily handled in taxize. A function called *tax_agg()* will aggregate community data to a specific taxonomic level. In this example, one can take the data for three species and aggregate them to family level. Again one can specify whether they want to use data from ITIS or NCBI. The rows in the *data.frame* are different communities.

```
R> data(dune, package = 'vegan')
R> df <- dune[ , c(1,3:4)]
R> colnames(df) <- c("Bellis perennis", "Juncus bufonius",
                     "Juncus articulatus")
R> head(df)

#    Bellis perennis Juncus bufonius Juncus articulatus
# 2                3               0                  0
# 13               0               3                  0
# 4                2               0                  0
# 16               0               0                  3
# 6                0               0                  0
# 1                0               0                  0


R> agg <- tax_agg(df, rank = 'family', db = 'ncbi')
R> agg

#
#  Aggregated community data
#
# Level of Aggregation: FAMILY
# No. taxa before aggregation: 3
# No. taxa after aggregation: 2
# No. taxa not found: 0
```

```
R> head(agg$x)

#    Asteraceae Juncaceae
# 2           3         0
# 13          0         3
# 4           2         0
# 16          0         3
# 6           0         0
# 1           0         0
```

The two *Juncus* species are aggregated to the family Juncaceae and their abundances are summed. There was only a single species in the family Asteraceae, so the data for *Bellis perennis* are carried over.

CONCLUSIONS

Taxonomic information is increasingly sought by biologists as we take phylogenetic and taxonomic approaches to science. Taxonomic data are becoming more widely available on the web, yet scientists require programmatic access to this data for developing reproducible workflows. taxize was created to bridge this gap - to bring taxonomic data on the web into R, where the data can be easily manipulated, visualized, and analyzed in a reproducible workflow.

We have outlined a suite of use cases in taxize that will likely fit real use cases for many biologists. Of course we have not thought of all possible use cases, so we hope that the biology community can give us feedback on what use cases they want to see available in taxize. One thing we could change in the future is to make functions that fit use cases, and then allow users to select the data source as a parameter in the function. This could possibly make the user interface easier to understand.

taxize is currently under development and will be for some time given the large number of data sources knitted together in the package, and the fact that APIs for each data source can change, requiring changes in taxize code. Contributions to taxize are strongly encouraged, and can be easily done using GitHub here: https://github.com/ropensci/taxize. We hope taxize will be taken up by the community and developed collaboratively, making it progressively better through time as new use cases arise, bug reports are squashed, and contributions are merged.

REFERENCES

Benton, M. J. (2000). "Stems, nodes, crown clades, and rank-free lists: is Linnaeus dead?" *Biological Reviews* 75 (4), 633–648.

Boyle, B., N. Hopkins, Z. Lu, J. A. Raygoza Garay, D. Mozzherin, T. Rees, N. Matasci, M. L. Narro, W. H. Piel, S. J. Mckay, and et al. (2013). "The taxonomic name resolution service: an online tool for automated standardization of plant names". *BMC Bioinformatics* 14 (1), 1.

Carvalho, G. H., M. V. Cianciaruso, and M. A. Batalha (2010). "Plantminer: a web tool for checking and gathering plant species taxonomic information". *Environmental Modelling & Software* 25 (6), 815–816.

Chamberlain, S. and K. Ushey (2013). *rsnps: Interface to SNP data on the web.* R package version 0.0.4. URL: https://github.com/ropensci/rsnps.

Federhen, S. (2012). "The NCBI Taxonomy database". *Nucleic Acids Research* 40 (D1), D136–D143.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, et al. (2004). "Bioconductor: open software development for computational biology and bioinformatics". *Genome biology* 5 (10), 1.

Lang, D. T. (2013a). *RCurl: General network (HTTP/FTP/...) client interface for R.* R package version 1.95-4.1. URL: http://CRAN.R-project.org/package=RCurl.

Lang, D. T. (2013b). *RJSONIO: Serialize R objects to JSON, JavaScript Object Notation*. R package version 1.0-3. URL: http://CRAN.R-project.org/package=RJSONIO.

Lang, D. T. (2013c). *XML: Tools for parsing and generating XML within R and S-Plus.* R package version 3.95-0.2. URL: http://CRAN.R-project.org/package=XML.

Lefeuvre, P. (2010). *BoSSA: a Bunch of Structure and Sequence Analysis*. R package version 1.2. URL: http://CRAN.R-project.org/package=BoSSA.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: analyses of phylogenetics and evolution in R language". *Bioinformatics* 20, 289–290.

Poff, N. L., J. D. Olden, N. K. Vieira, D. S. Finn, M. P. Simmons, and B. C. Kondratieff (2006). "Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships". *Journal of the North American Benthological Society* 25 (4), 730–755.

Rafferty, N. E. and A. R. Ives (2013). "Phylogenetic trait-based analyses of ecological networks". *Ecology* 94 (10), 2321–2333.

Statzner, B. and L. Bêche (2010). "Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems?" *Freshwater Biology* 55, 80–119.

Stodden, V. C. (2010). "Reproducible research: Addressing the need for data and code sharing in computational science". *Computing in Science & Engineering* 12 (5), 8–12.

Usseglio-Polatera, P., M. Bournaud, P. Richoux, and H. Tachet (2000). "Biological and ecological traits of benthic freshwater macroinvertebrates: relationships and definition of groups with similar traits". *Freshwater Biology* 43 (2), 175–205.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue (2002). "Phylogenies and community ecology". *Annual Review of Ecology and Systematics*, 475–505.

Webb, C. O. and M. J. Donoghue (2005). "Phylomatic: tree assembly for applied phylogenetics". *Molecular Ecology Notes* 5 (1), 181–183.

Weiser, M. D., B. J. Enquist, B. Boyle, T. J. Killeen, P. M. Jørgensen, G. Fonseca, M. D. Jennings, A. J. Kerkhoff, T. E. Lacher Jr, A. Monteagudo, and et al. (2007). "Latitudinal patterns of range size and species richness of New World woody plants". *Global Ecology and Biogeography* 16 (5), 679–688.

Wickham, H. (2011). "The Split-Apply-Combine Strategy for Data Analysis". *Journal of Statistical Software* 40 (1), 1–29.

Wickham, H. (2012a). *httr: Tools for working with URLs and HTTP*. R package version 0.2. URL: http://CRAN.R-project.org/package=httr.

Wickham, H. (2012b). *stringr: Make it easier to work with strings.* R package version 0.6.2. URL: http://CRAN.R-project.org/package=stringr.

Winter, D. (2013). *rentrez: Entrez in R*. R package version 0.2.1. URL: https://github.com/ropensci/rentrez.

Zimmer, C. (2012). "A Sharp Rise in Retractions Prompts Calls for Reform". *New York Times*. URL: http://www.nytimes.com/2012/04/17/science/rise-in-scientific-journal-retractions-prompts-calls-for-reform.html?_r=0.

6 GENERAL DISCUSSION AND OUTLOOK

STATISTICAL ECOTOXICOLOGY

The simulation study performed in chapter 2 clearly showed that common experimental designs exhibit unacceptably low statistical power (Szöcs and Schäfer, 2016; Van Der Hoeven, 1998). This underpins the criticism accumulated over the last 30 years towards the usage of NOEC as endpoint (Fox and Landis, 2016). Nevertheless, the NOEC is still one of the standard endpoint for mesocosm experiments in higher tier risk assessment (EFSA, 2013).

Recently, *a posteriori* calculations of statistical power have been proposed to counteract these limitations and aid the interpretation treatment-related effects in model ecosystems (Brock et al., 2015). The "minimum detectable difference" (MDD) estimates the difference between to means that must exist in order to produce a statistically significant result ($p < 0.05$ (Gelman and Stern, 2006)) and could be used to ínterpret NOEC. However, *a posteriori* calculations have been shown to have logical flaws when used for interpretation of non-significant results (Hoenig and Heisey, 2001; Nakagawa and Foster, 2004). However, conducting and report of *a priori* power calculations, as performed in chapter 2, might provide researchers important information to optimize their study designs, ensuring that their experimental designs have appropriate power (Johnson et al., 2015).

Moreover, similar simulations can not only be used to analyse data of factorial designs, but also from regression designs. Indeed, simulations could be used to determine optimal designs for dose-response models and EC_x determination, balancing precision and resources. Regression designs are generally more powerful and provide more information than factorial designs (Cottingham et al., 2005). Regression designs in mesocosm experiments, assigning the replicates to more tested concentrations, might also provide more insights. However, currently statistical tools to analyse a community dose-reponse relations, providing a $EC_{x,community}$ are not well explored. Separate dose-response models could be fit to each species (Ritz, 2010), leading to a EC_x for each species in a mesocosm study. Subsequently, this EC_x values could be combined and summarised

using Species Sensitivity Distributions (Posthuma et al., 2002), providing a hazardous concentration ($HC_{x,community}$) for x % of species affected (Maltby et al., 2005). Another possibility would be to use a logistic type of ordination (van den Brink et al., 2003). Reduced-Rank vector generalized linear models (RR-VGLM) could be used to fit such type of models (Yee, 2015; Yee and Hastie, 2003) but they have not been applied in ecotoxicology yet.

In a similar vein, community ecology is currently experiencing a shift towards new class of multivariate methods, incorporating statistical models for abundances across many taxa simultaneously (ter Braak and Šmilauer, 2015; Warton et al., 2015a; Warton et al., 2015b; Warton et al., 2012). However, this methods have not been applied frequently and their applicability to ecotoxicological data is currently unclear (Szöcs et al., 2015). All this models have in common, that the choice of statistical model is primarily based ...

LEVERAGING MONITORING DATA FOR ECOLOGICAL RISK ASSESSMENT

CHALLENGES TO UTILIZE 'BIG DATA' IN ECOLOGICAL RISK ASSESSMENT

CONCLUSIONS

REFERENCES

Brock, T. C. M., M. Hammers-Wirtz, U. Hommen, T. G. Preuss, H.-T. Ratte, I. Roessink, T. Strauss, and P. J. Van den Brink (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems". *Environmental Science and Pollution Research* 22 (2), 1160–1174.

Cottingham, K. L., J. T. Lennon, and B. L. Brown (2005). "Knowing when to draw the line: designing more informative ecological experiments". *Frontiers in Ecology and the Environment* 3 (3), 145–152.

EFSA (2013). "Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters". *EFSA Journal* 11 (7), 3290.

Fox, D. R. and W. G. Landis (2016). "Comment on ET&C perspectives, November 2015-A holistic view". *Environmental Toxicology and Chemistry* 35 (6), 1337–1339.

Gelman, A. and H. Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant". *The American Statistician* 60 (4), 328–331.

Hoenig, J. M. and D. M. Heisey (2001). "The abuse of power". *The American Statistician* 55 (1), 19–24.

Johnson, P. C. D., S. J. E. Barry, H. M. Ferguson, and P. Müller (2015). "Power analysis for generalized linear mixed models in ecology and evolution". *Methods in Ecology and Evolution* 6 (2), 133–142.

Maltby, L., N. Blake, T. C. M. Brock, and P. J. Van Den Brink (2005). "Insecticide species sensitivity distributions: Importance of test species selection and relevance to aquatic ecosystems". *Environmental Toxicology and Chemistry* 24 (2), 379–388.

Nakagawa, S. and T. M. Foster (2004). "The case against retrospective statistical power analyses with an introduction to power analysis". *acta ethologica* 7 (2), 103–108.

Posthuma, L., G. W. Suter, and T. P. Traas (2002). *Species sensitivity distributions in ecotoxicology*. Environmental and ecological risk assessment. Boca Raton and Fla: Lewis.

Ritz, C. (2010). "Toward a unified approach to dose-response modeling in ecotoxicology". *Environmental Toxicology and Chemistry* 29 (1), 220–229.

Szöcs, E., P. J. v. d. Brink, L. Lagadic, T. Caquet, M. Roucaute, A. Auber, Y. Bayona, M. Liess, P. Ebke, A. Ippolito, C. J. F. t. Braak, T. C. M. Brock, and R. B. Schäfer (2015). "Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods". *Ecotoxicology* 24 (4), 760–769.

Szöcs, E. and R. B. Schäfer (2016). "Statistical hypothesis testing—To transform or not to transform?" *Integrated Environmental Assessment and Management* 12 (2), 398–400.

Ter Braak, C. J. and P. Šmilauer (2015). "Topics in constrained and unconstrained ordination". *Plant Ecology* 216 (5), 683–696.

Van den Brink, P. J., N. W. van den Brink, and C. J. F. ter Braak (2003). "Multivariate analysis of ecotoxicological data using ordination: demonstrations of utility on the basis of various examples". *Australasian Journal of Ecotoxicology* 9. RS, 141–156.

Van Der Hoeven, N. (1998). "Power analysis for the NOEC: What is the probability of detecting small toxic effects on three different species using the appropriate standardized test protocols?" *Ecotoxicology* 7 (6), 355–361.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui (2015a). "So Many Variables: Joint Modeling in Community Ecology". *Trends in Ecology & Evolution* 30 (12), 766–779.

Warton, D. I., S. D. Foster, G. De'ath, J. Stoklosa, and P. K. Dunstan (2015b). "Model-based thinking for community ecology". *Plant Ecology* 216 (5), 669–682.

Warton, D. I., S. T. Wright, and Y. Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects". *Methods in Ecology and Evolution* 3 (1), 89–101.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models*. Springer Series in Statistics. New York, NY: Springer New York.

Yee, T. W. and T. J. Hastie (2003). "Reduced-rank vector generalized linear models". *Statistical modelling* 3 (1), 15–41.

SUPPLEMENTAL MATERIALS

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

SUPPLEMENTARY TABLES

Table A.1.: Count data simulations - Proportion of models converged. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p
3.00	2.00	1.00	0.33	1.00	1.00
3.00	4.00	1.00	0.53	1.00	1.00
3.00	8.00	1.00	0.79	1.00	1.00
3.00	16.00	1.00	0.94	1.00	1.00
3.00	32.00	1.00	0.99	1.00	1.00
3.00	64.00	1.00	1.00	1.00	1.00
3.00	128.00	1.00	1.00	1.00	1.00
6.00	2.00	1.00	0.63	1.00	1.00
6.00	4.00	1.00	0.85	1.00	1.00
6.00	8.00	1.00	0.98	1.00	1.00
6.00	16.00	1.00	1.00	1.00	1.00
6.00	32.00	1.00	1.00	1.00	1.00
6.00	64.00	1.00	1.00	1.00	1.00
6.00	128.00	1.00	1.00	1.00	1.00
9.00	2.00	1.00	0.76	1.00	1.00
9.00	4.00	1.00	0.95	1.00	1.00
9.00	8.00	1.00	1.00	1.00	1.00
9.00	16.00	1.00	1.00	1.00	1.00
9.00	32.00	1.00	1.00	1.00	1.00
9.00	64.00	1.00	1.00	1.00	1.00
9.00	128.00	1.00	1.00	1.00	1.00

Table A.2.: Count data simulations - Power to detect a treatment effect. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_{qp} = Poisson model, np = pairwise Wilcoxon test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_{p}	np	NA
3.00	2.00	0.13	0.17	0.17	0.08	0.36	0.04
3.00	4.00	0.14	0.18	0.17	0.10	0.54	0.06
3.00	8.00	0.19	0.36	0.24	0.21	0.78	0.09
3.00	16.00	0.23	0.49	0.33	0.29	0.95	0.14
3.00	32.00	0.31	0.57	0.38	0.35	0.99	0.16
3.00	64.00	0.32	0.58	0.38	0.34	1.00	0.18
3.00	128.00	0.35	0.61	0.42	0.37	1.00	0.19
6.00	2.00	0.26	0.30	0.29	0.22	0.49	0.21
6.00	4.00	0.36	0.48	0.44	0.40	0.78	0.32
6.00	8.00	0.48	0.64	0.57	0.53	0.94	0.44
6.00	16.00	0.59	0.76	0.70	0.65	0.99	0.54
6.00	32.00	0.68	0.82	0.76	0.73	1.00	0.63
6.00	64.00	0.72	0.85	0.80	0.77	1.00	0.64
6.00	128.00	0.73	0.84	0.80	0.76	1.00	0.63
9.00	2.00	0.34	0.40	0.42	0.35	0.64	0.31
9.00	4.00	0.56	0.69	0.66	0.63	0.91	0.54
9.00	8.00	0.70	0.82	0.79	0.76	0.98	0.68
9.00	16.00	0.81	0.91	0.89	0.88	1.00	0.79
9.00	32.00	0.89	0.95	0.94	0.92	1.00	0.87
9.00	64.00	0.92	0.96	0.95	0.95	1.00	0.89
9.00	128.00	0.94	0.97	0.96	0.95	1.00	0.91

Table A.3.: Count data simulations - Power to detect LOEC. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model, np = pairwise Wilcoxon test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p	np
3.00	2.00	0.05	0.01	0.02	0.02	0.00
3.00	4.00	0.08	0.09	0.08	0.15	0.00
3.00	8.00	0.11	0.22	0.12	0.30	0.00
3.00	16.00	0.13	0.30	0.18	0.42	0.00
3.00	32.00	0.17	0.35	0.22	0.50	0.00
3.00	64.00	0.19	0.37	0.23	0.51	0.00
3.00	128.00	0.18	0.37	0.23	0.53	0.00
6.00	2.00	0.14	0.11	0.09	0.15	0.06
6.00	4.00	0.17	0.23	0.19	0.30	0.12
6.00	8.00	0.28	0.39	0.32	0.52	0.20
6.00	16.00	0.33	0.48	0.39	0.59	0.23
6.00	32.00	0.40	0.54	0.47	0.64	0.28
6.00	64.00	0.44	0.56	0.48	0.61	0.29
6.00	128.00	0.44	0.57	0.49	0.56	0.29
9.00	2.00	0.19	0.20	0.18	0.26	0.13
9.00	4.00	0.29	0.37	0.31	0.48	0.27
9.00	8.00	0.40	0.52	0.46	0.62	0.35
9.00	16.00	0.51	0.63	0.57	0.70	0.45
9.00	32.00	0.57	0.69	0.63	0.68	0.52
9.00	64.00	0.61	0.72	0.66	0.65	0.53
9.00	128.00	0.65	0.73	0.68	0.61	0.58

Table A.4.: Count data simulations - Type 1 error to detect a global treatment effect. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_{pb} = negative binomial model with parametric bootstrap, GLM_p = Poisson model, np = Kruskal-Wallis test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_{pb}	GLM_p	np
3.00	2.00	0.07	0.04	0.02	0.07	0.21	0.03
3.00	4.00	0.05	0.07	0.03	0.05	0.37	0.01
3.00	8.00	0.04	0.12	0.05	0.05	0.58	0.02
3.00	16.00	0.05	0.14	0.05	0.05	0.84	0.02
3.00	32.00	0.04	0.13	0.03	0.04	0.94	0.01
3.00	64.00	0.05	0.16	0.05	0.05	0.99	0.03
3.00	128.00	0.05	0.13	0.05	0.06	1.00	0.02
6.00	2.00	0.04	0.05	0.04	0.06	0.20	0.03
6.00	4.00	0.05	0.08	0.05	0.05	0.36	0.04
6.00	8.00	0.06	0.09	0.05	0.06	0.58	0.04
6.00	16.00	0.05	0.08	0.05	0.05	0.80	0.04
6.00	32.00	0.06	0.08	0.05	0.06	0.94	0.04
6.00	64.00	0.05	0.09	0.05	0.05	0.98	0.04
6.00	128.00	0.05	0.09	0.04	0.05	1.00	0.04
9.00	2.00	0.06	0.06	0.05	0.07	0.20	0.05
9.00	4.00	0.04	0.08	0.05	0.06	0.36	0.04
9.00	8.00	0.05	0.08	0.05	0.06	0.58	0.04
9.00	16.00	0.04	0.07	0.04	0.05	0.81	0.04
9.00	32.00	0.04	0.06	0.04	0.06	0.94	0.05
9.00	64.00	0.04	0.07	0.05	0.05	0.99	0.04
9.00	128.00	0.05	0.07	0.05	0.06	1.00	0.04

Table A.5.: Count data simulations - Type 1 error to detect LOEC. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model, np = pairwise Wilcoxon.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p	np
3.00	2.00	0.05	0.02	0.02	0.02	0.00
3.00	4.00	0.04	0.08	0.04	0.14	0.00
3.00	8.00	0.05	0.11	0.06	0.24	0.00
3.00	16.00	0.03	0.11	0.04	0.36	0.00
3.00	32.00	0.04	0.15	0.05	0.55	0.00
3.00	64.00	0.05	0.16	0.06	0.61	0.00
3.00	128.00	0.04	0.13	0.05	0.68	0.00
6.00	2.00	0.04	0.04	0.02	0.07	0.02
6.00	4.00	0.03	0.06	0.03	0.15	0.02
6.00	8.00	0.04	0.08	0.05	0.26	0.03
6.00	16.00	0.04	0.08	0.05	0.37	0.03
6.00	32.00	0.04	0.08	0.04	0.52	0.03
6.00	64.00	0.05	0.10	0.05	0.61	0.04
6.00	128.00	0.04	0.08	0.04	0.66	0.05
9.00	2.00	0.03	0.05	0.04	0.08	0.03
9.00	4.00	0.04	0.06	0.05	0.15	0.04
9.00	8.00	0.04	0.05	0.04	0.27	0.04
9.00	16.00	0.04	0.07	0.04	0.38	0.03
9.00	32.00	0.03	0.05	0.04	0.49	0.03
9.00	64.00	0.04	0.06	0.04	0.61	0.04
9.00	128.00	0.04	0.06	0.04	0.67	0.04

Table A.6.: Binomial data simulations - Power to detect a global treatment effect. N = sample sizes, p_E = probability in effect treatments, LM = Linear model after transformation, GLM = binomial model, np = Kruskal-Wallis test.

N	p_E	LM	GLM	np
3.00	0.60	0.97	1.00	0.87
3.00	0.65	0.90	0.99	0.76
3.00	0.70	0.78	0.95	0.60
3.00	0.75	0.60	0.84	0.41
3.00	0.80	0.36	0.64	0.22
3.00	0.85	0.20	0.41	0.10
3.00	0.90	0.11	0.17	0.05
3.00	0.95	0.06	0.06	0.03
6.00	0.60	1.00	1.00	1.00
6.00	0.65	1.00	1.00	1.00
6.00	0.70	1.00	1.00	1.00
6.00	0.75	0.97	1.00	0.97
6.00	0.80	0.85	0.93	0.82
6.00	0.85	0.53	0.62	0.48
6.00	0.90	0.17	0.24	0.15
6.00	0.95	0.04	0.08	0.03
9.00	0.60	1.00	1.00	1.00
9.00	0.65	1.00	1.00	1.00
9.00	0.70	1.00	1.00	1.00
9.00	0.75	1.00	1.00	1.00
9.00	0.80	0.98	0.99	0.97
9.00	0.85	0.75	0.82	0.73
9.00	0.90	0.26	0.32	0.23
9.00	0.95	0.05	0.07	0.04

Table A.7.: Count data simulations - Power to detect LOEC. N = sample sizes, p_E = probability in effect treatments, LM = Linear model after transformation, GLM = binomial model, np = pairwise Wilcoxon.

N	p_E	LM	GLM	np
3.00	0.60	0.86	0.70	0.00
3.00	0.65	0.74	0.57	0.00
3.00	0.70	0.59	0.40	0.00
3.00	0.75	0.41	0.17	0.00
3.00	0.80	0.23	0.04	0.00
3.00	0.85	0.11	0.01	0.00
3.00	0.90	0.05	0.00	0.00
3.00	0.95	0.01	0.00	0.00
6.00	0.60	0.98	0.95	0.97
6.00	0.65	0.97	0.93	0.91
6.00	0.70	0.93	0.90	0.82
6.00	0.75	0.82	0.78	0.62
6.00	0.80	0.60	0.55	0.36
6.00	0.85	0.33	0.19	0.16
6.00	0.90	0.08	0.01	0.03
6.00	0.95	0.01	0.00	0.00
9.00	0.60	0.97	0.95	0.97
9.00	0.65	0.98	0.96	0.98
9.00	0.70	0.97	0.96	0.96
9.00	0.75	0.94	0.93	0.89
9.00	0.80	0.82	0.81	0.73
9.00	0.85	0.46	0.43	0.35
9.00	0.90	0.13	0.08	0.08
9.00	0.95	0.01	0.00	0.00

Table A.8.: Binomial data simulations - Type 1 error to detect a global treatment effect. N = sample sizes, p = probability, LM = Linear model after transformation, GLM = binomial model, np = Kruskal-Wallis test.

N	p	LM	GLM	np
3.00	0.60	0.05	0.06	0.02
3.00	0.65	0.06	0.06	0.02
3.00	0.70	0.04	0.05	0.02
3.00	0.75	0.06	0.05	0.02
3.00	0.80	0.05	0.07	0.02
3.00	0.85	0.06	0.07	0.02
3.00	0.90	0.05	0.08	0.01
3.00	0.95	0.06	0.07	0.02
6.00	0.60	0.06	0.06	0.04
6.00	0.65	0.04	0.05	0.03
6.00	0.70	0.04	0.05	0.04
6.00	0.75	0.05	0.05	0.03
6.00	0.80	0.06	0.06	0.04
6.00	0.85	0.04	0.06	0.04
6.00	0.90	0.06	0.06	0.04
6.00	0.95	0.05	0.08	0.03
9.00	0.60	0.05	0.05	0.04
9.00	0.65	0.06	0.06	0.05
9.00	0.70	0.06	0.05	0.05
9.00	0.75	0.05	0.05	0.05
9.00	0.80	0.06	0.07	0.06
9.00	0.85	0.04	0.05	0.04
9.00	0.90	0.06	0.07	0.05
9.00	0.95	0.06	0.06	0.04

Table A.9.: Binomial data simulations - Type 1 error to detect LOEC. N = sample sizes, p = probability, LM = Linear model after transformation, GLM = binomial model, np = pairwise Wilcoxon.

N	p_E	LM	GLM	np
3.00	0.60	0.03	0.03	0.00
3.00	0.65	0.04	0.03	0.00
3.00	0.70	0.04	0.03	0.00
3.00	0.75	0.04	0.03	0.00
3.00	0.80	0.03	0.01	0.00
3.00	0.85	0.04	0.01	0.00
3.00	0.90	0.03	0.00	0.00
3.00	0.95	0.05	0.00	0.00
6.00	0.60	0.05	0.06	0.02
6.00	0.65	0.03	0.04	0.01
6.00	0.70	0.05	0.04	0.02
6.00	0.75	0.03	0.03	0.02
6.00	0.80	0.04	0.04	0.01
6.00	0.85	0.03	0.02	0.01
6.00	0.90	0.05	0.01	0.01
6.00	0.95	0.05	0.00	0.01
9.00	0.60	0.04	0.04	0.04
9.00	0.65	0.04	0.03	0.04
9.00	0.70	0.05	0.04	0.05
9.00	0.75	0.03	0.04	0.02
9.00	0.80	0.04	0.04	0.03
9.00	0.85	0.04	0.03	0.03
9.00	0.90	0.04	0.03	0.03
9.00	0.95	0.05	0.00	0.01

WORKED R EXAMPLES

Count data example

Introduction

In this example we will analyse data from (Brock et al., 2015). The data are count of mayfly larvae in Macroinvertebrate Artificial Substrate Samplers in 18 mesocosms at one sampling day. There are 5 treatments and one control group.

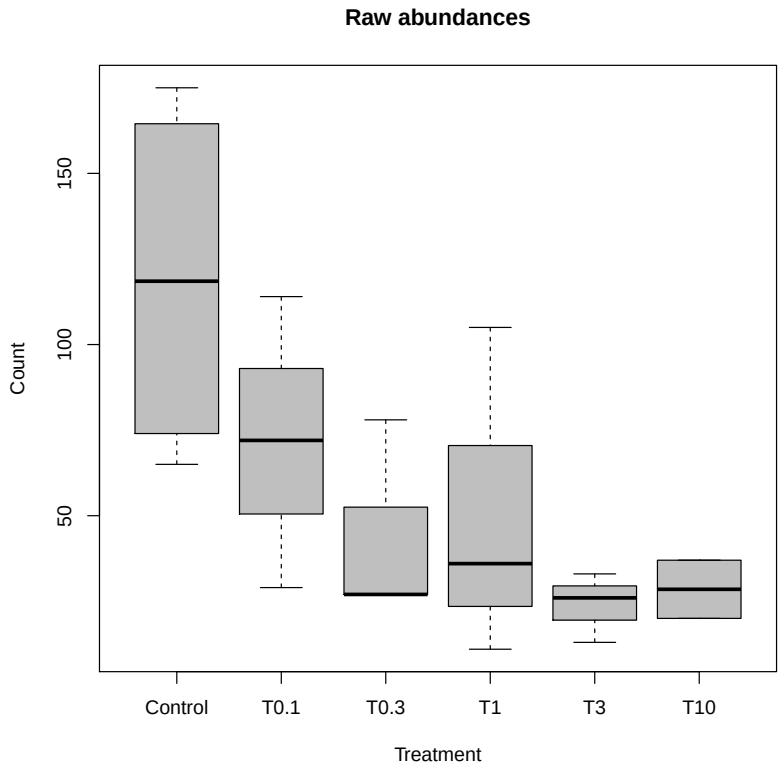
First, we load the data, bring it to the long format and remove NA values.

```
R> df <- read.table(header = TRUE,
                 text = 'Control  T0.1 T0.3  T1  T3  T10
                          175 29  27  36  26  20
                          65  114 78  11  13  37
                          154 72  27  105 33  NA
                          83  NA  NA  NA  NA  NA')
R> require(reshape2)
R> dfm <- melt(df, value.name = 'abu', variable.name = 'treatment')
R> dfm <- dfm[!is.na(dfm['abu']), ]
R> head(dfm)

##   treatment abu
## 1   Control 175
## 2   Control  65
## 3   Control 154
## 4   Control  83
## 5      T0.1  29
## 6      T0.1 114
```

This results in a table with two columns - one indicating the treatment and one with the measured abundances.

Let's have a first look at the data:

```
R> boxplot(abu ~ treatment, data = dfm, xlab = 'Treatment',
            ylab = 'Count', col = 'grey75', main = 'Raw abundances')
```



We clearly see a treatment related response. Moreover, we may note that variances are increasing with increasing abundances.

Assuming a normal distribution of transformed abundances

Data transformation

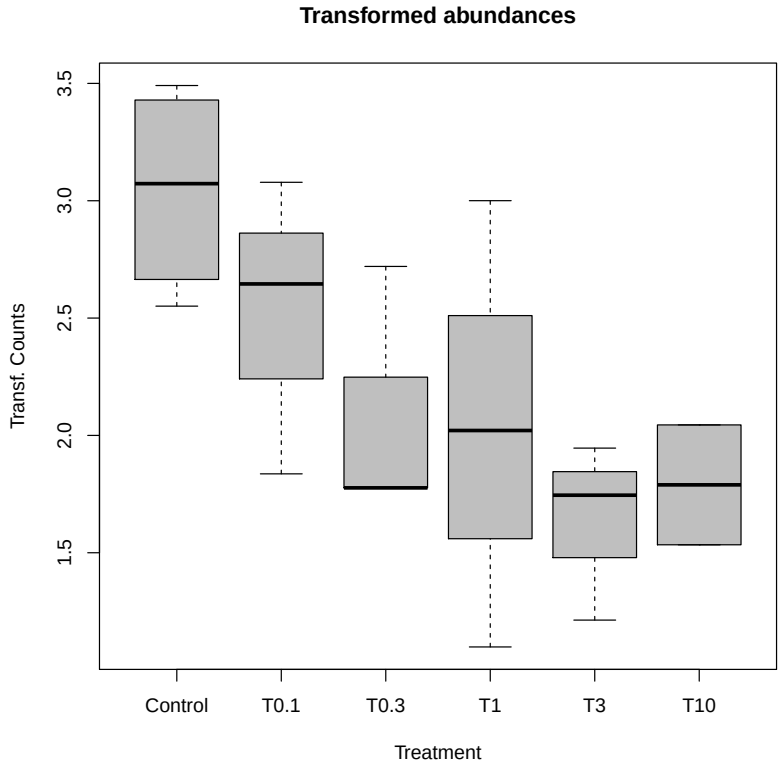
Next we transform the data using a ln(Ax + 1) transformation. A is chosen so that the term Ax equals two for the lowest non-zero abundance. We add these transformed abundances as extra column to our table.

```
R> A <- 2 / min(dfm$abu[dfm$abu != 0])
R> A

## [1] 0.1818182

R> dfm$abu_t <- log(A * dfm$abu + 1)
R> head(dfm)
```

```
##   treatment abu    abu_t
## 1   Control 175 3.490983
## 2   Control  65 2.550865
## 3   Control 154 3.367296
## 4   Control  83 2.778254
## 5      T0.1  29 1.836211
## 6      T0.1 114 3.078568
```

It looks like the transformation does a good job in equalizing the variances:

```
R> boxplot(abu_t ~ treatment, data = dfm,
           xlab = 'Treatment', ylab = 'Transf. Counts',
           col = 'grey75', main = 'Transformed abundances')
```



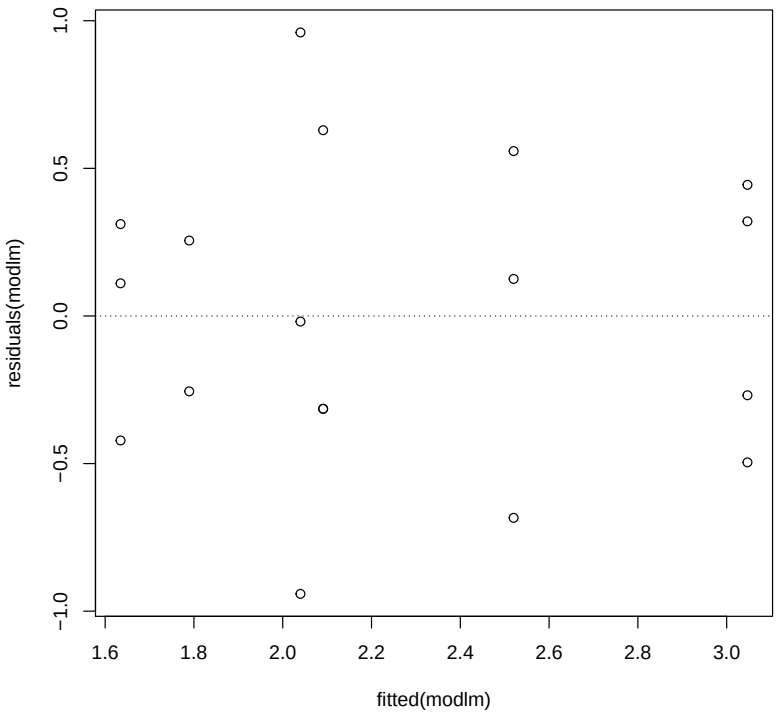
Model fitting

The model from eqn. 2 can be easily fitted using the `lm()` function:

```
R> modlm <- lm(abu_t ~ treatment, data = dfm)
```

The residuals vs. fitted values diagnostic plot show no problematic pattern, though it might be difficult to decide with such a small sample size

```
R> plot(residuals(modlm) ~ fitted(modlm))
R> abline(h = 0, lty = 'dotted')
```



The `summary()` gives the estimated coefficients with standard errors and Wald t tests:

```
R> summary(modlm)

##
## Call:
## lm(formula = abu_t ~ treatment, data = dfm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.94133 -0.31454  0.04576  0.31813  0.96033
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.0468     0.2970  10.260 2.71e-07 ***
## treatmentT0.1  -0.5267     0.4536  -1.161  0.26814
## treatmentT0.3  -0.9558     0.4536  -2.107  0.05682 .
## treatmentT1    -1.0069     0.4536  -2.220  0.04646 *
## treatmentT3    -1.4121     0.4536  -3.113  0.00897 **
## treatmentT10   -1.2575     0.5144  -2.445  0.03089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5939 on 12 degrees of freedom
## Multiple R-squared:  0.5167,Adjusted R-squared:  0.3154
## F-statistic: 2.566 on 5 and 12 DF,  p-value: 0.08406
```

Inference on general treatment effect

Or, if you want to have the ANOVA table with an F-test:

```
R> summary.aov(modlm)

##             Df Sum Sq Mean Sq F value Pr(>F)
## treatment    5  4.526  0.9052   2.566 0.0841 .
## Residuals   12  4.233  0.3528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this output we might infer that we cannot detect any treatment effect (F = 2.566, p = 0.084).

Inference on LOEC

Let's move on to the LOEC determination. This can be easily done using the multcomp package (Hothorn et al., 2008):

Here we perform a one-sided (`alternative = 'less'`) using Dunnett contrasts of treatment (`mcp(treatment='Dunnett')`). Moreover, we adjust for multiple testing using Holm's method (`test = adjusted('holm')`):

```
R> require(multcomp)
R> summary(glht(modlm, linfct = mcp(treatment = 'Dunnett'),
                alternative = 'less'),
                test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = abu_t ~ treatment, data = dfm)
##
## Linear Hypotheses:
##                      Estimate Std. Error t value Pr(<t)
## T0.1 - Control >= 0  -0.5267     0.4536  -1.161 0.1341
## T0.3 - Control >= 0  -0.9558     0.4536  -2.107 0.0697 .
## T1 - Control >= 0    -1.0069     0.4536  -2.220 0.0697 .
## T3 - Control >= 0    -1.4121     0.4536  -3.113 0.0224 *
## T10 - Control >= 0   -1.2575     0.5144  -2.445 0.0618 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

Here only treatment 3 mg/L shows a statistically significant difference from control and is the determined LOEC. The column `'Estimate'` gives the estimated difference in means between treatments and control and `'Std. Error'` the standard errors of these estimates.

To determine the LOEC we could also use a Williams type contrast (Bretz et al., 2010).

Here I use a step-up Williams contrast. First we need to define a contrast matrix (see also `?contrMat()`):

```
# observations per treatment
R> n <- tapply(dfm$abu_t, dfm$treatment, length)
R> k <- length(n)
R> CM <- c()
R> for (i in 1:(k - 1)) {
      help <- c(-1, n[2:(i + 1)] / sum(n[2:(i + 1)]), rep(0 , k - i - 1))
      CM <- rbind(CM, help)
}
```

```
R> rownames(CM) <- paste("C", 1:nrow(CM))
R> CM

##             T0.1
## C 1 -1 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## C 2 -1 0.5000000 0.5000000 0.0000000 0.0000000 0.0000000
## C 3 -1 0.3333333 0.3333333 0.3333333 0.0000000 0.0000000
## C 4 -1 0.2500000 0.2500000 0.2500000 0.2500000 0.0000000
## C 5 -1 0.2142857 0.2142857 0.2142857 0.2142857 0.1428571
```

Then we supply this contrast matrix to `glht()`:

```
R> summary(glht(modlm, linfct = mcp(treatment = CM),
                alternative = 'less'),
                test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = abu_t ~ treatment, data = dfm)
##
## Linear Hypotheses:
##          Estimate Std. Error t value Pr(<t)
## C 1 >= 0  -0.5267     0.4536  -1.161 0.1341
## C 2 >= 0  -0.7413     0.3834  -1.934 0.0771 .
## C 3 >= 0  -0.8298     0.3569  -2.325 0.0576 .
## C 4 >= 0  -0.9754     0.3429  -2.845 0.0295 *
## C 5 >= 0  -1.0157     0.3367  -3.016 0.0268 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

This indicates a LOEC at 3 mg/L.

If we do not adjust for multiple testing (`test = adjusted('none')`), we end up with the same NOEC (0.1 mg/L) as Brock et al., (2015):

```
R> summary(glht(modlm, linfct = mcp(treatment = CM),
                alternative = 'less'),
                test = adjusted('none'))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = abu_t ~ treatment, data = dfm)
##
## Linear Hypotheses:
##          Estimate Std. Error t value  Pr(<t)
## C 1 >= 0  -0.5267     0.4536  -1.161 0.13407
## C 2 >= 0  -0.7413     0.3834  -1.934 0.03855 *
## C 3 >= 0  -0.8298     0.3569  -2.325 0.01921 *
## C 4 >= 0  -0.9754     0.3429  -2.845 0.00739 **
## C 5 >= 0  -1.0157     0.3367  -3.016 0.00537 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- none method)
```

Note, this multiple contrast test is different from the original Williams test (Williams, 1972) used by (Brock et al., 2015). See Bretz, (1999) for a comparison.

Assuming a Poisson distribution of abundances

Model fitting

We are dealing with count data, so a Poisson GLM might be a good choice. GLMs can be fitted using the `glm()` function and here we fit the model from eqn. 3:

```
R> modpois <- glm(abu ~ treatment, data = dfm,
                  family = poisson(link = 'log'))
```

Here `family = poisson(link = 'log')` specifies that we want to fit a poisson model using a log link between response and predictors.

The `summary` gives the estimated coefficients, standard errors and Wald Z tests:

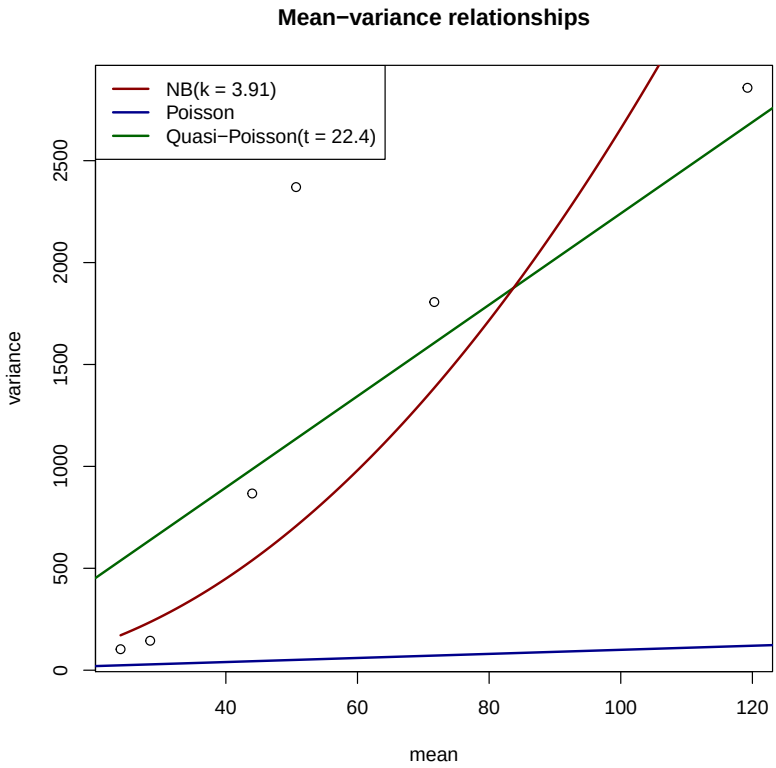
```
R> (sum_pois <- summary(modpois))

##
## Call:
```

```
## glm(formula = abu ~ treatment, family = poisson(link = "log"),
##     data = dfm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7625  -2.7621  -0.8219   2.7172   6.6602
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    4.78122    0.04579 104.423  < 2e-16 ***
## treatmentT0.1 -0.50920    0.08214  -6.199 5.69e-10 ***
## treatmentT0.3 -0.99703    0.09835 -10.138  < 2e-16 ***
## treatmentT1   -0.85595    0.09314  -9.190  < 2e-16 ***
## treatmentT3   -1.60317    0.12643 -12.680  < 2e-16 ***
## treatmentT10  -1.43132    0.14014 -10.213  < 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: 604.79  on 17  degrees of freedom
## Residual deviance: 273.77  on 12  degrees of freedom
## AIC: 387.63
##
## Number of Fisher Scoring iterations: 5
```

But is a poisson distribution appropriate here? A property of the poisson distribution is that its variance is equal to the mean. A simple diagnostic would be to plot group variances vs. group means:

```
R> require(plyr)
# mean and variance per treatment
R> musd <- ddply(dfm, .(treatment), summarise,
                 mu = mean(abu),
                 var = var(abu))
R> musd

##   treatment        mu      var
## 1   Control 119.25000 2857.583
## 2      T0.1  71.66667 1806.333
## 3      T0.3  44.00000  867.000
## 4        T1  50.66667 2370.333
```

```
## 5        T3  24.00000  103.000
## 6       T10  28.50000  144.500

# plot mean vs var
R> plot(var ~ mu, data = musd,
        xlab = 'mean', ylab = 'variance',
        main = 'Mean-variance relationships')
# poisson
R> abline(a = 0, b = 1, col = 'darkblue', lwd = 2)
# quasi-Poisson
R> abline(a = 0, b = 22.41, col = 'darkgreen', lwd = 2)
# negative binomial
R> curve(x + (x^2 / 3.91), from = 24, to = 119.25, add = TRUE,
          col = 'darkred', lwd = 2)
R> legend('topleft',
          legend = c('NB(k = 3.91)','Poisson','Quasi-Poisson(t=22.4)'),
          col = c('darkred', 'darkblue', 'darkgreen'),
           lty = c(1,1, 1),
           lwd = c(2,2, 2))
```



I also added the assumed mean-variance relationships of the Poisson, quasi-Poisson and negative binomial models (see below). We clearly see that the variance increases much more than would be expected under the poisson distribution (the data is overdispersed). Moreover, we could check overdispersion from the `summary`: If the ratio of residual deviance to degrees of freedom is >1 the data is overdispersed.

```
R> sum_pois$deviance / sum_pois$df.residual
```

```
## [1] 22.81412
```

Apply quasi-Poisson to deal with overdispersion

The plot above suggests that the variance may increasing stronger than the mean and a quasi-Poisson or negative binomial model might be more appropriate for this data.

Model fitting

Fitting a quasi-Poisson model (eqn. 4) is straight forward:

```
R> modqpois <- glm(abu ~ treatment, data = dfm, family = 'quasipoisson')
```

The summary gives the estimated coefficients:

```
R> summary(modqpois)

##
## Call:
## glm(formula = abu ~ treatment, family = "quasipoisson",
##     data = dfm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7625  -2.7621  -0.8219   2.7172   6.6602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.7812     0.2168  22.058 4.43e-11 ***
## treatmentT0.1  -0.5092     0.3889  -1.309   0.2149
## treatmentT0.3  -0.9970     0.4656  -2.142   0.0534 .
## treatmentT1    -0.8560     0.4409  -1.941   0.0761 .
## treatmentT3    -1.6032     0.5985  -2.679   0.0201 *
## treatmentT10   -1.4313     0.6634  -2.157   0.0519 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 22.411)
##
##     Null deviance: 604.79  on 17  degrees of freedom
## Residual deviance: 273.77  on 12  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

, with the dispersion parameter $\Theta = 22.41055$. Note, that the coefficients estimates are the same as from the Poisson model, only the standard errors are scaled/wider.

Inference on general treatment effect

An F-test can be performed using `drop1()`:

```
R> drop1(modqpois, test = 'F')

## Single term deletions
##
## Model:
## abu ~ treatment
##           Df Deviance F value  Pr(>F)
## <none>         273.77
## treatment  5   604.79  2.9019 0.06059 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we would reject that there is treatment effect (at alpha = 0.05).

Inference on LOEC

The LOEC can be determined with `multcomp`:

```
R> summary(glht(modqpois, linfct = mcp(treatment = 'Dunnett'),
                alternative = 'less'),
                test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm(formula = abu ~ treatment, family = "quasipoisson",
##          data = dfm)
##
## Linear Hypotheses:
##                      Estimate Std. Error z value Pr(<z)
## T0.1 - Control >= 0  -0.5092     0.3889  -1.309 0.0952 .
## T0.3 - Control >= 0  -0.9970     0.4656  -2.142 0.0619 .
## T1 - Control >= 0    -0.8560     0.4409  -1.941 0.0619 .
## T3 - Control >= 0    -1.6032     0.5985  -2.679 0.0185 *
## T10 - Control >= 0   -1.4313     0.6634  -2.157 0.0619 .
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

, which determines 3 mg/L as LOEC.

Assuming a negative binomial distribution of abundances

Model fitting

To fit a negative binomial GLM (eqn. 5) we could use `glm.nb()` from the MASS package (Venables and Ripley, 2002):

```
R> require(MASS)
R> modnb <- glm.nb(abu ~ treatment, data = dfm)
```

The estimated coefficients:

```
R> summary(modnb)

##
## Call:
## glm.nb(formula = abu ~ treatment, data = dfm,
##     init.theta = 3.905898474,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2554  -0.8488  -0.3020   0.5954   1.5899
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.7812     0.2571  18.596  < 2e-16 ***
## treatmentT0.1  -0.5092     0.3951  -1.289  0.19746
## treatmentT0.3  -0.9970     0.3988  -2.500  0.01241 *
## treatmentT1    -0.8560     0.3975  -2.153  0.03130 *
## treatmentT3    -1.6032     0.4066  -3.943 8.05e-05 ***
## treatmentT10   -1.4313     0.4601  -3.111  0.00186 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.9059)
##    family taken to be 1)
```

```
##
##     Null deviance: 39.057  on 17  degrees of freedom
## Residual deviance: 18.611  on 12  degrees of freedom
## AIC: 181.24
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  3.91
##           Std. Err.:  1.37
##
##  2 x log-likelihood:  -167.238
```

, with $\kappa = 3.91$.

Inference on general treatment effect (LR-test)

For an LR-Test we need to first fit a reduced model:

```
R> modnb.null <- glm.nb(abu ~ 1, data = dfm)
```

, so that the dispersion parameter κ is re-estimated for the reduced model. Then we can compare these two models with a LR-Test:

```
R> anova(modnb, modnb.null, test = 'Chisq')

## Likelihood ratio tests of Negative Binomial Models
##
## Response: abu
##       Model    theta Resid. df    2 x log-lik.   Test df LR stat.
## 1         1 1.861577        17       -181.2281
## 2 treatment 3.905898        12       -167.2383 1 vs 2  5 13.98985
##    Pr(Chi)
## 1
## 2 0.015674
```

, which suggests a treatment related effect on abundances.

Inference on general treatment effect (parametric bootstrap)

To test the LR statistic using paramtric bootstrap, we use two custom functions:

The first function `myPBrefdist` generates a boostrap sample and return the LR statistic for this sample:

```
#' PB of LR statistic
#' @param m1 Full model
#' @param m0 reduced model
#' @param  data data used in the models
#' @return LR of boostrap
# generate reference distribution
R> myPBrefdist <- function(m1, m0, data){
      # simulate from null
      x0 <- simulate(m0)
      # refit with new data
      newdata0 <- data
      newdata0[ , as.character(formula(m0)[[2]])] <- x0
      m1r <-  try(update(m1, .~., data = newdata0), silent = TRUE)
      m0r <- try(update(m0, .~., data = newdata0), silent = TRUE)
      # check convergence (otherwise return NA for LR)
      if(inherits(m0r, "try-error") | inherits(m1r, "try-error")){
        LR <- 'convergence error'
      } else {
        if(!is.null(m0r[['th.warn']]) | !is.null(m1r[['th.warn']])){
          LR <- 'convergence error'
        } else {
          LR <- -2 * (logLik(m0r) - logLik(m1r))
        }
      }
      return(LR)
}
```

The second one (`myPBmodcomp`) repeats `myPBrefdist` many time and returns a p-value:

```
#' generate LR distribution and return p value
#' @param m1 Full model
#' @param m0 reduced model
#' @param data data used in m1 and m0
#' @param npb number of bootstrap samples
#' @return p-value of boostrapped LR values
R> myPBmodcomp <- function(m1, m0, data, npb){
      ## calculate reference distribution
      LR <- replicate(npb, myPBrefdist(m1 = m1, m0 = m0, data = data),
                      simplify = TRUE)
      LR <- as.numeric(LR)
      nconv_LR <- sum(!is.na(LR))
```

```
    ## original stats
    LRo <- c(-2 * (logLik(m0) - logLik(m1)))
    ## p-value from parametric bootstrap
    p.pb <- mean(c(LR, LRo) >= LRo, na.rm = TRUE)
    return(list(nconv_LR = nconv_LR, p.pb = p.pb))
}
```

Sounds complicated, but we can easily apply this to the negativ binomial model using:

```
R> set.seed(1234)
R> myPBmodcomp(modnb, modnb.null, data = dfm, npb = 500)

## $nconv_LR
## [1] 499
##
## $p.pb
## [1] 0.042
```

Here, we specify to generate 500 bootstrap samples (`npb = 500`). Of these 500 samples, 499 converged (`nconv_LR`) (one did not and throws some errors) and gives a p-value of 0.042.

Inference on LOEC

This is similar to the other parametric models:

```
R> summary(glht(modnb, linfct = mcp(treatment = 'Dunnett'),
            alternative = 'less'),
            test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm.nb(formula = abu ~ treatment, data = dfm,
##     init.theta = 3.905898474,
##     link = log)
##
## Linear Hypotheses:
```

```
##                      Estimate Std. Error z value   Pr(<z)
## T0.1 - Control >= 0  -0.5092     0.3951  -1.289 0.098731 .
## T0.3 - Control >= 0  -0.9970     0.3988  -2.500 0.018615 *
## T1 - Control >= 0    -0.8560     0.3975  -2.153 0.031300 *
## T3 - Control >= 0    -1.6032     0.4066  -3.943 0.000201 ***
## T10 - Control >= 0   -1.4313     0.4601  -3.111 0.003727 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

which suggests a LOEC at the 0.3 mg/l treatment.

Non-parametric methods

Kruskal-Wallis Test

We can use the Kruskal-Wallies test to check if there is a difference between treatments:

```
R> kruskal.test(abu ~ treatment, data = dfm)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  abu by treatment
## Kruskal-Wallis chi-squared = 8.219, df = 5, p-value = 0.1446
```

Pairwise Wilcoxon test

To determine the LOEC we could use a Pairwise Wilcoxon test. The built-in `pairwise.wilcox.test()` compares by default all levels (Tukey-contrasts). We are only interested in a subset of these comparisons (Dunnett-contrast). Therefore, we use a custom function, which is a wrapper around `wilcox.exact` from the exactRankTests package:

```
#' pairwise wilcox.test with dunnett contrasrs
#' @param y numeric; vector of data values
#' @param g factor; grouping vector
#' @param dunnett logical; if TRUE dunnett contrast, otherwise
Tukey-contrasts
```

```
#' @param padj character; method for p-adjustment, see ?p.adjust.
#' @param ... other arguments passed to exactRankTests::wilcox.exact
R> pairwise_wilcox <- function(y, g, dunnett = TRUE, padj='holm',...){
      if(!require(exactRankTests)){
        stop('Install exactRankTests package')
      }
      tc <- t(combn(nlevels(g), 2))
      # take only dunnett comparisons
      if(dunnett){
        tc <- tc[tc[ , 1] == 1, ]
      }
      pval <- numeric(nrow(tc))
      # use wilcox.exact (for tied data)
      for(i in seq_len(nrow(tc))){
        pval[i] <- wilcox.exact(y[as.numeric(g) == tc[i, 2]],
                                y[as.numeric(g) == tc[i, 1]],exact=TRUE,
                                ...)$p.value
      }

      # adjust p-values
      pval <- p.adjust(pval, padj)
      names(pval) = paste(levels(g)[tc[,1]], levels(g)[tc[,2]],
                sep = ' vs. ')
      return(pval)
}
```

Here, we use one-sided Dunnett contrasts and adjust p-values using Holm's method:

```
R> pairwise_wilcox(y = dfm$abu, g = dfm$treatment,
                   dunnett = TRUE, p.adj = 'holm', alternative = 'less')

## Control vs. T0.1 Control vs. T0.3  Control vs. T1  Control vs. T3
##        0.2285714        0.2285714       0.2285714       0.1428571
##  Control vs. T10
##        0.2285714
```

This indicates no treatment effect at no level of concentration.

Binomial data example

Introduction

Here we will show how to analyse binomial data (*x out of n*). Data is provided in Newman, (2012) (example 5.1, page 223) and EPA, (2002). Ten fathead minnow (*Pimephales promelas*) larvals were exposed to sodium pentachlorophenol (NaPCP) and proportions of the total number alive at the end of the exposure reported.

First we load the data:

```
R> df <- read.table(header = TRUE, text = 'conc A B C D
0 1 1 0.9 0.9
32 0.8 0.8 1 0.8
64 0.9 1 1 1
128 0.9 0.9 0.8 1
256 0.7 0.9 1 0.5
512 0.4 0.3 0.4 0.2')
R> df

##   conc   A   B   C   D
## 1    0 1.0 1.0 0.9 0.9
## 2   32 0.8 0.8 1.0 0.8
## 3   64 0.9 1.0 1.0 1.0
## 4  128 0.9 0.9 0.8 1.0
## 5  256 0.7 0.9 1.0 0.5
## 6  512 0.4 0.3 0.4 0.2
```

The we do some house-keeping, reformat the data and convert concentration to a factor:

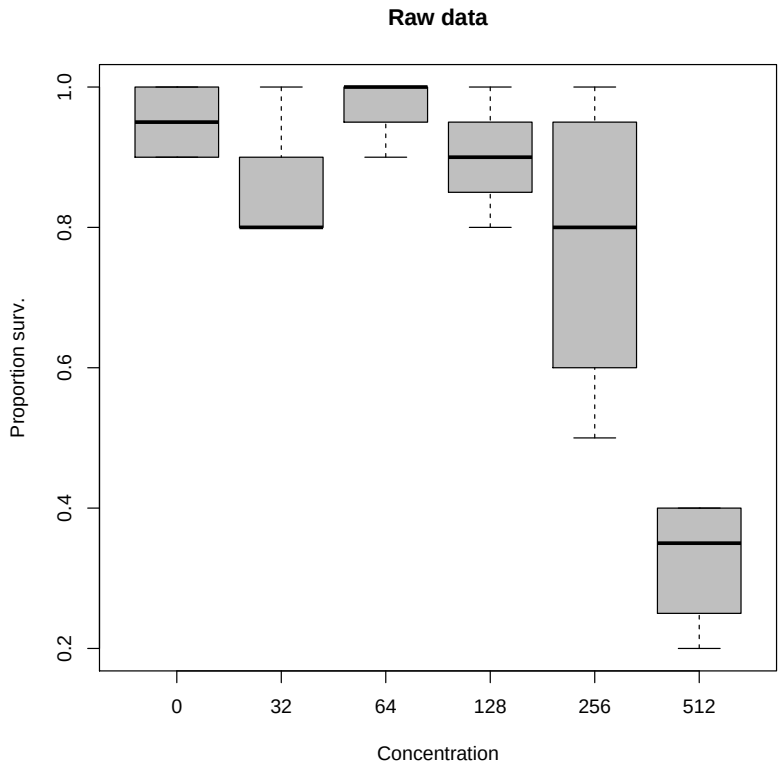
```
R> require(reshape2)
# wide to long
R> dfm <- melt(df, id.vars = 'conc', value.name = 'y',
                variable.name = 'tank')
# conc as factor
R> dfm$conc <- factor(dfm$conc)
```

So after data cleaning the data looks like

```
R> head(dfm)

##   conc tank   y
## 1    0    A 1.0
## 2   32    A 0.8
## 3   64    A 0.9
## 4  128    A 0.9
## 5  256    A 0.7
## 6  512    A 0.4
```

Let's have a first look at the data:

```
R> boxplot(y ~ conc, data = dfm,
           xlab = 'Concentration', ylab = 'Proportion surv.',
           main = 'Raw data', col = 'grey75')
```



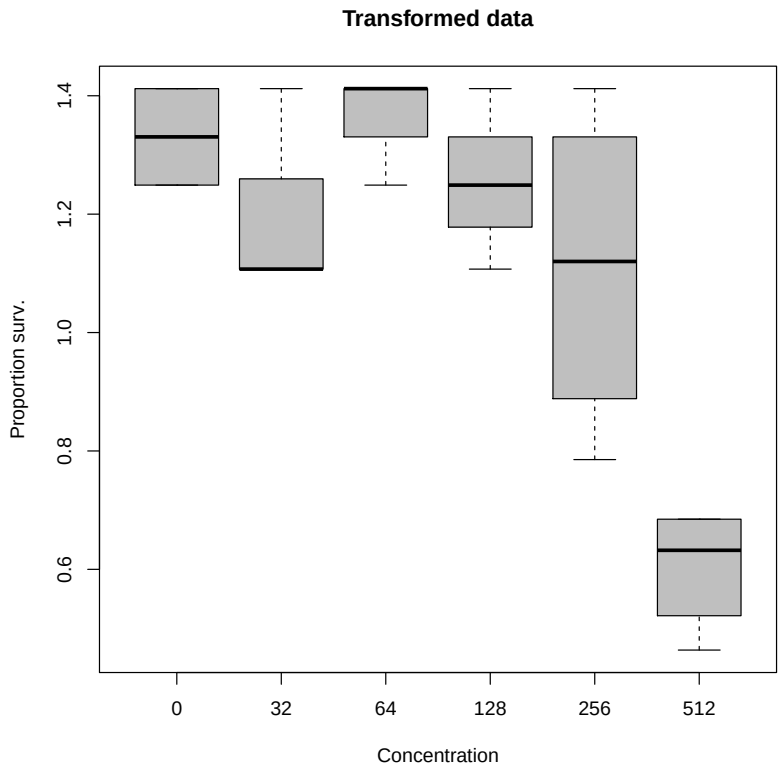
This plot indicates a strong effect at the highest concentration.

Assuming a normal distribution of transformed proportions

First, we arcsine transform (eqn. 6) the proportions:

```
R> dfm$y_asin <- ifelse(dfm$y == 1,
                    asin(1) - asin(sqrt(1/40)),
                    ifelse(dfm$y == 0,
                           asin(sqrt(1/40)),
                           asin(sqrt(dfm$y))
                           )
                    )
```

```
R> boxplot(y_asin ~ conc, data = dfm,
            xlab = 'Concentration', ylab = 'Proportion surv.',
            main = 'Transformed data', col = 'grey75')
```



Then, like in the count data example we fit the model using `lm()`:

```
R> modlm <- lm(y_asin ~ conc, data = dfm)
```

The summary gives the estimated coefficients:

```
R> summary(modlm)

##
## Call:
## lm(formula = y_asin ~ conc, data = dfm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32401 -0.08149 -0.00527  0.08150  0.30261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.33053    0.07693  17.295 1.16e-12 ***
## conc32      -0.14717    0.10880  -1.353   0.1929
## conc64       0.04074    0.10880   0.374   0.7124
## conc128     -0.07622    0.10880  -0.701   0.4925
## conc256     -0.22113    0.10880  -2.032   0.0571 .
## conc512     -0.72735    0.10880  -6.685 2.86e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1539 on 18 degrees of freedom
## Multiple R-squared:  0.7871,Adjusted R-squared:  0.7279
## F-statistic: 13.31 on 5 and 18 DF,  p-value: 1.612e-05
```

The F-test suggests a treatment related effect:

```
R> drop1(modlm, test = 'F')

## Single term deletions
##
## Model:
## y_asin ~ conc
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              0.42613 -84.746
## conc    5    1.5753 2.00142 -57.621  13.308 1.612e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And the LOEC is at the highest concentration:

```
R> summary(glht(modlm, linfct = mcp(conc = 'Dunnett'),
                alternative = 'less'),
                test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = y_asin ~ conc, data = dfm)
##
## Linear Hypotheses:
##               Estimate Std. Error t value   Pr(<t)
## 32 - 0 >= 0  -0.14717    0.10880  -1.353    0.289
## 64 - 0 >= 0   0.04074    0.10880   0.374    0.644
## 128 - 0 >= 0 -0.07622    0.10880  -0.701    0.493
## 256 - 0 >= 0 -0.22113    0.10880  -2.032    0.114
## 512 - 0 >= 0 -0.72735    0.10880  -6.685 7.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

Assuming a binomial distribution

The binomial model with a logit link (eqn. 7) between predictors and response can be fitted using the `glm()` function:

```
R> modglm <- glm(y ~ conc , data = dfm, family = binomial(link='logit'),
                 weights = rep(10, nrow(dfm)))
```

Here the weights arguments, indicates how many fish where exposed in each treatment (N=10, eqn .7).

The summary gives the estimated coefficients:

```
R> summary(modglm)

##
## Call:
## glm(formula = y ~ conc, family = binomial(link = "logit"),
```

```
##     data = dfm,
##     weights = rep(10, nrow(dfm)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8980  -0.5723   0.0000   0.7869   2.2578
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.9444     0.7255   4.059 4.94e-05 ***
## conc32       -1.2098     0.8499  -1.423   0.1546
## conc64        0.7191     1.2458   0.577   0.5638
## conc128      -0.7472     0.8967  -0.833   0.4047
## conc256      -1.7077     0.8183  -2.087   0.0369 *
## conc512      -3.6753     0.8002  -4.593 4.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 88.672  on 23  degrees of freedom
## Residual deviance: 23.889  on 18  degrees of freedom
## AIC: 72.862
##
## Number of Fisher Scoring iterations: 5
```

To perform a LR-test we can used the `drop1()` function:

```
R> drop1(modglm, test = 'Chisq')

## Single term deletions
##
## Model:
## y ~ conc
##        Df Deviance     AIC    LRT  Pr(>Chi)
## <none>      23.889  72.862
## conc    5   88.672 127.645 64.783 1.243e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Also with the binomial model the LOEC is at the highest concentration:

```
R> summary(glht(modglm, linfct = mcp(conc = 'Dunnett'),
                    alternative = 'less'),
                    test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm(formula = y ~ conc, family = binomial(link = "logit"),
##     data = dfm,
##     weights = rep(10, nrow(dfm)))
##
## Linear Hypotheses:
##              Estimate Std. Error z value   Pr(<z)
## 32 - 0 >= 0   -1.2098     0.8499  -1.423   0.2319
## 64 - 0 >= 0    0.7191     1.2458   0.577   0.7181
## 128 - 0 >= 0  -0.7472     0.8967  -0.833   0.4047
## 256 - 0 >= 0  -1.7077     0.8183  -2.087   0.0738 .
## 512 - 0 >= 0  -3.6753     0.8002  -4.593 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

REFERENCES

Bretz, F. (1999). "Powerful modifications of Williams' test on trend". Universität Hannover.

Bretz, F., T. Hothorn, and P. H. Westfall (2010). *Multiple comparisons using R*. London: Chapman /& Hall.

Brock, T. C. M., M. Hammers-Wirtz, U. Hommen, T. G. Preuss, H.-T. Ratte, I. Roessink, T. Strauss, and P. J. Van den Brink (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems". *Environmental Science and Pollution Research* 22 (2), 1160–1174.

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

Hothorn, T., F. Bretz, and P. Westfall (2008). "Simultaneous inference in general parametric models". *Biometrical Journal* 50 (3), 346–363.

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

Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

Williams, D. A. (1972). "The comparison of several dose levels with a zero dose control". *Biometrics*, 519–531.

B LARGE SCALE RISKS FROM PESTICIDES IN SMALL STREAMS

DATA CLEANING

Before combining into a common database, more than 30 datasets have been cleaned and homogenised separately. Cleaning steps comprised the following steps (Figure B.1 gives a graphical overview):

1. Structure: Datasets have been adjusted to the database structure.
2. Coordinates: Coordinates have been transformed to a common Coordinate Reference System (DHDN / 3-Grad Gauss-Krüger Zone 3 (EPSG:31467)) and duplicates merged.
3. Chemicals: Chemical names and identifiers have been unified using the webchem package (https://github.com/ropensci/webchem).
4. Identifiers: Unique identifiers have been assigned.
5. Units: All concentrations have been converted to µg/L. Values below limit of quantification were set to zero (and can be used to identify non-detects).
6. Other meta-data: meta-data has been standardised.
7. Temporal resolution: The temporal resolution of the database is 1 day. Samplings below this resolution have been aggregated by the maximum daily value.
8. Validity Checks: Simple rules for validity checks have been implemented.

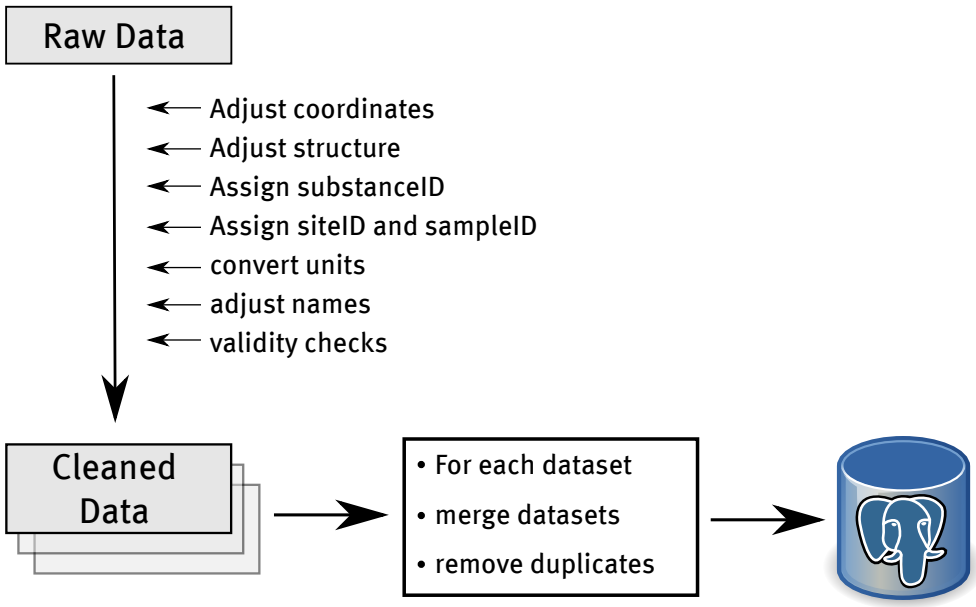


Figure B.1.: Overview on data cleaning steps. After cleaning, data have been stored in a relational spatial PostgreSQL database.

OVERVIEW ON COMPILED DATA

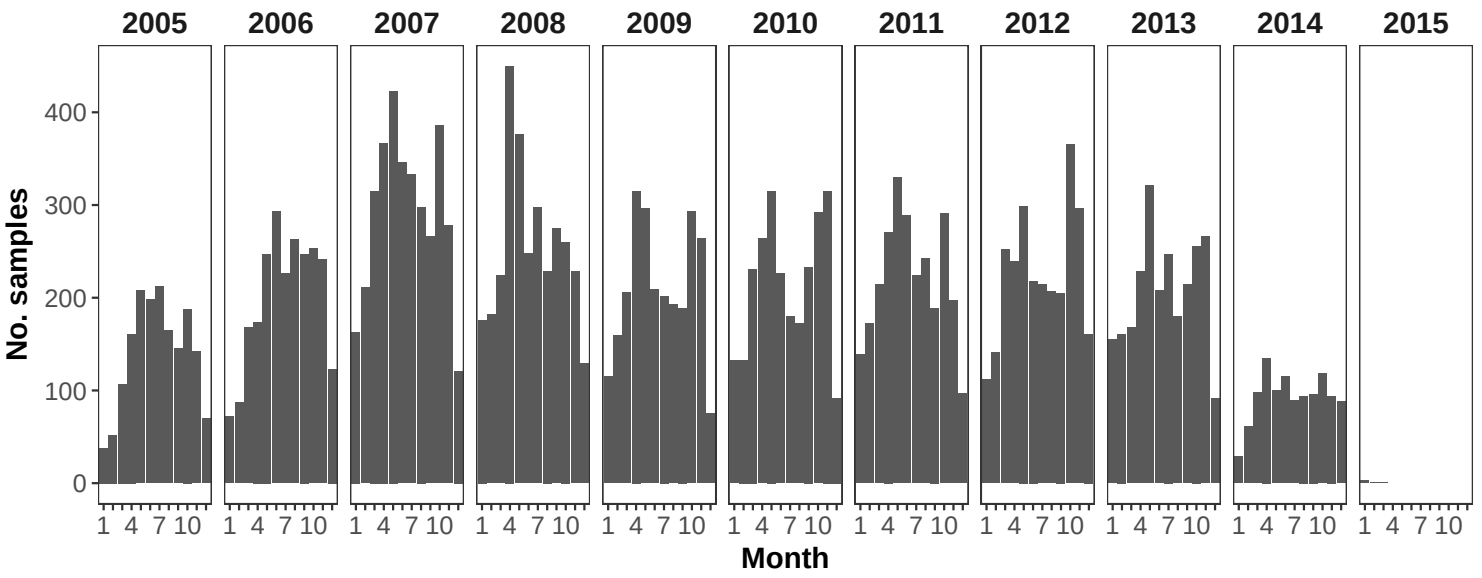


Figure B.2.: Number of sampling occasions per year and month.

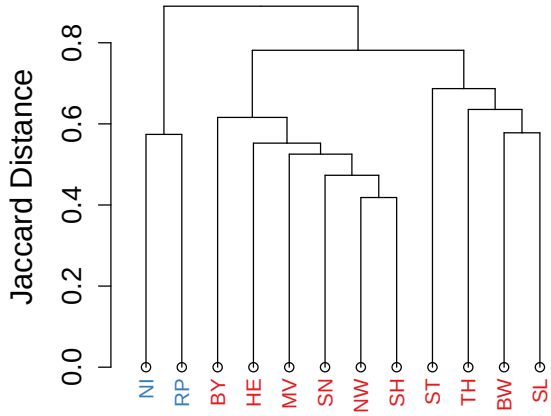


Figure B.3.: Complete Linkage Cluster Dendrogram of Jaccard Similarity of analysed compound spectra between federal states. Abbreviations of state names according to ISO 3166-2:DE.

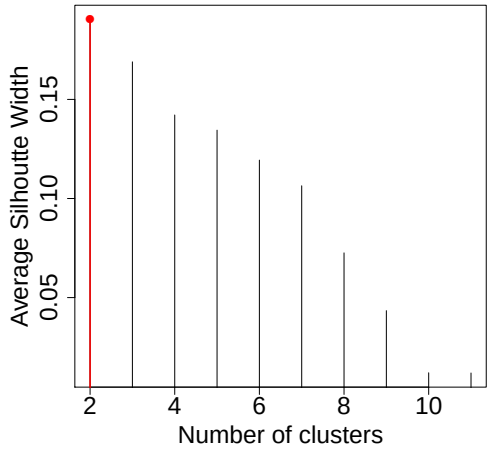


Figure B.4.: Average silhouette width for different cluster sizes of complete linkage clustering of jaccard similarity of analysed compound spectra between federal states. Two clusters showed the maximum silhouette width.

Table B.2.: Overview on pesticides (and metabolites) in the database. [a] Authorized in Germany (Source: German Federal Office of Consumer Protection and Food Safety (BVL) as at March 2015). [b] Authorized in the European union (Source: EU Pesticides database as at March 2015). [c] Regulatory Acceptable Concentration [µg/L] (Source: German Environment Agency (UBA) as at November 2015).

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
1	(E)7-(Z)9-Dodecadienylacetat	55774-32-8	other	x	x	
2	(Z)-9-Dodecenylacetat	16974-11-1	other	x	x	
3	1,3-cis-Dichlorpropen	10061-01-5	other			
4	1,3-trans-Dichlorpropen	10061-02-6	other			
5	1-(3,4-Dichlorphenyl)urea	2327-02-8	metabolite			
6	1-(4-Isopropylphenyl)urea	56046-17-4	metabolite			
7	1-Decanol	112-30-1	other	x	x	
8	1-Methylcyclopropen	3100-04-7	other	x	x	
9	2,4,5-T	93-76-5	herbicide			
10	2,4,6-Trichlorphenol	88-06-2	metabolite			
11	2,4-D	94-75-7	herbicide	x	x	1.10000
12	2,4-DB	94-82-6	herbicide		x	
13	2,4-Dichlorphenol	120-83-2	metabolite			
14	2,6-Dichlorobenzamid	2008-58-4	metabolite			
15	2-Hydroxydesethylatrazin	19988-24-0	metabolite			
16	3-Hydroxy Carbofuran	16655-82-6	metabolite			
17	4,6-Dinitro-o-Cresol	534-52-1	insecticide			
18	4-tert. Cyclobutylhexanon	98-53-3	metabolite			
19	AMPA	1066-51-9	metabolite			
20	Acequinocyl	57960-19-7	insecticide	x	x	9.00000
21	Acetamiprid	135410-20-7	insecticide	x	x	0.24000
22	Acetochlor	34256-82-1	herbicide			
23	Acetochlorsulfonsäure	187022-11-3	metabolite			
24	Acetochlorsäure	194992-44-4	metabolite			
25	Acifluorfen	50594-66-6	herbicide			
26	Aclonifen	74070-46-5	herbicide	x	x	1.06000
27	Acrinathrin	101007-06-1	insecticide		x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
28	Alachlor	15972-60-8	herbicide			
29	Aldicarb	116-06-3	insecticide			
30	Aldrin	309-00-2	insecticide			
31	Ametoctradin	865318-97-4	fungicide	x	x	
32	Ametryn	834-12-8	herbicide			
33	Amidosulfuron	120923-37-7	herbicide	x	x	
34	Aminopyralid	150114-71-9	herbicide	x	x	
35	Amisulbrom	348635-87-0	fungicide	x	x	
36	Anthranilsäure-isopropylamid	30391-89-0	metabolite			
37	Atraton	1610-17-9	herbicide			
38	Atrazin	1912-24-9	herbicide			
39	Atrazin, 2-Hydroxy	2163-68-0	metabolite			
40	Avermectin B1a	71751-41-2	insecticide	x	x	
41	Azadirachtin (Neem)	11141-17-6	insecticide	x	x	
42	Azinphos-ethyl	2642-71-9	insecticide			
43	Azinphos-methyl	86-50-0	insecticide			
44	Aziprotryn	4658-28-0	herbicide			
45	Azoxystrobin	131860-33-8	fungicide	x	x	0.55000
46	Azoxystrobin-CA		metabolite			
47	Beflubutamid	113614-08-7	herbicide	x	x	
48	Benalaxyl	71626-11-4	fungicide	x	x	20.00000
49	Benazolin	3813-05-6	herbicide			
50	Bensulfuron-methyl	83055-99-6	herbicide		x	
51	Bentazon	25057-89-0	herbicide	x	x	535.00000
52	Benthiavalicarb	413615-35-7	fungicide	x	x	
53	Benzoesäure	65-85-0	fungicide	x	x	
54	Betacypermethrin	65731-84-2	insecticide		x	
55	Bifenazate	149877-41-8	insecticide	x	x	
56	Bifenox	42576-02-3	herbicide	x	x	
57	Bifenthrin	82657-04-3	insecticide		x	0.00050
58	Bixafen	581809-46-3	fungicide	x	x	0.46000
59	Boscalid	188425-85-6	fungicide	x	x	12.50000
60	Bromacil	314-40-9	herbicide			
61	Bromadiolon	28772-56-7	other		x	
62	Bromocyclen	1715-40-8	insecticide			
63	Bromoxynil	1689-84-5	herbicide	x	x	3.30000
64	Bupirimat	41483-43-6	fungicide		x	
65	Buturon	3766-60-7	herbicide			
66	Captan	133-06-2	fungicide	x	x	5.00000
67	Carbendazim	10605-21-7	fungicide			0.15000
68	Carbetamid	16118-49-3	herbicide		x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
69	Carbofuran	1563-66-2	insecticide			
70	Carboxin	5234-68-4	fungicide		x	
71	Carfentrazone-ethyl	128639-02-1	herbicide	x	x	0.31000
72	Chloramben	133-90-4	herbicide			
73	Chlorantraniliprole	500008-45-7	insecticide	x	x	0.35500
74	Chlorbromuron	13360-45-7	herbicide			
75	Chlordan	57-74-9	insecticide			
76	Chlorfenac	85-34-7	herbicide			
77	Chlorfenvinphos	470-90-6	insecticide			
78	Chlorfluazuron	71422-67-8	insecticide			
79	Chloridazon	1698-60-8	herbicide	x	x	56.00000
80	Chlormequat	7003-89-6	other	x	x	
81	Chloroxuron	1982-47-4	herbicide			
82	Chlorpropham	101-21-3	herbicide	x	x	
83	Chlorpyrifos	2921-88-2	insecticide		x	0.00050
84	Chlorpyriphos methyl	5598-13-0	insecticide		x	
85	Chlorsulfuron	64902-72-3	herbicide			
86	Chlorthalonil	1897-45-6	fungicide	x	x	
87	Chlorthalonil-SA		metabolite			
88	Chlortoluron	15545-48-9	herbicide	x	x	2.30000
89	Cinidon-ethyl	142891-20-1	herbicide			
90	Clethodim	99129-21-2	herbicide	x	x	
91	Clodinafop	114420-56-3	herbicide	x	x	
92	Clodinafop-propargyl	105512-06-9	herbicide			
93	Clofentezin	74115-24-5	insecticide		x	
94	Clomazon	81777-89-1	herbicide	x	x	5.70000
95	Clopyralid	1702-17-6	herbicide	x	x	1080.00000
96	Cloquintocet-mexyl	99607-70-2	other		x	
97	Clothianidin	210880-92-5	insecticide	x	x	0.00700
98	Codlemone (Codlelure)	33956-49-9	other	x	x	
99	Coumaphos	56-72-4	insecticide			
100	Crimidin	535-89-7	other			
101	Cyanazin	21725-46-2	herbicide			
102	Cyazofamid	120116-88-3	fungicide	x	x	
103	Cyclanilide	113136-77-9	other			
104	Cycloat	1134-23-2	herbicide			
105	Cycloxidim	101205-02-1	herbicide	x	x	
106	Cyflufenamid	180409-60-3	fungicide	x	x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
107	Cyfluthrin (Summe Isomere)	68359-37-5	insecticide			
108	Cyhalothrin (Summe Isomere)	91465-08-6	insecticide	x	x	
109	Cymoxanil	57966-95-7	fungicide	x	x	4.40000
110	Cypermetryn	52315-07-8	insecticide	x	x	0.00100
111	Cyproconazol	94361-06-5	fungicide	x	x	
112	Cyprodinil	121552-61-2	fungicide	x	x	0.75000
113	Cyromazin	66215-27-8	insecticide		x	
114	Daminozid	1596-84-5	other	x	x	
115	Deiquat	2764-72-9	herbicide	x	x	
116	Deltamethrin	52918-63-5	insecticide	x	x	
117	Demeton-O	298-03-3	insecticide			
118	Demeton-S	126-75-0	insecticide			
119	Demeton-S-methyl	919-86-8	insecticide			
120	Demeton-S-methylsulfon	17040-19-6	insecticide			
121	Desaminometribuzin	35045-02-4	metabolite			
122	Desethyl-2-hydroxyterbuthylazin	66753-06-8	metabolite			
123	Desethylatrazin	6190-65-4	metabolite			
124	Desethylsebuthylazin	37019-18-4	metabolite			
125	Desethylsimazin	6190-65-4	metabolite			
126	Desethylterbuthylazin	30125-63-4	metabolite			
127	Desisopropylatrazin	1007-28-9	metabolite			
128	Desmedipham	13684-56-5	herbicide	x	x	
129	Desmethyldiuron	3567-62-2	metabolite			
130	Desmethylisoproturon	34123-57-4	metabolite			
131	Desmetryn	1014-69-3	herbicide			
132	Desphenyl-Chloridazon	6339-19-1	metabolite			
133	Diazinon	333-41-5	insecticide			
134	Dicamba	1918-00-9	herbicide	x	x	180.00000
135	Dichlobenil	1194-65-6	herbicide			
136	Dichlofluanid	1085-98-9	fungicide			
137	Dichlorprop	120-36-5	herbicide			
138	Dichlorprop-P	15165-67-0	herbicide	x	x	
139	Dichlorvos	62-73-7	insecticide			
140	Diclofop	40843-25-2	herbicide		x	
141	Dicofol	115-32-2	insecticide			
142	Dieldrin	60-57-1	insecticide			
143	Difenacoum	56073-07-5	other		x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
144	Difenoconazol	119446-68-3	fungicide	x	x	0.36000
145	Diflubenzuron	35367-38-5	insecticide		x	
146	Diflufenican	83164-33-4	herbicide	x	x	0.02500
147	Dimefuron	34205-21-5	herbicide			0.83000
148	Dimethachlor	50563-36-5	herbicide	x	x	3.50000
149	Dimethachlor-CA		metabolite			
150	Dimethachlor-sulfonsäure		metabolite			
151	Dimethachlorsäure		metabolite			
152	Dimethenamid	87674-68-8	herbicide			
153	Dimethenamid-CA		metabolite			
154	Dimethenamid-P	163515-14-8	herbicide	x	x	1.35000
155	Dimethenamid-SA		metabolite			
156	Dimethenamid-sulfonsäure		metabolite			
157	Dimethoat	60-51-5	insecticide	x	x	4.00000
158	Dimethomorph	110488-70-5	fungicide	x	x	5.60000
159	Dimoxystrobin	149961-52-4	fungicide	x	x	0.03160
160	Diniconazol	83657-24-3	fungicide			
161	Dinoseb	88-85-7	herbicide			
162	Dinotefuran	165252-70-0	insecticide			
163	Dinoterb	1420-07-1	herbicide			
164	Disulfoton	298-04-4	insecticide			
165	Dithianon	3347-22-6	fungicide	x	x	0.78000
166	Diuron	330-54-1	herbicide		x	0.79000
167	Dodin	2439-10-3	fungicide	x	x	5.33000
168	Endosulfan, alpha	959-98-8	insecticide			
169	Endosulfan, beta	33213-65-9	insecticide			
170	Endosulfansulfat	1031-07-8	metabolite			
171	Endrin	72-20-8	insecticide			
172	Epoxiconazol	133855-98-8	fungicide	x	x	0.53750
173	Esfenvalerat	66230-04-4	insecticide	x	x	
174	Etaconazol	60207-93-4	fungicide			
175	Ethidimuron	30043-49-3	herbicide			
176	Ethirimol	23947-60-6	fungicide			
177	Ethofenprox	80844-07-1	insecticide	x	x	
178	Ethofumesat	26225-79-6	herbicide	x	x	24.00000
179	Etrimfos	38260-54-7	insecticide			
180	Famoxadone	131807-57-3	fungicide	x	x	
181	Fenamidon	161326-34-7	fungicide	x	x	
182	Fenamiphos	22224-92-6	insecticide		x	
183	Fenarimol	60168-88-9	fungicide			

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
184	Fenazaquin	120928-09-8	insecticide	x	x	
185	Fenhexamid	126833-17-8	fungicide	x	x	10.10000
186	Fenitrothion	122-14-5	insecticide			
187	Fenoprop	93-72-1	herbicide			
188	Fenoxaprop	95617-09-7	herbicide			
189	Fenoxaprop-p	113158-40-0	herbicide	x	x	
190	Fenoxaprop-p-ethyl	71283-80-2	herbicide			
191	Fenoxycarb	72490-01-8	insecticide		x	
192	Fenpropidin	67306-00-7	fungicide	x	x	
193	Fenpropimorph	67564-91-4	fungicide	x	x	0.19500
194	Fenpyroximat	134098-61-6	insecticide	x	x	
195	Fenthion	55-38-9	insecticide			
196	Fenuron	101-42-8	herbicide			
197	Fipronil	120068-37-3	insecticide		x	0.00077
198	Flamprop	58667-63-3	herbicide			
199	Flazasulfuron	104040-78-0	herbicide	x	x	
200	Flonicamid	158062-67-0	insecticide	x	x	310.00000
201	Florasulam	145701-23-1	herbicide	x	x	
202	Fluazifop	69335-91-7	herbicide			146.00000
203	Fluazifop-P	83066-88-0	herbicide	x	x	146.00000
204	Fluazifop-P-butyl	79241-46-6	herbicide			7.70000
205	Fluazifop-butyl	69806-50-4	herbicide			7.70000
206	Fluazinam	79622-59-6	fungicide	x	x	0.26000
207	Fluchloralin	33245-39-5	herbicide			
208	Fludioxonil	131341-86-1	fungicide	x	x	0.50000
209	Flufenacet	142459-58-3	herbicide	x	x	2.40000
210	Flufenacet-SA		metabolite			
211	Flufenoxuron	101463-69-8	insecticide			
212	Flumioxazin	103361-09-7	herbicide	x	x	
213	Fluometuron	2164-17-2	herbicide		x	
214	Fluopicolide	239110-15-7	fungicide	x	x	
215	Fluopyram	658066-35-4	fungicide	x	x	5.12000
216	Fluoxastrobin	361377-29-9	fungicide	x	x	
217	Flupyrsulfuron	150315-10-9	herbicide	x	x	
218	Fluquinconazole	136426-54-5	fungicide	x	x	0.80000
219	Flurochloridon	61213-25-0	herbicide		x	
220	Fluroxypyr	69377-81-7	herbicide	x	x	16.00000
221	Fluroxypyr-methylheptyl	81406-37-3	herbicide			
222	Flurtamone	96525-23-4	herbicide	x	x	0.99000
223	Flusilazol	85509-19-9	fungicide			1.10000
224	Flutolanil	66332-96-5	fungicide	x	x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
225	Flutriafol	76674-21-0	fungicide		x	
226	Fluxapyroxad	907204-31-3	fungicide	x	x	
227	Folpet	133-07-3	fungicide	x	x	
228	Foramsulfuron	173159-57-4	herbicide	x	x	0.09500
229	Fosetyl	15845-66-6	fungicide	x	x	
230	Fosthiazat	98886-44-3	other	x	x	
231	Fuberidazol	3878-19-1	fungicide	x	x	
232	Furalaxyl	57646-30-7	fungicide			
233	Furmecyclox	60568-05-0	fungicide			
234	Glufosinat	51276-47-2	herbicide	x	x	
235	Glyphosate	1071-83-6	herbicide	x	x	100.00000
236	HCH, gamma (Lindan)	58-89-9	insecticide			
237	Haloxyfop	69806-34-4	herbicide			
238	Haloxyfop-P	95977-29-0	herbicide	x	x	
239	Haloxyfop-ethoxyethyl	87237-48-7	herbicide			
240	Heptachlor	76-44-8	insecticide			
241	Heptachlorepoxid	1024-57-3	metabolite			
242	Heptenophos	23560-59-0	insecticide			
243	Hexachlorbenzen	118-74-1	fungicide			
244	Hexachlorophen	70-30-4	other			
245	Hexaconazol	79983-71-4	fungicide			
246	Hexaflumuron	86479-06-3	insecticide			
247	Hexazinon	51235-04-2	herbicide			
248	Hexythiazox	78587-05-0	insecticide	x	x	
249	Hymexazol	10004-44-1	fungicide	x	x	
250	Icaridinsäure		metabolite			
251	Imazalil	35554-44-0	fungicide	x	x	
252	Imazamox	114311-32-9	herbicide	x	x	
253	Imazapic	104098-48-8	herbicide			
254	Imazaquin	81335-37-7	herbicide		x	
255	Imazethapyr	81335-77-5	herbicide			
256	Imazosulfuron	122548-33-8	herbicide	x	x	
257	Imidacloprid	138261-41-3	insecticide	x	x	0.00900
258	Indoxacarb	173584-44-6	insecticide	x	x	
259	Iodosulfuron	185119-76-0	herbicide	x	x	0.07900
260	Iodosulfuron-methyl	144550-06-1	herbicide			
261	Iodosulfuron-methyl-sodium	144550-36-7	herbicide			
262	Ioxynil	1689-83-4	herbicide	x		2.70000
263	Iprodion	36734-19-7	fungicide	x	x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
264	Iprovalicarb	140923-17-7	fungicide	x	x	189.00000
265	Isodrin	465-73-6	insecticide			
266	Isophenphos	25311-71-1	insecticide			
267	Isoproturon	34123-59-6	herbicide	x	x	1.30000
268	Isopyrazam	881685-58-1	fungicide	x	x	
269	Isoxaben	82558-50-7	herbicide	x	x	
270	Isoxaflutole	141112-29-0	herbicide	x	x	
271	Karbutylat	4849-32-5	herbicide			
272	Kresoxim-methyl	143390-89-0	fungicide	x	x	1.00000
273	Kresoximsäure		metabolite			
274	Lenacil	2164-08-1	herbicide	x	x	0.65000
275	Linuron	330-55-2	herbicide		x	
276	MCPA	94-74-6	herbicide	x	x	9.00000
277	MCPB	94-81-5	herbicide		x	
278	Malathion	121-75-5	insecticide		x	
279	Mancozeb	8018-01-7	fungicide	x	x	0.21900
280	Mandipropamid	374726-62-2	fungicide	x	x	7.60000
281	Maneb	12427-38-2	fungicide	x	x	
282	Mecoprop	93-65-2	herbicide		x	160.00000
283	Mefenpyr-diethyl	135591-00-3	other	x		
284	Mepanipyrim	110235-47-7	fungicide	x	x	
285	Mepiquat	15302-91-7	other	x	x	
286	Mepronil	55814-41-0	fungicide			
287	Meptyldinocap	131-72-6	fungicide		x	
288	Mesosulfuron	400852-66-6	herbicide	x	x	
289	Mesotrion	104206-82-8	herbicide	x	x	
290	Metaflumizone	139968-49-3	insecticide	x	x	
291	Metalaxyl	57837-19-1	fungicide		x	46.00000
292	Metalaxyl-CA	75596-99-5	metabolite			
293	Metalaxyl-CA2	104390-56-9	metabolite			
294	Metalaxyl-M	70630-17-0	fungicide	x	x	46.00000
295	Metaldehyd	108-62-3	other	x	x	
296	Metamitron	41394-05-2	herbicide	x	x	38.00000
297	Metamitron-Desamino	36993-94-9	metabolite			
298	Metazachlor	67129-08-2	herbicide	x	x	0.88000
299	Metazachlor-dicarbonsäure		metabolite			
300	Metazachlor-sulfonsäure	172960-62-2	metabolite			
301	Metazachlorsäure	1231244-60-2	metabolite			
302	Metconazol	125116-23-6	fungicide	x	x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
303	Methabenzthiazuron	18691-97-9	herbicide			
304	Methamidophos	10265-92-6	insecticide			2.60000
305	Methidathion	950-37-8	insecticide			
306	Methiocarb	2032-65-7	insecticide	x	x	0.01000
307	Methobromuron	3060-89-7	herbicide		x	2.00000
308	Methomyl	16752-77-5	insecticide		x	
309	Methoprotryn	841-06-5	herbicide			
310	Methoxychlor	72-43-5	insecticide			
311	Methoxyfenozid	161050-58-4	insecticide	x	x	
312	Methyldesphenyl-Chloridazon	17254-80-7	metabolite			
313	Metiram	9006-42-2	fungicide	x	x	
314	Metolachlor	51218-45-2	herbicide			
315	Metolachlor-sulfonsäure	171118-09-5	metabolite			
316	Metolachlorsäure	152019-73-3	metabolite			
317	Metosulam	139528-85-1	herbicide	x	x	
318	Metoxuron	19937-59-8	herbicide			
319	Metrafenon	220899-03-6	fungicide	x	x	
320	Metribuzin	21087-64-9	herbicide	x	x	0.58400
321	Metsulfuron	79510-48-8	herbicide	x	x	
322	Metsulfuron-methyl	74223-64-6	herbicide			
323	Mevinphos	7786-34-7	insecticide			
324	Milbemectin	51596-11-3	insecticide	x	x	
325	Mirex	2385-85-5	insecticide			
326	Monolinuron	1746-81-2	herbicide			
327	Monuron	150-68-5	herbicide			
328	Myclobutanil	88671-89-0	fungicide	x	x	2.40000
329	Napropamid	15299-99-7	herbicide	x	x	6.70000
330	Neburon	555-37-3	herbicide			
331	Nicosulfuron	111991-09-4	herbicide	x	x	0.08500
332	Nitenpyram	120738-89-8	insecticide			
333	Nitrofen	1836-75-5	herbicide			
334	Norflurazon	27314-13-2	herbicide			
335	Omethoat	1113-02-6	insecticide			
336	Orysastrobin	248593-16-0	fungicide			
337	Oxadiazon	19666-30-9	herbicide		x	
338	Oxadixyl	77732-09-3	fungicide			
339	Oxamyl	23135-22-0	insecticide		x	
340	Oxydemeton-methyl	301-12-2	insecticide			1.10000
341	Paclobutrazol	76738-62-0	other	x	x	
342	Parathion-ethyl	56-38-2	insecticide			

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
343	Parathion-methyl	298-00-0	insecticide			
344	Pelargonsäure	112-05-0	herbicide	x	x	
345	Penconazol	66246-88-6	fungicide	x	x	3.20000
346	Pencycuron	66063-05-6	fungicide	x	x	
347	Pendimethalin	40487-42-1	herbicide	x	x	0.63000
348	Penflufen	494793-67-8	fungicide		x	
349	Penoxsulam	219714-96-2	herbicide	x	x	
350	Permethrin	52645-53-1	insecticide			
351	Pethoxamid	106700-29-2	herbicide	x	x	1.77200
352	Phenmedipham	13684-63-4	herbicide	x	x	
353	Phoxim	14816-18-3	insecticide			0.00700
354	Picloram	1918-02-1	herbicide	x	x	
355	Picolinafen	137641-05-5	herbicide	x	x	0.03600
356	Picoxystrobin	117428-22-5	fungicide	x	x	0.60000
357	Pinoxaden	243973-20-8	herbicide	x		
358	Pirimicarb	23103-98-2	insecticide	x	x	0.09000
359	Pirimicarb-desmethyl	30614-22-3	metabolite			
360	Pirimiphos-ethyl	23505-41-1	insecticide			
361	Pirimiphos-methyl	29232-93-7	insecticide	x	x	
362	Primisulfuron-methyl	86209-51-0	herbicide			
363	Prochloraz	67747-09-5	fungicide	x	x	5.00000
364	Procymidon	32809-16-8	fungicide			
365	Profoxydim	139001-49-3	herbicide		x	
366	Prohexadion	88805-35-0	other	x	x	
367	Prometryn	7287-19-6	herbicide			
368	Propachlor	1918-16-7	herbicide			
369	Propamocarb	24579-73-5	fungicide	x	x	
370	Propanil	709-98-8	herbicide			
371	Propaquizafop	111479-05-1	herbicide	x	x	
372	Propazin	139-40-2	herbicide			
373	Propetamphos	31218-83-4	insecticide			
374	Propham	122-42-9	herbicide			
375	Propiconazol	60207-90-1	fungicide	x	x	2.00000
376	Propoxur	114-26-1	insecticide			
377	Propoxycarbazone	145026-81-9	herbicide	x	x	
378	Propyzamid	23950-58-5	herbicide	x	x	34.00000
379	Proquinazid	189278-12-4	fungicide	x	x	
380	Prosulfocarb	52888-80-9	herbicide	x	x	3.80000
381	Prosulfuron	94125-34-5	herbicide	x	x	
382	Prothioconazol	178928-70-6	fungicide	x	x	1.71000

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
383	Prothioconazol-desthio	120983-64-4	metabolite			
384	Pymetrozin	123312-89-0	insecticide	x	x	
385	Pyraclostrobin	175013-18-0	fungicide	x	x	
386	Pyraflufen	129630-17-7	herbicide	x	x	
387	Pyrazophos	13457-18-6	fungicide			
388	Pyrethrum	8003-34-7	insecticide	x	x	0.01400
389	Pyridaben	96489-71-3	insecticide		x	
390	Pyridat	55512-33-9	herbicide	x	x	
391	Pyrifenox	88283-41-4	fungicide			
392	Pyrimethanil	53112-28-0	fungicide	x	x	8.00000
393	Pyroxsulam	422556-08-9	herbicide	x	x	
394	Quinalphos	13593-03-8	insecticide			
395	Quinmerac	90717-03-6	herbicide	x	x	316.00000
396	Quinoclamin	2797-51-5	herbicide	x	x	
397	Quinoxyfen (5,7-dichloro-4-(p-fluorophenoxy)quinoline)	124495-18-7	fungicide	x	x	
398	Quintozen	82-68-8	fungicide			
399	Quizalofop	76578-12-6	herbicide			
400	Quizalofop-ethyl	76578-14-8	herbicide			
401	Rimsulfuron	122931-48-0	herbicide	x	x	0.46000
402	Saflufenacil	372137-35-4	herbicide			
403	Sebuthylazin	7286-69-3	herbicide			
404	Secbumeton	26259-45-0	herbicide			
405	Silthiofam	175217-20-6	fungicide	x	x	
406	Simazin	122-34-9	herbicide			
407	Simazin, 2-Hydroxy	2599-11-3	metabolite			
408	Spinosad	168316-95-8	insecticide	x	x	0.06200
409	Spirodiclofen	148477-71-8	insecticide	x	x	
410	Spiromesifen	283594-90-1	insecticide		x	
411	Spiroxamin	118134-30-8	fungicide	x	x	0.13000
412	Sulcotrion	99105-77-8	herbicide	x	x	
413	Sulfosulfuron	141776-32-1	herbicide		x	
414	Sulfurylfluorid	2699-79-8	insecticide	x	x	
415	Tebuconazol	107534-96-3	fungicide	x	x	0.57800
416	Tebufenozid	112410-23-8	insecticide	x	x	
417	Tebufenpyrad	119168-77-3	insecticide	x	x	
418	Tebutam	35256-85-0	herbicide			
419	Teflubenzuron	83121-18-0	insecticide		x	
420	Tefluthrin	79538-32-2	insecticide	x	x	
421	Telodrin	297-78-9	insecticide			

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
422	Tembotrione	335104-84-2	herbicide	x	x	
423	Tepraloxydim	149979-41-9	herbicide	x	x	
424	Terbumeton	33693-04-8	herbicide			
425	Terbuthylazin	5915-41-3	herbicide	x	x	1.20000
426	Terbutryn	886-50-0	herbicide			
427	Terbutylazin-Metabolit CGA 324007	309923-18-0	metabolite			
428	Terbutylazin-Metabolit SYN 545666		metabolite			
429	Tetraconazol	112281-77-3	fungicide	x	x	
430	Thiabendazol	148-79-8	fungicide	x	x	
431	Thiacloprid	111988-49-9	insecticide	x	x	0.00400
432	Thiacloprid-SA		metabolite			
433	Thiamethoxam	153719-23-4	insecticide	x	x	0.04300
434	Thiencarbazon-methyl	317815-83-1	herbicide	x	x	
435	Thifensulfuron-methyl	79277-27-3	herbicide			
436	Thifenylsulfuron	79277-67-1	herbicide	x	x	
437	Thiometon	640-15-3	insecticide			
438	Thiophanat-methyl	23564-05-8	fungicide	x	x	
439	Thiram	137-26-8	fungicide	x	x	0.11000
440	Tolclofos-methyl	57018-04-9	fungicide	x	x	
441	Tolylfluanid	731-27-1	fungicide			
442	Topramezone	210631-68-8	herbicide	x		0.90000
443	Tralkoxydim	87820-88-0	herbicide		x	
444	Triadimefon	43121-43-3	fungicide			
445	Triadimenol	55219-65-3	fungicide	x	x	3.40000
446	Triallat	2303-17-5	herbicide		x	
447	Triasulfuron	82097-50-5	herbicide	x	x	
448	Triazophos	24017-47-8	insecticide			0.03000
449	Triazoxid	72459-58-6	fungicide	x	x	
450	Tribenuron	106040-48-6	herbicide	x	x	
451	Tribenuron-methyl	101200-48-0	herbicide			
452	Trichlorfon	52-68-6	insecticide			
453	Triclopyr	55335-06-3	herbicide	x	x	
454	Trifloxystrobin	141517-21-7	fungicide	x	x	0.08620
455	Trifloxystrobin-CA2		metabolite			
456	Triflumizol	99387-89-0	fungicide		x	
457	Triflumuron	64628-44-0	insecticide		x	

Table B.2 Continued.

	Name	CAS	Group	Auth. GER[a]	Auth. EU[b]	RAC [c]
458	Trifluralin	1582-09-8	herbicide			
459	Triflusulfuron	135990-29-3	herbicide	x	x	
460	Triforin	26644-46-2	fungicide			
461	Trinexapac-ethyl	95266-40-3	other	x	x	
462	Triticonazol	131983-72-7	fungicide	x	x	
463	Tritosulfuron	142469-14-5	herbicide	x	x	
464	Valifenalate	283159-90-0	fungicide	x	x	
465	Vinclozolin	50471-44-8	fungicide			
466	Warfarin	81-81-2	other			
467	Zoxamid	156052-68-5	fungicide	x	x	
468	alpha-Cypermethrin	67375-30-8	insecticide	x	x	
469	cis-Chlordan	5103-71-9	insecticide			
470	gamma-Cyhalothrin	76703-62-3	insecticide	x	x	
471	o,p-DDE	3424-82-6	metabolite			
472	o,p-DDT	789-02-6	insecticide			
473	oxi-Chlordan	27304-13-8	metabolite			
474	p,p-DDD (p,p TDE)	72-54-8	insecticide			
475	p,p-DDE	72-55-9	metabolite			
476	p,p-DDT	50-29-3	insecticide			
477	tau-Fluvalinat	102851-06-9	insecticide	x	x	0.03300
478	trans-Chlordan	5103-74-2	insecticide			

THRESHOLDS FOR AGRICULTURAL LAND USE AND CATCHMENT SIZE

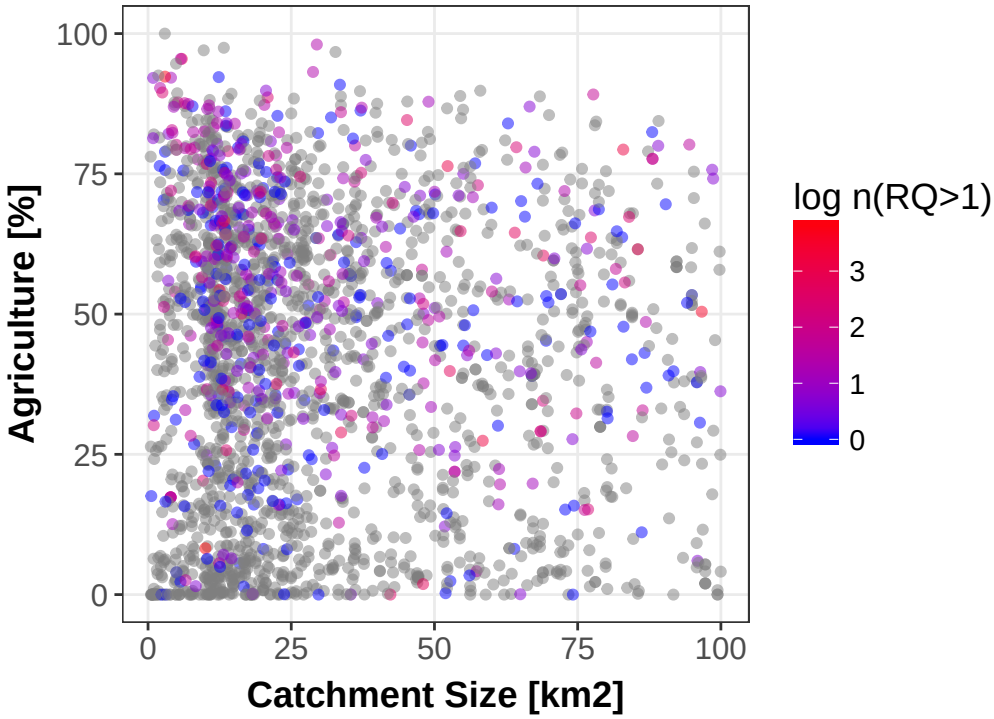


Figure B.5.: Raw data used for the model in equation 2 and Figure 3 of the main article. Color codes the number of RAC exceedances (on a log-scale). Grey points denote sites without any exceedance.

EFFECT OF PRECIPITATION AND SEASON ON RQ

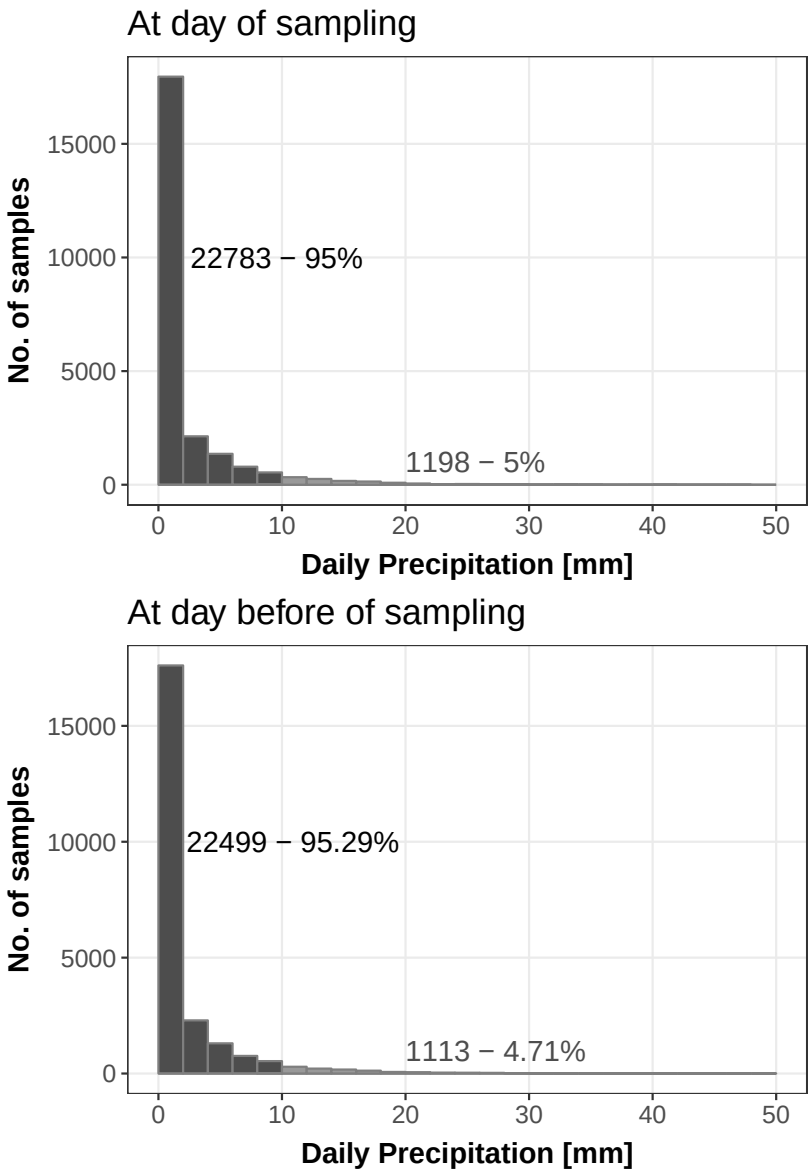


Figure B.6.: Distribution of precipitation at sampling occasions. top: at sampling date. bottom: at the day before sampling.

Table B.3.: 23 pesticides for which we modelled the relationship between RQ and precipitation and seasonality, respectively. Order is the same as in Figure 5 of the main text. See Table B.4 for model coefficients.

	Name	CAS	Group	%>LOQ	no. > LOQ	total no.
1	Azoxystrobin	131860-33-8	fungicide	9.58	644	6723
2	Bentazon	25057-89-0	herbicide	19.43	2313	11905
3	Boscalid	188425-85-6	fungicide	23.00	2175	9455
4	Carbendazim	10605-21-7	fungicide	16.10	582	3615
5	Chlorpyrifos	2921-88-2	insecticide	6.17	865	14026
6	Clothianidin	210880-92-5	insecticide	6.30	141	2237
7	Diflufenican	83164-33-4	herbicide	12.63	1867	14784
8	Dimoxystrobin	149961-52-4	fungicide	6.83	216	3164
9	Diuron	330-54-1	herbicide	12.07	2138	17708
10	Ethofumesat	26225-79-6	herbicide	5.10	998	19552
11	Flufenacet	142459-58-3	herbicide	5.97	772	12923
12	Glyphosate	1071-83-6	herbicide	40.73	1389	3410
13	Imidacloprid	138261-41-3	insecticide	5.88	176	2992
14	Isoproturon	34123-59-6	herbicide	21.84	3984	18239
15	MCPA	94-74-6	herbicide	12.81	1567	12237
16	Mecoprop	93-65-2	herbicide	12.21	1463	11984
17	Metazachlor	67129-08-2	herbicide	9.23	1930	20907
18	Nicosulfuron	111991-09-4	herbicide	5.33	263	4934
19	Propiconazol	60207-90-1	fungicide	5.67	772	13622
20	Quinmerac	90717-03-6	herbicide	13.46	939	6974
21	Tebuconazol	107534-96-3	fungicide	6.08	968	15924
22	Terbuthylazin	5915-41-3	herbicide	14.59	3142	21540

Table B.4.: Coefficients and CI from per compound models. Bold values denote coefficients where the CI for precipitation encompasses zero. Coefficients are on the link scale (log for μ and logit for ν).

	Compound	effect	$\log \text{precip}_0$	$\log \text{precip}_{-1}$	Quarter 1	Quarter 2	Quarter 3	Quarter 4
1	Azoxystrobin	μ	**0.23**	0.04	-3.39	-3.02	-3.16	-3.47
			(0.15 - 0.31)	(-0.03 - 0.12)	(-3.56 - -3.22)	(-3.14 - -2.89)	(-3.29 - -3.03)	(-3.63 - -3.3)
2	Bentazon	μ	-0.03	0.02	-9.46	-8.97	-9.14	-9.46
			(-0.07 - 0)	(-0.02 - 0.05)	(-9.53 - -9.38)	(-9.02 - -8.92)	(-9.2 - -9.07)	(-9.53 - -9.39)
3	Boscalid	μ	**0.06**	**0.1**	-6.72	-6.42	-6.51	-6.58
			(0.02 - 0.1)	(0.06 - 0.13)	(-6.79 - -6.64)	(-6.49 - -6.36)	(-6.58 - -6.45)	(-6.65 - -6.5)
4	Carbendazim	μ	**-0.1**	**0.16**	-2.42	-1.95	-2.11	-2.32
			(-0.16 - -0.03)	(0.09 - 0.22)	(-2.58 - -2.27)	(-2.05 - -1.84)	(-2.22 - -2)	(-2.46 - -2.18)
5	Chlorpyrifos	μ	**0.08**	-0.03	0.85	1	0.9	0.94
			(0.04 - 0.13)	(-0.08 - 0.01)	(0.77 - 0.93)	(0.93 - 1.06)	(0.82 - 0.98)	(0.86 - 1.03)
6	Clothianidin	μ	0.08	-0.1	0.94	0.67	1.02	1.55
			(-0.04 - 0.19)	(-0.21 - 0.02)	(0.77 - 1.12)	(0.49 - 0.84)	(0.8 - 1.25)	(1.32 - 1.78)
7	Diflufenican	μ	-0.02	**0.05**	-0.56	-1.01	-1.08	-0.71
			(-0.06 - 0.02)	(0.02 - 0.09)	(-0.62 - -0.5)	(-1.07 - -0.94)	(-1.16 - -1)	(-0.77 - -0.65)
8	Dimoxystrobin	μ	**0.35**	0.02	-1.17	-0.42	-0.07	-0.02
			(0.2 - 0.5)	(-0.15 - 0.19)	(-1.44 - -0.89)	(-0.64 - -0.2)	(-0.39 - 0.25)	(-0.35 - 0.31)
9	Diuron	μ	0	**0.07**	-2.72	-2.43	-2.48	-2.64
			(-0.03 - 0.03)	(0.04 - 0.1)	(-2.83 - -2.61)	(-2.47 - -2.39)	(-2.53 - -2.44)	(-2.71 - -2.58)
10	Ethofumesat	μ	**0.12**	0.01	-6.11	-5.49	-6.18	-6.1
			(0.06 - 0.17)	(-0.05 - 0.06)	(-6.27 - -5.96)	(-5.56 - -5.42)	(-6.29 - -6.08)	(-6.24 - -5.95)
11	Flufenacet	μ	0.03	**0.05**	-3.71	-3.7	-3.29	-3.63
			(-0.02 - 0.08)	(0.01 - 0.1)	(-3.79 - -3.62)	(-3.81 - -3.59)	(-3.44 - -3.15)	(-3.68 - -3.57)
12	Glyphosate	μ	-0.04	**0.14**	-6.3	-6.08	-5.73	-6.11
			(-0.09 - 0.01)	(0.09 - 0.19)	(-6.46 - -6.13)	(-6.16 - -6)	(-5.8 - -5.66)	(-6.21 - -6.01)
13	Imidacloprid	μ	0	-0.01	0.61	1.15	1.4	1.24
			(-0.08 - 0.09)	(-0.09 - 0.07)	(0.33 - 0.88)	(1.02 - 1.27)	(1.28 - 1.53)	(1.06 - 1.42)
14	Isoproturon	μ	0.02	**0.21**	-3.29	-3.01	-3.43	-2.79
			(-0.02 - 0.05)	(0.17 - 0.24)	(-3.37 - -3.22)	(-3.06 - -2.96)	(-3.5 - -3.35)	(-2.84 - -2.73)
15	MCPA	μ	0.04	**0.09**	-5.07	-4.25	-4.48	-4.7
			(-0.01 - 0.09)	(0.04 - 0.14)	(-5.27 - -4.87)	(-4.32 - -4.19)	(-4.57 - -4.4)	(-4.81 - -4.58)

Table B.4 Continued.

	Compound	effect	log precip$_0$	log precip$_{-1}$	Quarter 1	Quarter 2	Quarter 3	Quarter 4
16	Mecoprop	μ	0.04	**0.05**	-8.36	-7.59	-7.77	-8.07
			(-0.01 - 0.09)	(0.01 - 0.1)	(-8.49 - -8.22)	(-7.65 - -7.52)	(-7.85 - -7.69)	(-8.18 - -7.97)
17	Metazachlor	μ	**-0.07**	**0.09**	-2.97	-2.94	-2.21	-2.77
			(-0.12 - -0.02)	(0.04 - 0.13)	(-3.06 - -2.88)	(-3.04 - -2.85)	(-2.28 - -2.14)	(-2.84 - -2.7)
18	Nicosulfuron	μ	**0.23**	**-0.28**	-0.98	-0.2	-0.07	-0.97
			(0.12 - 0.34)	(-0.39 - -0.18)	(-1.22 - -0.74)	(-0.36 - -0.03)	(-0.25 - 0.11)	(-1.16 - -0.78)
19	Propiconazol	μ	**0.08**	0.01	-3.99	-3.63	-3.82	-3.63
			(0.02 - 0.14)	(-0.05 - 0.07)	(-4.15 - -3.83)	(-3.71 - -3.55)	(-3.91 - -3.72)	(-3.74 - -3.53)
20	Quinmerac	μ	0.02	0.05	-9.08	-9.12	-8.46	-8.64
			(-0.05 - 0.09)	(-0.01 - 0.12)	(-9.19 - -8.96)	(-9.24 - -9)	(-8.59 - -8.33)	(-8.72 - -8.55)
21	Tebuconazol	μ	-0.01	**0.09**	-2.17	-1.93	-2.2	-2.15
			(-0.06 - 0.03)	(0.04 - 0.14)	(-2.28 - -2.06)	(-2 - -1.86)	(-2.28 - -2.11)	(-2.24 - -2.06)
22	Terbuthylazin	μ	**0.09**	**0.11**	-3.65	-2.78	-3.25	-3.52
			(0.06 - 0.13)	(0.08 - 0.15)	(-3.73 - -3.56)	(-2.84 - -2.73)	(-3.3 - -3.19)	(-3.59 - -3.44)
23	Azoxystrobin	ν	0	**0.24**	-3.5	-2.33	-2.14	-3.2
			(-0.13 - 0.13)	(0.11 - 0.37)	(-3.76 - -3.25)	(-2.54 - -2.13)	(-2.36 - -1.92)	(-3.45 - -2.95)
24	Bentazon	ν	0	0.05	-2.26	-1.53	-1.88	-2.25
			(-0.08 - 0.08)	(-0.03 - 0.13)	(-2.44 - -2.09)	(-1.65 - -1.4)	(-2.02 - -1.74)	(-2.4 - -2.11)
25	Boscalid	ν	-0.01	**0.45**	-1.99	-1.22	-1.24	-1.81
			(-0.1 - 0.08)	(0.37 - 0.54)	(-2.16 - -1.82)	(-1.36 - -1.07)	(-1.38 - -1.09)	(-1.96 - -1.65)
26	Carbendazim	ν	0.09	**0.19**	-2.72	-1.49	-1.26	-2.31
			(-0.04 - 0.22)	(0.06 - 0.32)	(-3 - -2.44)	(-1.69 - -1.28)	(-1.48 - -1.04)	(-2.56 - -2.06)
27	Chlorpyrifos	ν	**0.11**	0.1	-3.27	-2.63	-3.22	-3.42
			(0.01 - 0.21)	(0 - 0.19)	(-3.45 - -3.1)	(-2.79 - -2.48)	(-3.39 - -3.05)	(-3.61 - -3.23)
28	Clothianidin	ν	-0.05	0.19	-2.66	-2.58	-3.19	-3.93
			(-0.3 - 0.2)	(-0.07 - 0.44)	(-3.06 - -2.26)	(-2.97 - -2.19)	(-3.69 - -2.69)	(-4.46 - -3.41)
29	Diflufenican	ν	0.06	**0.26**	-1.89	-2.45	-3.14	-2.09
			(-0.02 - 0.14)	(0.17 - 0.34)	(-2.03 - -1.75)	(-2.59 - -2.31)	(-3.3 - -2.98)	(-2.22 - -1.95)
30	Dimoxystrobin	ν	0.19	**0.23**	-3.37	-2.25	-3.14	-3.58
			(-0.02 - 0.41)	(0.01 - 0.46)	(-3.78 - -2.96)	(-2.58 - -1.91)	(-3.55 - -2.72)	(-4.02 - -3.15)

Table B.4 Continued.

	Compound	effect	log precip$_0$	log precip$_{-1}$	Quarter 1	Quarter 2	Quarter 3	Quarter 4
31	Diuron	ν	0.05	**0.28**	-3.88	-1.67	-1.74	-2.72
			(-0.01 - 0.12)	(0.22 - 0.35)	(-4.09 - -3.67)	(-1.76 - -1.58)	(-1.84 - -1.63)	(-2.85 - -2.6)
32	Ethofumesat	ν	0.09	**0.21**	-4.39	-2.23	-3.49	-4.23
			(-0.01 - 0.18)	(0.12 - 0.3)	(-4.63 - -4.16)	(-2.35 - -2.11)	(-3.66 - -3.32)	(-4.44 - -4.01)
33	Flufenacet	ν	**0.16**	**0.59**	-2.57	-3.8	-4.17	-1.76
			(0.06 - 0.27)	(0.49 - 0.69)	(-2.75 - -2.39)	(-4.01 - -3.58)	(-4.44 - -3.89)	(-1.88 - -1.64)
34	Glyphosate	ν	0.11	**0.29**	-1.79	-0.12	0.34	-0.53
			(0 - 0.23)	(0.18 - 0.4)	(-2.09 - -1.48)	(-0.3 - 0.05)	(0.17 - 0.51)	(-0.73 - -0.32)
35	Imidacloprid	ν	-0.01	-0.1	-4.68	-3.04	-2.83	-4.07
			(-0.26 - 0.25)	(-0.34 - 0.15)	(-5.35 - -4)	(-3.41 - -2.68)	(-3.21 - -2.45)	(-4.56 - -3.58)
36	Isoproturon	ν	0.04	**0.31**	-1.82	-1.19	-2.11	-0.8
			(-0.02 - 0.09)	(0.25 - 0.36)	(-1.93 - -1.7)	(-1.27 - -1.12)	(-2.22 - -2.01)	(-0.88 - -0.72)
37	MCPA	ν	-0.06	**0.35**	-3.79	-1.27	-1.81	-2.77
			(-0.13 - 0.02)	(0.28 - 0.42)	(-4.04 - -3.54)	(-1.37 - -1.18)	(-1.93 - -1.68)	(-2.92 - -2.62)
38	Mecoprop	ν	0.07	**0.35**	-3.04	-1.56	-1.89	-2.71
			(-0.01 - 0.15)	(0.27 - 0.42)	(-3.23 - -2.84)	(-1.67 - -1.45)	(-2.02 - -1.76)	(-2.86 - -2.56)
39	Metazachlor	ν	0.06	**0.21**	-2.81	-3.22	-2.11	-2.05
			(-0.01 - 0.13)	(0.14 - 0.27)	(-2.94 - -2.67)	(-3.36 - -3.09)	(-2.22 - -2.01)	(-2.16 - -1.95)
40	Nicosulfuron	ν	**0.2**	**0.26**	-3.87	-2.96	-2.99	-3.23
			(0.01 - 0.39)	(0.07 - 0.45)	(-4.27 - -3.48)	(-3.26 - -2.66)	(-3.3 - -2.68)	(-3.56 - -2.9)
41	Propiconazol	ν	-0.02	**0.39**	-4.05	-2.72	-2.88	-3.43
			(-0.13 - 0.09)	(0.29 - 0.5)	(-4.32 - -3.78)	(-2.88 - -2.57)	(-3.06 - -2.7)	(-3.63 - -3.24)
42	Quinmerac	ν	-0.03	**0.32**	-2.23	-2.58	-2.49	-1.2
			(-0.13 - 0.08)	(0.22 - 0.42)	(-2.43 - -2.02)	(-2.76 - -2.41)	(-2.69 - -2.29)	(-1.34 - -1.06)
43	Tebuconazol	ν	**0.1**	**0.3**	-3.41	-2.66	-2.9	-3.17
			(0.01 - 0.2)	(0.21 - 0.39)	(-3.61 - -3.2)	(-2.8 - -2.53)	(-3.06 - -2.75)	(-3.34 - -3)
44	Terbuthylazin	ν	**0.06**	**0.28**	-2.92	-1.45	-1.48	-2.47
			(0.01 - 0.12)	(0.22 - 0.33)	(-3.05 - -2.79)	(-1.53 - -1.37)	(-1.57 - -1.39)	(-2.58 - -2.37)

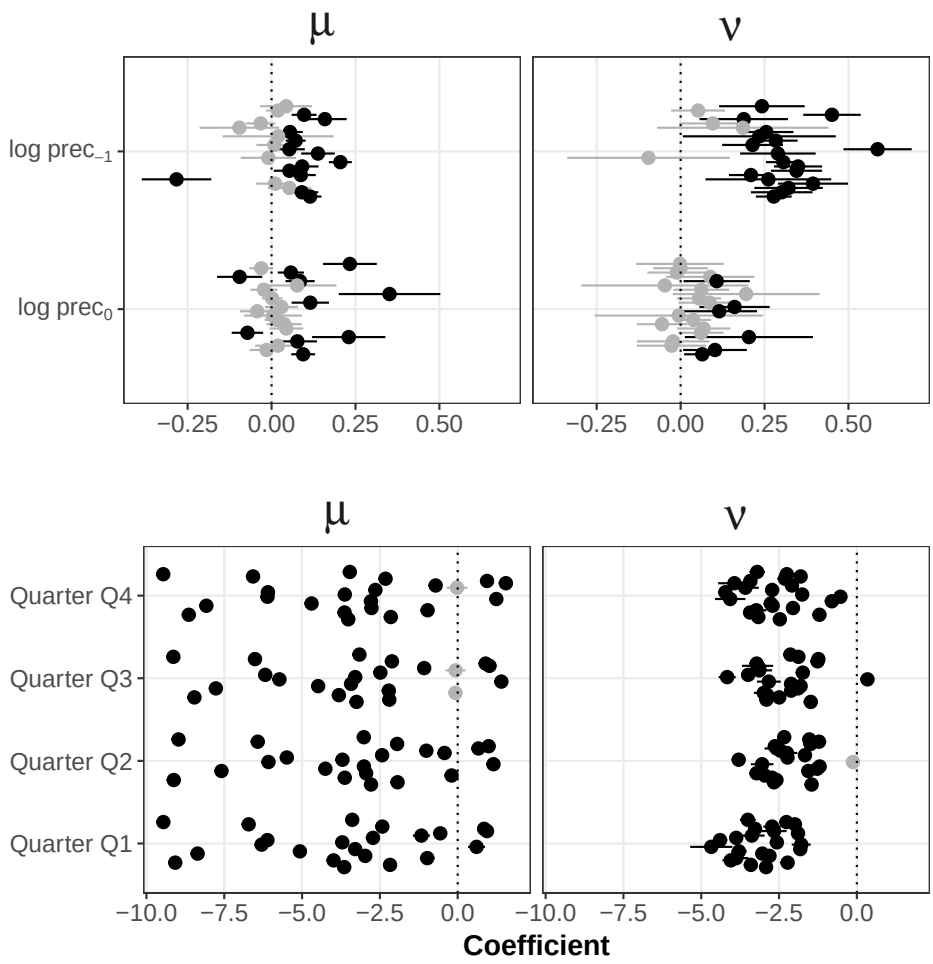


Figure B.7.: Graphical representation of coefficients from table B.4. Top row: Effect of precipitation at the day before sampling and at day of sampling. Bottom row: estimates for the four Quarters. Each dot represents one compound (in the order described in table B.3). Coefficients where the CI encompasses zero are shown in gray colour. Coefficients are shown on the link scale (log for μ and logit for ν).

PESTICIDES IN SMALL STREAMS

Table B.5.: Overview on RAC exceedances of the 78 compounds with more than 1000 measurements. No. = number of measurements; % RQ >1 = RAC exceedances; % RQ >1 | >LOQ= RAC exceedances as fraction of detects.

Name	No.	No. >LOQ	% >LOQ	No. RQ >1	% RQ >1	% RQ >1 | >LOQ
2,4-D	12290	284	2.3	10	0.1	3.5
Aclonifen	9861	67	0.7	4	0.0	6.0
Azoxystrobin	7059	690	9.8	6	0.1	0.9
Benalaxyl	6964	10	0.1	0	0.0	0.0
Bentazon	12429	2421	19.5	0	0.0	0.0
Bifenthrin	1353	0	0.0	0	0.0	
Boscalid	9886	2296	23.2	0	0.0	0.0
Bromoxynil	9451	78	0.8	0	0.0	0.0
Carbendazim	3851	654	17.0	12	0.3	1.8
Chloridazon	15724	511	3.2	0	0.0	0.0
Chlorpyrifos	14704	954	6.5	838	5.7	87.8
Chlortoluron	18286	371	2.0	2	0.0	0.5
Clomazon	9268	440	4.7	0	0.0	0.0
Clopyralid	5520	107	1.9	0	0.0	0.0
Clothianidin	2409	154	6.4	123	5.1	79.9
Cypermetryn	1428	5	0.4	1	0.1	20.0
Cyprodinil	9779	118	1.2	0	0.0	0.0
Dicamba	7641	76	1.0	0	0.0	0.0
Difenoconazol	1644	11	0.7	2	0.1	18.2
Diflufenican	15457	1932	12.5	273	1.8	14.1
Dimefuron	7833	5	0.1	0	0.0	0.0
Dimethachlor	8858	344	3.9	0	0.0	0.0
Dimethoat	14423	185	1.3	1	0.0	0.5
Dimethomorph	2316	91	3.9	0	0.0	0.0
Dimoxystrobin	3370	232	6.9	49	1.5	21.1
Diuron	18560	2336	12.6	40	0.2	1.7
Epoxiconazol	16454	621	3.8	7	0.0	1.1
Ethofumesat	20430	1078	5.3	0	0.0	0.0
Fenhexamid	2690	42	1.6	0	0.0	0.0
Fenpropimorph	12850	199	1.5	5	0.0	2.5
Fluazifop	3022	57	1.9	0	0.0	0.0
Fluazifop-P	4033	14	0.3	0	0.0	0.0
Fluazifop-P-butyl	1728	0	0.0	0	0.0	
Fluazifop-butyl	1287	0	0.0	0	0.0	
Fludioxonil	3203	42	1.3	1	0.0	2.4
Flufenacet	13509	798	5.9	1	0.0	0.1
Fluquinconazole	6762	117	1.7	0	0.0	0.0
Fluroxypyr	8096	378	4.7	0	0.0	0.0

Table B.5 Continued.

Name	No.	No. >LOQ	% >LOQ	No. RQ >1	% RQ >1	% RQ >1 | >LOQ
Flurtamone	16958	638	3.8	2	0.0	0.3
Flusilazol	5257	53	1.0	1	0.0	1.9
Glyphosate	3557	1455	40.9	1	0.0	0.1
Imidacloprid	3169	192	6.1	169	5.3	88.0
Ioxynil	8114	20	0.2	0	0.0	0.0
Isoproturon	19112	4164	21.8	92	0.5	2.2
Kresoxim-methyl	6929	14	0.2	0	0.0	0.0
Lenacil	13837	183	1.3	0	0.0	0.0
MCPA	12773	1687	13.2	2	0.0	0.1
Mecoprop	12521	1552	12.4	0	0.0	0.0
Metalaxyl	14460	299	2.1	0	0.0	0.0
Metamitron	15390	613	4.0	0	0.0	0.0
Metazachlor	21906	2015	9.2	55	0.3	2.7
Methamidophos	1303	0	0.0	0	0.0	
Methobromuron	14968	24	0.2	1	0.0	4.2
Metribuzin	15411	192	1.2	15	0.1	7.8
Napropamid	9914	269	2.7	1	0.0	0.4
Nicosulfuron	5172	288	5.6	77	1.5	26.7
Penconazol	4846	159	3.3	0	0.0	0.0
Pendimethalin	16997	328	1.9	4	0.0	1.2
Pethoxamid	3102	37	1.2	0	0.0	0.0
Phoxim	1492	0	0.0	0	0.0	
Picolinafen	8901	11	0.1	2	0.0	18.2
Picoxystrobin	3620	7	0.2	0	0.0	0.0
Pirimicarb	11330	232	2.0	27	0.2	11.6
Prochloraz	5795	33	0.6	0	0.0	0.0
Propiconazol	14250	818	5.7	7	0.0	0.9
Propyzamid	11937	453	3.8	0	0.0	0.0
Prosulfocarb	5001	126	2.5	0	0.0	0.0
Pyrimethanil	8136	122	1.5	0	0.0	0.0
Quinmerac	7291	989	13.6	0	0.0	0.0
Rimsulfuron	1240	2	0.2	0	0.0	0.0
Spiroxamin	2469	109	4.4	1	0.0	0.9
Tebuconazol	16584	1024	6.2	26	0.2	2.5
Terbuthylazin	22568	3370	14.9	35	0.2	1.0
Thiacloprid	3540	85	2.4	85	2.4	100.0
Thiamethoxam	1853	39	2.1	7	0.4	17.9
Triadimenol	3067	51	1.7	0	0.0	0.0
Triazophos	3588	2	0.1	1	0.0	50.0
Trifloxystrobin	3674	10	0.3	1	0.0	10.0

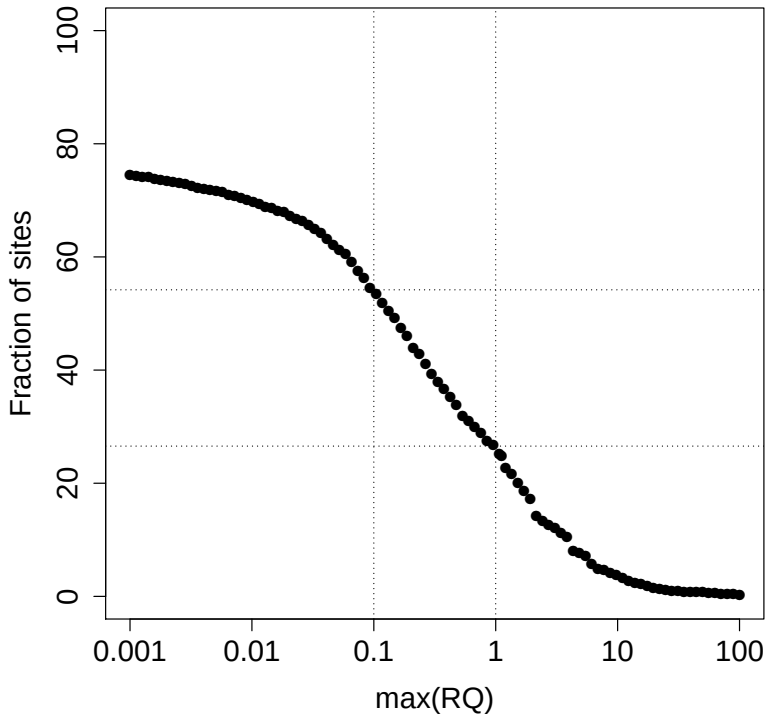


Figure B.8.: Cumulative distribution of sites exceeding RAC. Dotted lines indicate fraction of sites exceeding a RQ of 1 and 0.1. 23% of the sites showed no detection of compounds with available RAC values and are not shown due to logarithmic x-axis.

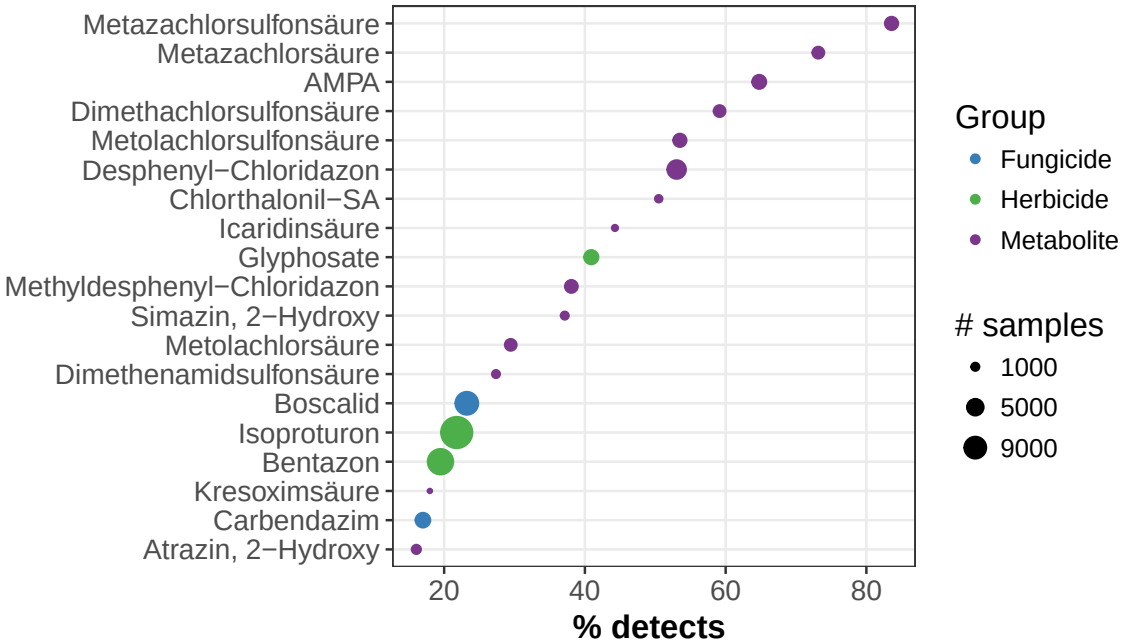


Figure B.9.: Proportion of samples with detects in small streams. Only Compounds with more than 100 samples and 15% of detects are shown.

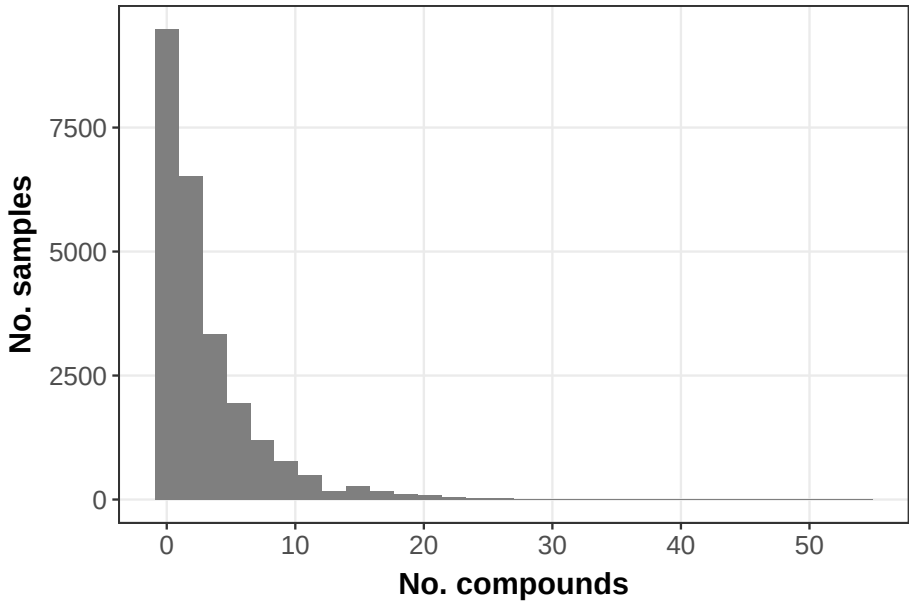


Figure B.10.: Distribution of the number of quantified compounds in the samples.

CATCHMENT SIZE - STREAM WIDTH RELATIONSHIPS

We studied the relationship between catchment size based on three datasets containing this informations: Data delivered by the federal state Thuringia, Voß et al., (2015) and Fernández et al., (2015) (both from Rhineland-Palatinate). We fitted to each dataset separately and to the combined dataset a power-function. The resulting models are shown in Figure B.11.

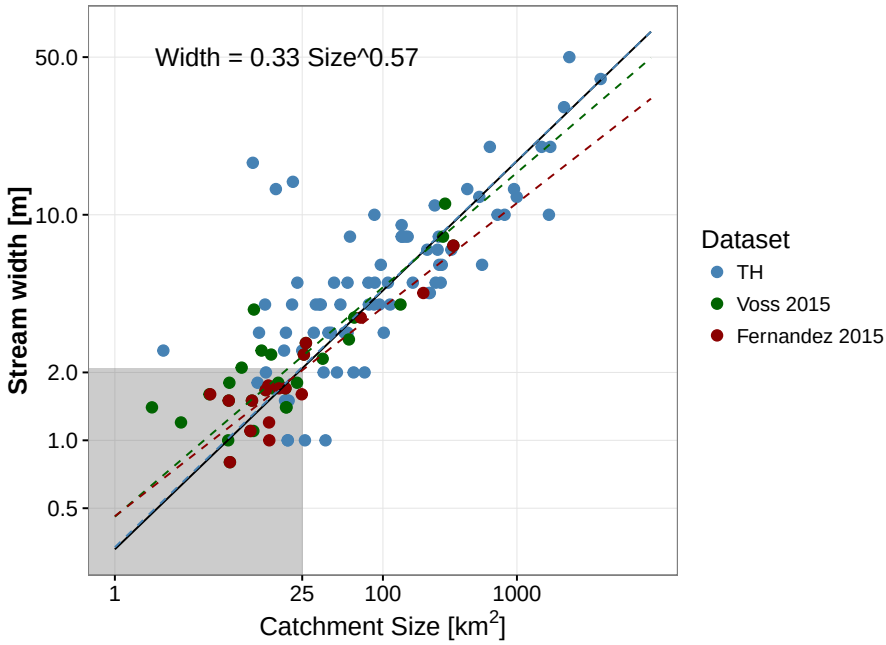


Figure B.11.: Relationship between catchment size and stream width. A power function has been fitted to each dataset separately and the combined dataset (black line and equation). The gray rectangle marks the estimated with for a catchment size of 25km².

REFERENCES

Fernández, D., K. Voss, M. Bundschuh, J. P. Zubrod, and R. B. Schäfer (2015). "Effects of fungicides on decomposer communities and litter decomposition in vineyard streams". *Science of The Total Environment* 533, 40–48.

Voß, K., D. Fernández, and R. Schäfer (2015). "Organic matter breakdown in streams in a region of contrasting anthropogenic land use". *Science of The Total Environment* 527-528, 179–184.

C SUPPLEMENTAL MATERIAL FOR: TAXIZE: TAXONOMIC SEARCH AND RETRIEVAL

A COMPLETE REPRODUCIBLE WORKFLOW - FROM A SPECIES LIST TO A PHYLOGENY, AND DISTRIBUTION MAP.

If you aren't familiar with a complete workflow in R, it may be difficult to visualize the process. In R, everything is programmatic, so the whole workflow can be in one place, and be repeated whenever necessary. The following is a workflow for taxize, going from a species list to a phylogeny.

First, install taxize

```
R> install.packages("taxize")
```

Then load it into R

```
R> library(taxize)
```

Most of us will start out with a species list, something like the one below. Note that each of the names is spelled incorrectly.

```
R> splist <- c("Helanthus annuus", "Pinos contorta",
               "Collomia grandiflorra", "Rosa california",
               "Mimulus bicolour", "Nicotiana glauca", "Maddia sativa")
```

There are many ways to resolve taxonomic names in taxize. Of course, the ideal name resolver will do the work behind the scenes for you so that you don't have to do things like fuzzy matching. There are a few services in taxize like this we can choose from: the Global Names Resolver service from EOL (see function *gnr_resolve*) and the Taxonomic Name Resolution Service from iPlant (see function *tnrs*). In this case let's use the function *tnrs*.

```
# The tnrs function accepts a vector of 1 or more
R> splist_tnrs <- tnrs(query = splist, getpost = "POST",
                       source_ = "iPlant_TNRS")
# Remove some fields
R> (splist_tnrs <- splist_tnrs[, !names(splist_tnrs) %in%
                       c("matchedName", "annotations",
                       "uri")])
```

```
#           submittedName         acceptedName    sourceId score
# 5      Helanthus annuus    Helianthus annuus iPlant_TNRS  0.98
# 1        Pinos contorta       Pinus contorta iPlant_TNRS  0.96
# 7 Collomia grandiflorra Collomia grandiflora iPlant_TNRS  0.99
# 6       Rosa california     Rosa californica iPlant_TNRS  0.99
# 4      Mimulus bicolour      Mimulus bicolor iPlant_TNRS  0.98
# 3      Nicotiana glauca     Nicotiana glauca iPlant_TNRS     1
# 2         Maddia sativa         Madia sativa iPlant_TNRS  0.97

# Note the scores. They suggest that there were no perfect matches,
# but they were all very close, ranging from 0.77 to 0.99
# (1 is the highest).
# Let's assume the names in the 'acceptedName' column
# are correct (and they should
# be).
# So here's our updated species list
R> (splist <- as.character(splist_tnrs$acceptedName))

# [1] "Helianthus annuus"  "Pinus contorta"  "Collomia grandiflora"
# [4] "Rosa californica"  "Mimulus bicolor"  "Nicotiana glauca"
# [7] "Madia sativa"
```

Another thing we may want to do is collect common names for our taxa.

```
R> tsns <- get_tsn(searchterm = splist, searchtype = "sciname",
                   verbose = FALSE)
R> comnames <- lapply(tsns, getcommonnamesfromtsn)
# Unfortunately, common names are not standardized like species
# names, so there are multiple common names for each taxon
R> sapply(comnames, length)

# [1] 3 3 3 3 3 3 3

# So let's just take the first common name for each species
R> comnames_vec <- do.call(c, lapply(comnames,
            function(x) as.character(x[1, "comname"])))
# And we can make a data.frame of our scientific and common names
R> (allnames <- data.frame(spname = splist, comname = comnames_vec))

#                 spname                      comname
# 1    Helianthus annuus              common sunflower
# 2       Pinus contorta                lodgepole pine
# 3 Collomia grandiflora        largeflowered collomia
```

```
# 4     Rosa californica           California wildrose
# 5      Mimulus bicolor yellow and white monkeyflower
# 6     Nicotiana glauca                  tree tobacco
# 7         Madia sativa                coast tarweed
```

Another common task is getting the taxonomic tree upstream from your study taxa. We often know what family or order our taxa are in, but it we often don't know the tribes, subclasses, and superfamilies. taxize provides many avenues to getting classifications. Two of them are accessible via a single function (*classification*): the Integrated Taxonomic Information System (ITIS) and National Center for Biotechnology Information (NCBI); and via the Catalogue of Life (see function *col_classification*):

```
# As we already have Taxonomic Serial Numbers from ITIS, let's just
# get classifications from ITIS.Note that we could use uBio instead.
R> class_list <- classification(tsns)
R> sapply(class_list, nrow)

# [1] 12 11 12 12 12 12 12

# And we can attach these names to our allnames data.frame
R> library(plyr)
R> gethiernames <- function(x) {
        temp <- x[, c("rankName", "taxonName")]
        values <- data.frame(t(temp[, 2]))
        names(values) <- temp[, 1]
        return(values)
R> }
R> class_df <- ldply(class_list, gethiernames)
R> allnames_df <- merge(allnames, class_df, by.x = "spname",
                        by.y = "Species")
# Now that we have allnames_df, we can start to see some
# relationships among species simply by their shared taxonomic names
R> allnames_df[1:2, ]

#                 spname                comname Kingdom     Subkingdom
# 1 Collomia grandiflora largeflowered collomia Plantae Viridaeplantae
# 2    Helianthus annuus       common sunflower Plantae Viridaeplantae
#   Infrakingdom     Division     Subdivision Infradivision
# 1 Streptophyta Tracheophyta Spermatophytina  Angiospermae
# 2 Streptophyta Tracheophyta Spermatophytina  Angiospermae
#   Class         Superorder     Order        Family      Genus
# 1  Magnoliopsida Asteranae  Ericales Polemoniaceae   Collomia
# 2  Magnoliopsida Asteranae Asterales    Asteraceae Helianthus
```

```
# Ah, so Abies and Bartlettia are in different infradivisions, but
# share taxonomic names above that point.
```

However, taxonomy can only get you so far. Shared ancestry can be reconstructed from molecular data, and phylogenies created. Phylomatic is a web service with an API that we can use to get a phylogeny.

```
# Fetch phylogeny from phylomatic
R> phylogeny <- phylomatic_tree(taxa = as.character(allnames$spname),
            taxnames = TRUE,
            get = "POST", informat = "newick", method = "phylomatic",
            storedtree = "R20120829",
            taxaformat = "slashpath", outformat = "newick", clean = "true",
            parallel = TRUE)
# Format teeth-labels
R> phylogeny$tip.label <- capwords(phylogeny$tip.label,
                          onlyfirst = TRUE)
# plot phylogeny
R> plot(phylogeny)
```

Using the species list, with the corrected names, we can now search for occurrence data. The Global Biodiversity Information Facility (GBIF) has the largest collection of records data, and has a API that we can interact with programmatically from R. First, we need to install rgbif.

```
# Install rgbif from github.com
R> install.packages("devtools")
R> library(devtools)
R> install_github("rgbif", "ropensci")
```

Now we can search for occurrences for our species list and make a map.

```
R> library(rgbif)
R> library(ggplot2)
# get occurences
R> occurr_list <- occurrencelist_many(as.character(allnames$spname),
            coordinatestatus = TRUE,
            maxresults = 100, removeZeros = TRUE,
            fixnames = "changealltorig")
# Make a map
R> p <- gbifmap_list(occurr_list) +
            guides(col = guide_legend(title = "", nrow = 3,
            byrow = TRUE)) + theme(legend.position = "bottom",
```

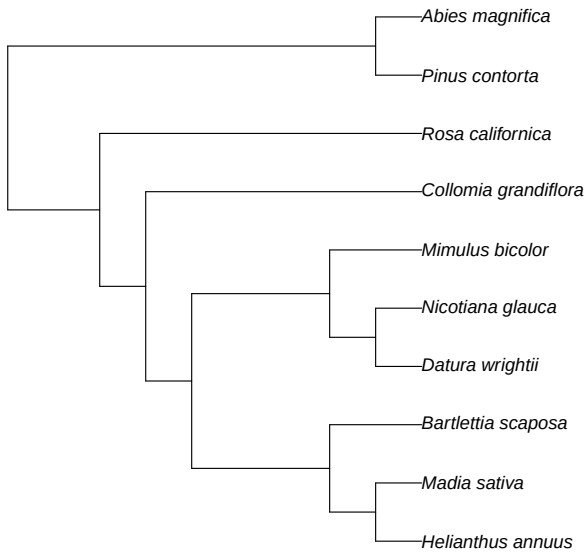


Figure C.1.: A phylogeny created using taxize.

```
            legend.key = element_blank()) +
            coord_equal()
R> p
```

MATCHING SPECIES TABLES WITH DIFFERENT TAXONOMIC RESOLUTION

Trait-based approaches are a promising tool in ecology. Unlike taxonomy-based methods, traits may not be constrained to biogeographic boundaries (Baird et al., 2011) and have potential to disentangle the effects of multiple stressors (Statzner and Bêche, 2010).

To analyse trait-composition abundance data must be matched with trait databases like (Usseglio-Polatera et al., 2000). However these two datatables may contain species information on different taxonomic levels and perhaps data must be aggregated to a joint taxonomic level.

taxize can help in this data-cleaning step, providing a reproducible workflow. Here we illustrate this on a small fictitious example.

Suppose we have fuzzy coded trait table with 2 traits with 3 respectively 2 modalities:

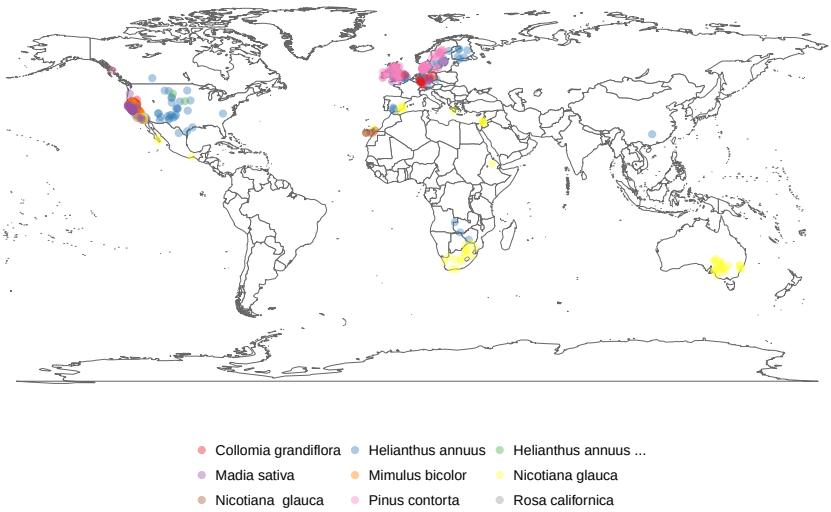


Figure C.2.: A map created using taxize.

```
(traits <- read.table(header = TRUE, sep = ';', stringsAsFactors=FALSE,
                      text = 'taxon;T1M1;T1M2;T1M3;T2M1;T2M2
Gammarus sp.;0;0;3;1;3
Potamopyrgus antipodarum;1;0;3;1;3
Coenagrion sp.;3;0;1;3;1
Enallagma cyathigerum;0;3;1;0;3
Erythromma sp.;0;0;3;3;1
Baetis sp.;0;0;0;0;0
'))
```

```
                   taxon T1M1 T1M2 T1M3 T2M1 T2M2
1            Gammarus sp.    0    0    3    1    3
2 Potamopyrgus antipodarum    1    0    3    1    3
3          Coenagrion sp.    3    0    1    3    1
4   Enallagma cyathigerum    0    3    1    0    3
5          Erythromma sp.    0    0    3    3    1
6              Baetis sp.    0    0    0    0    0
```

And want to match this to a table with abundances:

```
(abundances <- read.table(header = TRUE, sep = ';', stringsAsFactors=FALSE,
                          text = 'taxon;abundance;sample
Gammarus roeseli;5;1
Gammarus roeseli;6;2
Gammarus tigrinus;7;1
Gammarus tigrinus;8;2
Coenagrionidae;10;1
Coenagrionidae;6;2
Potamopyrgus antipodarum;10;1
xxxxx;10;2
'))
```

```
                    taxon abundance sample
1         Gammarus roeseli         5      1
2         Gammarus roeseli         6      2
3        Gammarus tigrinus         7      1
4        Gammarus tigrinus         8      2
5           Coenagrionidae        10      1
6           Coenagrionidae         6      2
7 Potamopyrgus antipodarum        10      1
8                   xxxxx        10      2
```

First we do some basic data-cleaning and create a lookup-table, that will link taxa in trait table with the abundance table.

```
# first we remove ' sp.' from out trait table:
traits$taxon_cleaned <- tolower(gsub(" sp.", "", traits$taxon))
# since abundance tables can be very long with repeating taxa, we look only
# at unique taxon names This will be a lookup-table linking taxon names
# between both tables
lookup <- data.frame(taxon = tolower(unique(abundances$taxon)),
            stringsAsFactors = FALSE)
```

The we query the taxonomic hierarchy for both tables, this will be the backbone of this procedure:

```
library(taxize)
traits_classi <- classification(get_uid(traits$taxon_cleaned))
lookup_classi <- classification(get_uid(lookup$taxon))
```

First we look if we can find any direct matches between taxon names:

```
# first search for direct matches
direct <- match(lookup$taxon, traits$taxon_cleaned)
# and add the matched name to our lookup table
lookup$traits <- tolower(traits$taxon[direct])
lookup$match <- ifelse(!is.na(direct), "direct", NA)
lookup
```

```
                    taxon                   traits  match
1         gammarus roeseli                     <NA>   <NA>
2        gammarus tigrinus                     <NA>   <NA>
3           coenagrionidae                     <NA>   <NA>
4 potamopyrgus antipodarum potamopyrgus antipodarum direct
5                   xxxxx                     <NA>   <NA>
```

We found a direct match - *potamopyrgus antipodarum* - so nothing to do here.

Next we look for species which are on a higher taxonomic resolution than our trait table. For these species we will take directly the trait-data since no better information is available.

```
# look for cases where taxonomic resolution in abundance data is higher
# than in trait data: here we take the trait-values for the lower
# resolution.
for (i in which(is.na(lookup$traits))) {
    if (is.data.frame(lookup_classi[[i]])) {
        matches <- tolower(lookup_classi[[i]]$ScientificName) %in%
        traits$taxon_cleaned
        if (any(matches)) {
            lookup$traits[i] <- tolower
            (lookup_classi[[i]]$ScientificName[matches])
            lookup$match[i] <- lookup_classi[[i]]$Rank[matches]
        }
    }
}
lookup
```

```
                    taxon                     traits  match
1         gammarus roeseli                  gammarus  genus
2        gammarus tigrinus                  gammarus  genus
3           coenagrionidae                      <NA>   <NA>
4 potamopyrgus antipodarum potamopyrgus antipodarum direct
5                   xxxxx                      <NA>   <NA>
```

So our abundance data has two *Gammarus* species, however trait data is only on genus level.

The next step is to search for species were we have to aggregate trait-data, since our abundance data is on a lower taxonomic level. We are walking the taxonomic latter for the species in our trait-

data upwards and search for matches with out abundance data. Since we'll have many taxa in the trait-data belonging to one taxon, we'll take the median modality scores as an approximation. Of course also other methods may be used here, e.g. weighting by genetic distance.

```
# look for cases taxonomic resolution in abundance data is lower than in
# trait data, here we need to aggregate the trait-values (eg. median value
# for modality)
for (i in which(is.na(lookup$traits))) {
    # find matches
    agg <- sapply(traits_classi, function(x) any(
    tolower(x$ScientificName) %in%
        lookup$taxon[i]))
    if (sum(agg) > 1) {
        # add taxon as aggregate to trait-table
        traits <- rbind(traits, c(paste0(lookup$taxon[i], "-aggregated"),
        apply(traits[agg,
            2:6], 2, median), paste0(lookup$taxon[i], "-aggregated")))
        # fill lookup table
        lookup$traits[i] <- paste0(lookup$taxon[i], "-aggregated")
        lookup$match[i] <- "aggregated"
    }
}
lookup

#                     taxon                    traits      match
# 1         gammarus roeseli                 gammarus      genus
# 2        gammarus tigrinus                 gammarus      genus
# 3           coenagrionidae coenagrionidae-aggregated aggregated
# 4 potamopyrgus antipodarum  potamopyrgus antipodarum     direct
# 5                   xxxxx                      <NA>       <NA>
```

Finally we have only one taxon left - clearly an error. We remove this from our dataset:

```
abundances <- abundances[!abundances$taxon == lookup$taxon[is.na(
                      lookup$traits)],
    ]
```

Now we can create *species x sites* and *traits x species* matrices, which could be plugged into methods to analyse trait responses [28].

```
# species (as matched with trait table) by site matrix
abundances$traits_taxa <- lookup$traits[match(tolower(abundances$taxon),
                lookup$taxon)]
```

```
library(reshape2)
# reshape data to long format and name rows by samples
L <- dcast(abundances, sample ~ traits_taxa, fun.aggregate = sum,
          value.var = "abundance")
rownames(L) <- L$sample
L$sample <- NULL
L

#   coenagrionidae-aggregated gammarus potamopyrgus antipodarum
# 1                        10       12                       10
# 2                         6       14                        0

# traits by species matrix
Q <- traits[, 2:7][match(names(L), traits$taxon_cleaned), ]
rownames(Q) <- Q$taxon_cleaned
Q$taxon_cleaned <- NULL
Q

#                           T1M1 T1M2 T1M3 T2M1 T2M2
# coenagrionidae-aggregated    0    0    1    3    1
# gammarus                     0    0    3    1    3
# potamopyrgus antipodarum     1    0    3    1    3

# check
all(rownames(Q) == colnames(L))

# [1] TRUE
```

This is just an example how taxonomic APIs (via taxize) could be used to search for matches up- and downwards the taxonomic ladder. We are looking forward to integrate other databases into taxize, which will facilitate trait-based analyses in R.

REFERENCES

Baird, D. J., C. J. O. Baker, R. B. Brua, M. Hajibabaei, K. McNicol, T. J. Pascoe, and D. de Zwart (2011). "Toward a knowledge infrastructure for traits-based ecological risk assessment". *Integrated Environmental Assessment and Management* 7 (2), 209–215.

Statzner, B. and L. Bêche (2010). "Can biological invertebrate traits resolve effects of multiple stressors on running water ecosystems?" *Freshwater Biology* 55, 80–119.

Usseglio-Polatera, P., M. Bournaud, P. Richoux, and H. Tachet (2000). "Biological and ecological traits of benthic freshwater macroinvertebrates: relationships and definition of groups with similar traits". *Freshwater Biology* 43 (2), 175–205.

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

CONTRIBUTIONS: Szöcs (85%) Designed research and simulations, analysed data, discussed results, wrote manuscript

Schäfer (15%) Designed research, discussed results, edited manuscript

ARTICLE II

TITLE: Large scale risks from pesticides in small streams

AUTHORS: Eduard Szöcs, Marvin Brinke, Bilgin Karaoglan, and Ralf B. Schäfer

STATUS: Submitted to *Environmental Science & Technology* in 2016

CONTRIBUTIONS: Szöcs (75%) Designed research

Brinke (5%) helped with data, commented on manuscript

Karaoglan (5%) provided data (RACs), commented on manuscript

Schäfer (15%) Designed research, discussed results, edited manuscript

ARTICLE III

TITLE: webchem: An R Package to Retrieve Chemical Information from the Web.

AUTHORS: Eduard Szöcs and Ralf B. Schäfer

STATUS: Accepted in 2016 in *Journal of Statistical Software*

CONTRIBUTIONS: Szöcs (90%) Designed, programmed and tested software, wrote manuscript

Schäfer (10%) discussed results, edited manuscript

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*

CONTRIBUTIONS: Chamberlain (50%) Designed, programmed and tested software, wrote manuscript

Szöcs (50%) Designed, programmed and tested 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,
15. November 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

12.2013	**Research Assistant**, *University of Koblenz-Landau*, Landau. Development of a PostgreSQL-database of german physico-chemical data.
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.
2011 – 2015	**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	LaTeX, git, PostgreSQL, GrassGIS, PostGIS, shell, regex, xml, xpath, cloud computing
Beginner	CDK, RDKit, openbabel, NetLogo, Python, C++

(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 **webchem** package to retrieve chemical information from the web.
- The **taxize** package (together with Scott Chamberlain) allows taxonomic search, retrieval and handling in R.
- Contributions to the **vegan** package.
- The **rspear** package calculates $SPEAR_{pesticides}$ 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 and environmental sciences.

All software is freely available from my github account (`https://github.com/EDiLD`), homepage (`https://edild.github.io`) or The Comprehensive R Archive Network (CRAN).

Publications and Conference contributions

Articles

[1] L. Lagadic, R. B. Schäfer, M. Roucaute, E. **Szöcs**, S. Chouin, J. de Maupeou, C. Duchet, E. Franquet, B. Le Hunsec, C. Bertrand, S. Fayolle, B. Francés, Y. Rozier, R. Foussadier, J.-B. Santoni, and C. Lagneau (2016). "No association between the use of Bti for mosquito control and the dynamics of non-target aquatic invertebrates in French coastal and continental wetlands". en. *Science of The Total Environment* 553, 486–494.

[2] V. C. Schreiner, E. **Szöcs**, A. K. Bhowmik, M. G. Vijver, and R. B. Schäfer (2016). "Pesticide mixtures in streams of several European countries and the USA". *Science of The Total Environment* 573, 680–689.

[3] E. **Szöcs** and R. B. Schäfer (2016). "Statistical hypothesis testing—To transform or not to transform?" en. *Integrated Environmental Assessment and Management* 12 (2), 398–400.

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

[5] E. **Szöcs**, P. J. v. d. Brink, L. Lagadic, T. Caquet, M. Roucaute, A. Auber, Y. Bayona, M. Liess, P. Ebke, A. Ippolito, C. J. F. t. Braak, T. C. M. Brock, and R. B. Schäfer (2015). "Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods". *Ecotoxicology* 24 (4), 760–769.

[6] E. **Szöcs** and R. B. Schäfer (2015). "Ecotoxicology is not normal: A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology". en. *Environmental Science and Pollution Research* 22 (18), 13990–13999.

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

[8] S. A. Chamberlain and E. **Szöcs** (2013). "taxize: taxonomic search and retrieval in R [v2; ref status: indexed, http://f1000r.es/24v]". *F1000Research* 2 (191).

[9] R. B. Schäfer, M. Bundschuh, D. A. Rouch, E. **Szöcs**, P. C. von der Ohe, V. Pettigrove, R. Schulz, D. Nugegoda, and B. J. Kefford (2012). "Effects of pesticide toxicity, salinity and other environmental variables on selected ecosystem functions in streams and the relevance for ecosystem services". *Science of the Total Environment* 415 (1), 69–78.

[10] E. **Szöcs**, B. J. Kefford, and R. B. Schäfer (2012). "Is there an interaction of the effects of salinity and pesticides on the community structure of macroinvertebrates?" *Science of the Total Environment* 437 (1), 121–126.

Poster

[1] E. **Szöcs** and R. B. Schäfer (2015). "Ecotoxicology is not normal." Poster. SETAC Europe; Barcelona.

[2] E. **Szöcs** and S. A. Chamberlain (2014). "taxize: taxonomic search and retrieval in R". Poster. International Statistical Ecology Conference 2014; Montpellier.

[3] E. **Szöcs**, B. J. Kefford, V. Pettigrove, and R. B. Schäfer (2011). "Einfluss von Pestiziden und Salinität auf Makroinvertebratengemeinschaften". Poster. SETAC GLB; Landau.

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, November 12, 2016