technische universität
dortmund

fakultät
statistik

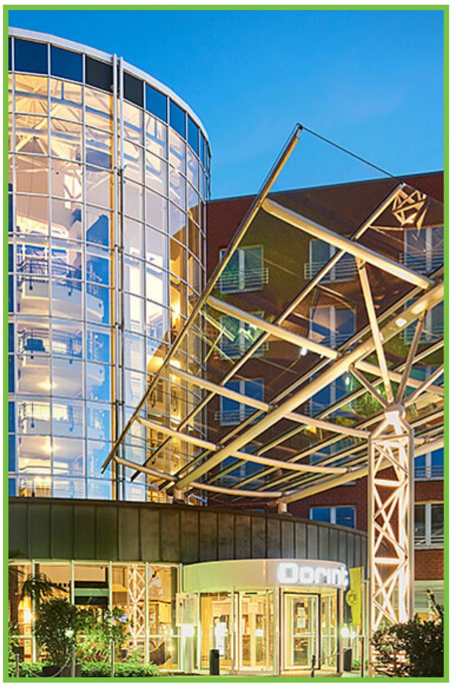
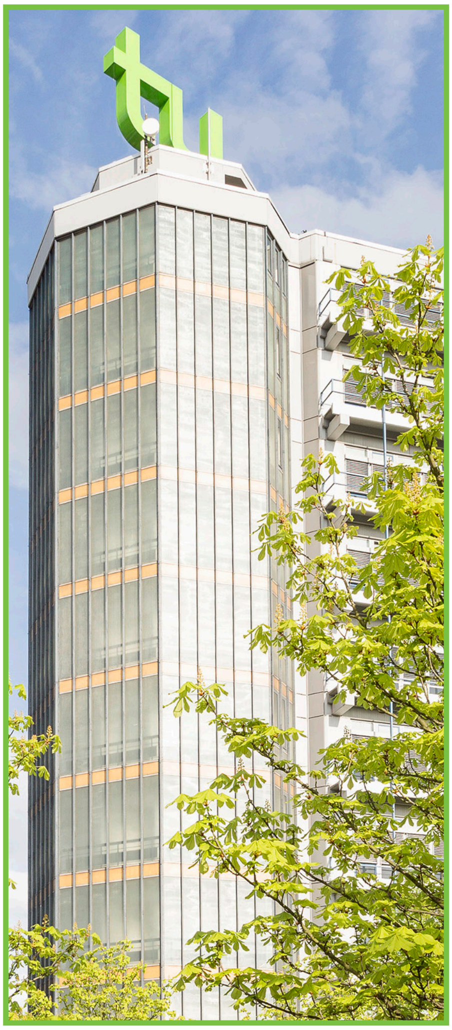
IWSM 2023

37th International Workshop on Statistical Modelling

16.07. – 21.07.2023

Dortmund

Proceedings book

Proceedings of the 37th International Workshop on Statistical Modelling

July 17-21, 2023 - Dortmund, Germany

Editors
Elisabeth Bergherr
Andreas Groll
Andreas Mayr

TU Dortmund University

International Workshop on Statistical Modelling (37°. 2023. Dortmund)

Proceedings of the 37th International Workshop on Statistical Modelling : July 17-21, 2022
Dortmund, Germany / Elisabeth Bergherr, Andreas Groll, Andreas Mayr (editors). – Dortmund : TU Dortmund University, 2023. – 1 copy online : PDF (693 S. : ill.)

ISBN: 978-3-947323-42-5

Authors:

Bergherr, Elisabeth
Groll, Andreas
Mayr, Andreas

Topics:

1. Statistics congress. 2. Econometrics models congress
330.015195 = Mathematical statistics

Editors:

ELISABETH BERGHERR
University of Göttingen, Chair of Spatial Data Science and Statistical Learning

ANDREAS GROLL
TU Dortmund University, Department of Statistics

ANDREAS MAYR
University of Bonn, Department of Medical Biometry, Informatics and Epidemiology

ISBN 978-3-947323-42-5 (online)

TU Dortmund University
Department of Statistics
Vogelpothsweg 78
44227 Dortmund
Germany

https://ub.tu-dortmund.de/
https://statistik.tu-dortmund.de/

Scientific Committee

Ruggero Bellio

University of Udine (Italy)

Elisabeth Bergherr (Co-Chair)

University of Göttingen (Germany)

Fernanda De Bastiani

University of Pernambuco (Brazil)

María L. Durbán Reguera

University of Madrid (Spain)

Jan Gertheiss

Helmut Schmidt University, Hamburg (Germany)

Andreas Groll (Chair)

TU Dortmund (Germany)

Thomas Kneib

University of Göttingen (Germany)

Dae-Jin Lee

IE University, School of Science and Technology, Madrid (Spain)

Andreas Mayr (Co-Chair)

University of Bonn (Germany)

Fulvia Pennoni

University of Milano-Bicocca (Italy)

María Xosé Rodríguez Álvarez

University of Vigo (Spain)

Gunther Schauberger

TU München (Germany)

Nicola Torelli

University of Trieste (Italy)

Lola Ugarte

University of Navarra (Spain)

Nikolaus Umlauf

University of Innsbruck, (Austria)

Helga Wagner

University of Linz, (Austria)

Local Organising Committee

Chiara Balestra
TU Dortmund University

Elisabeth Bergherr (Co-Host)
University of Göttingen

Guillermo B. Sánchez
TU Dortmund University

Jennifer Engel
TU Dortmund University

Alexander Gerharz
TU Dortmund University

Colin Griesbach
University of Göttingen

Andreas Groll (Host)
TU Dortmund University

Tobias Hepp
University Erlangen-Nürnberg

Hannah Klinkhammer
University of Bonn

Andreas Mayr (Co-Host)
University of Bonn

Hendrik van der Wurp
TU Dortmund University

Preface

Dear Participants,

we are more than happy to host the 37th International Workshop on Statistical Modelling in Dortmund, Germany! This is the second year to meet in person after the COVID break, and we hope to have a wonderful time like we did last year in Trieste.

This year we will have 54 contributed talks and more than 60 posters, and it was a tough challenge to pick among the many excellent submissions we had! Thanks again to the scientific committee for putting so much work into the selection process. But we obviously also want to give our thanks to all the researchers, who contributed with their great submissions and made it possible to put together such an excellent set of presentations. Having a special focus on students is a tradition of the Statistical Modelling Society, hence we are especially happy to welcome such a large number of younger researchers contributing to the conference. We are already excited to find out who will win the awards for best student paper, best student presentation and best student poster! The Statistical Modelling Society furthermore awarded travel grants to two students.

We will also have five great invited talks, from different areas in statistics: Brian Reich, Maria Iannario, Alexander Gerharz together with Matthias Kolodziej, Gillian Heller and Simon Wood agreed to give keynotes at the workshop. Furthermore, Andreas Bender and Fabian Scheipl will provide a short course about Piece-wise Exponential (Additive) Models (PEMs / PAMs) before the conference starts.

As always, the IWSM is a one-track conference, leading to a familiar atmosphere and to the possibility for communication between the different fields of statistical modelling.

Looking back at all the years we were participating in great workshops, hosted at so many different universities and all the amazing people we got to meet there, we are both humble and exited to welcome you all to enjoy the conference and your stay at the river Ruhr area with its long tradition of coal mining and steal production, beer brewing and, of course, its omnipresent football vibe.

Andreas Groll, Elisabeth Bergherr and Andreas Mayr

Dortmund, Göttingen and Bonn, July 2023

Contents

Part I

1 Data science meets football 20
Alexander Gerharz, Mathias Kolodziej

2 Robust regression modelling for ordinal categorical data 28
Maria Iannario

3 Back to the future: model what you measure 38
Gillian Heller

4 Modeling extremal streamflow using deep learning approximations and a flexible spatial process 47
Reetam Majumder, Brian Reich, Banjamin Shaby

5 On Covid, dynamic models and inferring smooth functions 57
Simon Wood

Part II

6 State-switching decision trees 69
Timo Adam, Marius Ötting, Rouven Michels

7 Efficient stochastic learning of graphical structures for large-scale mixed data surveys 74
Giuseppe Alfonzetti, Ruggero Bellio, Yunxiao Chen, Irini Moustaki

8 Flexible habitat selection analysis with generalized additive models 80
Rafael Arce Guillen, Jennifer Pohle, Björn Reineking, Ulrike Schlägel

9 An information-theoretic perspective on double descent in flooded boosting 86
Chiara Balestra, Andrés Madariaga, Emmanuel Müller, Christian Staerk, Andreas Mayr

10 Adaptive random forests for high-dimensional regression 91
Moritz Berger, Christian Staerk

11 Evolutionary algorithm for the estimation of discrete latent variables models 97
Luca Brusa, Fulvia Pennoni, Francesco Bartolucci

12 Coherent cause-specific mortality forecasting via constrained penalized regression models 103
Carlo G. Camarda, María Durbán

13 The influence of resolution on the predictive power of spatial heterogeneity measures as a biomarker of disease severity 109
Jari Claes, Annelies Agten, Alfonso Blázquez-Moreno, Marjolein Crabbe, Marianne Tuefferd, Hinrich Goehlmann, Helena Geys, Thomas Neyens, Christel Faes

14 A multi-state model for the natural history of prostate cancer; using data from a screening trial 115
Ilse Cuevas Andrade, Ardo van den Hout, Nora Pashayan

15 Bayesian smoothing for joint extremes 121
Miguel de Carvalho, Junho Lee

16 Semi-parametric estimation of growth curves 125
Chiara Di Maria, Vito M. R. Muggeo

17 Modelling time-of-day variation in hidden Markov models using cyclic P-splines 131
Carlina C. Feldmann, Sina Mews, Roland Langrock

18 Bayesian inference of dynamic models emulated with a time series Gaussian process 135
Yuzhang Ge, Arash Rabbani, Hao Gao, Dirk Husmeier

19 Gradient boosting for parsimonious additive covariance matrix modelling 141
Vincenzo Gioia, Matteo Fasiolo, Ruggero Bellio

20 Functional multilevel modelling of the influence of the menstrual cycle on the performance of female cyclists **147**

Steven Golovkine, Tom Chassard, Alice Meignié, Emmanuel Brunet, Jean-Francois Toussaint, Juliana Antero

21 Confidence intervals for finite mixture regression based on resampling techniques **152**

Colin Griesbach, Tobias Hepp

22 Component-wise boosting for mixture distributional regression models **157**

Tobias Hepp, Jakob Zierk, Elisabeth Bergherr

23 Fusion, smoothing and model selection for item-on-item regression **163**

Aisouda Hoshiyar, Jan Gertheiss

24 Induced nonparametric ROC surface regression **169**

Vanda Inácio, María Xosé Rodríguez-Álvarez

25 Assessing spatial trends in health outcomes using primary care registry data **173**

Arne Janssens, Pieter Libin, Gijs Van Pottelbergh, Jonas Crèvecoeur, Bert Vaes, Thomas Neyens

26 Statistical inference for high-dimensional logistic regression: Variable selection and levels fusion for categorical covariates **178**

Lea Kaufmann, Maria Kateri

27 Advanced statistical modelling for polygenic risk scores based on large cohort data **183**

Hannah Klinkhammer, Christian Staerk, Carlo Maj, Peter M. Krawitz, Andreas Mayr

28 Sparse modality regression **188**

Chris Kolb, Bernd Bischl, Christian L. Müller, David Rügamer

29 On prediction via equal-tailed intervals with an application to sensor data analytics **193**

Michele Lambardi di San Miniato, Ruggero Bellio, Luca Grassetti, Paolo Vidoni

30 Asymmetry issues with non-penalized parameters in Laplace P-splines models 199
Philippe Lambert, Oswaldo Gressani

31 Local moment matching with Gamma mixtures and automatic smoothness selection 204
Oskar Laverny, Philippe Lambert

32 Linear mixed modelling of federated data when only the mean, covariance, and sample size are available 208
Marie Analiz April Limpoco, Christel Faes, Niel Hens

33 Feedforward neural networks from a statistical-modelling perspective 214
Andrew McInerney, Kevin Burke

34 Modelling medical claims data using Markov-modulated marked Poisson processes 219
Sina Mews, Bastian Surmann, Lena Hasemann, Svenja Elkenkamp

35 Estimating what is under the tip of gender-based violence: A statistical modelling approach 225
Isabel Millán, Amanda Fernández-Fontelo, Pere Puig, David Moriña

36 A bivariate Poisson regression model for radiation dose estimation 231
Dorota Młynarczyk, Pedro Puig, Carmen Armero, Virgilio Gómez-Rubio, Jayne Moquet

37 Bayesian spatio-temporal conditional overdispersion models proposals 237
Mabel Morales-Otero, Vicente Núñez-Antón

38 Lasso-based order selection in hidden Markov models: a case study using stock market data 243
Marius Ötting, Roland Langrock

39 Bayesian survival analysis using pseudo-observations 247
Léa Orsini, Caroline Brard, Emmanuel Lesaffre, David Dejardin, Gwénaël Le Teuff

40 Clustering anterior cruciate ligament reconstruction patients using functional walking biomechanics 253

Garritt L. Page, Matthew K. Seeley, Brian G. Pietrosimone

41 Forecasting insect abundance using time series embedding and environmental covariates 258

Gabriel R. Palma, Rodrigo F. Mello, Wesley A. Godoy, Eduardo Engel, Douglas Lau, Charles Markham, Rafael A. Moral

42 Studying animal interactions with Markov-switching step-selection models 262

Jennifer Pohle, Johannes Signer, Jana A. Eccard, Melanie Dammhahn, Ulrike E. Schlägel

43 Prediction-based variable selection for component-wise gradient boosting 267

Sophie Potts, Elisabeth Bergherr, Constantin Reinke, Colin Griesbach

44 Computationally efficient ranking of groundwater monitoring locations 273

Peter Radvanyi, Claire Miller, Craig Alexander, Marnie Low, Wayne R. Jones, Luc Rock

45 A distributional regression approach for Gaussian process responses 279

Hannes Riebl, Nadja Klein, Thomas Kneib

46 Multi-state models for double transitions associated with parasitism in biological control 285

Idemauro Antonio Rodrigues de Lara, Gabriel Rodrigues Palma, Victor José Bon, Carolina Reigada, Rafael de Andrade Moral

47 Bias reduced predictions for black-box models 290

Philipp Sterzinger, Ioannis Kosmidis

48 Autoregressive hidden Markov models for high-resolution animal movement data 294

Ferdinand V. Stoye, Roland Langrock

49 Complexity reduction via deselection for boosting distributional copula regression **300**
Annika Strömer, Nadja Klein, Christian Staerk, Hannah Klinkhammer, Andreas Mayr

50 Bayesian nowcasting with Laplacian-P-splines **305**
Bryan Sumalinab, Oswaldo Gressani, Niel Hens, Christel Faes

51 Boosting distributional soft regression trees **311**
Nikolaus Umlauf, Johannes Seiler, Mattias Wetscher, Nadja Klein

52 A one-step spatial+ approach to mitigate spatial confounding in multivariate spatial areal models **317**
Arantxa Urdangarin, Thomás Goicoa, María Dolores Ugarte

53 Extending central statistical monitoring to electronic patient-reported outcomes in clinical trials **321**
Lawson Wang, Sebastiaan Höppner, Laura Trotta

54 Ordinal compositional data and time series **325**
Christian H. Weiß

55 Stagewise boosting distributional regression **331**
Mattias Wetscher, Johannes Seiler, Reto Stauffer, Nikolaus Umlauf

56 Gaussian process models: From astrophysics to industrial data **337**
Jamie Wilson, Kevin Burke, Norma Bargary

57 A multilevel multivariate response model for data with latent structures **343**
Yingjuan Zhang, Jochen Einbeck, Reza Drikvandi

58 Flexible modelling of time-varying training exposures on the risk of recurrent injuries in football **349**
Lore Zumeta-Olaskoaga, Andreas Bender, Dae-Jin Lee

Part III

59 Modelling single-nucleotide polymorphism to assess genetic contribution to disease progression 355
Mazin Aouf, Kenan M. Matawie

60 Spatially adaptive Bayesian P-splines 361
Paul Bach, Nadja Klein

61 A weighted curve clustering approach for analyzing pass rush routes in american football 366
Robert Bajons, Kurt Hornik

62 Playful introduction to data competencies for economic students 371
Julia Berginski, Alexander Silbersdorff

63 Accounting for clustering in automated variable selection using hospital data: A comparison of different LASSO approaches 377
Stella Bollmann, Andreas Groll, Michael M. Havranek

64 An active deep learning method for high out-of-sample predictive performance in image classification 383
Ludwig Bothmann, Lisa Wimmer, Omid Charrakh, Tobias Weber, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, Annette Menzel

65 A smooth Laplace regression model 387
Kevin Burke

66 `TwoTimeScales`: an R-package for smoothing hazards with two time scales 390
Angela Carollo, Jutta Gampe, Paul Eilers, Hein Putter

67 A new statistical methodology to detect earnings management 394
M. Chavent, V. Darmendrail, D. Feral, H. Lorenzo, F. Pourtier, J. Saracco

68 Automatic effect selection for generalized additive models 400
Claudia Collarin, Matteo Fasiolo, Claudio Agostinelli

69 A multifidelity framework for wind speed data 405
Pietro Colombo, Claire Miller, Ruth O'Donnell, Xiaochen Yang

70 Group penalized models with an adaptive non-convex penalty function 409
Daniele Cuntrera, Vito M.R. Muggeo, Luigi Augugliaro

71 Gradient boosting for GAMLSS using adaptive step lengths 414
Alexandra Daub, Andreas Mayr, Boyao Zhang, Elisabeth Bergherr

72 Mixture confidence sequences for regression coefficients in generalized linear models 420
Claudia Di Caterina, Alessandra Salvan, Nicola Sartori

73 On the nature of one–inflation in microbial diversity studies 424
Davide Di Cecco, Andrea Tancredi

74 Prediction of record performances in sports in a record-values model 430
Christina Empacher, Udo Kamps

75 Competing risk modelling for in-hospital length of stay 436
Juan Carlos Espinosa-Moreno, Fernando García-García, Dae-Jin Lee, María J. Legarreta-Olabarrieta, Susana García-Gutiérrez, Naia Mas

76 Mixed nonlinear modelling in food engineering: determination of the salting time of boneless dry-cured Cerretan hams 440
Xavier Espuña, Lesly Acosta, Josep A. Sanchez-Espigares, Xavier Tort-Martorell

77 Learning Gaussian Bayesian networks from incomplete data - the Bayesian way 445
Marco Grzegorczyk

78 Grouped regression modeling of proteins 451
Jonas Heiner, Jan Hengstler, Andreas Groll

79 A new scalar-on-function generalized additive model for partially observed curves: an application to aneurysm patients 457
Pavel Hernández-Amaro, María Durbán, M. Carmen Aguilera-Morillo

80 Detecting heterogeneity of treatment effect between centers in multicenter randomized clinical trials 463
Sebastiaan Höppner, Marc Buyse, Laura Trotta

81 Rate of return to education of compliers: Estimation based on Rubin causal models 467
Caizhu Huang, Jierui Du, Claudia Di Caterina

82 Understanding the role of conditional residual distances from simulated envelopes in half normal plots 472
Darshana Jayakumari, Jochen Einbeck, John Hinde, Rafael A. Moral

83 Targeted bias reduction for generalised additive models 477
Oliver Kemp, Ioannis Kosmidis

84 A novel gradient boosting framework for generalised additive mixed models 482
Lars Knieper, Elisabeth Bergherr, Torsten Hothorn, Nadia Müller-Voggel, Colin Griesbach

85 Interval-censored covariates in regression models 488
Klaus Langohr, Andrea Toloba López-Egea, Guadalupe Gómez Melis

86 Bayesian regularisation for tail index regression 493
M.W. Lee, M. de Carvalho, D. Paulin, S. Pereira, R. Trigo, C. Da Camara

87 Best subset selection for principal components analysis and partial least square models using continuous optimization 497
Benoit Liquet, Sarat Moka, Samuel Muller

88 The consequences of not completing the generational cohort in estimating age-at-menopause 502
Rui Martins, Bruno de Sousa, Thomas Kneib, Maike Hohberg, Nadja Klein, Elisa Duarte, Vítor Rodrigues

89 Information retrieval models with GPT-3: Techniques for improving ranking performance through text enhancement 507
Kenan M. Matawie, Sargon Hasso

90 Analysis of climatological drivers of low-flow events in hydrological Bavaria using large ensemble climate projections 513
Theresa Meier, Nikita Paschan, Andrea Böhnisch, Henri Funk, Alexander Sasse, Helmut Küchenhoff

91 Modeling women's football scores with bivariate distributions from the Sarmanov family 519
Rouven Michels, Marius Ötting, Dimitris Karlis

92 Using measures of effect size and decision trees for variable selection 525
Annette Möller, Ann Cathrice George, Jürgen Groß

93 A comparison of time series forecasting models on industrial process data 531
Jack Moore, Jamie Wilson, Norma Bargary, Kevin Burke

94 Comparing trial and variable association in contingency table data using multinomial models for clustered data 536
Darcy Steeg Morris, Andrew M. Raim

95 Covariate-adjusted association of sensor outputs using a non-parametric estimate of the conditional covariance 543
Lizzie Neumann

96 Bayesian probit models for preference classification: An analysis of chess players' propensity for risk-taking 549
Lennart Oelschläger, Dietmar Bauer

97 Kriging wind on pressure levels to enrich the statistical modelling of aircraft trajectories 554
Rémi Perrichon, Xavier Gendre, Thierry Klein

98 Wind speed/direction in complex alpine terrain and snow avalanche accidents in the western part of Austria 560
Christian Pfeifer

99 Wastewater analysis in the light of Covid-19: A GAMLSS approach 564
Roman Pfeiler, Helga Wagner, Hans Peter Stüger, Karin Weyermair, Sabrina Kuchling, Patrick Hyden

100 Evaluating academic performance using nonparametric regression 570
Hildete P. Pinheiro, Fernando H.S. Barreto

101 Bayesian effect selection in structured piecewise additive joint models using the NBPSS prior **575**
Anja Rappl, Elisabeth Bergherr

102 Multivariate survival trees for prediction of lower limb injuries in professional male and female football players **579**
Jone Renteria, Lore Zumeta-Olaskoaga, Eder Bikandi, Jon Larruskain, Dae-Jin Lee

103 Focussed information criteria for model selection - a Bayesian perspective **584**
Bijit Roy, Emmanuel Lesaffre

104 Spatio-temporal modelling using an opportunistically sampled open-survey data: a simulation study based on the Belgian Great Corona Study **590**
Alejandro Rozo, Thomas Neyens, Christel Faes

105 Meta-analysis of variability in survival outcomes in precision oncology trials **595**
Maximilian Schuessler, Elizaveta Skarga, Pascal Geldsetzer, Ying Lu, Maike Hohberg

106 Challenges in statistical consulting for animal science **601**
Sabine K. Schnabel

107 Neural additive quantile regression **605**
Quentin E. Seifert, Elisabeth Bergherr, Benjamin Säfken

108 Mixed effects neural networks for longitudinal k-inflated count responses **611**
Nastaran Sharifian, Kevin Burke

109 A flexible non-mixture cure model for recurrent gap time data **615**
Ivo Sousa-Ferreira, Ana Maria Abreu, Cristina Rocha

110 A tool to detect nonlinearity and interactions in generalized regression models **620**
Nikolai Spuck, Matthias Schmid, Moritz Berger

111 Variable selection for statistical fine-mapping and prediction modelling of polygenic traits 624
Christian Staerk, Carlo Maj, Oleg Borisov, Hannah Klinkhammer, Peter Krawitz, Andreas Mayr

112 Long-term foehn reconstruction combining unsupervised and supervised learning 628
Reto Stauffer, Georg J Mayr, Achim Zeileis

113 Asymmetry model and its properties for square contingency tables 632
Kouji Tahata, Yusuke Kori

114 A superiority test for comparing sensitivity, specificity, and predictive values of two diagnostic tests 636
Kanae Takahashi, Kouji Yamamoto

115Individual participant data meta-analysis: pooled effect of EEF funded educational trials on low baseline attaining group 640
Germaine Uwimpuhwe, A. Singh, N. Akhter, B. Ashraf, T. Coolen-Maturi, T. Robinson, S. Higgins, J. Einbeck

116 Estimating short-term air pollution effects on health via spectral methods 644
Massimo Ventrucci, Garritt L. Page

117 Examining quantiles of sensor outputs in structural health monitoring 648
Frederike Vogel

118 Change point regression for estimated time series: An application to COVID-19 hospitalization data 653
Maximilian Weigert, Kai Becker, Helmut Küchenhoff

119 A consistent way to define p-values 658
Paul Wilson, Jochen Einbeck

120 Modelling SHM sensor outputs: A functional data approach 664
Philipp Wittenberg, Jan Gertheiss

121 Analyzing blended learning education with eye tracking and deep learning methods 669
Hilal Yagimli, Julia Berginski, Alexander Silbersdorff

122 Crime predicting models in the São Paulo state of Brazil 675
Wellington Yuanhe Zhao, Luis Gustavo Nonato, Cibele M. Russo

123 A scalable and embedded diachronic sense change model 681
Schyan Zafar, Geoff K. Nicholls

124 Bayesian nonparametric inference for the three-class covariate-specific overlap coefficient 687
Zhaoxi Zhang, Vanda Inácio

On the nature of one–inflation in microbial diversity studies

Davide Di Cecco[1], Andrea Tancredi[1]

[1] Sapienza University of Rome, Italy

E-mail for correspondence: davide.dicecco@uniroma1.it

Abstract: The phenomenon of one–inflation is gaining more and more attention in the recent literature on species abundance and capture–recapture analysis. When analysing frequency count distribution, the excess of singletons is often ascribed to the erroneous inclusion of spurious cases. Various works propose to estimate the true number of singletons relying on the higher, supposedly error–free, counts ("discounting" approach). We argument that, in the case of microbial diversity studies, the generating process of the spurious singletons can be described in terms of false negative record linkage errors. Errors in sequencing the RNA genomes result in chimeric sequences that cannot be associated to the correct species, and constitute missing links that are added to the true singletons. In this scenario, none of the observed frequency counts is assumed to be error–free, and we propose an ABC algorithm to estimate the true frequency counts. The number of true singletons estimated in this way may differ considerably from the discounting approach. This implies different estimates of the diversity as measured, e.g., by Shannon's index. However, curiously, the total population count estimates under the two approaches coincide.

Keywords: Species problem; Biodiversity; Linkage Errors; Approximate Bayesian Computing.

1 Introduction

The problem of estimating the number of species in a population given a sample arises in many applications in the natural sciences, in linguistics and computer science. Our focus is on applications in microbial ecology. The spread of next generation high-throughput sequencing technology allowed to analyse an unprecedented amount of data on microbial communities. In order to study the biodiversity in a microbial community, an environmental sample is processed to detect, amplify and sequence RNA genomes. The

This paper was published as a part of the proceedings of the 37th International Workshop on Statistical Modelling (IWSM), Dortmund, Germany, 16–21 July 2023. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

sequences are clustered into distinct species (or Operational Taxonomic Units) on the basis of a similarity score. The diversity analysis is then conducted on the abundance frequency counts, i.e., the counts $\{n_j\}_{j=1,2,\ldots}$ representing the number of species with j captured occurrences. In most microbial studies, the distribution $\{n_j\}_{j=1,2,\ldots}$ is characterized by an unexpected number of low–abundance species, in particular singletons, accompanied by a low number of very common species. The nature of these singletons has been debated at length, and the presence of spurious singletons resulting from sequencing errors has been confirmed in various ways (e.g., Quince et al. 2011, Haas et al. 2011). While bioinformatics focuses on avoiding the formation of the so–called chimera sequences, or removing them in a pre–processing step, various statistical contributions attempt to estimate ex–post their number.

The study of one–inflation in frequency count distribution is gaining more and more attention also in the recent capture–recapture literature on human and animal population, which shares many methodological aspects with the species abundance problem, (see, e.g., Godwin and Böhning 2017, Böhning et al. 2019, Tuoto et al. 2022). The possible sources of one–inflation can be categorized as:

- a behavioural effect, where certain units, once captured, avoid subsequent captures;
- the presence of out-of-scope units, which enter the sample for a peculiar error mechanism and should be excluded;
- the presence of missing links in the record linkage procedure employed to create the frequency counts.

Various authors adopted a “discounting” approach to the problem of one–inflation. That is, they propose to ignore the data affected by errors, i.e., the observed singletons, and re-estimate their number on the basis of the counts n_j, $j \geq 2$, (see, e.g, Willis and Bunge 2015, Willis 2016, Chiu and Chao 2016). We argument that this approach is consistent for the second mechanism listed above: a model where out-of-scope singletons are added to the baseline distribution of the true counts. We believe that the nature of the spurious cases can alternatively be described by linkage errors. That is, we assume that random errors occurring in sequencing result in the impossibility of a correct classification of the specimen, which cannot be associated to the right existing species. Therefore, we can describe these cases as false negative linkage errors (or missing links), which are added to the true singletons. This approach implies a re–estimation of the “real” frequency counts for all the abundances, not just the singletons. We found that treating the excess of singletons in this way leads to significant differences in the diversity estimates with respect to the discounting approach.

In this work we adopt a secondary approach to the linkage problem, i.e., we try to estimate the linkage errors solely on the basis of the vector

$\{n_j\}_{j=1,2,...}$ and our distributional assumptions, as we do not have access to the actual linkage process. Modeling linkage errors in this secondary setting, appears quite complex from a computational point of view. We fix some simplifying assumptions on the type of error in order to tackle the issue, but we still resorted to a Bayesian likelihood–free approach as the most convenient approach.

2 One–inflation models

Say we get n species in our sample with abundances $y_1, ..., y_n$, and abundance frequency counts $\{n_j\}_{j\geq 1}$. Under an out–of–scope singletons model, the distribution of the abundances (whether the species are observed or not, spurious or not) results in the following mixture of a baseline distribution $\widetilde{f}$ of the non–spurious counts, and a Dirac measure over one:

$$P(Y_i = j \; ; \; \widetilde{f}, \psi) = \begin{cases} (1-\psi)\widetilde{f}_1 + \psi & \text{if } j = 1; \\ (1-\psi)\widetilde{f}_j & \text{otherwise,} \end{cases} \tag{1}$$

where ψ denotes the portion of spurious cases over the total population count. Let $\widetilde{n}_j$ denote the number of species with j non spurious captures. Then, since we assumed $\widetilde{n}_j = n_j$ for $j \geq 2$, we just have to estimate the number of unsampled species $\widetilde{n}_0$, and the number of non–spurious singletons $\widetilde{n}_1$ as a portion of n_1. The estimate of the total number of distinct species $\widetilde{N}$ will result as:

$$\sum_{j\geq 0} \widetilde{n}_j = \widetilde{n}_0 + n - n_1 + \widetilde{n}_1.$$

A Bayesian estimation of this model presents no difficulties under various parametric families choices for $\widetilde{f}$. A simple Gibbs sampler scheme is the following: under a Beta prior for ψ, its posterior is easily updated. Then, a value for $\widetilde{n}_1$ is generated from a Binomial with parameters $1-\psi$ and n_1. Steps to update the values of $\widetilde{n}_0$ and of the parameters of $\widetilde{f}$ are easily found in literature (see, e.g., Tuoto et al. 2022).

Under our missing links proposal, we assume that each sequence has the same probability μ of being missclassified as a singleton independently from the other. Denote the true number of sampled distinct species as n^*, ($n^* < n$). For each species i with X_i^* true captures, we have M_i missing links, such that the registered abundance is reduced from X_i^* to $X_i = X_i^* - M_i$. M_i has the following distribution:

$$P(M_i = m_i \mid X_i^* = x_i^*) = \binom{x_i^*}{m_i} \mu^{m_i} (1-\mu)^{x_i^* - m_i}, \quad i = 1, ..., n^*. \tag{2}$$

Let f^* be the baseline distribution of the X_i^*. The distribution of the X_i results as a thinning process where a portion μ of captures disappear. Let

n_j^* denote the true number of species with j captures, and as $N^* = \sum_{j\geq 0} n_j^*$ the total number of distinct species according to the missing links model. Unlike the spurious singletons model, in this case all values $\{n_j^*\}_{j\geq 0}$ have to be estimated, as they will be, in general, different from the observed values. Denote as θ the parameters defining f^*. We adopted an ABC rejection algorithm with the following scheme:

1. generate values for (θ, N^*) from the priors $\pi(\theta)$ and $\pi(N^*)$;

2. generate values $(n_0^*, n_1^*, n_2^*, ...)$ conditional on N^* and θ;

3. generate a value for μ from the Beta prior $\pi(\mu)$ (independent from all the rest);

4. generate missing links at random according to the distribution described in (2), given $(n_0^*, n_1^*, n_2^*, ...)$ and μ. Each missing link modifies the observed count, and increments accordingly the number of singletons, thus obtaining the fictitious data D^*;

5. retain the current generated values if a measure of distance ρ between the generated data D^* and the observed data D is below a certain threshold ϵ:

   $$\rho(D^*, D) < \epsilon.$$

   In our application we utilized the euclidean distance.

As the simple ABC rejection scheme can have a low acceptance rate, we further adopted a sequential ABC to accelerate the procedure, as described in Marin et al. 2012.

A simulation study confirmed the correctness of the ABC algorithm under a Poisson, Geometric, and finite mixture of Poisson distributions for f^*. Our first finding in a further simulation study comparing the spurious cases and the missing links proposal, has been the substantial identity of the estimates of the total number of species under the two competing models. That is, if we choose f^* and $\widetilde{f}$ in the same family, despite the fact that the estimates of the true abundance frequencies differ under the two models (i.e., $\widetilde{n}_j \neq n_j^*$ for all j), we have $N^* = \widetilde{N}$.

To demonstrate this identity, consider the baseline distribution f^* of the values X_i^* introduced above. It is easily demonstrated that, under various parametric family for f^*, (notably, if f^* is any mixed Poisson), the distribution of the X_i belongs to the same parametric family. Then, under identifiability of that family, if we use model (1) and take $\widetilde{f}$ in the same family as f^*, $\widetilde{f}$ will be identified as the distribution of the x_i, for all $x_i > 0$, and ψ would represent the portion of missing links over the total population count. Let r_0 be the number of captured species whose occurrences where all missclassified, i.e., such that $M_i = X_i^*$. Let M be the total number of missing links: $M = \sum_{i=1}^{n^*} M_i$. Then we have

$$n^* = n - M + r_0 \quad \text{and} \quad \widetilde{n}_1 = n_1 - M.$$

The missing links mechanism does not affect the number of undetected species n_0^*, but under $\widetilde{f}$ the r_0 values are included in $\widetilde{n}_0$, i.e., we have $\widetilde{n}_0 = n_0^* + r_0$. Finally, we can write

$$\widetilde{N} = \widetilde{n}_0 + \widetilde{n}_1 + n - n_1 = \widetilde{n}_0 + n - M \ = \ n_0^* + r_0 + n - M = n_0^* + n^* = N^*.$$

As we have said, even if the estimates of the total number of species coincides under the two models, the abundance distribution will differ, and consequently, the estimated diversity will differ. To illustrate the effect of (ignoring) a missing links mechanism on the estimation of diversity, we utilized a simulation study. As a measure of diversity we considered Shannon's diversity H (see, e.g., Chiu and Chao 2016) calculated as:

$$H = \exp\left(-\sum_{j\geq 1} n_j \frac{j}{s} \ln \frac{j}{s}\right). \tag{3}$$

We generated various datasets under Poisson and Geometric baseline distributions, then simulated the effect of missing links to simulate from our model. Then, we estimated Shannon's diversity on the observed data (that is, ignoring any one–inflation mechanism), on the "adjusted" counts as derived from the spurious cases model (that is, trimming the observed number of singletons) and as derived from the ABC procedure for the missing links model. Note that in our Bayesian approach we can easily estimate the posterior distribution of (3). First, we concluded that ignoring an existing one–inflating mechanism, implies a severe overestimation of the diversity. Second, utilizing model (1) when missing links are the true source of error, reduces sensibly the overestimation, but still leads to different results than what can be achieved with an ABC simulating the actual generating process.

References

Böhning, D., Kaskasamkul, P., van der Heijden, P. G. (2019). A modification of Chao's lower bound estimator in the case of one-inflation. *Metrika*, **82(3)**, 361 – 384.

Chiu, C. H., Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, **4**, e1634.

Godwin, R. T., Böhning, D. (2017). Estimation of the population size by using the one–inflated positive Poisson model. *Journal of the Royal Statistical Society. Series C*, 425 – 448.

Haas, B.J., Gevers, D., Earl, A.M. et al. (2011) Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced pcr amplicons. *Genome research*, **21(3)**, 494 – 504.

Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J (2012). Approximate Bayesian computational methods. *Statistics and computing*, **22(6)**, 1167-1180.

Quince, C., Lanzen, A., Davenport, R. J. (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, **12(1)**, 1 – 18.

Tuoto, T., Di Cecco, D., Tancredi, A. (2022). Bayesian analysis of one-inflated models for elusive population size estimation. *Biometrical Journal*, **64(5)**, 912 – 933.

Willis, A., Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71(4)**, 1042 – 1049.

Willis, A. (2016). Species richness estimation with high diversity but spurious singletons. *arXiv preprint arXiv:1604.02598.*