International Conference

APPLIED STATISTICS 2023

Program and Abstracts

September 24–26, 2023

Koper, Slovenia

International Conference

APPLIED STATISTICS 2023

Program and Abstracts

September 24–26, 2023

Koper, Slovenia

Kataložni zapis o publikaciji (CIP) pripravili v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 165713411
ISBN 978-961-94283-3-7 (PDF)

International Conference
APPLIED STATISTICS 2023
Program and Abstracts

Scientific Advisory Committee

Lara Lusa (Chair), Slovenia
Mihael Perman, Slovenia
Janez Stare, Slovenia
Vladimir Batagelj, Slovenia
Andrej Blejec, Slovenia
Matevž Bren, Slovenia
Maurizio Brizzi, Italy
Anuška Ferligoj, Slovenia
Herwig Friedl, Austria
Dario Gregori, Italy
Katarina Košmelj, Slovenia
Irena Križman, Slovenia
Stanislaw Mejza, Poland
Jože Rovan, Slovenia
Tamas Rudas, Hungary
Vasja Vehovar, Slovenia

Scientific Program Committee

Lara Lusa (Chair), Slovenia
Andrej Kastrin, Slovenia

Organizing Committee

Irena Vipavc Brvar
Jerneja Čuk
Bogdan Grmek
Andrej Kastrin
Dean Lipovac
Lara Lusa
Ana Zalokar

Edited by	Andrej Kastrin and Lara Lusa
Published by	Statistical Society of Slovenia
	Litostrojska cesta 54
	1000 Ljubljana, Slovenia
Publication year	2023
Publication format	PDF
Electronic version:	`https://as.mf.uni-lj.si/pdf/as2023book.pdf`

PROGRAM and ABSTRACTS

PROGRAM

		Room 1	Room 2
Sunday	15.00 – 17.00	Workshop	
Monday	9.00 – 9.10	Opening	
	9.10 – 10.00	Invited lecture	
	10.00 – 10.30	Break	
	10.30 – 12.15	Biostatistics 1	Social sciences
	12.15 – 13.30	Lunch	
	13.30 – 15.15	Biostatistics 2	Measurement and modeling
	15.15 – 15.40	Break	
	15.40 – 16.30	Mathematical statistics and modeling	
	15.50 – 16.40		Student session 1
	16.30 – 16.40	Posters	
	16.40 – 16.50	Break	
	16.50 – 17.40	Invited lecture	
	18.00 – 19.30	Sightseeing	
	19.45 – 21.00	Reception	
Tuesday	9.00 – 9.50	Invited lecture	
	9.50 – 10.20	Break	
	10.20 – 12.05	Network analysis	Modeling and simulation
	12.20 – 13.30	Lunch	
	13.30 – 15.15	Official Statistics	Student session 2
	14.30 – 15.20		Student session 3
	15.20 – 15.40	Break	
	15.40 – 17.40	Invited session	
	17.40 – 18.00	Closing	

15.00–17.00 **Workshop**
Room 1

1. **Micro-econometric techniques for program evaluation in the social sciences**
   Anže Burger

9.00–9.10 **Opening**
Room 1 *Chair: Lara Lusa*

9.10–10.00 **Invited lecture**
Room 1 *Chair: Lara Lusa*

1. **Analysis of multivariate longitudinal and survival data: From joint models to random forests**
   Cécile Proust-Lima

10.00–10.30 **Break**

10.30–12.15 **Biostatistics 1**
Room 1 *Chair: Cécile Proust Lima*

1. **A semi-mechanistic dose-finding design in oncology using pharmacokinetic/pharmacodynamic modeling**
   Xiao Su, Yisheng Li, Peter Mueller, Chia-Wei Hsu, Haitao Pan and Kim-Anh Do
2. **Comparing the survival probability in breast cancer screening programmes**
   Maja Pohar Perme and Bor Vratanar
3. **Measures of lifetime difference in relative survival**
   Erik Langerholc
4. **Regression modelling in relative survival and its application in multi-state models**
   Damjan Manevski
5. **Assessing the effectiveness of cardiac rehabilitation in patients after myocardial infarction from large administrative reimbursement claim databases**
   Borut Jug, Janez Bijec and Dalibor Gavrić
6. **COVID-19 pandemic: Impact on clinical research**
   Jay Mandrekar

10.30–12.15 **Social sciences**
Room 2 *Chair: Gregor Sočan*

1. **Propensity score matching, regression with unbalanced covariates, or both? An application to rehabilitation**
   Gaj Vidmar, Miha Rutar and Helena Burger
2. **Research on consumer segments and their expectations regarding failures and recovery strategies in online stores**
   Samo Kropivnik and Ema Zdešar
3. **Residential renovation in Slovenia: Comparison of survey and registry data**
   Ana Slavec
4. **Integral territorial governance model for urban tourism destinations**
   David Klepej, Naja Marot and Irena Ograjenšek
5. **A network approach to team leadership**
   Helena Kovačič, Hajdeja Iglič and Barbara Lužar

6. **Career-relevance of statistics and research methods revisited**
   Irena Ograjenšek and Iddo Gal

12.15–13.30 **Break**

13.30–15.15 **Biostatistics 2**
Room 1 *Chair: Janez Stare*

1. **Identifying typical trajectories in longitudinal data: A case of study for patients affected by diabetic kidney disease**
   Veronica Distefano, Maria Mannone and Irene Poli
2. **Climate change and the quality of wine: The case of Collio**
   Barbara Campisi, Gaetano Carmeci, Gianluigi Gallenti, Giovanni Millo, Matteo Carzedda and Paolo Bogoni
3. **External validation of the OAC3-PAD Risk Score and its underlying survival model**
   Nataša Kejžar, Kevin Pelicon, Klemen Petek, Anja Boc, Vinko Boc and Tjaša Vižintin Cuderman
4. **Patient reported outcome measures of patients undergoing a primary knee or hip arthroplastics**
   Eva Podovšovnik and Vesna Levašič
5. **Management of patients with acute coronary syndrome during the COVID-19 pandemic in Slovenia**
   Tjaša Furlan, Janez Bijec, Dalibor Gavrić, Petra Došenović Bonča, Irena Ograjenšek and Borut Jug

13.30–15.15 **Measurement and modeling**
Room 2 *Chair: Irena Ograjenšek*

1. **Falling for the leading questions**
   Vanja Erčulj and Ajda Šulc
2. **Measuring concordance and discordance of student reading literacy data around the world**
   Simona Korenjak-Černe and Barbara Japelj Pavešić
3. **Assessing reliability and measurement error for continuous measurements**
   Nina Ružić Gorenjec and Nataša Kejžar
4. **Non-linear stochastic model for dopamine cycle**
   Nenad Šuvak, Marija Milošević and Jasmina Djordjević
5. **Stochastic SEIPHAR model for epidemic of the SARS-CoV-2 virus**
   Jasmina Djordjević, Ivan Papić and Nenad Šuvak
6. **Inverse problem for parameters identification in a modified SIRD epidemic model using ensemble neural networks**
   Marian Petrica and Ionel Popescu

15.15–15.40 **Break**

15.40–16.30 **Mathematical statistics and modeling**
Room 1 *Chair: Ana Zalokar*

1. **Adaptive applicability of the Random Environment INAR models**
   Aleksandar Nastić
2. **Approximate Bayesian algorithm for tensor robust PCA using relative entropy**
   Andrej Srakar
3. **Improvements in parameter estimation for some class of the INAR models**
   Miodrag Djordjević

15.50–16.40 **Student session 1**
Room 2 *Chair: Marjan Cugmas*

1. **Guitar tablature transcription with convolutional neural networks**
   Matija Marolt, Igor Nikolaj Sok and Igor Grabec
2. **A package for generating and grading exams**
   Jakob Peterlin
3. **Land take, land use and environmental issues. Is the Kuznets Curve valid? The case of Italy**
   Giuseppe Borruso, Andrea Gallo, Francesco Magris and Nicola Pontarollo
4. **Application of machine learning to fundamental analysis of securities**
   Aleksandr Panteleev
5. **Linguistic analysis of suicide related questions in the online counselling service This is Me**
   Vili Smolič, Sara Atanasova and Marjan Cugmas

16.30–16.40 **Posters**
Room 1 *Chair: Ana Zalokar*

1. **The application of NESTOREv1.0 to forecast strong aftershocks in the North-eastern Italy and Western Slovenia**
   Piero Brondi, Stefania Gentili and Rita Di Giovambattista
2. **Applying walkthrough method for researching the moral references of the Signal application**
   Kristina Rakinić
3. **Reconstruction of sea surface temperature with spectral convolution**
   Matic Klopčič, Matej Kristan and Matjaž Ličer
4. **Data mining of chlorophyll-a satellite data (CMEMS) enables reconstruction of phytoplankton blooms in the Adriatic Sea on large temporal and spatial scales**
   Nejc Prinčič, Martin Vodopivec, Patricija Mozetič and Janja Francé

16.40–16.50 **Break**

16.50–17.40 **Invited lecture**
Room 1 *Chair: Andrej Blejec*

1. **Data science ethics: Some stories from the trenches**
   Richard De Veaux

18.00-19.30 **Sightseeing**

19.45–21.00 **Reception**

9.00–9.50 **Invited lecture**
Room 1 *Chair: Mihael Perman*

1. **Capture-recapture methods with applications in health and society**
   Dankmar Böhning

9.50–10.20 **Break**

10.20–12.05 **Network analysis**
Room 1 *Chair: Anuška Ferligoj*

1. **3D visualization of multiway networks**
   Vladimir Batagelj
2. **The network effects of international sanctions: A temporal blockmodeling study of trade sanctions**
   Fabio Ashtar Telarico
3. **Some considerations regarding blockmodeling of dynamic networks**
   Aleš Žiberna
4. **Patterns of scientific collaboration in doctoral education: An analysis of mentor-mentee relationships**
   Marjan Cugmas, Luka Kronegger and Franc Mali
5. **Ranking genes based on gene spreading strength and mutation neighbor influence in network**
   Peter Juma Ochieng, József Dombi, Tibor Kalmár and Miklós Krész
6. **Visibility graph analysis of MODIS satellite evapotranspiration time series of olive groves in southern Italy: Revealing Xylella Fastidiosa induced phytopathogenic status**
   Luciano Telesca and Rosa Lasaponara

10.20–12.05 **Modeling and simulation**
Room 2 *Chair: Dankmar Boehning*

1. **The linear model vs. the proportional odds model for analysing ordinal and continuous outcomes: Simulation study**
   Georg Heinze, Michael Kammer, Daniel Kraemmer and Daniela Dunkler
2. **Choosing among three proportional odds models for ordinal and count outcomes—a matter of taste? A simulation study**
   Andreas Klinger, Daniela Dunkler, Mariella Gregorich and Georg Heinze
3. **Bayesian state-space modeling of indoor radon concentration and entry rate**
   Marek Brabec
4. **Modelling extreme bivariate data using R software**
   Maria Manuela Neves and Helena Penalva
5. **To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets**
   Hana Šinkovec, Rok Blagus, Georg Heinze and Angelika Geroldinger

12.20–13.30 **Break**

13.30–15.15 **Official Statistics**
Room 1 *Chair: Mojca Bavdaž*

1. **The use of the Earth observation data for the monitoring of permanent grassland and soil moisture**
   Črt Šuštar
2. **Using scraped data in calculating inflation**
   Matevž Postrašija
3. **Using machine learning for assisted classification of articles**
   Črt Grahonja
4. **Using the cell key method for protection of grids at Statistics Slovenia**
   Manca Golmajer
5. **Analysis of protection of Census hypercubes with the cell key method**
   Janez Bijec

13.30–14.20 **Student session 2**
Room 2 *Chair: Maja Pohar Perme*

1. **Prediction models to support adult IgA vasculitis diagnosis and assessing renal involvement**
   Ana Markež, Matija Bajželj, Alojzija Hočevar, Katja Lakota and Rok Blagus
2. **A new statistical index for evaluating variability in patient state index during pediatric anesthesia**
   Noor Muhammad Khan, Claudia Maria Bonardi, Angela Amigoni and Dario Gregori
3. **Bridging the gap: Integrating efficacy and quality of life in colorectal cancer observational study using the win ratio approach**
   Maria Vittoria Chiaruttini, Giulia Lorenzoni, Gaya Spolverato and Dario Gregori
4. **Modeling and forecasting mortality with economic, environmental and lifestyle variables**
   Matteo Dimai
5. **A study on the use and the necessity of machine learning dimensionality reduction and clustering methods in actuarial sciences: Defining the right methodology for business planning under the requirements of IFRS 17**
   Mateo Antonac

14.30–15.20 **Student session 3**
Room 2 *Chair: Nataša Kejžar*

1. **The challenge of multiple testing: Case of emission coupons trading**
   Anja Žavbi Kunaver, Marjan Cugmas and Irena Ograjenšek
2. **Corrections to Bland–Altman analysis for repeated measures data—are they always essential?**
   Maša Kušar
3. **Handling separation in generalized linear mixed effects models with a random intercept**
   Tina Košuta, Rok Blagus, Georg Heinze and Nina Ružić Gorenjec

4. **An investigation into overrepresentation of COVID-related genes in pathway enrichment analysis for RNA-seq data**
   Sara Ahsani-Nasab, Daniele Sabbatini, Elena Pegoraro, Dario Gregori and Luca Vedovelli
5. **Enhancing survey design and analysis: Leveraging machine learning for post-stratification**
   Mingmeng Geng and Roberto Trotta

15.20–15.40 **Break**

15.40–17.40 **Invited session**
Room 1 *Chair: Georg Heinze*

1. **Phases of methodological research in biostatistics: Building the evidence base for new methods**
   Georg Heinze, Anne-Laure Boulesteix, Michael Kammer, Tim Morris and Ian White
2. **Initial data analysis: Making the effort worthwhile**
   Lara Lusa, Carsten Oliver Schmidt, Georg Heinze and Marianne Huebner
3. **Analysis of time-to-event for observational studies: Guidance to the use of intensity models**
   Maja Pohar Perme
4. **Correctly accounting for misclassification when linking latent groups with external variables**
   Cécile Proust-Lima, Maris Dussartre, Viviane Philipps, Cécilia Samieri, Paul Gustafson and Pamela A. Shaw

17.40–18.00 **Closing**
Room 1 *Chair: Lara Lusa*

ABSTRACTS

Workshop

Micro-econometric techniques for program evaluation in the social sciences

Anže Burger

University of Ljubljana, Ljubljana, Slovenia

The workshop provides a brief overview of both theoretical and applied tools for implementation of modern micro-econometric methods for program evaluation in the social sciences. First part of the course presents statistical setup and a short description of identification issues in estimating causal effects of a program or some other policy change under different assumptions about the selection into the program. The second part presents the most common evaluation techniques discussed in the literature, such as the regression adjustment, matching, difference-in-differences, instrumental variables, regression discontinuity design, synthetic control method and will be offered a series of practical guidelines for the selection and application of the most suitable approach to implement under differing policy contexts. Each method presented will be illustrated with an application from a published scientific article.

Invited lecture

Analysis of multivariate longitudinal and survival data: From joint models to random forests

Cécile Proust-Lima

University of Bordeaux, Bordeaux, France

Health studies usually involve the collection and analysis of variables repeatedly measured over time. This includes exposures (e.g., treatment, blood pressure, nutrition) and markers of progression (e.g., brain volumes, blood tests, cognitive functioning, tumor size). When interested in modeling how these variables are associated with clinical endpoints such as death in survival models, some statistical challenges arise. First, they constitute inaccurate measures of the underlying continuous-time processes of interest: they are measured with error and at sparse visit times. Neglecting such characteristics may lead to biased associations with clinical endpoints. A dedicated solution is the joint analysis of the longitudinal processes and the time-to-event in so-called joint models [1]. This methodology, now available in many software programs, can be easily applied. However, it reaches numerical limits when the number of repeated variables substantially increases [2]. In this talk, I first introduce the specificity of longitudinal data collected in health studies and show through simulations how naive techniques may lead to incorrect inference. Then I describe the methodology of the joint models for longitudinal and survival data [1, 3]. Finally, I present how this methodology can be incorporated into random survival forests to account for a large dimension of longitudinal variables when interested in predicting a time-to-event (potentially with multiple causes) and identifying the most important predictors [4]. Throughout the talk, I illustrate the methods with examples from epidemiological cohorts, notably in cerebral aging research to predict the risk of Alzheimer's disease.

References

[1] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: with applications in R.* Boca Raton: CRC Press, 2012.

[2] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint modelling of time-to-event and multivariate longitudinal outcomes: Recent developments and issues," *BMC Medical Research Methodology*, vol. 16, no. 1, p. 117, 2016. DOI: 10.1186/s12874-016-0212-5.

[3] D. Rustand, J. Van Niekerk, E. T. Krainski, H. Rue, and C. Proust-Lima, "Fast and flexible inference for joint models of multivariate longitudinal and survival data using integrated nested Laplace approximations," *Biostatistics*, kxad019, 2023. DOI: 10.1093/biostatistics/kxad019.

[4] A. Devaux, C. Helmer, R. Genuer, and C. Proust-Lima, *Random survival forests with multivariate longitudinal endogenous covariates*, arXiv:2208.05801 [stat], 2023. DOI: 10.48550/arXiv.2208.05801.

Biostatistics 1

A semi-mechanistic dose-finding design in oncology using pharmacokinetic/pharmacodynamic modeling

Xiao Su[1], Yisheng Li[2], Peter Mueller[3], Chia-Wei Hsu[4], Haitao Pan[4] and Kim-Anh Do[2]

[1]PlayStation, San Mateo, CA, United States
[2]University of Texas MD Anderson Cancer Center, Houston, TX, United States
[3]University of Texas at Austin, Austin, TX, United States
[4]St. Jude Children's Research Hospital, Memphis, TN, United States

While a number of phase I dose-finding designs in oncology exist, the commonly used ones are either algorithmic or empirical model-based. We propose a new framework for modeling the dose-response relationship, by systematically incorporating the pharmacokinetic (PK) data collected in the trial and the hypothesized mechanisms of the drug effects, via dynamic PK/PD modeling, as well as modeling of the relationship between a latent cumulative pharmacologic effect and a binary toxicity outcome. This modeling framework naturally incorporates the information on the impact of dose, schedule and method of administration (e.g., drug formulation and route of administration) on toxicity. The resulting design is an extension of existing designs that make use of pre-specified summary PK information (such as the area under the concentration-time curve [AUC] or maximum serum concentration [Cmax]). Our simulation studies show, with moderate departure from the hypothesized mechanisms of the drug action, that the performance of the proposed design on average improves upon those of the common designs, including the continual reassessment method (CRM), Bayesian optimal interval (BOIN) design, modified toxicity probability interval (mTPI) method, and a design called PKLOGIT that models the effect of the AUC on toxicity. In case of considerable departure from the underlying drug effect mechanism, the performance of the design is shown to be comparable with that of the other designs. We illustrate the proposed design by applying it to the setting of a phase I trial of a γ-secretase inhibitor in metastatic or locally advanced solid tumors. We also provide R code to implement the proposed design.

Comparing the survival probability in breast cancer screening programmes

Maja Pohar Perme and Bor Vratanar

University of Ljubljana, Ljubljana, Slovenia

Cancer screening is a programme for medical screening of asymptomatic people who are at risk of developing cancer. In Slovenia, women between 50 and 70 are invited biannually to mammography screening. The programme has been running since 2008, we wish to evaluate its effectiveness. The most direct way to evaluate a screening programme is through survival analysis—we wish to know whether patients who participated in the programme have better chances of survival than those who did not take part. However, it turns out that any straightforward comparison of survival probabilities results in important biases that should not be neglected. In our work [1], we split the complex problem into simpler building blocks and show how survival can be compared in each of these blocks. While some of the issues can be solved non-parametrically, parametric assumptions may be needed for others. We have formulated a general theory and we adapt it to the particular issues of Slovene breast screening programme.

References

[1] B. Vratanar and M. Pohar Perme, "Evaluating cancer screening programs using survival analysis," *Biometrical Journal*, p. 2 200 344, 2023. DOI: 10.1002/bimj.202200344.

Measures of lifetime difference in relative survival

Erik Langerholc

University of Ljubljana, Ljubljana, Slovenia

Assessing the impact of a condition on a person's lifetime is a problem in survival analysis which requires a reference group. A reference population is one from which we have no sample available, but the conditional distribution of its lifetime given demographic covariates is known. An example is the general population—tables recording overall mortality of people depending on their sex, age, year of birth and location often serve as a reference. Comparison of lifetime in a sampled cohort versus a reference population is straightforward in a right-censored setting, but surprisingly difficult in a left-truncated right-censored one. This talk will introduce what a lifetime difference measure is in a truncated and censored setting and provide some examples of such measures.

Regression modelling in relative survival and its application in multi-state models

Damjan Manevski

University of Ljubljana, Ljubljana, Slovenia

In survival analysis, when analysing the event of interest (death), one is commonly interested in the cause of death as well. Often this information is not provided in the data and the field of relative survival deals with estimating the proportion of deaths dying due to the disease in question or due to other (population) causes. In this work, the focus will be on the extended case when other (intermediate) events may be observed in the data (e.g. relapse, remission, transplantation). Multi-state models are a common tool for considering such intermediate events in the analysis. In previous work, we have already considered an extended multi-state model where cause of is accounted for using relative survival. Based on this model, non-parametric estimates are obtained. The next step is to implement a regression framework for modelling covariate effects and obtaining subject-specific predictions based on the model. We will consider two possible regression approaches: a Cox-type multiplicative model and an Aalen additive hazards model. During the presentation, we will delve into both regression approaches, discussing their respective strengths, limitations, and practical considerations.

Assessing the effectiveness of cardiac rehabilitation in patients after myocardial infarction from large administrative reimbursement claim databases

Borut Jug[1], Janez Bijec[2] and Dalibor Gavrić[3]

[1]University Medical Centre Ljubljana, Ljubljana, Slovenia, Ljubljana, Slovenia
[2]University of Ljubljana, Ljubljana, Slovenia
[3]Health Insurance Institute of Slovenia, Ljubljana, Slovenia

Observational analyses of large administrative healthcare datasets represent an opportunity to appraise the effectiveness of healthcare interventions. Outcomes research can expand the generalizability (external validity) of interventions to real-life populations when compared to selected and controlled settings of randomized trials, although at the expense of causal inference/internal validity (allocation to intervention is observed, not random). Propensity score-based methods are one potential tool to adjust for non-random intervention allocation. In the framework of this analysis, we sought to estimate the effectiveness of participation to cardiac rehabilitation (CR) on all-cause mortality in patients after myocardial infarction. Data on all patients hospitalized for myocardial infarction (ICD I20.0 and I21.x) in Slovenia between 2015 and 2021 were collected by merging (using unique patient identifiers) hospital management, cardiac rehabilitation, medication, and survival status databases for reimbursement claims from the National Institute of Healthcare Insurance of Slovenia (which provides universal healthcare coverage for the whole country/nation-wide population). Propensity scores were calculated using logistic regression (for CR vs. no CR participation) with covariates plausibly associated with CR allocation and mortality—demographic characteristics (age and sex), hospital episode intensity, recorded co-morbidities, and medication prescription at discharge. Survival time was analyzed using double-robust Cox proportional hazards regression (with propensity score weighting and covariates adjustment) to estimate hazard ratios (HR) and 95% confidence intervals (CI, from robust standard errors). The obtained results indicate that participation in CR is associated with reduced all-cause mortality. Our analysis thus seem to provide robust inference on the benefits of CR participation for patients after myocardial infarction as well as a useful framework for appraising real-life effectiveness of healthcare interventions from available datasets for evidence-based decision-making guidance.

COVID-19 pandemic: Impact on clinical research

Jay Mandrekar

Mayo Clinic, Rochester, MN, United States

Research in academic medical centers offer opportunities to collaborate on clinical projects that require novel application of both common and uncommon statistical methods. In this talk we will focus on the impact of COVID-19 pandemic. In the first part of the talk, we will discuss some of the opportunities brought forward by the COVID-19 pandemic and in the second part we will discuss the challenges encountered while conducting clinical research during and after the pandemic. Examples from diverse clinical areas such as clinical microbiology, infectious diseases, nursing research etc. will be discussed. The talk will focus on various topics such as novel statistical approaches, setting up databases, recruitment and engagement of participants, missing data issues, effect on grant funding and new work environment.

Social sciences

Propensity score matching, regression with unbalanced covariates, or both? An application to rehabilitation

Gaj Vidmar, Miha Rutar and Helena Burger

University Rehabilitation Institute Republic of Slovenia, Ljubljana, Slovenia

Achieving mobility and maximum possible level of functioning and participation are the principal goals of rehabilitation of patients after lower limb amputation (LLA). To that end, a prosthesis is usually fitted, but it requires sufficient physical and cognitive capacity. We wanted to assess how the Mini Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) cognitive screening tests are associated with rehabilitation outcomes in patients after LLA. Data from 155 patients (119 men; median age 69 years) after LLA who completed rehabilitation at our institute between 9/2017 and 2/2018 were analysed. Until 1/2018, they had been cognitively tested with MMSE ($n = 66$), and afterwards with MoCA ($n = 89$). The primary study outcome was successful prosthesis fitting (binary); secondary outcomes were ability to independently don the prosthesis, ability to independently climb stairs using the prosthesis, walking aid type (all binary) and result of the 6-minute-walk-test (numerical). Propensity-score-matching and (logistic and linear) regression models with matching variables as additional covariates were applied to adjust parameter estimates for group-imbalance in terms of sex, age, education level, level and cause of amputation, additional neurological disease and total number of diseases. The degree of imbalance before and after matching was assessed using standardised differences. All the patients tested with MMSE were successfully matched with their MoCA counterparts, so the matched sample comprised 132 patients, whereby the distribution of propensity score was practically identical in both groups. The regression models using either matched or total sample showed no statistically significant difference between the groups regarding association with the outcomes (the corresponding interaction terms were not statistically significant). MMSE/MoCA score was statistically significantly positively associated with prosthesis fitting and result of 6-minute-walk-test. Hence, both modelling approaches led to the same conclusions, and both cognitive screening test proved useful for predicting the primary rehabilitation outcome in patients after LLA.

Research on consumer segments and their expectations regarding failures and recovery strategies in online stores

Samo Kropivnik[1] and Ema Zdešar[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]Vrije Universiteit Amsterdam, Amsterdam, Netherlands

With increasing digitalization, shopping has shifted to online environment for many consumers. The online environment requires service providers to adapt their already established practises, including their recovery strategies. These help to resolve failures consumers face during service and thus enable to maintain their satisfaction and long-term loyalty to the provider. In the physical environment, the process of recovery is face-to-face, which makes it much easier to adapt to consumers in an intuitive or experiential way than in the online environment, where we have to rely on data and models that have not been sufficiently researched and tested. In this article, we therefore present established recovery strategies and make connections to justice theory, most commonly used by consumers to evaluate the process and the outcome. We emphasize the importance of identifying the expectations of online service users and offer a typology of Slovenian online service users. Indicators of the perception of distributive, procedural and interactional justice, as well as numerous other consumers' characteristics, were derived from previous research and adapted to the online environment. We collected data (2022, purposive sample, $N = 372$) with an online questionnaire. The questions were tested and adapted to Slovenian consumers based on preliminary interviews and a focus group. By using a multivariate method for hierarchical agglomerative clustering, we have uncovered consumer segments (types) at the general as well as at the specific level, for two different degrees of failure and linked typologies to each other and to additional consumer characteristics. In this way, we have contributed to filling in the research gap in this area and provided a basis for choosing the most efficient recovery strategy for Slovenian online stores.

Residential renovation in Slovenia: Comparison of survey and registry data

Ana Slavec

InnoRenew CoE, Izola/Isola, Slovenia

Data on residential renovations are crucial for assessing their energy efficiency and living conditions of residents. Based on the real estate register data in Slovenia, we have information about the year of construction of buildings, as well as the year of renovation of roofs, facades, windows, and installations. However, the data for the period after 2007, when the last real estate census was conducted, is incomplete. Moreover, official surveys also have very limited data on this topic. To get more detailed insights into the renovation procedures of Slovenian households, we conducted our own survey on a web panel, which provided insights into renovation characteristics not captured by existing data sources. This contribution compares survey data on renovations with registry data and attempts to evaluate the quality of survey data.

Acknowledgment: The author acknowledge receiving funding from the Slovenian Research Agency for the project Using questionnaires to measure attitudes and behaviours of building users [Z5-1879] and from the European Cooperation in Science and Technology for the InnoRenew project [grant agreement #739574] under the H2020 Spreading Excellence and Widening Participation Horizon2020 Widespread-Teaming program.

Integral territorial governance model for urban tourism destinations

David Klepej, Naja Marot and Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia

Cities increasingly promote development of tourism for the positive economic impacts in line with the growth paradigm, leading to (negative) impacts of this activity becoming ever more evident. Yet analysis of current strategic tourism and spatial planning documents for less and more prominent urban tourist destinations shows an alarming lack of consideration for wider impacts tourism has on urban destination and their natural, societal and economic environment. To address this gap, we developed an integral model of urban tourism's territorial governance based on two pillars: stakeholders (individuals, civil society, public sector, private sector) and policies (tourism, spatial, other). The model aims to assess the strength of ties among and between the policy areas and stakeholders, thereby identifying segments that need improved integration. In the first step of the destination-specific integral model of urban tourism's territorial governance building, relevant policies and stakeholders at the destination should be identified and placed into the relevant model segment. Assessment of individual segments is next, followed by the assessment of the quality of inter-segment collaborations in the processes of urban tourism's territorial governance within the (given) destination. The model also supports assessment of relationships between any chosen segment and urban tourism by differentiating among strong working relationships, relationships with room for improvement, defunct relationships and non-existent relationships. The model has been validated on Slovenian urban tourist destinations using both secondary resources (e.g., policy documents, written reports, clippings) and primary data (from interviews with stakeholders). This confirmed its potential use for identification of areas in the current destination governance (model segments) needing improvement in terms of policy formulation or stakeholder collaboration (or both) in urban destinations, regardless of their size and current tourism development stage.

A network approach to team leadership

Helena Kovačič, Hajdeja Iglič and Barbara Lužar

University of Ljubljana, Ljubljana, Slovenia

Our work presents a survey-based study of social networks in work organisations. The main objective of our study was to enhance our understanding of the relational aspect of team leadership. We use a network approach that allows for a more comprehensive conceptualisation of formal leadership. Our network analysis involves studying the entire population (team) through a holistic network approach. Collecting data on social networks of members in organisations through survey presented several challenges. We used the full-roster approach to collecting network data. The data was collected between September 2009 and May 2012. The final sample consisted of 5 teams from two Finnish organisations with a team member response rate of 80.7% and 18 teams from 8 Slovenian organisations with a team member response rate of 89%. We conducted a cluster analysis of the social network data of 23 teams and team leaders from Slovenian and Finnish organisations to develop a classification of the structural characteristics of teams' and leaders' networks. We chose cluster analysis instead of using dimension reduction methods to obtain a manageable set of variables, as it is an exploratory data analysis tool suitable for identifying patterns of relationships and classification. The results suggest four distinct social structures within which team leaders perform their leadership roles. We have shown that leadership roles can be distinguished from teams and team leaders based on the characteristics of social networks. We shift the discussion from leadership styles to leadership roles, which we understand as stable patterns of relationships. Not only did the analysis show that leaders differ along network characteristics, but also that this tendency is not randomly distributed across teams. We believe this is an important contribution to the field of leadership in general and team leadership in particular.

Career-relevance of statistics and research methods revisited

Irena Ograjenšek[1] and Iddo Gal[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]University of Haifa, Haifa, Israel

We already reported on an innovative study based on some revised and several newly proposed measurement scales to capture diverse beliefs and attitudes related to statistics; value of research methods; legitimacy of qualitative and quantitative research in the business and management domain; importance of both quantitative and qualitative reasoning; and more. While most studies to date focused on students of introductory statistics service courses, we aimed to capture beliefs and attitudes of graduate business and management students with varying degrees of work experience, who have already entered professional and leadership positions in the labor market or will do so upon their graduation.Our preliminary empirical results pointed to various gaps and interesting patterns. They indicated an intriguing gender difference, i.e., how male students have a more positive view of quantitative research for their career. They also showed that in general, graduate students of business and management do not necessarily view quantitative methods as more relevant for their future career when they take more quantitative research courses, contrary to expectations that students will develop positive views of statistics and research as they improve their knowledge in this regard. In this paper, we present some additional intriguing empirical findings and further elaborate on the implications of our study not only for those teaching statistics and research methods (both in the framework of graduate business and management studies and in general), but also for designers of study and training programs developing future corporate leaders.

Biostatistics 2

Identifying typical trajectories in longitudinal data: A case of study for patients affected by diabetic kidney disease

Veronica Distefano[1,2], Maria Mannone[2,3] and Irene Poli[2]

[1]University of Salento, Lecce, Italy
[2]Ca' Foscari University of Venice, Venice, Italy
[3]University of Palermo, Palermo, Italy

Diseases presenting a high heterogeneity make the choice of therapeutic treatments difficult, as in the case of diabetic kidney disease, a complication of type-2 diabetes. The issue can be faced with precision medicine. Repeated measures of kidney-filtering efficiency over time for each patient, and their comparison with other clinical features, can help retrieve information about the disease progress and the effect of drug combinations for specific subgroups of patients. We focus on the DC-ren longitudinal dataset, analyzing estimated glomerular filtration rate (eGFR) trajectories over time points corresponding to yearly visits. Finding subgroups of similar patients according to their eGFR trajectories allows one to associate characteristics at the baseline, disease progress, and effectiveness of drug combinations. We compute clusters of patients according to the shape similarity of their eGFR trajectories, measured via the Fréchet distance, also computing the mean trajectory in each cluster. We compare the obtained results with the information provided by the Traj method, associating trajectories according to several parameters. We also propose our results using visualization and sonification strategies, in line with the most recent developments of multi-sensory information representation in data sciences and signal processing. The results can be fed into a decisional system to help physicians select the best drug combination for each patient.

Climate change and the quality of wine: The case of Collio

Barbara Campisi, Gaetano Carmeci, Gianluigi Gallenti, Giovanni Millo, Matteo Carzedda and Paolo Bogoni

University of Trieste, Trieste, Italy

We address the influence of climate on wine quality, with the aim of predicting the effects of climate change in the medium term. We estimate an Ashenfelter equation on an unbalanced panel sample of 14 white and red varieties from 61 producers observed in the Collio region between 1985 and 2021. For the first time in the literature. We document the different reaction of wine varieties within a same terroir to temperature and rainfall.

External validation of the OAC3-PAD Risk Score and its underlying survival model

Nataša Kejžar[1], Kevin Pelicon[2], Klemen Petek[2], Anja Boc[1,2], Vinko Boc[2] and Tjaša Vižintin Cuderman[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]University Medical Centre Ljubljana, Ljubljana, Slovenia

The OAC3-PAD Risk Score was published in 2022 [1]. The Risk Score tries to predict major bleeding events for patients one year after hospitalization for symptomatic peripheral arterial disease (PAD). It was built upon the Cox proportional hazards model. The authors used the data of more than 81 thousand patients from a German insurance fund to create the statistical model and the Risk Score with 8 risk factors; model performance was supported by internal validation. Every risk prediction model needs to be validated externally. In this work, we conduct an external validation of the model and the Risk Score using data from a Slovenian retrospective cohort of 1500 PAD patients who underwent successful revascularisation at the University Medical Centre Ljubljana. To assess its performance, we present mean and moderate calibration results of the published statistical model. Furthermore, we examine the discrimination performance of the model and the Risk Score, providing insights into the ability to distinguish between patient groups with different bleeding risks. We point out the challenges and limitations that stem from the statistical model itself as well as the available data. Lastly, we try to evaluate the potential clinical utility of the model by conducting a clinical decision curve analysis.

References

[1] C.-A. Behrendt *et al.*, "The oac3-pad risk score predicts major bleeding events one year after hospitalisation for peripheral artery disease," *European Journal of Vascular and Endovascular Surgery*, vol. 63, no. 3, pp. 503–510, 2022. DOI: 10.1016/j.ejvs.2021.12.019.

Patient reported outcome measures of patients undergoing a primary knee or hip arthroplastics

Eva Podovšovnik and Vesna Levašič

Valdoltra Orthopedic Hospital, Ankaran, Slovenia

Patient Related Outcome Measures (PROMs) are very important to determine the success of the medical procedure. Measuring quality of life is one of the key elements for ensuring a healthy society, as stated by the World Health Organisation. In our case, we present results of PROMs questionnaires of patients undergoing primary knee or hip arthroplastics from September to December 2022 at the Orthopaedic Hospital Valdoltra, using Oxford Knee Score or Oxford Hip Score and EQ-5D-5L questionnaires, before surgery and 3 months after surgery. Data are collected in the National Arthroplasty Registry of Slovenia, which is managed at the Orthopaedic Hospital of Valdoltra. Both questionnaires were administred as self-reported surveys, in the case before surgery, and telephone surveys, in the case three months after surgery. Our research hypothesis states there are no differences in the PROMs results before surgery and 3 months after surgery. 232 patients undergoing primary hip arthroplastics, and 193 patients undergoing primary knee arthrtoplastics participated in our study. To test the differences in the overall score of the Oxford Knee Score or Oxford Hip Score and the overall evaluation of today's health (derived from EQ-5D-5L questionnaire), Related-Samples Wilcoxon Signed Rank Test was used. Results show statistically significant differences, at the 0.05 level, in score before the surgery and three months after surgery. The overall score on the Oxford Knee Score or Oxford Hip Score and the evaluation of today's health improved significantly. As such, we can conclude that patients reported an increase in their quality of life after the surgery—knee or hip arthroplastics.

Management of patients with acute coronary syndrome during the COVID-19 pandemic in Slovenia

Tjaša Furlan[1], Janez Bijec[2], Dalibor Gavrić[3], Petra Došenović Bonča[2], Irena Ograjenšek[2] and Borut Jug[4]

[1]Trbovlje General Hospital, Trbovlje, Slovenia
[2]University of Ljubljana, Ljubljana, Slovenia
[3]Health Insurance Institute of Slovenia, Ljubljana, Slovenia
[4]University Medical Centre Ljubljana, Ljubljana, Slovenia

In this study, we assessed the impact of the COVID-19 pandemic on hospital presentation and quality of care for patients with acute coronary syndrome. Data on 21 001 patients were used in the analysis (7057 with STEMI, 7649 with NSTEMI, and 6295 with unstable angina pectoris). We included patients admitted to the hospital for acute coronary syndrome between 2014 and 2021. We used unique identifying numbers to merge the national hospital database, national medicines reimbursement database and population mortality registry. We compared pre- and post-COVID-19 time trends for hospital admission and quality of hospital care indicators: reperfusion interventions, secondary preventive medication uptake and mortality. Data were fitted to segmented regression models for interrupted time series. Segmentation was based on the stringency index, and March 2020 was selected as the breakpoint. The pandemic caused no significant change in STEMI hospitalisations (92 patients; +1 patient per month, $p = 0.783$), but a significant decrease in NSTEMI (81 patients; −21 patients per month, $p = 0.015$) and unstable angina pectoris hospitalisations (47 patients; −28 patients per month, $p = 0.025$). We additionally analysed patients with STEMI, the only stable pre/post-COVID-19 cohort. There was no difference in reperfusion procedures (0.29%, [95%CI] −1.5%, 2.1%, $p = 0.755$) and in-hospital mortality (0.1%, [95%CI] −0.9%, 1.1%, $p = 0.815$), but we observed a significant negative trend for secondary preventive medication uptake (−0.12%, [95%CI] −0.23%, −0.01%, $p = 0.034$). Our findings indicate that the COVID-19 pandemic significantly affected hospitalisations of patients with NSTEMI and unstable angina pectoris. Mortality of hospitalised patients and reperfusion procedures remained unchanged, but we observed a steady decrease in the uptake of secondary preventive medication.

Measurement and modeling

Falling for the leading questions

Vanja Erčulj and Ajda Šulc

University of Maribor, Maribor, Slovenia

It is commonly known that leading questions in surveys should be avoided at all costs. Such questions are posed in a way to gauge the answer from the respondent, which a researcher wants to hear. There are several types of leading questions, from hypothetical questions, assumption questions to interlinked questions or questions using unbalanced scales. The answers obtained using leading questions distort the reality and inflate the measurement error. There are some people, however, that resist the leading questions. The objective of our research was to examine whether personality traits are associated with suggestibility. For this purpose, a pilot study was conducted among the students of Faculty of Criminal Justice and Security using a mixed research design approach. The results from analysis of interviews with students indicated that some students did not change their initial answers after asked a leading question. The (pilot) survey among the students lead to similar findings. No association, however, was found between personality traits and suggestibility.

Measuring concordance and discordance of student reading literacy data around the world

Simona Korenjak-Černe[1] and Barbara Japelj Pavešić[2]

[1]University of Ljubljana; Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
[2]Educational Research Institute, Ljubljana, Slovenia

Like similarity, which measures the agreement between two objects, concordance can be defined more formally and used as such in models. In 2019, Diday introduced such measures, called S-concordance and S-discordance with the prefix "S", as they were defined for symbolic data where the objects represent aggregations of individuals. S-concordance (S-discordance) measures the similarity or agreement (or disagreement, respectively) between an object and a collection of objects, and thus cannot have the symmetric property required for similarity measures. While similarity only considers the proximity between two objects, concordance also includes information about the distribution of values of all objects. Therefore, similarity and concordance provide a different view of the data and a different kind of knowledge about the data. We will illustrate some possible practical applications of these measures using the PIRLS dataset. The PIRLS dataset is based on an international, large-scale assessment that measures student achievement in traditional reading on paper and reading on digital devices in several countries around the world. As online reading has become a very important skill for younger generations and its importance has increased tremendously during the Covid-19 period, it is also important to support the study of measured literacies with new methodological approaches to allow extended understanding of the differences between these two skills. We examine the application of S-concordance and S-discordance measures to these data from several perspectives: (i) with S-concordance, we examine the agreements of high reading achievement between the single country and all other countries from this study; we also examine how well the aggregate data for the country represent the classes/teachers within the country; (ii) with S-discordance, we determine in which countries the observed categorical value is characteristic. In addition, we illustrate the use of these measures for comparisons between the two types of reading.

Assessing reliability and measurement error for continuous measurements

Nina Ružić Gorenjec and Nataša Kejžar

University of Ljubljana, Ljubljana, Slovenia

Any measurement used in practise or research must be reliable and measurement error should be quantified. Depending on a field, there are several different terms used for reliability, e.g. repeatability, reproducibility, agreement, consistency, concordance, precision, variability, dependability, stability. We will focus on measurements that are continuous variables. The reliability of such measurements (i.e. the extent to which repeated measurements for patients who have not changed are the same under different conditions) can be addressed in terms of intra-rater (same observer), inter-rater (different observers), or test-retest (over time) reliability, but all of them can be evaluated with the same statistical methods. Reliability for continuous measurements can typically be assessed by intraclass correlation coefficient (ICC) and measurement error by standard error of measurement or limits of agreement (using Bland and Altman method). There are several different types of ICC (and several different names for them), and the appropriate choice depends on our goal. We will highlight the distinction between reliability and measurement error, explain the differences between the various types of ICC, and provide guidelines on assessing reliability and measurement error for continuous measurements. Moreover, we will illustrate the aforementioned differences using simulated and real data, and show how (inappropriate) reporting can be misleading.

Non-linear stochastic model for dopamine cycle

Nenad Šuvak[1], Marija Milošević[2] and Jasmina Djordjević[2]

[1]University of Osijek, Osijek, Croatia
[2]University of Niš, Niš, Serbia

Dopamine is a crucial neurotransmitter that plays a central role in various aspects of brain functions, including reward processing, motivation, learning, and movement control. Its intricate involvement in these biological processes has made it a subject of extensive research across multiple disciplines, ranging from neuroscience and psychology to computational modeling. In this paper the non-linear stochastic model that describes synthesis, storage, release, uptake and metabolism of dopamine in dopaminergic nerve terminal of the rat striatum is introduced. The model is driven by 9-dimensional Brownian motion with constant coordinate intensities. Existence and uniqueness of a positive global solution $x(t) = [x_1(t), \ldots, x_9(t)]^\tau$ on a time-interval $[0, T]$ are proved and lower and upper bounds for specific moments of coordinate processes $(x_i(t), 0 \leq t \leq T)$ are derived. These results are used for calculation of bounds for intensities of driving processes and time horizons ensuring that expected values of coordinate processes stay in the same time-interval as the corresponding expected starting values. Furthermore, classical Euler-Maruyama scheme is adapted to ensure the positivity of numerical solution of the observed SDE system. This modified numerical scheme is used to simulate the coordinate processes in order to illustrate the theoretical results.

Stochastic SEIPHAR model for epidemic of the SARS-CoV-2 virus

Jasmina Djordjević[1], Ivan Papić[2] and Nenad Šuvak[2]

[1]University of Niš, Niš, Serbia
[2]University of Osijek, Osijek, Croatia

A refined version of the classical SEIR (susceptible-exposed-infected-recovered) model for the epidemic of the SARS-CoV-2 virus is proposed. The compartment of infected individuals is divided into four disjoint classes: symptomatic infected individuals (I), superspreaders (P), hospitalized infected individuals (H) and asymptomatic infected individuals (A). The model differentiates the spread of the virus via regular infected individuals (transmission coefficient β) and via superspreaders (transmission coefficient β'). The model is based on the system of ordinary differential equations describing the dynamics of epidemic, where stochastic component emerges from perturbing transmission coefficients β and β' by two independent Brownian motions with different intensities. The resulting system of stochastic differential equations (SDEs) is called the stochastic SEIPHAR model. The results include proof of existence and uniqueness of the positive global solution of the corresponding system of SDEs as well as the conditions for the extinction of the virus and its persistence in population (persistence in mean). Theoretical results are illustrated via simulations based on the data from the early phase of the epidemic in Wuhan (January 4 to March 9, 2020).

Inverse problem for parameters identification in a modified SIRD epidemic model using ensemble neural networks

Marian Petrica and Ionel Popescu

University of Bucharest, Bucharest, Romania

In this talk, we propose a parameter identification methodology of the SIRD model, an extension of the classical SIR model, that considers the deceased as a separate category. In addition, our model includes one parameter which is the ratio between the real total number of infected and the number of infected that were documented in the official statistics. Due to many factors, like governmental decisions, several variants circulating, the typical assumption that the parameters of the model stay constant for long periods of time is not realistic. Thus our objective is to create a method which works for short periods of time. In this scope, we approach the estimation relying on the previous seven days of data and then use the identified parameters to make predictions. To perform the estimation of the parameters we propose the average of an ensemble of neural networks. Each neural network is constructed based on a database built by solving the SIRD for seven days, with random parameters. In this way, the networks learn the parameters from the solution of the SIRD model. Lastly we use the ensemble to get estimates of the parameters from the real data of Covid-19 in Romania and then we illustrate the predictions for different periods of time, from 10 up to 45 days, for the number of deaths. The main goal was to apply this approach on the analysis of COVID-19 evolution in Romania, but this was also exemplified on other countries like Hungary, Czech Republic and Poland with similar results. The results are backed by a theorem which guarantees that we can recover the parameters of the model from the reported data. We believe this methodology can be used as a general tool for dealing with short term predictions of infectious diseases or in other compartmental models.

Mathematical statistics and modeling

Adaptive applicability of the Random Environment INAR models

Aleksandar Nastić

University of Niš, Niš, Serbia

In the past few years, several papers were published in which certain types of r-states random environment integer-valued autoregressive (RrINAR) processes were introduced and studied in detail. All of the observed processes showed exceptional performance in the modeling of different kinds of counting, i.e. nonnegative integer-valued time series. However, possibly due to the first glance complexity of the construction of these processes, unfortunately these RrINAR models have not been studied or further developed by other authors. Therefore, here I will try to present the basic concept of these autoregressive processes in the simplest possible way. This r-state Random environment INAR process {Xn} will be explained as a "part by part" standard INAR process, which parameters values of the marginal distribution, the thinning operator, as well as the order of the model, are all specified by another Markov process {Zn}. The only purpose of this {Zn} sequence of random states is to model the conditions of the environment in which we observe the counting process. Consequently, and considering the construction of the process, it ensures the adaptability of the main RrINAR model to all the process changes, so finally making it almost always very competitive and successful in data fitting. Moreover, unlike the standard INAR models, here one model can be used to describe many different types of count data.

Approximate Bayesian algorithm for tensor robust PCA using relative entropy

Andrej Srakar

University of Ljubljana, Ljubljana, Slovenia

Matrix and tensor completion methods are gaining interest. They allow to study un- and semistructured datasets of some contemporary endeavours such as citizen science initiatives. Recently proposed Tensor Robust Principal Component Analysis (TRPCA) aims to exactly recover the low-rank and sparse components from their sum. We extend an own, recently published Bayesian approximate inference algorithm for TRPCA, based on regression adjustment methods and compare it to earlier studies using variational Bayes inference. As the estimation is set in a high-dimensional context this leads to known bottlenecks which we solve using a novel proposal to use functional Bregman divergence between posterior distributions as a measure of the posterior surrogate loss. Definition (Functional Bregman Divergence): let $\phi : L^p(\nu) \rightarrow \mathbb{R}$ be a strictly convex, twice-continuously Fréchet-differentiable functional. The Bregman divergence $d_\phi : \mathcal{A} \times \mathcal{A} \rightarrow [0, \infty)$ is defined for all $f, g \in \mathcal{A}$ as $d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g]$, where $\delta\phi[g; \bullet]$ is the Fréchet derivative of ϕ at g. Namely, we use relative entropy as divergence measure combined with more general regression adjustment perspective. We provide proofs of the asymptotic properties of the approach as well as a simulation study. In a short application, we study two datasets of citizen science initiatives, Great Greenhouse Gas Grass Off initiative where individuals collected grass samples to infer the atmospheric fossil fuel CO2 mole fraction; and COVID-19 Sledilnik/Tracker with data on pandemic variables throughout the COVID-19 pandemic.

Improvements in parameter estimation for some class of the INAR models

Miodrag Djordjević

University of Niš, Niš, Serbia

During the second decade of the 21st century, a couple of new INAR time series models, based on the thinning operators, were introduced reaching whole set of integers, both non-negative and negative values. The elements of those time series were equal in distribution to a difference of two independent latent components, i.e., two independent random variables originating from the same distribution class such as Poisson (Freeland, 2010) or negative binomial (Barreto-Souza & Bourguignon, 2015; Nastic et al., 2016; Djordjevic, 2017) distribution. Those models are facing problems in parameter estimation phase. In this paper, solutions for some of the mentioned problems are proposed. The solutions rely on the latent components of the models and their properties.

Student session 1

Guitar tablature transcription with convolutional neural networks

Matija Marolt, Igor Nikolaj Sok and Igor Grabec

University of Ljubljana, Ljubljana, Slovenia

Automatic music transcription is still a rather prominent problem in music information retrieval and multimedia. In the past there have been many advances in the field, but only recently a direct transcription from guitar music to tablatures, which, unlike notes, are not uniform, has been developed. The goal of our presentation is to demonstrate the performance of the program, that is capable of estimating a playable and accurate tablature from a given recording of guitar music. For this process, we utilise convolutional neural networks, trained on the GuitarSet dataset. By further combining a convolutional and recurrent neural network architecture, we achieve rather good results for the task of automatic guitar tablature transcription.

A package for generating and grading exams

Jakob Peterlin

University of Ljubljana, Ljubljana, Slovenia

Creating and grading exams comprised of multiple-choice questions can be a daunting task. To alleviate this, we present our innovative open-source package designed to simplify and enhance the process significantly. Our toolkit facilitates constructing LaTeX-rendered exams, incorporating figures, tables, special characters, and equations. It further offers the convenience of reordering the exam questions, making it challenging for students to rely on their peers merely. This package steps beyond the traditional approach of recording answers on a predefined sheet of paper. Instead, it introduces a streamlined method for inputting answers, enabling the development of a tailored solutions sheet specific to a given exam. The student is then able to input their identification number conveniently. Upon completion of the exam scoring, the package can produce a report integrating basic statistical analysis, providing a comprehensive understanding of the exam's performance. The package's foundation is the Julia programming language, augmented with OpenCV for solution recognition and Flux.jl's neural networks. In its fully realized state, the package aims to maximize accessibility, potentially featuring a browser-based user interface and an easily deployable Docker image.

Land take, land use and environmental issues. Is the Kuznets Curve valid? The case of Italy

Giuseppe Borruso, Andrea Gallo, Francesco Magris and Nicola Pontarollo

University of Trieste, Trieste, Italy

Land take represents a relevant issue in all economies, with important consequences on the quality of the environment. These include soil sealing, increasing hydrogeological risk, reducing pollutants and warehouse gasses uptake, a competition among human activities for the "best" parcels of land, and is strictly related to other phenomena of urban and non-urban sprawl and sprinkling. In this paper we investigate the issue of the increasing land take and changes in land use, studying the relationship with economic development following the Kuznets Curve hypothesis. We analyze the Italian case considering provinces as the intermediate administrative scale of analysis, considering the evolution of land take/land use phenomenon during the last 20 years and that of economic development (GDP). We do so by making innovative use of spatial panel econometric techniques as well as of local indicator of spatial association (LISA) to examine local effects and spatial clustering. Spatial patterns of land take proximity arise in the years considered, highlighting areas of major presence of land take in neighbouring locations. The findings can be significant for local, regional and national policy makers in targeting coordinated policies, in terms of tax management and planning urban development and regeneration reducing extensive built-up land expansion.

Application of machine learning to fundamental analysis of securities

Aleksandr Panteleev

University of Ljubljana, Ljubljana, Slovenia

This research aims to apply machine learning to the time-honored investment philosophy of Benjamin Graham, as articulated in his 1940 book, "Security Analysis: Principles and Techniques." Graham's philosophy, further popularized by his notable former student, Warren Buffet, has had a significant influence on the world of investing. Traditionally, applications of machine learning in stock analysis have primarily been focused on two areas. One approach concentrates solely on the movement of stock prices as a predictor of future performance. The other uses fundamental data, such as net income, asset size, and profit margin, to calculate a firm's intrinsic value. While these methods have their merits, they often overlook key principles highlighted by Graham and Buffet. These include the prioritization of identifying highly undervalued stocks over merely predicting the direction of trends, the exclusion of certain economic sectors that are highly vulnerable to external influences (e.g., oil and real estate prices, interest rates), and the need for differentiation in acceptable price-to-earnings ratios among sectors. This study aims to address these gaps by aligning machine learning applications with Graham's principles. This research is ongoing, and future findings promise to offer a fresh perspective on the intersection of traditional investment philosophy and modern technological advancements.

Linguistic analysis of suicide related questions in the online counselling service This is Me

Vili Smolič, Sara Atanasova and Marjan Cugmas

University of Ljubljana, Ljubljana, Slovenia

There is an increasing trend in depressive symptoms among young people in recent years (Keyes et al., 2019), which can potentially lead to suicidality (Oexle et al., 2019). Numerous studies have shown that suicidality and mental distress have a negative impact on the desire to seek help (Wilson et al., 2005; Rickwood et al., 2007; Stewart, 2009; Dey & Jorm, 2017), however, due to its anonymity and convenience, the internet has proven to be an efficient tool in overcoming barriers, such as shame and stigma, that would normally inhibit young people's help-seeking (Stewart, 2009; Lamont-Mills et al., 2022). Especially, online health communities and online counselling services can be an important source of information and support, including when it comes to mental health issues and problems (Rickwood et al., 2007; Chuang & Yang, 2010; Saha & Sharma; 2020; Liu & Wang, 2021). Unfortunately, due to a lack of face-to-face interaction and limited availability of non-verbal and visual cues, there is a decrease in mutual awareness of concerns and interests between patients and health professionals (Johnston et al., 2013). To mitigate these limitations and optimize our effectiveness in providing patient care, it is beneficial to conduct a thorough text analysis of posts within health community forums. Therefore, the linguistic text analysis of questions and answers in the online counselling service for young people will be presented. Special attention will be given to compare the sentiment of questions and answers among different sub-themes within the suicide questions, controlling for some users' characteristics such as gender and age. The data were obtained by the Slovenian largest and oldest online counselling service for young people This is Me for the period between 2012 and 2021.

Posters

The application of NESTOREv1.0 to forecast strong aftershocks in the Northeastern Italy and Western Slovenia

Piero Brondi[1], Stefania Gentili[1] and Rita Di Giovambattista[2]

[1]National Institute of Oceanography and Applied Geophysics, Udine, Italy
[2]National Institute of Geophysics and Volcanology, Rome, Italy

During a seismic sequence, a strong earthquake can be followed by another event of similar magnitude. This subsequent event can aggravate the effects of the First Strong Earthquake (FSE), leading to the collapse of already damaged buildings and a higher death toll. To reduce the seismic risk during the occurrence of a seismic sequence, the forecasting of a Strong Subsequent Event (SSE) would be of strategic importance. Recently, the NESTORE (Next STrOng Related Earthquake) algorithm has been developed to forecast clusters in which a FSE of magnitude Mm is followed by an SSE of magnitude greater than Mm-1. In this case, the cluster is referred to as "A-type"; otherwise, it is referred to as "B-type." The distinction between these two classes is based on nine seismicity parameters (features) calculated for the cluster at increasing time intervals after the FSE related to spatial distribution, source area, magnitude, and energy trend over time. The NESTORE software is based on a machine learning approach and its first MATLAB version (NESTOREv1.0) was recently released on GitHub. The package consists of four modules: The first module extracts clusters from a seismic catalogue, the second learns to distinguish A-type from B-type clusters on a training dataset, the third and the last module compute the Bayesian probability that it is an A-type cluster for an independent database and an ongoing sequence, respectively. We applied NESTOREv1.0 to northeastern Italy and western Slovenia using the OGS network catalogue. In particular, we trained the algorithm on seismicity between 1977 and 2009 and tested its performance for the period 2010–2021. We found that six hours after the FSE the 94% of the clusters were correctly classified, which supports the application of NESTOREv1.0 in the studied area.

Acknowledgment: Funded by a grant from the Italian Ministry of Foreign Affairs and International Cooperation.

Applying walkthrough method for researching the moral references of the Signal application

Kristina Rakinić

University of Ljubljana, Ljubljana, Slovenia

The walkthrough method was developed by application and programme engineers to understand how programme codes work and to correct and improve them for easier and more user-friendly use (e.g., Fagan, 1976). Over the years it has evolved to analyse users' experiences with applications and to evaluate the usability of applications (Cavagnuolo et al., 2022). More recently, the method has been used to analyse cultural and social practices, ideas and ideologies mediated by applications (e.g., Duguay, 2017; Cabalquinto & Wood-Bradley, 2020). Using a walkthrough method, we analysed the functioning of the application Signal. Signal is a messaging application in which privacy is of paramount importance. Using the walkthrough method, we analysed the terms and conditions, features and marketing aspects of the application. We were also interested in what moral aspects are conveyed in the application according to moral foundations theory. Our analysis revealed that the fairness/cheating foundation predominates, as Signal positions itself as a fair application that values user privacy and does not steal and sell user data. The Loyalty/Betrayal foundation is also prominent, as the community part of the application is strongly emphasised, with addressing users as an important factor in the overall functioning of the application. The walkthrough method proved to be a useful way of analysing moral references embedded in the application.

Reconstruction of sea surface temperature with spectral convolution

Matic Klopčič[1], Matej Kristan[1] and Matjaž Ličer[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]National Institute of Biology, Ljubljana, Slovenia

Sea surface temperaure (SST) is crucial for accurate weather forecasting. However, clouds above the sea are preventing sensors on satellites to measure SST beneath them. Therefore, some measurements are missing. Current state of the art method DINCAE2 consists of autoencoder with refinement step. It reconstructs the missing data with bilinear convolution. We proposed new method for SST reconstruction, which is based on fast Fourier convolution (FFC). In spectral space, every frequency covers the entire image. Consequently, using convolution in spectral space, the receptive field covers the entire image in the first layer of network. We have analyzed the effect of different number of FFC blocks, skip connections, refinement step, feature fusion module and different loss functions. Our best method is AEFFC, which is an autoencoder with 9 FFC blocks without refinement step. State of the art method DINCAE2 has 0.5% lower error on the entire sea surface. Nevertheless, AEFFC has 2% lower error on the reconstructed surface.

Data mining of chlorophyll-a satellite data (CMEMS) enables reconstruction of phytoplankton blooms in the Adriatic Sea on large temporal and spatial scales

Nejc Prinčič[1], Martin Vodopivec[2], Patricija Mozetič[2] and Janja Francé[2]

[1]University of Primorska, Koper/Capodistria, Slovenia
[2]National Institute of Biology, Ljubljana, Slovenia

The rise of average global air temperature reflects in the rise in the temperature of the oceans as well. The latter play an important role in oxygen production, nutrient cycling, food availability and absorption of anthropogenic greenhouse gases. One way of studying the effect of climate change on the marine ecosystems is through the main primary producers—phytoplankton. We focused on satellite data for the area of the Adriatic Sea, obtained from the Copernicus Marine Environment Monitoring Service (CMEMS) for the 1997–2022 period. These data were used to identify changes in phytoplankton phenology. We investigated the start of the phytoplankton blooms from year to year and timing of peaks by using the satellite-derived concentration of chlorophyll a. Three different statistical methods were used. One was rate of change method (ROC), where we used the fast Fourier transform to obtain Fourier coefficients that showed us how many peaks occurred each year and their timing. The second one was the threshold method (TH) where we set a yearly median that represents the threshold value which indicates the start of the phytoplankton bloom. The last method used was the cumulative sum method (CS) where the cumulative biomass of chlorophyll exceeded a preset threshold value. We failed to detect statistically significant trends in most locations, which can also be attributed to relatively short period (in climatological sense) of data acquisition and many gaps in the dataset.

Invited lecture

Data science ethics: Some stories from the trenches

Richard De Veaux

Williams College, Williamstown, MA, United States

The ethical use of data may sound simple, but data can have unintended consequences, sometimes causing great harm. In this talk I will present several real examples of breaches of data ethics, both intentional and unintentional. From these, I hope we can all learn how to think about the impact of data in our own work and arrive at a set of principles that can guide us better in the ethical use of data.

Invited lecture

Capture-recapture methods with applications in health and society

Dankmar Böhning

University of Southampton, Southampton, England

The talk will focus on how capture-recapture methods based upon uni-list approaches can tackle questions of population size in health and society [1]. Starting with typical count data modelling approaches such as zero-truncation of Poisson, geometric or negative-binomial count densities the talk will continue introducing the more flexible ratio regression approach for capture-recapture applications. Among case studies considered will be the completeness of contact-tracing of Covid-19 in Thailand, the amount of drink-driving in Great Britain and the completeness of a meta-analysis on completed suicide after a specific surgery to reduce obesity. Proper estimation of uncertainty will also be given priority in the talk.

References

[1] D. Böhning, P. G. M. v. d. Heijden, and J. Bunge, Eds., *Capture-recapture methods for the social and medical sciences*. Boca Raton: CRC Press/Taylor & Francis Group, 2018.

Network analysis

3D visualization of multiway networks

Vladimir Batagelj

Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
University of Primorska, Koper/Capodistria, Slovenia

A multiway network $N = (V, L, w)$ is based on a list of k ways (sets of nodes) $V = (V_1, V_2, \ldots, V_k)$, a set of links L, and a weight $w : L \rightarrow \mathbb{R}$. To each link $e \in L$ corresponds a k-tuple of its (end)nodes $(e(1), e(2), \ldots, e(k))$, $e(i) \in V_i$. Visualization has an important role in the exploration of multiway networks and in the interpretation and presentation of the obtained results. For a selected 3D slice of a multiway network, we will present an approach based on a 3D layout described using the X3D format. Such a description can be dynamically inspected using some X3D viewer (such as view3dscene) or as a part of the web page supported by the javascript library X3DOM. The approach will be illustrated with applications to some real-life multiway networks. The code and data are available at GitHub `https://github.com/bavla/ibm3m`.

The network effects of international sanctions: A temporal blockmodeling study of trade sanctions

Fabio Ashtar Telarico

University of Ljubljana, Ljubljana, Slovenia

International trade was of the first topics in which network scienists got interested while approaching international relations. Already in the 1980s, network approaches showed that international trade is structured into clusters of "core' and "peripheral" countries performing distinct functions in accordance with dependency and world-system theory. Although there have been enormous advancements in network analysis as well as renewed geopolitical competition and the emergence of economic warfare, the topic is mostly absent from most recent studies. In particular, few authors are using cutting-edge techniques that were not available until a few years ago to analyse the international trade network. Hereby, a clustering techniques well-suited to identify such core-periphery structures called "blockmodeling" is used to analyse the dynamic evolution over time of international trade within the framework of world-system theory. In short, by focusing on countries under international sanctions regimes (Iran, Venezuela, Cuba) and specifically Russia. Practically, the analysis shows that sanctions can affect the target countries' position in the world economy. But this does not happen as often as some economists argue, especially for commodity producers. Finally, there seem to be a shift in the pattern of trade away from sanctioning countries and towards neutral/friendly ones. Crucially, these findings support the result of empirical srudies not using network methods. But more research is needed to test the working hypotheses on the mechanisms explaining how sanctioned countries established and drop ties in the international trade network.

Some considerations regarding blockmodeling of dynamic networks

Aleš Žiberna

University of Ljubljana, Ljubljana, Slovenia

While blockmodeling of dynamic networks might look like a very specific problem, it encompasses a broad range of approaches. This is due to several versions of dynamic (or temporal) networks and due to the vast number of different approaches to blockmodeling or, more generally partitioning. A dynamic network could be a network where we know for each tie the time of creation, occurrence or duration, or we could have a collection of snapshot of a network at different points in time, or aggregation of a number of "ties" that occurred (or whether at least one occurred) in a given time period. These different types of networks naturally require different blockmodeling approaches, as they must consider the nature of the data. Furthermore, the number of units could be stable or change in time. In addition, types of research questions and the nature of the data may influence the requirement of the blockmodeling approaches. For example, one might want the partition to be fixed in time or dynamic (changing in time), and the same could be also said for the blockmodel (model of connections among clusters). The talk will also try to highlight possible interesting research questions. A bit more emphasis will be placed on blockmodeling dynamic co-authorship networks, which have several characteristics that make them especially interesting.

Patterns of scientific collaboration in doctoral education: An analysis of mentor-mentee relationships

Marjan Cugmas, Luka Kronegger and Franc Mali

University of Ljubljana, Ljubljana, Slovenia

Doctoral study is central to the scientific socialization of young people, paving the way for their future careers, whether inside or outside academia. The bond between mentors and doctoral students is an essential part of this process, and the relationship manifests itself in various forms and degrees of intensity. Understanding the characteristics of mentoring collaborations is essential for developing successful higher education strategies for attracting potential doctoral students, and for developing effective academic policies. In a recent research study, we examined the different patterns of scientific collaboration between mentors and their mentees, focusing on bibliographic publications as an indicator of such collaborations. To identify patterns of collaboration, we applied a symbolic data clustering approach. We then used discriminant analysis to explain the clusters obtained. We considered several explanatory variables such as scientific field, age of mentee and mentor, gender homophily, year of completion of doctoral studies, number of mentors, and whether the Young Researcher Program provided financial support to the doctoral students. We obtained the data from the Slovenian information systems COBISS and SiCRIS, covering the period from 1991 to 2020. The most common pattern of collaboration is characterized by students being isolated from the scientific community at the beginning of their studies and being well integrated into the scientific community and highly productive researchers after finishing a doctoral study. This type of collaboration is more frequent in the years closer to the first years of the analyzed period. On the other hand, the type of mentor-mentee relationship limited to doctoral studies seems to become more frequent. This could be an indicator of several phenomena, such as the saturation of doctors in academia, the good receptivity of the nonacademic labor market, and the pursuit of a doctoral study for pragmatic reasons, such as for promotion.

Ranking genes based on gene spreading strength and mutation neighbor influence in network

Peter Juma Ochieng[1], József Dombi[1], Tibor Kalmár[1]and Miklós Krész[2]

[1]University of Szeged, Szeged, Hungary
[2]InnoRenew CoE, Izola/Isola, Slovenia

Understanding the intricate relationships between genes and their functions is essential for gaining insights into complex biological processes and diseases. In this study, we propose a novel network-based approach for ranking significantly related genes based on their spreading strength and neighbor influence in networks. Our method leverages network-based analysis to capture the interconnectedness of genes and their potential impact on cellular processes. First, we compute a mutation score for the genes based on the type of mutations they have. Then we consider the gene spreading strength, which measures the ability of a gene to propagate its influence to other genes, and the mutation neighbor influence, which quantifies the importance of a gene in the context of mutations, we can identify genes that play critical roles in the network. To evaluate the effectiveness of our approach, we applied it to a comprehensive dataset of genetic interactions and mutation data. The results demonstrated that our method successfully identified genes that are both highly connected and influential within the network. These genes represent potential targets for further investigation in the context of disease etiology and therapeutic interventions. Furthermore, we compared the performance of our approach with existing prioritization methods and found that it outperformed them in terms of precision and Discounted Cumulative Gain (DCG). By incorporating both gene spreading strength and mutation influence, our method offers a more comprehensive and informative approach to prioritizing genes of biological significance. In conclusion, our study presents a novel network-based approach for prioritizing significantly related genes based on their gene-spreading strength and mutation influence in networks. This approach has the potential to enhance our understanding of gene function, disease mechanisms, and the development of targeted therapies.

Visibility graph analysis of MODIS satellite evapotranspiration time series of olive groves in southern Italy: Revealing Xylella Fastidiosa induced phytopathogenic status

Luciano Telesca and Rosa Lasaponara

National Research Council, Tito, Italy

The visibility graph (VG) has become a statistical method widely employed to characterize the dynamical properties of time series. Developed by Lacasa et al. (2008) the VG converts time series into networks, whose nodes represent the series values linked between each other by their reciprocal "visibility". Recently, the VG has been used for the statistical investigation of time series in a wide variety of research fields. In this paper, we use the VG to analyse the topological properties of MODIS satellite evapotranspiration time series of areas covered by olive groves in southern Italy to reveal the presence of *Xylella Fastidiosa*, a very dangerous phytobacterium capable to induce a severe disease in olive trees, known as olive quick decline syndrome. For several hundreds pixels of infected and healthy sites different VG-based network metrics (mean connectivity degree, closeness, betweenness, diameter) were calculated. Our results suggest that by applying the VG a good discrimination between infected and healthy sites can be carried out, envisaging the use of this network analysis method as an operational tool for early diagnosis of plant deterioration due to *Xylella Fastidiosa*.

Modeling and simulation

The linear model vs. the proportional odds model for analysing ordinal and continuous outcomes: Simulation study

Georg Heinze, Michael Kammer, Daniel Kraemmer and Daniela Dunkler

Medical University of Vienna, Vienna, Austria

Ordinal outcomes occur frequently in medicine. A popular model for ordinal data is the cumulative logit link model with a proportional odds assumption (CLPO), which is a robust alternative to linear regression if model assumptions of the latter are in doubt. In a simulation study, we aimed at comparing the CLPO model to linear regression when data was generated by CLPO, a linear model or a multinomial model with non-proportional odds. We assumed that outcomes depended on three binary and three continuous covariates with moderate correlation. For analysis, we considered these three models but also linear regression with splines to model the continuous predictors. We targeted the performance in prediction and in detecting a non-zero effect of a binary covariate. When data followed the linear model, CLPO's predictive accuracy in independent test data was as good as that of linear regression. When data followed the CLPO, linear regression performed much worse than CLPO. The use of splines improved the poor performance, but without attaining the performance of CLPO. While CLPO achieved almost nominal type-I-error rates under the linear model, the linear model's type-I-error rates were seriously inflated under CLPO data generation. The multinomial model performed well for prediction under proportional odds, but was less powerful in detecting a covariate effect than CLPO even under non-proportional odds. Regression coefficients of the linear model and CLPO were correlated but not in agreement. Ordinal outcomes should not be analysed with linear regression, but continuous outcomes can be analysed with ordinal regression models without any loss in predictive accuracy and with negligible inflation of type-I-error rates. Multinomial regression is as accurate for prediction as CLPO but less powerful and should be used if the proportional odds assumption is seriously in doubt. Despite their satisfactory performance, ordinal models are still underused in medicine.

Choosing among three proportional odds models for ordinal and count outcomes—a matter of taste? A simulation study

Andreas Klinger, Daniela Dunkler, Mariella Gregorich and Georg Heinze

Medical University of Vienna, Vienna, Austria

Ordinal outcomes are frequently observed in medicine. Despite the flexibility of models for ordinal data, they still seem underused for the statistical analysis of medical studies. We review differences between three popular models for ordinal data: the cumulative logit link model, the adjacent categories model and the continuation ratio model. These models incorporate the assumption of proportional odds, which can be relaxed in order to obtain a multinomial model. We present results from a simulation study where we compared the three models under different data generating mechanisms for ordinal outcomes and a mechanism that produces negative-binomially distributed count-type outcomes. The data generating mechanism involved three binary and three continuous predictors of the outcome. We targeted the performance in prediction and in detecting a non-zero effect of a binary covariate. We generated outcome variables following a cumulative logit (CLPO), adjacent categories (ACPO) and continuous ratio (CRPO) proportional odds model and used the hit rate as performance measure. When the outcome variable was generated by a CLPO or ACPO model, the CLPO model achieved slightly better predictions, while the ACPO model was slightly better at detecting a non-zero effect of a binary covariate. When the outcome variable was generated by the CRPO model, the CRPO model performed marginally better in both categories. When the outcome variable followed a negative binomial distribution, the negative binomial regression model clearly outperformed the proportional odds models with a lower mean squared prediction error and higher power to detect a non-zero effect of a binary covariate. Performance of the three different proportional odds models is very similar when fitted on ordinal outcomes. Slight performance differences depend on the exact distribution of the ordinal outcome variable. For count-type outcomes, the proportional odds model performs clearly worse than negative binomial regression and should be avoided.

Bayesian state-space modeling of indoor radon concentration and entry rate

Marek Brabec

Institute of Computer Science of Czech Academy of Sciences, Prague, Czech Republic

In this paper, we will derive a structured, physically motivated dynamical model of indoor radon concentrations. Since not only concentrations, but also time-varying radon entry rates are of interest in real situations (e.g., for evaluation effectiveness of mitigation strategies), measurements are bivariate—consisting of radon and tracer gas concentrations. Being motivated by underlying physical model based on a nonlinear system of two differential equations, we formulate a structured state-space model allowing for estimation of all (functional) quantities of interest. Then, we postulate a fully Bayesian model and obtain smoother estimates of radon concentration, radon entry rate and ventilation rate from posterior. We illustrate the approach on real measurement campaign data using Stan MCMC computations.

Modelling extreme bivariate data using R software

Maria Manuela Neves[1] and Helena Penalva[1,2]

[1]University of Lisbon, Lisbon, Portugal
[2]Polytechnic Institute of Setúbal, Setúbal, Portugal

Extreme value theory and methods are intensively used in many research fields, such as finance, insurance, health, climate, and environmental studies. Statistical inference about extreme events is interested in the estimation of the probability of occurrence of events more extremes than any that have already been observed. In the univariate analysis of extreme values it is of great importance the model assumptions on the tail of the distribution function underlying the data. However many problems involving extreme events are inherently multivariate. In this talk we will begin to show the challenges that arise when using bivariate data. Bivariate extreme value distributions contain parameters of two types: those that define the marginal distributions, and parameters defining the dependence between suitably standardized variates. Typically, first the margins are dealt with and second, after a transformation standardizing the margins to a common scale, the dependence structure is studied. There currently exist a variety of statistical methods for modelling bivariate extremes. Here a set of steps for performing a bivariate data analysis of extreme values in R environment is discussed. Traditional and more recent procedures are compared through some packages and functions existing and also constructed in R. Several applications to real data sets are shown.

Acknowledgment: Research partially supported by National Funds through Fundação para a Ciência e a Tecnologia, project UIDB/00006/2020 (CEA/UL).

To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets

Hana Šinkovec[1], Rok Blagus[1], Georg Heinze[2] and Angelika Geroldinger[2]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]Medical University of Vienna, Vienna, Austria

For finite samples with binary outcomes penalized logistic regression such as ridge logistic regression (RR) has the potential of achieving smaller mean squared errors (MSE) of coefficients and predictions than maximum likelihood estimation. There is evidence, however, that RR can result in highly variable calibration slopes in small or sparse data situations, thus, Firth's correction (FC) may be preferable. Motivated by a study relating dependence in daily activities to nine risk factors we demonstrate that estimating the complexity parameter in RR from the data by minimizing some measure of the out-of-sample prediction error or information criterion may be difficult. We elaborate this issue further by performing a comprehensive simulation study, investigating the performance of RR in terms of coefficients and predictions and compare it to FC. In addition to tuned RR where the penalty strength is estimated from the data, we also considered RR with pre-specified degree of shrinkage. We included "oracle" models in the simulation study to show what the best possible performance of RR could be if the true underlying data generating mechanism was known. We observed that complexity parameter values optimized in small or sparse datasets are negatively correlated with optimal values and suffer from substantial variability which translates into large MSE of coefficients and large variability of calibration slopes. In contrast, in our simulations pre-specifying the degree of shrinkage prior to fitting led to accurate coefficients and predictions even in non-ideal settings such as encountered in the context of rare outcomes or sparse predictors.

Official Statistics

The use of the Earth observation data for the monitoring of permanent grassland and soil moisture

Črt Šuštar

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia

This study presents country-scale procedures for official statistics production related to grassland use and soil moisture (irrigation detection) parameters in Slovenia using machine-learning models applied to freely available satellite imagery. We obtained reliable results through robust procedures for mowing detection and the assessment of the age of permanent grassland, owing to clear spectral dependencies and ample reference data. For mowing detection, a combination of drops in the vegetation index (NDVI) for optical (Sentinel-2) data and coherences for radar (Sentinel-1) data proved effective. By integrating both data sources, we enhanced the overall data quality. To estimate the age of permanent grassland, we applied bare soil detection algorithm on Sentinel and Landsat imagery, allowing us to deduce the age of grassland for up to 20 years in the past. However, detecting fertilisation, grassland renewal, and grazing presence presented more challenges due to limited reference data and unclear spectral dependencies. On the other hand, the pilot study on irrigation detection delivered promising results, yet it would require stratified processing and increased availability of reference data for reliable country-scale results. The developed procedures enable high-resolution, large-scale mapping of attributes, as well as the monitoring of parameters for past periods, a task challenging for conventional statistical methods. However, automated processing of such extensive data on a country scale poses significant challenges. The main obstacle to achieving fully reliable results is the scarcity of accurate and representative reference data. Therefore, we recommend improving reference databases and fostering collaborations with organisations and institutions to enhance data sharing and partnership opportunities. This project contributes to advancing statistics production using Earth observation data and provides valuable insights for further implementation of such procedures. By overcoming current limitations, the study lays the foundation for efficient country-scale monitoring of grassland parameters and irrigation, with potential applications in other regions as well.

Using scraped data in calculating inflation

Matevž Postrašija

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia

Consumer habits are changing, new technology is coming into general use, and new data collection and processing tools are available. For some time now, online shopping has been the primary way of purchasing certain types of products. Multiplication of online stores has not only benefited customers but also analysts. In traditional data collection, data collectors go from store to store and collect the prices of products, but the Internet offers the possibility of collecting data on prices centrally and automatically for all online stores. Online price collection or online scraping made it possible to optimize the inflation calculation process in terms of time, price and quality. For some products, we now collect prices faster, on a larger scale and using fewer resources. Although there are already tools that enable web scraping, we have concluded that for effective work we need a tool that is completely adapted to our needs and desires. We are not only interested in prices, but also in product characteristics and some metadata. With a very large amount of collected data, new options for calculating indices emerged. By modernising data collection, we also upgraded the methodology, as we changed the type of index used in the calculation. Rather than comparing products on a one-to-one basis (i.e. like-for-like comparison between two months), we now compare groups of products. In the background, there is of course a demanding process of data cleaning and preparation, but in this way, we avoid other time-consuming and more subjective methods, such as making quality adjustments when replacing products in the sample. In the process, we are aided by big data processing tools. By following current trends (the shift to online shopping) and using an updated method of data collection (web scraping), we developed an improved index (more up-to-date data being used).

Using machine learning for assisted classification of articles

Črt Grahonja

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia

We present the results of an introduction of machine learning to a linking process at the Statistical Office of the Republic of Slovenia (SURS). SURS has been collecting scanner data of food, beverage and tobacco articles from main Slovenian stores for a few years. Until now, we linked each store's own article groups to COICOP codes using automatic text rules and manual checks. Such methods are not completely reliable and mistakes happen. Furthermore, the process is labor intensive and it would take a long time to classify the data fully, which we cannot afford due to time constraints in the process. We introduced machine learning techniques as a replacement of these methods. The training data include names and GTIN values, which bring a lot of predictive power into the process. We tested two machine learning types of classifiers: decision trees and one-vs-all based on logistic regressions. In both cases, the resulting accuracy is very high (around 95%). The classification is also much faster than the previous method and seems promising for future use in production stages.

Using the cell key method for protection of grids at Statistics Slovenia

Manca Golmajer

Statistical Office of the Republic of Slovenia, Ljubljana, Slovenia

The cell key method (CKM) is a perturbative method used for tabular data protection. The value of each table cell is randomly changed. The random change depends on the record keys and the perturbation table. The resulting table is not additive, even though the unprotected table is. For grid data protection, from reference year 2021 on the Statistical Office of the Republic of Slovenia has been using the CKM. For the CKM, the R software is used. Record keys are generated in R using the `runif()` function (random uniform distribution). The perturbation table is generated in R using the `ptable` package. The record keys and the perturbation table are the same for different grids for the same reference year. For the selection of record keys, additional conditions are imposed, such as the perturbation of total population is zero, and record keys from previous reference year are taken into account. Due to a large number of grids, during the process of protection, partitions of grids and microdata are made. Grid data for Statistics Slovenia for reference years 2021 and 2022 have already been published in the special application called STAGE. Various statistics for grids with 1 km, 500 m, and 100 m sides are prepared.

Analysis of protection of Census hypercubes with the cell key method

Janez Bijec

University of Ljubljana, Ljubljana, Slovenia

In recent years, statistical disclosure control (SDC) has seen a lot of development of so-called perturbation methods, which are used for statistical disclosure control of aggregate data. The main characteristic of such methods is that they slightly change the values of cells. The main objective of this paper is to present and analyse the results of the cell key method (CKM), which is the method we want to use for statistical protection of 2021 Census hypercubes. The cell key method perturbs every single cell in the table by a special algorithm and thus protects the table by creating doubt in the cell values presented. Eurostat proposed targeted record swapping and CKM as the preferred SDC methods for protecting 2021 Census data. We test the feasibility of statistical disclosure control with CKM in view of the fact the Census tables are linked to the annual demographic tables, which need to be sent annually to Eurostat in accordance with EU regulation 2017/712 (CIR-2). In years when the Census is conducted, the data source is the same, so protection of the tables needs to be harmonized. We implement the CKM method in the R programing language, more specifically in RStudio working environment. The main objective of our work is to analyse the possibilities of using CKM for protecting Census hypercubes. Our analysis shows that information loss using CKM is small compared to other methods such as suppression. The method is also quite easy to implement using R. The main takeaway of the analysis is that SDC using only CKM is not sufficient, since we can not create enough doubt in the cell values at lower levels (cohesion and statistical regions), if their value at the highest geographical level of Slovenia is 1.

Student session 2

Prediction models to support adult IgA vasculitis diagnosis and assessing renal involvement

Ana Markež[1], Matija Bajželj[1,2,3], Alojzija Hočevar[3], Katja Lakota[2,3] and Rok Blagus[1]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]University of Primorska, Koper/Capodistria, Slovenia
[3]University Medical Centre Ljubljana, Ljubljana, Slovenia

Immunoglobulin A vasculitis (IgAV) is a small vessel vasculitis, characterized by variable clinical presentation. The severe form of the disease with renal involvement may lead to significant morbidity, especially in adults. Clinicians lack biomarkers to support the diagnosis of adult IgAV. Additionally, it is not clear if clinically used markers for IgAV could predict patients at risk for renal involvement. The first aim was to build the prediction model using the data for 59 IgAV patients and 22 age and sex-matched healthy controls (HC). Model was built based on the measurement of 15 serum analytes potentially involved in the disease pathogenesis (Osteopontin, LBP, ANGPTL4, IL-15, FABP-4, CCL19, Kalikrein5, CCL3, Leptin, IL-18, MMP1, CXCL10, AdipoQ, CXCL5, SERPINA12) and measured using Luminex Assay. The second aim was to build a prediction model using the data for 21 IgAV patients with skin-limited disease and 19 patients with renal involvement using: routinely gathered clinical and laboratory data (70 variables, Model 1) and additional 15 analytes (85 variables, Model 2). We used two classifiers that were shown to perform well with this type of data, random forests and stochastic EasyEnsemble. The predictive ability was evaluated by performing 100 splits to a training and a test set (80:20 ratio). The area under the ROC curve (AUC) was considered as a primary performance measure. Models 1 and 2 were formally compared by DeLong's test; 95% confidence interval for difference in their respective AUCs were also calculated by using bootstrap. Based on a model built using 15 analytes it is possible to distinguish between IgAV patients and HC (AUC 0.98 [0.93, 1]), but the discrimination between the patients with renal involvement from patients with skin-limited disease is limited and not enhanced by additional analytes. Validation of models on the larger IgAV patient's cohort is warranted.

A new statistical index for evaluating variability in patient state index during pediatric anesthesia

Noor Muhammad Khan[1], Claudia Maria Bonardi[2], Angela Amigoni[2] and Dario Gregori[1]

[1]University of Padua, Padua, Italy
[2]University Hospital of Padua, Padua, Italy

The Patient State Index (PSI) is widely used tool for monitoring sedation levels in pediatric anesthesia, providing an indication of the patient's level of consciousness and depth of anesthesia. However, the detection of time points when the PSI level changes from a stable sedation level remains a challenge, since the distribution of PSI is usually non normal and is expected to have outliers, a robust method is required to evaluate the phases. The anesthesia period is clinically divided into well-defined five phases; however, recent studies have unexpectedly detected large variations of PSI have even within same phase during pediatric anesthesia. This study proposes the Variability Ratio Index (VARI), a simple statistical tool based on the deviation of PSI from its stationary process, to evaluate sedation phases. The VARI is calculated as the ratio between the total number of change points and the total time points within each phase. Change points are detected using the pruned exact linear time algorithm. VARI showed robust behavior in both parametric bootstrapping in Bayesian paradigm and Monte Carlo simulation. To demonstrate the practical application of VARI, a single-center retrospective study was conducted using PSI data from pediatric patients undergoing cardiac surgery with extracorporeal circulation. The study included twenty patients monitored using the Sedline monitor at 124 699 time points. VARI successfully identified the hypothermic phase with the lowest value and the awakening phase with the highest value of it, highlighting its potential in assessing sedation depth during anesthesia. Furthermore, an R package called `varifinder` has been developed to facilitate the use of VARI. Further research and data are necessary to fully explore the utility of VARI in different clinical settings.

Bridging the gap: Integrating efficacy and quality of life in colorectal cancer observational study using the win ratio approach

Maria Vittoria Chiaruttini, Giulia Lorenzoni, Gaya Spolverato and Dario Gregori

University of Padua, Padua, Italy

Traditional oncological studies prioritize efficacy outcomes, overlooking crucial aspects of patient quality of life and safety. The win ratio (WR) approach addresses this gap by considering both efficacy and quality of life/safety outcomes, providing a comprehensive evaluation of treatment benefits and drawbacks. However, achieving comparability between different treatment arms in observational studies poses a challenge. This analysis explores the application of the WR approach in observational study of colorectal cancer, taking into consideration the comparability of patient pairs to draw reliable conclusions. This multicentric study enrolled consecutive colorectal cancer patients who underwent local excision (LE) or chose the watch and wait approach (WW). Employing the WR approach, the WW was compared to LE as the treatment of interest. Primary efficacy outcomes—ranked by importance—included local recurrence, distant recurrence, and overall survival. Additionally, quality of life outcomes such as ostomy presence and rectum preservation were evaluated. WRs and 95% confidence intervals were calculated in unmatched and matched datasets by propensity score matching. Sensitivity analysis examined WR considering efficacy outcomes alone versus incorporating quality of life endpoints with varying relative importance. The study included 184 complete cases: 113 patients underwent LE, and 71 the WW approach. In the matched dataset, each arm consisted of 60 patients. WRs for efficacy outcomes alone were 0.24 [0.12, 0.48] in the unmatched dataset and 0.27 [0.07, 0.58] in the matched dataset. Considering ostomy presence and rectum preservation, WRs were 0.55 [0.31, 1.00] and 0.78 [0.45, 1.34] in the unmatched dataset (when postponed and anteponed to efficacy outcomes, respectively), and 0.5 [0.21, 0.98] and 0.74 [0.34, 1.46] in the matched dataset. The WR is able to comprehensively evaluates efficacy and quality of life in observational studies. While LE initially appears superior, incorporating quality of life endpoints reveals no significant difference between treatment arms.

Modeling and forecasting mortality with economic, environmental and lifestyle variables

Matteo Dimai

University of Trieste, Trieste, Italy

Traditional stochastic mortality models tend to extrapolate, to focus on identifying trends in mortality without explaining them. Those that do link mortality with other variables usually limit themselves to GDP—but we can only account for what is in the model and focusing on the impact of economic growth on life expectancy ignores factors that have been identified as important at the micro level. This article presents a novel stochastic mortality model that incorporates a wide range of variables related to economic, environmental, and lifestyle factors to predict mortality. The model uses principal components derived from these variables in an extension of the Niu and Melenberg (2014) model and is applied to 37 countries from the Human Mortality Database. Model fit is superior to the Lee-Carter model for 18 countries. The forecasting accuracy of the proposed model is better than that of the Niu-Melenberg model for half of the countries analyzed under various jump-off years. The model is designed to facilitate scenario building and policy planning, providing insights into the interplay between different factors that affect mortality. Furthermore, an extension of the model to the multipopulation case is presented.

A study on the use and the necessity of machine learning dimensionality reduction and clustering methods in actuarial sciences: Defining the right methodology for business planning under the requirements of IFRS 17

Mateo Antonac

University of Primorska, Koper/Capodistria, Slovenia

The insurance industry is getting close to a new chapter, as automatization, big data, and improvement of prediction models are around the corner. In recent times, much progress has been made regarding the use of machine learning methods in the insurance industry. Moreover, clustering methods are often a topic in actuarial journals. However, not a lot of attention is given to the part of the process that precedes the clustering analyses, which are often dimensionality reduction models. This article is covering the areas of dimensionality reduction and clustering of insurance portfolios. Here, the focus is on the combination of both dimensionality reduction and clustering models. The article's main contribution is that it treats an actual challenge in the actuarial industry where procedures, such as partitioning the current portfolio by identifying similarities in the data and automatizing the task of generating a predicted new portfolio, are needed. The presented results show that the combination of the factor analysis of mixed data model and the hierarchical clustering model, is efficient when dealing with complex datasets. The described and tested combination of dimension reduction and clustering methods provides the foundation for the methodology of multi-year business planning, introduced in this article, satisfying the requirements of the International Financial Reporting Standard 17.

Student session 3

The challenge of multiple testing: Case of emission coupons trading

Anja Žavbi Kunaver, Marjan Cugmas and Irena Ograjenšek

University of Ljubljana, Ljubljana, Slovenia

In practice, we often face the challenge of multiple testing, which, without appropriate corrections, may lead to too many rejected true null hypothesis (type I error rate). The challenge also needs to be dealt with in the energy sector of the economy. One such case is trading with emission coupons, which we address in this study. The purpose of emission coupons is to reduce the emissions of carbon dioxide. Trading with coupons takes place at two levels: on the primary and secondary market. Here, companies which actively trade on the emission markets, strive to provide themselves with enough allowance for carbon dioxide emissions. On the primary market, there are auctions every working day at 11 o'clock. The secondary market is active ten hours a day every working day. There is trading going on continuously on the secondary market and the traders want to know which day and which hour are optimal for selling or buying emission coupons (when is the actual price higher or lower than daily or weekly average). There are various possible approaches for testing statistical significance of differences between average prices such as t-test, Wilcoxon signed-rank test, Tukey test and permutation test. In this study, we focus on comparison between paired t-test and a permutation test. Our simulations show that both tests are appropriate and have about the same test power. But the paired t-test is faster and easier to use, so we choose it as more appropriate. We also attempt to find out which multiple testing procedure is the most appropriate—with which procedure do we get the best test power and how big are differences between procedures. Finally, we move away from simulations to real-life data in order to propose a relevant trading strategy.

Corrections to Bland–Altman analysis for repeated measures data—are they always essential?

Maša Kušar

University of Ljubljana, Ljubljana, Slovenia

Simple Bland–Altman analysis assumes that all measurements by each of the compared methods are independent. At the same time, it is highly advisable to assess the repeatability of each separate method alongside a Bland–Altman analysis. This leads to the availability of multiple measurements per subject and method, which could and should be used in the analysis of method agreement. However, these measurements are not independent within the subject. The authors suggest three different corrections to the Bland–Altman analysis, to be used after consideration of the data structure. A review of the medical literature shows that these corrections are usually not used, or even considered. Exacerbating the problem, these methods are often not included in R packages that enable Bland–Altman analysis. Considering the widespread practice of using simple Bland–Altman analysis on repeated measures data, we sought to evaluate the error in the estimate of method agreement, if the corrections are not applied. This could help authors in deciding whether their data allows a simplified analysis, or would such an approach result in an excessively optimistic estimation of method agreement.

Handling separation in generalized linear mixed effects models with a random intercept

Tina Košuta[1], Rok Blagus[1], Georg Heinze[2] and Nina Ružić Gorenjec[1]

[1]University of Ljubljana, Ljubljana, Slovenia
[2]Medical University of Vienna, Vienna, Austria

The issue of separation in logistic regression arises when the events and non-events can be perfectly separated by a hyperplane in the covariate space. As a result, at least one of the maximum likelihood estimates diverges to infinity. For logistic regression, regularization methods, such as introducing a penalty into likelihood according to Firth (1993), have been successful in solving the issue of separation. However, satisfactory solutions are lacking in generalized linear mixed effect models (GLMM) so far. Our work aims to propose a solution for the separation in GLMMs with a random intercept through the regularization of parameter estimation by using pseudo-observations. In this presentation, we will explore the extension of regularization to the parameter estimation of the random effect variance by using a conditionally conjugate prior. The fixed effect estimates were penalized using Firth's penalty, which yields equivalent results to regularizing fixed effects using Jeffery's prior invariant. To incorporate the priors, we have expressed them as pseudo-observations and added them to our data. Then, we estimated the parameters with maximum likelihood estimation. This approach was compared to a regular hierarchical Bayesian model with identical priors to address the differences between both approaches. In a simulation study, we evaluated the effectiveness of our proposed approach in dealing with separation as well as several performance metrics such as bias and mean-squared error. We also illustrated the approach by means of the analysis of a real data example.

An investigation into overrepresentation of COVID-related genes in pathway enrichment analysis for RNA-seq data

Sara Ahsani-Nasab, Daniele Sabbatini, Elena Pegoraro, Dario Gregori and Luca Vedovelli

University of Padua, Padua, Italy

The COVID-19 pandemic has spurred an unprecedented surge in biomedical research, leading to the production of extensive amounts of RNA-related data. However, the haste in data generation, often with suboptimal experimental design and analysis, has raised concerns about the potential overrepresentation of COVID-related pathways in enrichment analysis. Overrepresentation of these could confound findings, leading to identification of COVID-related pathways in unrelated diseases. This may potentially distort our understanding of the molecular underpinnings of these diseases and hinder therapeutic progress. This study aims to investigate the potential overrepresentation of COVID-19-related genes in pathway enrichment analyses using multiple pathway analysis tools, and to detect the possible reasons behind this overrepresentation. Furthermore, we sought to explore the ramifications of overrepresentation in the context of other diseases, such as cancer. Utilizing an in silico approach, we evaluated a vast dataset of published RNA data from COVID-19 and non-COVID-19 research. Multiple pathway analysis tools were used to perform enrichment analysis identify false overrepresentation of COVID-19-related pathways in unrelated disease contexts. Statistical methods were applied to assess the validity of gene involvement in these pathways that are improperly overrepresented, compared to non-COVID-19-related genes. Our findings confirm the existence of overrepresentation of COVID-related genes across various pathway analysis tools. This overrepresentation can be attributed to factors such as the sheer volume of COVID-19 data, bias in data interpretation, and the innate overlap of gene networks across diseases. This issue is not unique to COVID-19 and could be observed in other diseases, indicating a systemic challenge in the bioinformatics field. These findings underline the importance of rigorous experimental design, unbiased data analysis and interpretation, and the need for novel algorithms that can effectively account for potential false overrepresentation in pathway analysis.

Enhancing survey design and analysis: Leveraging machine learning for post-stratification

Mingmeng Geng and Roberto Trotta

International School for Advanced Studies, Trieste, Italy

Surveys play a crucial role in social science research, and the representativeness of the survey samples often determines its quality. However, no matter how we design and distribute the survey, it's always challenging to predict its completers and raw survey data generally cannot be used directly to represent the population of interest. Therefore, post-stratification is usually a necessary and practical step, which allows us to reduce errors. The significance of post-stratification is obvious, but how can it be effectively implemented? Traditionally, many researchers have grouped the target population by age, gender, region, and other factors based on common sense. One of our goals is to find a more systematic path for post-stratification thanks to machine learning methods to predict opinions based on a set of observed features. This framework also proves suitable for predicting non-responses in surveys. Using some tree-based models, such as random forests and XGBoost, we explored the personal perspective on certain issues based on his background information. Our strategy has demonstrated efficacy on several datasets. There are various aspects of our model that could be further explored and discussed, for example, the influence of missing data, the effect of different encoding techniques for categorical data, the handling of continuous variables.

Invited session

Phases of methodological research in biostatistics: Building the evidence base for new methods

Georg Heinze[1], Anne-Laure Boulesteix[2], Michael Kammer[1], Tim Morris[3] and Ian White[3]

[1]Medical University of Vienna, Vienna, Austria
[2]Ludwig Maximilian University of Munich, Munich, Germany
[3]University College London, London, England

Although new biostatistical methods are published at a very high rate, many of these developments are not trustworthy enough to be adopted by the scientific community. We propose a framework to think about how a piece of methodological work contributes to the evidence base for a method. Similar to the well-known phases of clinical research in drug development, we propose to define four phases of methodological research. These four phases cover (i) proposing a new methodological idea while providing, for example, logical reasoning or proofs, (ii) providing empirical evidence, first in a narrow target setting, then (iii) in an extended range of settings and for various outcomes, accompanied by appropriate application examples, and (iv) investigations that establish a method as sufficiently well-understood to know when it is preferred over others and when it is not; that is, its pitfalls. We suggest basic definitions of the four phases to provoke thought and discussion rather than devising an unambiguous classification of studies into phases. Too many methodological developments finish before phase (iii/iv), but we give two examples with references. Our concept rebalances the emphasis to studies in phases (iii) and (iv), that is, carefully planned method comparison studies and studies that explore the empirical properties of existing methods in a wider range of problems. All authors of this paper are members of the international STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies). The proposed framework aims at refining the notion of evidence in methodological research that is central to STRATOS' efforts.

Initial data analysis: Making the effort worthwhile

Lara Lusa[1,2], Carsten Oliver Schmidt[3], Georg Heinze[4] and Marianne Huebner[5]

[1]University of Primorska, Koper/Capodistria, Slovenia
[2]University of Ljubljana, Ljubljana, Slovenia
[3]University of Greifswald, Greifswald, Germany
[4]Medical University of Vienna, Vienna, Austria
[5]Michigan State University, East Lansing, MI, United States

Initial data analysis (IDA) is the part of the data pipeline that takes place between the end of data retrieval and the beginning of data analysis that addresses the research question. Systematic IDA and clear reporting of the IDA findings is an important step towards reproducible research. A general framework of IDA for observational studies includes preparation of metadata, data cleaning, data screening, possible updates of pre-planned statistical analyses, and documentation and reporting of IDA results. In this talk we present our proposals on how to efficiently embed the data screening step of IDA in the data analysis process. Our proposals are based on checklists and on reproducible examples that facilitate the planning and performance of data screening. We also discuss how embedding IDA in statistical analysis plans can enhance the quality of planning and reporting of observational studies.

Analysis of time-to-event for observational studies: Guidance to the use of intensity models

Maja Pohar Perme

University of Ljubljana, Ljubljana, Slovenia

This talk will present a paper prepared as a part of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies), a work that provides guidance for researchers on the conduct of time-to-event analysis in observational studies based on intensity (hazard) models [1]. In this talk we use an example of peripheral arterial disease patients to illustrate various topics that play a vital issue in the analysis of survival data. We start by discussing basic concepts like time axis, event definition and censoring. Hazard models are introduced, with special emphasis on the Cox proportional hazards regression model. We mention the issue of immortal time bias and wrap up the basic survival topics by extending them to the competing risks setting.

References

[1] P. Kragh Andersen *et al.*, "Analysis of time-to-event for observational studies: Guidance to the use of intensity models," *Statistics in Medicine*, vol. 40, no. 1, pp. 185–211, 2021. DOI: 10.1002/sim.8757.

Correctly accounting for misclassification when linking latent groups with external variables

Cécile Proust-Lima[1], Maris Dussartre[1], Viviane Philipps[1], Cécilia Samieri[1], Paul Gustafson[2] and Pamela A. Shaw[3]

[1]University of Bordeaux, Bordeaux, France

[2]University of British Columbia, Vancouver, Canada

[3]Kaiser Permanente Washington Health Research Institute, Seattle, WA, United States

Latent groups (LGs) constitute a convenient solution to summarize complex multidimensional exposures such as lifestyle behaviors. Once the LG structure is defined and each individual is assigned to a group, predictors of the LGs or association between the LGs and health outcomes can be assessed in subsequent regression models. Yet, the quality of inference in the subsequent analyses may be altered by the inherent error of classification when assigning each individual to a specific LG. As part of the "measurement error and misclassification" topic group of the STRATOS Initiative (STRengthening Analytical Thinking for Observational Studies), our goal was to review the methods adopted in the literature to correct for this misclassification and, using simulations, compare their performance and potential biases to ultimately provide recommendations. Four methods were identified: (i) the naive approaches which directly use the class assignment in the subsequent regression, potentially weighted by posterior class membership probabilities; (ii) a bias-adjusted method that accounts for the assignment error in the subsequent regression using weights; (iii) a rewriting of the subsequent regression as a latent class model (LCM) with specifically determined class-membership probabilities that incorporate the assignment error (bias-adjusted LCM); (iv) a two-stage estimation of the LCM and the subsequent regression using their joint likelihood. In simulations exploring different levels of group separation and strengths of association with the outcome, we found that naive methods systematically showed substantial bias. The bias-adjusted weighted method showed residual bias with ambiguous classification. In contrast, bias-adjusted LCM and two-stage methods, which apply to various data, showed correct inference in all the scenarios provided the variance of the estimates correctly accounted for the sequential estimation procedures. The methods were further illustrated in an application assessing the association between groups of late-life lifestyle behavior and brain health outcomes in the population-based Three City cohort.

INDEX

Index

MY NOTES